DOCUMENT RESUME

ED 117 346 CE 005 967

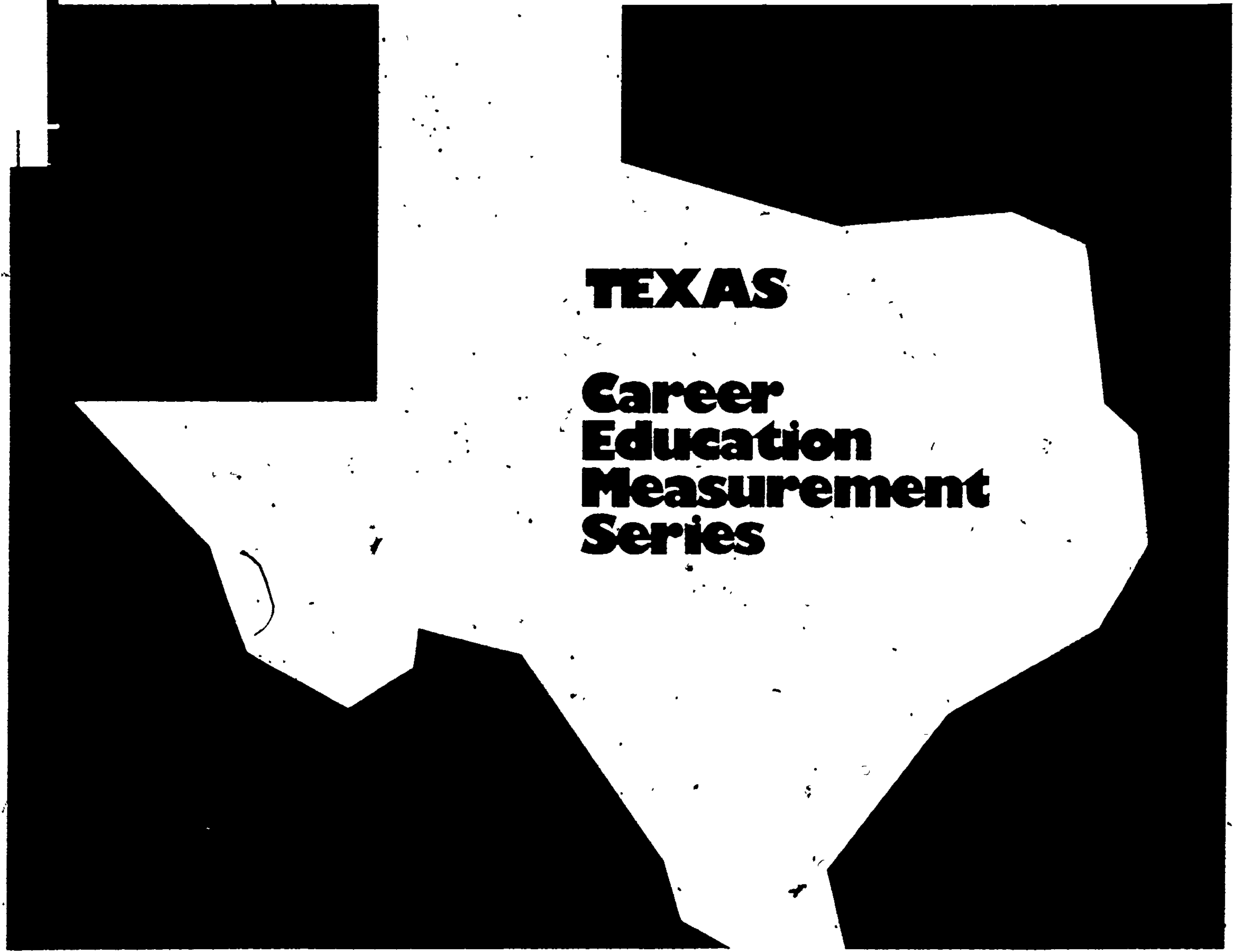
TITLE Development Report on the Texas Career Education Measurement Series.
INSTITUTION Westinghouse Learning Corp., Iowa City, Iowa.
SPONS AGENCY Partners in Career Education, Arlington, Tex.; Texas Education Agency, Austin.
PUB DATE 31 Aug 75
NOTE 108p.; For related documents, see CE 005 964-966

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage
DESCRIPTORS Career Education; Criterion Referenced Tests; Educational Objectives; *High School Students; *Item Analysis; *Measurement Instruments; Measurement Techniques; Secondary Education; State Surveys; Statistical Analysis; Student Characteristics; Test Results; *Test Validity; *Vocational Maturity
IDENTIFIERS Partners in Career Education Project; *Texas

ABSTRACT

A criterion-referenced measurement and diagnostic system for career education was developed using 79 of the 177 basic learner outcomes identified in Texas. Approximately 500 test items, referenced to the outcomes, were developed and submitted for student and professional review and statistical analyses following item tryouts and field testing of the instruments. A sample of schools was selected for each instrument at each of two levels, with 10 instruments at the lower level (grades 7 and 10) and 12 instruments at the upper level (Grades 8 and 11). In all, 506 classes were distributed among 130 campuses in 84 school districts. Various statistical procedures were used in item and instrument validation for item tryouts and field testing. Forty-four of the learner outcomes were tried out with students who had received instruction specifically designed to develop the behavior described by these outcomes. Data were obtained on 51 objectives measured by 215 items for the 44 learner outcomes. The test results were reported to give the student and school personnel diagnostic information about student performance on the outcomes by using the school curriculum--reference evaluation format. Over two-thirds of the document contains appended materials related to the processes involved in the study. (Author/EC)

DEVELOPMENT REPORT



TEXAS

Career Education Measurement Series

DEVELOPMENT REPORT ON THE TEXAS CAREER EDUCATION MEASUREMENT SERIES

PREPARED BY
WESTINGHOUSE LEARNING CORPORATION/
MEASUREMENT RESEARCH CENTER

FOR
THE TEXAS EDUCATION AGENCY
AND
THE PARTNERS IN CAREER EDUCATION PROJECT

AUGUST 31, 1975

ACKNOWLEDGEMENTS

REPORT COORDINATOR ....JAMES R. VEALE Ph.D., STATISTICAL CONSULTANT

Gratitude is expressed to the large number of teachers and students who contributed their thoughts and time to insure a high quality measurement system for Career Education.

ABSTRACT

Introduction.

A criterion-referenced measurement and diagnostic system for career education was developed using 79 of the 177 Basic Learner Outcomes identified in Texas. Approximately 500 test items, referenced to the outcomes, were developed by professional item writers with limited input from a select sample of Texas educators. These items were submitted to extensive student and professional review and statistical analyses following item tryouts and field testing of the instruments

Item Development and Validation.

The 79 outcomes were described in more detail by TEA and PARTNERS staff called "expansions. One to ten behavioral objectives (approximately 220 in all) for each of these expanded outcomes were written by WLC/MRC. Item development workshops with Texas educators were held to generate item ideas, and a total of 500 items from these ideas and the literature were written by SCORE test development specialists. These items were submitted to student and professional review. Professional reviews were based on the following criteria. (1) objective-item linkage, (2) reading level (6th grade), (3) non-offensiveness, (4) clarity, and scorability of items. Student reviews were conducted with groups of five eleventh-grade students with at least two representatives of each sex and a black, a brown and a white student. Approximately 400 items were reviewed at 34 schools.

Sampling Procedures.

Approximately 1,800 eighth and eleventh-grade Texas students in 60 classrooms from Education Service Centers (ESCs) IV, X, XI, XIII, and XX were selected for the first item tryout. The items were arranged into fifteen "packages" and each package was administered to four classrooms of students. one eighth-grade class from a campus of over 75% Mexican-American, one eighth-grade class from a campus of over 75% black, one eighth-grade class from a campus of over 75% anglo, and one eleventh-grade class from a campus of over 75% anglo. A second tryout focused on 200 additional items. For the spring (1975) field test, a statewide sample of approximately 13,000 students was selected (not twenty regional samples) using a stratified sampling procedure for drawing schools according to the "proportional allocation" of students from the following strata. (1) less than 33% Mexican-American, less than 33% black, (2) less than 33% Mexican-American, greater than 33% black, (3) greater than 33% Mexican-American, less than 33% black. A sample of schools was selected for each instrument at each of two levels, with ten instruments at the lower level (grades seven and ten) and twelve instruments at the upper level (grades eight and eleven). In all, 506 classes were distributed among 130 campuses in 84 school districts.

Statistical Procedures for Evaluation of Items and Instruments.

A variety of statistical procedures was used in item and instrument validation for item tryouts and field testing. *Item tryout* analysis focused on. (1) measures and tests related to item difficulty — the relative difficulty of the items as measured by p-values (the proportion or percent correctly answering the item) and the significance test for chance (guessing) level performance as determined by the "Z-test", (2) chi-square test for uniform foil response distribution—a test indicating the deviation from a uniform foil response distribution; and (3) variations of p-values and foil response distribution across ethnic groups (black, Mexican-Americans, and "others"). The statistical reports for the *field test* included the statistics used in the item tryouts and, in addition. (1) measures of internal consistency. point biserial correlation coefficient — a measure of the extent to which the students' performance on the item is correlated with performance on the outcome, (2) measures of instrument reliability — the Kuder-Richardson "formula 20", (3) cultural validity analysis. (a) chi-square test for detecting heterogenous foil response distributions across cultural groups or "cultural variation," (b) Cramér's V — a measure of cultural variation which incorporates the sample size, (c) measures of cultural variation with probabilistic interpretations which are especially useful for items with a small number of incorrect responses, (d) content analysis which describes "bad" foils, ethnic bias, sex bias, and/or diagnostic items, and (4) regression analysis — a statistical technique using p-values, number of foils, and z-scores for placement of items at appropriate grade level.

Sensitivity-to-Instruction.

Forty-four of the learner outcomes were tried out with a special group of students who had received instruction specifically designed to develop the behavior described by these outcomes. Learning modules were prepared for students in the eighth and eleventh grades for objectives believed to be amenable to instruction over a relatively short period of time. About 138 teachers in 36 schools volunteered to function as experimental and control groups. The students in the experimental group were pretested, instructed, and posttested utilizing WLC/MRC test items, the students in the control groups were pretested, received no instruction, and were posttested utilizing the same items. The following statistical procedures were used in analyzing the data (1) the Internal Sensitivity Index (ISI) measuring item quality from the perspective of the total test's discriminating power, (2) the External Sensitivity Index (ESI) and the Roudabush "S" measuring an individual item's ability to reflect learning (independent of the test), (3) the Objective Sensitivity Index (OSI) measuring the total test's ability to discriminate between learners and non-learners, and (4) statistical tests of significance for detecting differences between sensitivity indices for experimental and control groups. Data were obtained on 51 objectives measured by 215 items for the 44 learner outcomes.

Systems for Reporting Field Test Results to Teachers.

The test results were reported to give the student and school personnel diagnostic information about student performance on the outcomes by using a modified version of the SCORE (WLC/MRC) report which contains data on. (1) whether each student mastered each outcome, (2) the percent of outcomes mastered by each student, and (3) the percent of students mastering each outcome. A TEA-designed report which contains concise statements reflecting the degree of outcome mastery rather than the mastery/nonmastery format used in the SCORE reporting system was also utilized. Teachers favored the SCORE format, although the response to the questionnaire was low due to the fact that it was sent out rather late in the school year.

Statistical Procedures for Development of the Survey Instrument.

A survey instrument was developed to diagnose the need for further measurement of student performance by using one or more of the 22 category tests. A stepwise regression analysis was employed to select one or two items which correlate highly with the "outcome" scores.

Implications

Some of the implications of this effort are. (1) benefits occur as a result of using objectives that have been developed from a large-scale study of the views of students, educators, and those outside of the field of education, (2) objectives should be organized in appropriate form before selection/development of items, (3) design of reporting strategies should begin with the initial development procedures, (4) special attention should be given to item development activities for an area such as career education, (5) from 30% to 50% of the items in an objective-based system will be discarded during a rigorous review by students, (6) student review of items is productive, (7) advances have been made in the kinds of statistical analyses that are available for item and test construction in an objective-based measurement system, (8) additional benefits accrue when a state department of education, a regionally-based project, and a contractor work together.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Background

In 1972 the Texas State Board of Education identified career education as one of several top priorities for development. An initial implementing activity of this priority designation was a statewide survey conducted by the Division of Program Planning and Needs Assessment of the Texas Education Agency (TEA) and the Partners in Career Education Project (PARTNERS)[1] to find out what the citizens of Texas believed student development should be in terms of career education. The specific research question considered for the survey was. what skills, capabilities, knowledge, attitudes or other characteristics are considered to be basic requirements for 17-year-old Texas students? A listing of 279 possible student outcomes was prepared for the survey based upon the following:

- an extensive review of all available career education literature
- visits and consultations with career education practitioners both in Texas and in other states
- the products generated during a series of more than thirty work-group conferences with students, educators, parents and representatives of the business and industrial community.

More than 6,000 individuals (parents, students, educators and representatives of business and industry) from every region of the state reviewed the listing and rated the outcomes as either "basic," "desirable," or "inappropriate" for Texas students. Of the 279 outcomes utilized for the survey, 177 were rated as "basic" and 102 as "desirable." None were rated as "inappropriate" for Texas students. To assist in organizing the basic outcomes, they were arranged into nine categories.

A Request for Proposal (RFP) was issued by TEA detailing the requirements of a career education measurement system for Texas. The measurement system was to contain test items designed to measure student development in terms of the previously validated basic learner outcomes. In February of 1974 WLC/MRC entered into an agreement with PARTNERS and with TEA for the development of a criterion-referenced measurement and diagnostic system for career education.

Selection of Outcomes to be Measured

Reduction of the 177 basic learner outcomes to a more manageable number prior to commencing test item development was a first step in the developmental process. A series of activities involving staff of TEA, WLC/MRC, and PARTNERS, knowledgeable educators, and representatives of the business and industrial community reduced the number of basic learner outcomes to 79. WLC/MRC was instructed to develop test items for the measurement of this reduced number.

Item Development and Reviews

Following identification of the outcomes to be measured, WLC/MRC developed some 220 parallel behavioral objectives to be used as guides in the creation of test items. Upon acceptance of the behavioral objectives, WLC/MRC, PARTNERS and TEA personnel conducted an initial test item development program in two stages. In the first stage, groups of Texas educators consisting primarily of counselors and career education specialists were brought together in four regional education service centers (ESCs). After an initial orientation session, the greater part of one day was spent in generating items to measure specifically assigned objectives. Participants were urged to continue with the creation of test items during the following two week period. Items generated in this fashion were sent to WLC/MRC for refinement and editing. The second stage of the initial item development effort involved the creation of approximately 450 test items by the WLC/MRC professional staff. PARTNERS and TEA coordinated stringent review sessions with Texas educators, through the ESCs, across the state. The review process required the objective classification of items according to a specially prepared evaluation form. Another aspect of the item review which yielded valuable results utilized panels of students who were encouraged to give their opinions freely about the intelligibility, appropriateness for various grade levels, and the relevancy of the items.

[1] Partners in Career Education is a five-year cooperative project for the development and dissemination of a career education learning system It is funded by the Texas Education Agency and sponsored by Dallas and Fort Worth Independent School Districts and Education Service Centers Regions X and XI

After the item reviews were completed, the items were edited, revised, or deleted according to the composite recommendations of the reviewing groups. Additional reviews of the items and objectives were then conducted by PARTNERS and TEA personnel and by consulting career education professionals in preparation for an initial tryout of the items with students.

Item Tryouts and Analyses

Items found to be acceptable were then prepared in test format for a tryout with a broader sampling of Texas students. This sample was carefully selected to include students from all geographic areas of the state. All substantial minorities and all sizes of schools were represented. Test items utilized in this tryout (Phase I) were administered to more than 1,700 eighth and eleventh-grade students.

Simultaneously, 52 of the 220 behavioral objectives prepared by WLC/MRC were chosen for use in a sensitivity-to-instruction study. The students involved were pretested, instructed toward the particular objectives selected, and posttested. The instructional materials used were PARTNERS/teacher developed learning activity packages. The pretests and posttests were identical. This was the only phase of the Item tryout testing in which students were actually instructed toward objectives which the test items were designed to measure. A control group of students who had not received instruction toward the objectives was also used. WLC/MRC statisticians conducted tests of statistical significance for the observed differences in the proportion of gainers' (those who failed the pretest and passed the posttest) between experimental and control groups. Moreover, various sensitivity-to-instruction indices were computed and tests of statistical significance conducted on the difference in index values between experimental and control groups.

Completion of the Phase I tryouts marked a major milestone in the item development stage and a thorough reexamination of the WLC/MRC objectives prepared for each outcome and the items tried out for each objective was undertaken. PARTNERS, WLC/MRC and TEA personnel reviewed the relationship of these major components of the system for the purpose of assuring that there was a clear and significant link between each outcome, its objectives and the test items. Approximately 25% of the objectives were revised as a result of this reexamination. A similar percentage of the items were either revised or discarded. Also considered during this stage was the practicality of test administration. A decision was reached to convert a number of items from matching or open-end response patterns to a multiple-choice format. It should be noted that both PARTNERS and TEA personnel retained a willingness to utilize types of items which called for more difficult administrative modes in order to obtain more valid measurement. A number of short-answer items and attitudinal surveys were retained, as were teacher-completed longitudinal surveys of individual student behaviors. "Comic-strip" type items and videotape stimuli were continued as a part of the item bank.

Because of changes to existing objectives and new objectives being developed, new items were also needed. These new items were developed by WLC/MRC, by PARTNERS, and by TEA personnel. Two reviews of these new items were conducted, one to verify the item-to-objective match, and another for item content validity. Item reviewers had available all of the previously accumulated review information. The reviewed and refined items were then tried out (Phase II) in essentially the same manner that Phase I was conducted. Some of the Phase I items were again tried out during Phase II to gain additional response information. The number of students involved in Phase II was somewhat smaller than for Phase I, with approximately 1,600 individuals participating.

Analysis of the results of both Phase I and Phase II item tryouts was conducted by WLC/MRC. The analysis focused on three major concerns. (1) the relative difficulty of the items as measured by p-values and significance tests for chance performance (student guessing), (2) statistics measuring deviation from a uniform foil response distribution, and (3) variation of p-values and foil response distributions across ethnic groups (blacks, Mexican-Americans, and "others"). In addition, a technique utilizing professional judgment and regression analysis was developed for determining the appropriate grade level for the items tried out for each outcome. A three-day review session involving members of PARTNERS, TEA (including the Assessment of Career Education Steering Committee), and WLC/MRC was conducted using the accumulated data and subjective judgment as to content analysis. A number of the items tried out were dropped, some were passed as tried out, and some were passed subject to editing and/or revision.

Field Test

An extensive field test of the refined items initiated the final developmental stage. Twenty-two instruments utilizing 382 items — from an original bank of more than 500 — for the measurement of 200 objectives were designed. Items were sequenced on each instrument in the order of outcome difficulty within each category. A sample of 13,000 Texas students was selected from grades 7, 8, 10, and 11. The WLC/MRC Report Coordinator, working in close consultation with TEA, designed the sampling procedures and selected the sample of schools.

Analysis of Field Test Results and Instrument Design

The statistical procedures and software for scoring and analyzing the field tests were developed by the report coordinator and WLC/MRC programming staff. Statistical reports designed for the field tests included those statistics used in the Phase I and Phase II tryouts and the following additional components:

1. a measure of internal consistency (point biserial) and a statistic which measures the extent of influence of the p-value on the point biserial
2. a separate item analysis for each group corresponding to various cultural variables, such as ethnic origin, sex, and educational emphasis in the home, etc.
3. statistical indicators of "cultural variation," i.e., the degree to which foil response distributions (excluding correct response) vary across cultural groups

Some of the above procedures were developed during the course of the project in an attempt to deal more effectively with questions concerning item and instrument validity for criterion-referenced tests. For example, the procedures mentioned in (3) above, were found to be useful in detecting culturally related problems with items, such as bias, bad foils, bad format, etc.

The results of the field test were analyzed by personnel of TEA, PARTNERS and WLC/MRC. The statistical data were then used to determine which items should be dropped, revised, edited, or used without modification. This revision session resulted in sixteen instruments with a total of 273 items for use in measuring the nine categories of learner outcomes. Of the 273 items, 187 were judged to be acceptable in that they passed the review with minor or no modification. Using this pool of acceptable items, a stepwise regression analysis was conducted to determine which items were most appropriate for inclusion in a survey instrument intended for use in screening students prior to administration of the more detailed category tests. Based upon these statistical procedures and the judgment of TEA professionals, the survey test was developed. It will be tried out with a statistically controlled sample of Texas students during the fall of 1975 for a statewide needs assessment study.

CHAPTER II

ITEM DEVELOPMENT AND VALIDATION

Objectives

Activities required by the WLC/MRC contract began following selection of the 79 priority outcomes to be used for the assessment of career education in the state of Texas. Selection was made by TEA and PARTNERS personnel based upon the votes of a large number of Texas educators and other professional groups. Each of these 79 outcomes was then described in greater detail by TEA and PARTNERS staff personnel in paragraph format. These expansions were descriptors of the intent of each outcome.

Based upon the expanded outcomes, WLC/MRC prepared from one to ten objectives for each outcome. The objectives were stated in behavioral terms and formed the bases for test item development.

The objectives (approximately 220 in number) were reviewed by TEA and PARTNERS to assure that each one represented an element of the outcome for which it was written. The review also evaluated the sufficiency with which the objectives addressed all of the elements of each outcome.

Item Development

Once the objectives were developed and reviewed, the plans for item development began. Two processes were simultaneously initiated. One was to assign sets of outcomes and objectives to career education specialists and counselors in the Iowa City area and request that items be developed. The other process was to conduct four regional workshops in Texas for the purpose of training Texas educators and specialists to develop items.

The workshops consisted of a one-day meeting with about 20 to 30 people being trained in each workshop. In the morning, item development procedures and techniques were discussed. Included in the discussion was a review of item formats and the procedures for appropriately matching format to an objective. In the afternoon, the participants divided into groups of three to six to work on item development. During that time, the WLC/MRC representative circulated and critiqued the work being done. This item development work continued for about three hours at which time some of the work was collected.

At the end of the day, each participant was assigned specific outcomes/objectives and requested to attempt to develop additional items on an individual basis over a period of two to three weeks. These completed items were sent to WLC/MRC for review and refinement prior to inclusion in the measurement system.

Phase I item development was completed utilizing the experienced test development specialists who had been involved with the WLC/MRC SCORE program. Objectives were assigned to professionals from this program and within one month over 500 test items were delivered to TEA and PARTNERS.

Review

As the items were developed, they were submitted for review by

- a WLC/MRC career education specialist,
- the TEA staff and
- the PARTNERS staff.

The purpose of these reviews was to find out if the

- items measured the keyed objectives.
- language of the item was at a reading level of sixth grade or below.
- item communicated its intent.
- item measured was non-offensive.
- format was simple and clear.
- item was scorable.
- instructions for administration were clear.
- item was technically correct in use of terms.

The results of these professional reviews were submitted to WLC/MRC for inclusion in the revision/recommendations file.

The second phase of the review process involved students. In April, 1974, guidelines were developed to obtain the candid reactions of eleventh-grade students to the test items proposed for the career education measurement system. Early in May, 1974, a plan was finalized for obtaining evaluation data from students. This plan was outlined in the "Criteria for Item Acceptability." (See Appendix L.) Guidelines for student reviews as described in the plan were:

- Each item would be submitted to student review.
- Item reviews would be conducted by a person not employed by the school.
- The person conducting the review would serve as a facilitator and recorder of student reactions.
- Schools selected as review sites would contain students with different ethnic backgrounds (Mexican-American, black, and anglo) and both sexes.
- The review teams would include five eleventh-grade students with at least two representatives of each sex and a black, a Mexican-American, and an anglo student.

The student reviews were conducted by a TEA or a PARTNERS staff member. When the student review team at a particular school had been assembled, the individual conducting the review described the procedures, assured the students that they were not being tested but that they were being asked to critique new test items, stated how their input would be used, and explained that the test items had been written by a third party (the contractor). The last comment seemed to make the students feel free to comment on the items.

A list of questions was developed to guide the student review sessions (Appendix A). These questions dealt with item readability, appropriateness, structure, bias, and non-offensiveness. Students were asked to read a career education outcome and the item that was proposed for measuring it. Open discussion followed, with the recorder documenting student reactions for as many of the above areas as possible. After 15-30 minutes, direct questioning was used to fill gaps in the areas of concern listed above. On the average, students reviewed eight items in a two-hour session. Students tended not to tire as readily when they were asked to review items of differing formats.

In the judgment of TEA and PARTNERS staff members who conducted the sessions, the reviews were productive and fully justified the time and effort expended. The students were generally open in their comments about items. They saw implications that the staff and educator reviews did not see. Approximately 400 items were reviewed at 34 school campuses. Schools ranged in location from those in large cities to those in rural areas. A majority of reviews were conducted in metropolitan areas.

The results of the student review sessions are summarized as follows:

Conditions	Number of Items	Percentage
Acceptable	129	30%
Need Revision	267	63%
Rejected	30	7%
TOTALS	426	100%

In most instances, item writers had files of student suggestions for improvement as well as reasons for their recommended revisions. The results of the student review of items indicated that the obtaining of student inputs is a necessary step in the development of an objective-based instrument. Although statistical analyses of item tryout data will yield information pertinent to certain item characteristics, student interviews seem to be the most feasible and economical method of determining answers to questions such as:

- Do students understand the intent of the item?
- Is the item too advanced or too simple for the target age-group of students?
- Do certain words or phrases offend the target age-group?
- Why do many students feel that there is more than one correct answer?

Finally, each of the twenty ESCs was requested to provide a sufficient amount of staff time to conduct teacher/educator review sessions to obtain a critique of the items from classroom teachers, counselors and administrators. One-half day was allocated for these sessions.

The twenty regions were divided into four characteristic classes. (1) Mexican-American, (2) black, (3) rural white (anglo), and (4) big city suburban white (anglo). Each item set (about sixteen items) was submitted to one review group of five educators in each of the four classes using the review form contained in Appendix B. In this way, every item was seen by four different groups of people. It was anticipated that this would provide input on every item from representatives of every major population group in the state.

All of the information obtained from the four phases of review (career education specialists, TEA and PARTNERS, students, and educators) was compiled and summarized by WLC/MRC staff. When a disagreement or discrepancy in decision existed for an item, the WLC/MRC professional staff reviewed all of the inputs from the various groups and disposed of the item in a manner considered to be most consistent with the reviewing groups positions. When understandability was in question, emphasis was placed upon student input. If, however, the problem was one of administration or clarity of the scoring guide, emphasis was placed upon educator input.

The summary information obtained from a detailed analysis of the review data was referred to the professional writers for use in item revision in preparation for item tryouts. Every item that was not passed by all review groups as suitable for use was referred to an item writer for possible revision. In a few instances, items were not changed because of insufficient information from the reviewing groups. Tryout data were required to determine final revision on these items.

Tryouts

The trying out of items was an important step in the overall development of the Career Education Measurement System because the process provided a substantial amount of insight about each item prior to its becoming part of an instrument. This information was gathered from approximately 1,800 eighth and eleventh-grade Texas students and, in some instances, teachers. The particulars gathered about each item included appropriateness, readability, acceptability, and clearness of direotions.

Although information similar to this was obtained through student and professional reviews, the item tryouts presented the items visually in test context and format. Inputs from the large number of students who actually responded to these tests provided real life information about the test items. From the data obtained, the following decisions could be made:

- include an item in the instruments being designed,
- exclude an item,
- revise an item prior to inclusion in an instrument, and
- determine the range of additional items needed for satisfactory measurement of an outcome.

For the initial tryouts, the test items were organized into approximately fifteen booklets or packages by category and mode of administration. The classroom was the smallest unit of sampling for the item tryouts. Approximately sixty classrooms were used. Each item package was administered to four classrooms of students as follows:

- one eighth-grade class from a campus over 75% Mexican-American,
- one eighth-grade class from a campus over 75% black,
- one eighth-grade class from a campus over 75% anglo, and
- one eleventh-grade class from a campus over 75% anglo.

Administration time was 45 minutes or more for each package. Each student was asked to complete student identification information questions. In addition, approximately 20% of the students from each classroom were randomly selected for individual interviews of about ten minutes following completion of the test. The tryout administration extended over a two-hour period, in most instances. Personnel from either an ESC, PARTNERS, WLC/MRC, or TEA administered the test packages in cooperation with the teacher in charge of the class. The package administrator conducted the personal interviews with the selected students.

The tryout data were used for determining the extent to which each item met the following criteria for acceptability:

- Not more than 10% of the students will indicate difficulty in understanding the item.
- Not more than 10% of the (student) responses may indicate offensiveness or bias.
- Not more than 10% of the students in item tryouts will indicate difficulty with understanding item directions as determined by interview.
- No more than 5% of the teachers should express any difficulties in scoring the items.
- Questions were asked of educator-administrators about ease of administration and clearness of directions. No more than 15% of the responses should indicate any difficulty.

Additional Item Development and Tryouts

As a result of the reviews described above, because of changes to objectives and due to new objectives being developed, many new items were needed. Of the 400 items tried out, approximately 25% were discarded for various reasons.

Because of the limited time available, PARTNERS sent a staff of five people to Iowa City to work on the review and revision of the new items with WLC/MRC staff members. Items were routed to a review/revision committee as they were written and the necessary changes, revisions made immediately. More than 200 items were completed.

As revisions were completed, items were typed in a camera ready format of about fifteen items in each test booklet for a second tryout phase.

The Phase II item tryouts were accomplished utilizing students in ESCs I, X and XI. Administration of the test booklets was accomplished by the PARTNERS staff in cooperation with the participating classroom teachers. The procedures followed were essentially the same for both Phase I and Phase II.

A final tryout was conducted which included trying out Phase I items for which there had been an insufficient number of respondents during Phase I. This tryout was also conducted by the PARTNERS staff. Results of all three tryouts were utilized in a final review session attended by personnel from WLC/MRC, PARTNERS, and TEA representatives, including the ACE Committee. Decisions were made about which items would become a part of field test.

Preparation for Field Tests

By March of 1975 the information accumulated from three phases of item tryouts, a sensitivity-to-instruction study (see Chapter V for details), and four workshop conferences had been subjected to detailed examination and penetrating analyses. Many of the original test items had been abandoned, most of the remainder had been revised in some fashion, and a number of new items had been written. The total number of items available for the field test was 382. (See Appendix M for materials used in Texas with ESCs and local school districts during the field test). These were prepared in 22 separate instruments for administration to students in four grades at two levels. The level one instruments were for grades seven and ten and the level two instruments were for grades eight and eleven. Sampling procedures for the field test are discussed in Chapter III.

The following considerations guided the design of the 22 instrument battery of tests:

- a standardized format
- clarity of instructions for administration and scoring
- item readability
- item simplification
- item arrangement within each instrument
- grade level appropriateness

Grade level appropriateness was determined by a regression analysis technique which is discussed in Chapter IV.

Post-Field Test Reviews

As a result of the field trials in the Spring of 1975, item analyses were provided to TEA and PARTNERS. Some tentative guidelines for item validation were proposed by WLC/MRC statisticians. (See Chapter IV for a discussion of the statistical procedures.)

Two teams of reviewers were formed, each having representation from the three organizations (PARTNERS, TEA, and WLC/MRC). The teams reviewed the findings using the following. (1) the statistical analyses (summary sheets prepared by TEA), (2) a content analysis examining the quality of the content of the item in relationship to the outcome it purported to measure, as well as the vocabulary level of the items, and (3) teacher input from a questionnaire obtained from the spring field test. Each item was then categorized as acceptable, editable with minor revisions, or inappropriate for the measurement system.

Assembling the Category and Survey Tests

Assembling the final tests consisted of selecting appropriate formats and organizing the items into sixteen category instruments and one survey instrument. The organization of items for the category instruments was based upon the general category, the sub-category, and the outcome for which sets of items had been developed. The order or sequence of items within an instrument was determined by the content dimension of each item. The resulting arrangement was according to difficulty, specificity, and item length. Also considered was the relationship of items within a set or group which measured a sub-category, the stimulus for each item, and the response patterns of linked items.

The survey instrument was developed to diagnose student performance in relation to the various categories and sub-categories as measured by the sixteen category tests. The items found to be the most appropriate (representative) from each sub-category were selected to provide indicators of probable student performance on the outcomes contained within a particular sub-category. Forty-five items were selected for the survey instrument to represent the 26 sub-categories into which the nine general categories were divided. Performance on the survey test will be utilized to determine whether administration of one or more of the category tests to a student (or groups of students) is indicated.

CHAPTER III

SAMPLING PROCEDURES

Item Tryout Sample

Approximately 1,800 eighth and eleventh-grade Texas students were selected for the first tryout sample. Approximately 60 classrooms were used (the classroom was the smallest unit of sampling for the item tryouts). The items were arranged into fifteen packages" and each package was administered to four classrooms of students. one eighth-grade class from a campus of over 75% Mexican-American, one eighth-grade class from a campus of over 75% black, one eighth-grade class from a campus of over 75% anglo, and one eleventh-grade class from a campus of over 75% anglo. Because of their high ethnic concentration, ESCs IV, X, XI, XIII, and XX were selected as item tryout sites for grades eight and eleven. A sample of campuses in these districts was selected proportional to student enrollment.

This was *not* a random sample. No statistical controls were deemed necessary here since the purpose of item tryouts was to try out items, not to make statewide inferences. A list of campuses participating is given in Appendix C by district and region.

Schools participating in the Phase II item tryouts were located in ESCs X and XI. These were selected from the sample used for Phase I. In addition, four schools were added in ESC I to include a greater number of Mexican-American students. Each Phase II package was tried out with six classrooms:

- eighth and eleventh-grade blacks;
- eighth and eleventh-grade Mexican-Americans;
- eighth and eleventh-grade "others."

Field Test Sample

A random sample of approximately 13,000 students was selected for the field test which was administered in the spring, 1975. This sample was smaller than originally planned. Additional refinement of the instruments was considered to be essential prior to attempting a larger statewide field trial. Moreover, because of the developmental stage of the measurement instruments neither state nor regional inferences were considered. Nevertheless, statistical controls were applied in an attempt to obtain a sample that would yield unbiased estimates with reasonably good precision. The main purpose of the field test was, however, to secure information to be used for further refining the measurement instruments.

A stratified sampling procedure was utilized for selecting schools from the following strata.

- less than 33% Mexican-American, less than 33% black;
- less than 33% Mexican-American, greater than 33% black,
- greater than 33% Mexican-American, less than 33% black.

A fourth category, "greater than 33% Mexican-American, greater than 33% black," contained only a few schools; these were randomly allocated to strata two and three above.

A sample (of schools) was selected for each instrument in four grades at two levels. grades seven and ten for lower level instruments and grades eight and eleven for upper level instruments. The number of schools selected within each stratum was determined by "proportional allocation", with respect to the number of students within each stratum. In other words, the number of schools selected within each stratum (for each instrument) is proportional to the number of students in each stratum, the more students the more schools are sampled.

The information relevant to the allocation of schools to strata is given in the table below.

Grade	Stratum	Population	Proportion[1]	Number of Schools Selected
8 (7)	1	146,456	0.70	8
	2	28,872	0.14	2
	3	32,399	0.16	2
11 (10)	1	136,053	0.73	8
	2	19,296	0.10	1
	3	31,227	0.17	2

[1] Proportion equals the number of students in the stratum divided by the total number of students in all strata (at a grade level)

The reason for taking n = 12 schools in grades seven and eight, and n = 11 schools in grades ten and eleven was to obtain an allocation which was closer to the values given in the "proportion" column.

The schools in each stratum were then selected with probability proportional to size (p.p.s.). That is, larger schools were more likely to be selected than smaller schools, and their relative likelihoods were proportional to their relative sizes. This process may be illustrated as follows. Suppose there are five schools in a certain stratum and two are to be selected. The school populations are 20, 30, 100, 150, and 200, respectively. The populations may be represented graphically as ranges or distances between points as plotted in Figure 1 below. For example, school 3 falls in the range of 50 to 150 which corresponds to a population of 100 students.

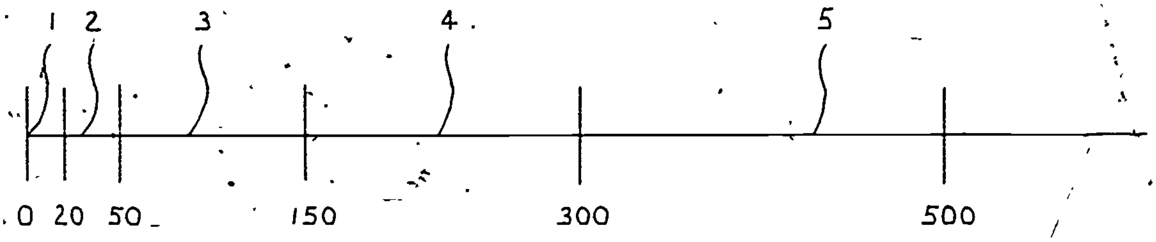


Figure 1. Graph representing populations of schools 1-5.

The probabilities of selection are thus proportional to the lengths of the line segments corresponding to the populations. If one thinks of each unit on the line in Figure 1 as representing one student, it is clear that each student has an equal chance of being selected. This is as it should be, since a sample representative of students in the population is desired. (cf., Cochran, 1963.)

Finally, one classroom was volunteered from each school. The classroom selected was "typical" according to ethnic and other cultural considerations. There were, prior to field testing, ten instruments at the upper level. Since the seventh and eighth-grade samples each had twelve classrooms per instrument and the tenth and eleventh-grade samples each had eleven classrooms per instrument, there were

$$(10\times11) + (12\times11) + (10\times12) + (12\times12) = 506$$

classrooms selected altogether. These 506 classes were distributed among 84 school districts and included 130 campuses. Since the average class size was thought to be around 30, a sample of around 506 × 30 = 15,180 was anticipated. (The number of students actually selected was somewhat lower than this number.) A list of the schools selected and information concerning their involvement in the project is provided in Appendix C.

Estimation Procedures

An attractive by-product of the sampling method discussed in the preceding section is that "self-weighting" procedures may be employed to estimate p-values, percent mastering objectives, point biserials, and KR-20 reliability coefficients, obviating the computation of more complicated weighted estimates. The theoretical basis for using "self-weighting" estimators is given in Appendix K.

CHAPTER IV

STATISTICAL PROCEDURES FOR EVALUATION OF ITEMS AND INSTRUMENTS

Introduction

This chapter contains a description of the statistical procedures used in item and instrument validation, along with some examples of how the techniques were applied to actual test data.

Various statistical procedures were employed to secure a scientific evaluation of the items and instruments which comprise the Texas Career Education Measurement System. Some of these procedures, such as p-values, point biserials, and KR-20 reliability coefficients, are classical test construction statistics. During the course of the project, however, new approaches and procedures to statistical validation of items were developed. For example, the techniques for measuring the cultural validity of items were developed through valuable interaction between the WLC/MRC project coordinator and Keith Cruse of TEA. The test for chance (guess) level of functioning, a Z-test, was developed in order to test whether or not the p-value for a sample was above or below that which would be expected by chance if the students were guessing. All of the above statistics were computed by the WLC/MRC Instrument Analysis program package. (See Appendix D.)

Measures and Tests of Item/Instrument Appropriateness

1. Measures and tests related to item difficulty (p-values and Z-test):
The difficulty of an item is traditionally measured by the proportion (or percent) correctly answering the item or p-value, denoted $\hat{p}$. This may be adjusted to account for guessing (cf., Lord and Novick, 1968, and Magnusson, 1967). In addition, WLC/MRC statisticians proposed a Z-test to test the hypothesis that the students, as a group, are at the chance (guess) level of functioning on a given item. This test is conducted by the following formula:

$$Z = \frac{\hat{P} - \left(\frac{1}{f+1}\right)}{\sqrt{\left(\frac{1}{f+1}\right)\left(\frac{f}{f+1}\right)/n}}$$

when $\hat{p}$ is the p-value, f is the number of foils, and n is the number of respondents (sample size). If the hypothesis that $p = 1/(f + 1)$ is true, i.e., the population sampled is functioning at the chance or guess level, the above statistic has (approximately) a standard normal distribution (for large n, say $n > 50$). If Z is positive and statistically significant, one may conclude that the students are operating above the chance level. On the other hand, if Z is negative and statistically significant, one concludes that the students are operating "below the chance level." This may be an indication that the item is wrongly keyed or that the item format is inappropriate. If Z is not statistically significant, one concludes that the students are guessing.

2. Chi-square test for uniform foil response distribution:
Ideally, one would hope that the foils in a multiple-choice item would draw about equally. To test this hypothesis (conditional on a given total number of foil responses), one may compute the chi-square statistic:

$$\chi^2 = \sum_{i=1}^{f} (O_i - E_i)^2 / E_i$$

where f is the number of foils, O_i is the observed number of responses to foil i, and $E_i = \Sigma O_i/f$, the "expected" number of responses to foil i under the uniform foil response hypothesis, $i = 1,2,\ldots,f$.

3. Measures of internal consistency (point biserial correlation coefficient):
The classical measure of "internal consistency" of a test, i.e., the degree to which the items measure the same thing, is the point biserial correlation coefficient, denoted r_{pb} (cf., Lord and Novick, 1968, and Magnusson, 1966).

In the Texas Career Education Measurement System (CEMS) items are grouped or clustered around learner outcomes so that each outcome in effect becomes a subtest of a larger instrument. The point biserial is the degree to which performance on an item is correlated with performance on the learner outcome, i.e., the consistency with which students correctly or incorrectly answer an item in relation to its outcome score. Moreover, point biserials were computed for each cultural (ethnic and sex) group.

The p-value influences the value of the point biserial. In particular when $\hat{p}$ becomes close to 0 or 1, r_{pb} becomes close to zero. The "WLC/MRC Instrument Analysis" computes a statistic called "maximum" r_{pb}, which is simply the value r_{pb} would achieve if $\hat{p}$ were equal to 1/2. It may be obtained from r_{pb} as follows.

$$\max r_{pb} = \frac{r_{pb}}{2\sqrt{\hat{p}(1-\hat{p})}}$$

This statistic, when contrasted with the value of r_{pb}, provides an indication of the extent to which the p-value is influencing the point biserial. Thus, if r_{pb} is quite low, and max r_{pb} is not low, this may be due to a low (or high) p-value, and not (necessarily) due to lack of internal consistency.

4. Measures of instrument reliability (KR-20):
The Kuder-Richardson "Formula 20" or KR-20 was used to measure test reliability (cf., Lord and Novick, 1968, and Magnusson, 1966). The KR-20 is an internal consistency measure of reliability. Thus, like the point biserial, it measures the degree to which the items all measure the same thing. Unlike the point biserial, the KR-20 provides *one* measure for any given instrument. KR-20's were computed for each outcome instrument. Overall, 37% of the outcomes had KR-20's greater than 0.50.

Cultural Validity Analysis

Are the items and instruments measuring what they are intended to measure for students in each cultural group? The question of the cultural validity of items and instruments is investigated using an approach developed by the coordinator and others. (cf., Veale and Foreman, 1975). The approach focuses on the foil response distribution broken down by cultural group. Three cultural variables were considered in the cultural validity analysis of the Texas career education test items. (1) ethnic origin (Mexican-American, black, and other), (2) sex (male, female), and (3) "educational emphasis index" (high, medium, and low). The data available from the "Student Information Sheet" given to each student at field test time were utilized to obtain the aforementioned cultural information. (See Appendix E.) Only the first two (ethnic and sex) cultural variables are considered in the discussion which follows. The extent of variation in foil responses across cultural groups is said to measure "cultural variation" which may be evidence of cultural bias.

1. Description of the statistical techniques:
The following example serves to illustrate the approach and statistical technique. Suppose that the total number in the sample is 500, with 125 blacks and 375 non-blacks. Suppose further that 75 blacks and 225 non-blacks answer the item correctly, yielding identical p-values of 0.6, and that the foil distribution is as in the table below:

Item Data With Equal p-value and
Heterogeneous Foil Response Distributions

	A	B	C	**Totals**
Black	40	10	0	50
Non-black	50	50	50	150
Totals	90	60	50	200

Clearly, blacks are strongly attracted to foil A, while non-blacks are uniformly attracted to the three foils. This may be an indication of cultural bias, i.e., because of cultural factors only (or primarily) blacks are drawn to foil A. If this differential attraction to foil A were not present, the p-values for blacks and non-blacks might have been quite different.

On the other hand, it may be that foil A is a more reasonable response than B or C among students who have been instructed to the objective being measured. In this case, it might be that blacks have been instructed (and thus find that foil A is more attractive than B or C) while non-blacks have not been instructed (and thus are uniformly attracted to the foils due to guessing). Another possibility is that A is a "bad" or "tricky" foil. suppose that blacks had not been instructed to the objectives, while the non-blacks had been instructed. In this case, blacks may be drawn to the "tricky" foil simply because they have not been instructed. In these cases, no cultural bias can be claimed. *Cultural variation does not imply cultural bias.* The approach may thus yield valuable diagnostic information about the group or about the item (other than bias), as well as information about cultural bias (Appendix F). Several statistical techniques were employed to measure the degree of cultural variation in foil response distributions. One of these is the chi-square statistic based on the foil responses for the various cultural groups. (Formally speaking, this statistic tests the statistical hypothesis that cultural groups and foil response are independent or uncorrelated.) For example, the chi-square for the data in the previous table is 37.04 which is statistically significant at the .001 level. A measure of the *degree* of cultural variation is Cramér's V statistic which is found to be 0.43 in this example. Other statistics which have probabilistic interpretations and operational significance irrespective of the sample size (in this context, the total number of foil responses) were utilized to measure the extent of cultural variability, especially in cases where the chi-square does not apply. For a more detailed description of the statistical procedures used to measure the cultural variation of items, see Appendix G.

In addition to measures of cultural variation, conventional item analysis statistics (such as point biserials) were used as *supplementary* indicators of possible cultural bias. For example, if the chi-square and Cramér's V statistics manifest a high degree of cultural variation for an item and, moreover, the point biserials *vary* across cultural groups, the item is probably culturally biased. (However, variation in the point biserials alone, without corresponding cultural variation in foil responses, does not constitute clear evidence of cultural bias.)

A computer program has been written at WLC/MRC to compute the various statistics used to measure cultural variation. The data from the field tests were analyzed according to the aforementioned techniques. Some tentative "cut-off" values (of chi-square, V, etc.) were suggested by WLC/MRC statisticians, but were used only as rough guidelines. Flexibility of application was strongly encouraged.

2. Content analysis:

The content analysis is handled by grade (and grade combinations). Appendix H consists of a set of tables for the upper and lower grade samples in which a probable cause of cultural variability (of foil responses) is presented for each item by booklet number and test item number. Following the tables are several sample items which manifest cultural variation (statistically) and a brief explanation of the probable cause of the variability (bias, diagnostic foils, bad foil, bad format). In some cases, the variability existed at two grade levels and is discussed for both grade levels together. Some items seemed to have more than one possible source of variability. These items are discussed under separate combination headings.

It should be made clear that the discussion of these items in Appendix H constitute (data-based) content hypotheses of one specialist.

Item and Instrument Analysis: A 'Global' View

In order to take maximum advantage of the available statistical data, a flexible, 'global' approach is recommended. Pre-assigned "cut-offs" were used as rough guidelines only. Rigid application of such systems (however tempting for expedient decision making) was strongly discouraged.

All of the statistics discussed in the previous sections should be considered in making decisions about items. The following three examples serve to illustrate how this process should work.

Example 1. (Item 12, Booklet 11, Grade 7)

Item: Grace wants a job where she does not have to deal with strangers.
Which career do you feel would *BEST* match Grace's goal?

(A) receptionist
(B) bookkeeper
(C) public librarian
(D) salesperson

```
FORM 01                              SAMPLE OF ITEM PRINT OUT

                                        OBJECTIVE 0104000

                                                                            SIG    95%          95%       CHI    SIG  PT-BI  MAX-PT
ITEM  GR  SP1   SP2   N    A    B    C    D    E   F   INV  DM  OMIT  Z      LEV  (2 TAIL)     (1 TAIL)   SQ     LEV  SER    BI SER
12    07              339  29   56*  8    6            0    0   0     13 20  000  0 51  0 61   0 52       77 17  000  0 59   0 60
12    07  MA          80   38   35*  19   8            0    0   1   - 2 07   019  0 24  0 46   0 26       17 28  000  0 56   0 58
12    07  BL          55   36   40*  9    15           0    0   0     2 57   005  0 27  0 53   0 29       11 44  003  -0 39  0 40
12    07  OT          204  24   64*  4    3            0    0   0     14 39  000  0 62  0 75   0 63       56 08  000  0 58   0 62
          TOTAL       339  29   56*  8    6            0    0   1     13 20  000  0 51  0 61   0 52       77 17  000

      GRADE 07
          FOILS       A         C         D

          GROUP
        MA            30        15        6
        BL            20        5         8
        OT            49        8         6
          CHI-SQ -      9 906     SIG LEV =      042       V =  0 184     DF -      4 000
          T -         0 035     C  I -     -0 012   L* =      0 000     C  I =    0 000
          T* -        0 033                         L =      0 000
```

The p-value (overall) is .56, which yields a Z value well above chance level. It is noted, however, that the p-values are quite variable across ethnic groups, with minorities doing worse than anglos.

The chi-square for testing uniformity of foil responses is highly significant, due to the strong attraction to foil "A." The cultural validity indices are as follows. X^2 = 9.906 (significant at .05 level), V = .184, T = .035, T_{95} = −.012, L* = 0.000, L^*_{95} = .000. There is some degree of cultural variability present.

Minorities ("MA" and "BL") are more attracted to "C" and "D" than are "others." Moreover, blacks are more attracted to "D" while Mexican-Americans are more attracted to "C" (although "A" is the most popular foil). Finally, the point biserial (overall) is reasonably high (.58), indicating fairly good internal consistency. It is interesting to note, however, that the point biserial is only 0.39 for blacks, while it is 0.58 for "others." The conclusion is that the item is ethnically biased. Minorities simply have had less experience with these occupations.

Example 2. (Item 12, Booklet 72A, Grade 8)

Item: Which *ONE* of the following quotations reflects an individual's positive attitude toward participation in the economic system of the United States?

(A) "Big businesses cheat on their taxes, so I do too."
(B) "Irish wool is of better quality than local wool."
(C) "I've invested my savings in a local corporation."
(D) "I think that I should be able to get money any way I can."

```
FORM 16                              SAMPLE OF ITEM PRINT OUT

                                        OBJECTIVE 0713000

                                                                            SIG    95%          95%       CHI    SIG  PT-BI  MAX-PT
ITEM  GR  SP1   SP2   N    A    B    C    D    E   F   INV  DM  OMIT  Z      LEV  (2 TAIL)     (1 TAIL)   SQ     LEV  SER    BI SER
012   08              237  11   6    71*  11           0    0   0     16 31  000  0 65  0 77   0 66       4 67   096  0 40   0 44
012   08  MA          38   5    8    61*  24           0    0   2     5 06   000  0 45  0 76   0 47       6 13   046  0 53   0 54
012   08  BL          40   28   10   53*  10           0    0   0     4 02   000  0 37  0 68   0 39       5 14   076  0 36   0 36
012   08  OT          159  8    4    78*  9            0    0   0     15 43  000  0 70  0 86   0 71       2 52        0.29   0 35
          TOTAL       237  11   6    71*  11           0    0   1     16 31  000  0 65  0 77   0 66       4 67   096

      GRADE 08
          FOILS       A         B         D
      NUMBER OF RESPONSES
      CULTURAL GROUP
        MA            2         3         9
        BL            11        4         4
        OT            13        7         14
20% OF VALUES LESS THAN 5
          T =         0,071     C  I =     0 008    L* =      0 212     C  I =    0 050
          T* =        0 098                         L =       0 175
```

The overall p-value is .71, well above chance level. The numbers responding to the foils are not sufficient to perform chi-square for testing cultural validity. However, L* is high, 0.212, the lower 95% confidence interval is 0.05.

Note the differential attraction of "A" and "D" for Mexican-Americans and blacks. Finally, note that the point biserial varies from 0.53 (Mexican-American) to 0.36 (black) to 0.29 (other). There is some evidence of cultural bias in this item, although total number of respondents was low.

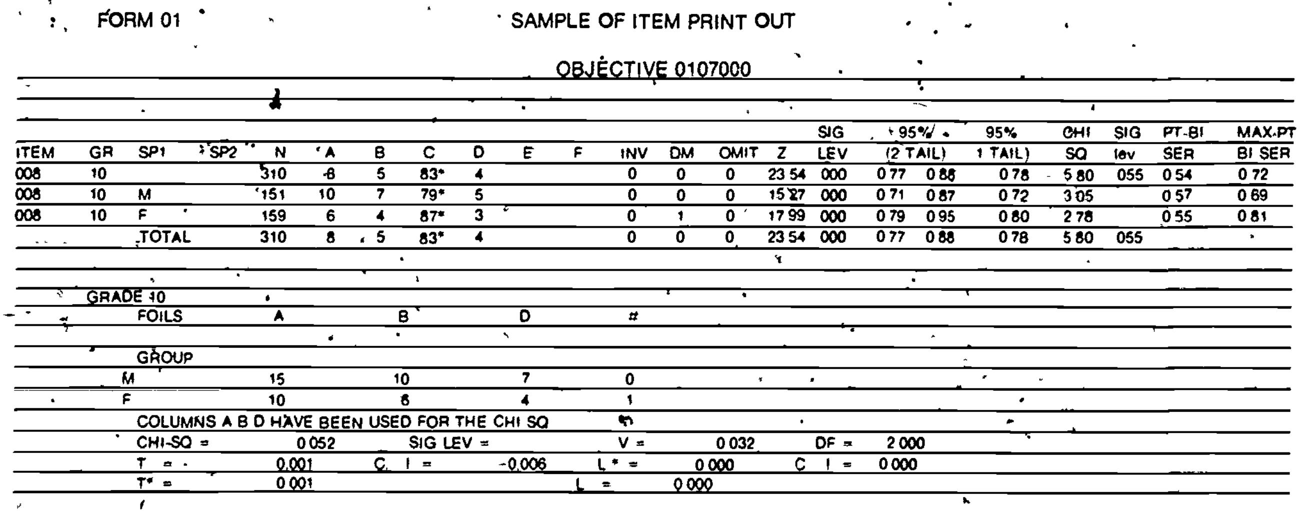
Even though there are sound statistical reasons for eliminating this item from the instrument, it may be argued that it is preferable to retain the item and use the diagnostic information to provide guidelines for instruction. The middle ground between throwing out the item and keeping it as it stands is to revise it. Perhaps an improved correct response (a more positive, constructive, creative idea for participating in the economic system) would help to reduce the cultural bias.

Example 3. (Item 8, Booklet 11, both grades).

Item: Graduation is coming soon. You have no idea of what you want to do when you leave school. You are fearful about your future and have stayed awake at night trying to decide what to do.

Below are actions that you might take in an effort to solve your problem. Identify the action that is *LEAST* helpful by darkening the appropriate letter on your Answer Sheet.

(A) talk with the school counselor
(B) write to universities, community colleges and trade schools to learn about opportunities
(C) find out what your best friend to going to do
(D) get information and advice from the local state employment office



FORM 01 SAMPLE OF ITEM PRINT OUT

OBJECTIVE 0107000

															SIG	95%		95%	CHI	SIG	PT-BI	MAX-PT
ITEM	GR	SP1	SP2	N	A	B	C	D	E	F	INV	DM	OMIT	Z	LEV	(2 TAIL)		1 TAIL)	SQ	lev	SER	BI SER
008	10			310	8	5	83*	4			0	0	0	23 54	000	0 77	0 88	0 78	5 80	055	0 54	0 72
008	10	M		151	10	7	79*	5			0	0	0	15 27	000	0 71	0 87	0 72	3 05		0 57	0 69
008	10	F		159	6	4	87*	3			0	1	0	17 99	000	0 79	0 95	0 80	2 78		0 55	0 81
		TOTAL		310	8	5	83*	4			0	0	0	23 54	000	0 77	0 88	0 78	5 80	055		

```
GRADE 10
        FOILS            A              B              D           #

        GROUP
      M                  15             10             7           0
      F                  10             6              4           1
        COLUMNS A B D HAVE BEEN USED FOR THE CHI SQ
        CHI-SQ =           0 052        SIG LEV =                  V =          0 032        DF =     2 000
        T   =              0.001        C. I =         -0.006      L * =        0 000        C  I =   0 000
        T* =               0 001                                   L  =        0 000
```

This item is working well according to *all* criteria. The p-value is significantly above chance, the foils are drawing uniformly, the point biserial is fairly high (.54) and all the cultural (both ethnic and sex) validity indices are low. A statistically sound item.

CHAPTER V

SENSITIVITY TO INSTRUCTION

Introduction

An important element of the item tryout program required the utilization of the WLC/MRC test items with a special group of students who had received instruction specifically designed to develop the behavior described by a selected number of the learner outcomes. This particular phase of the item tryouts — referred to as in-depth tryouts — was expected to provide information for determining whether the test items measured a dimension of knowledge that was sensitive to instruction. To accomplish this phase of the tryouts, the PARTNERS project was committed to the preparation of learning modules which directly addressed elements of forty-four of the seventy-nine basic outcomes for which test items were being prepared. Modules were to be prepared for students in the eighth and eleventh-grades in various subject areas.

Assumptions

The decision to conduct a study of the sensitivity to instruction of the WLC/MRC developed test items was based in part upon the following assumptions:

- Criterion-referenced test items should measure student development in terms of clearly stated objectives.
- Criterion-referenced test items should reflect changes which may take place in student capability with regard to objective attainment.
- The behaviors described by the WLC/MRC prepared objectives were elements of the basic learner outcomes and could be developed in students within the classroom.
- Learning modules could be developed that were adequate for the identified objectives and appropriate for the students to be instructed.

Procedures

The theories and procedures suggested by Roudabush (1973) provided the basis for this study. The statistical analyses proposed by Kosecoff and Klein (1974) were among those applied to the data developed. The following procedures utilized in conducting this study are presented in the approximate sequential order of occurrence.

- WLC/MRC behavioral objectives, derived from the basic learner outcomes, were selected which were believed to be amenable to instruction within a relatively short period of time.
- Schools were identified and teachers (classrooms) were selected to function as experimental groups. These groups of students were pretested, instructed, and posttested utilizing WLC/MRC test items. Participating teachers were volunteers.
- Personal interviews were conducted with participating teachers to identify the curriculum in use and the resources appropriate for infusion of necessary new material.
- Resources such as books, curriculum guides, etc., were obtained for the development of infused learning activities.
- Schools and classrooms were identified to function as comparison groups. Teachers in the comparison group classrooms were also volunteers. Students were not exposed to material contained in PARTNERS special curriculum modules.
- Learning modules were prepared to infuse the selected career education concepts into the ongoing curriculum.
- The learning modules were submitted to participating teachers for review and critical comment.
- An evaluation form was prepared to obtain teacher reactions to the modules.
- Career education test items (mini-tests), answer sheets, and scoring sheets were prepared by WLC/MRC.
- A manual was also provided by WLC/MRC to guide teachers in the administration of the mini-tests.
- Testing materials, learning modules, and curriculum resource materials were delivered to and collected from the teachers participating in the study.
- Students' answer sheets were scored and the data statistically analyzed by WLC/MRC.
- Teacher evaluation data were complied for use within the project.

Selection of Outcomes/Objectives

In the selection of basic learner outcomes and derived WLC/MRC behavioral objectives, toward which learning modules would be prepared, several factors were considered. First, some of the outcomes which describe attitudinal behavior were identified as not being amenable to instruction over the relatively short time span available. Second, those outcomes which had been identified previously as being more appropriately introduced and emphasized in the lower grades tended to be eliminated as inappropriate for instruction in the eighth and eleventh-grades. Finally, outcomes were selected for the tryout program which, in the judgment of the professional staff, could be at least partially (measurably) developed during the period allocated, i.e., approximately ten weeks. After screening the total number of outcomes for which test items were being developed, 52 objectives (elements of 44 outcomes) were selected for this in-depth item tryout study.

Selection of Schools and Teachers

Two factors of primary concern in the selection of schools for this study were the degree of willingness to participate displayed by the individuals contacted and the geographic location of the schools concerned. The appearance of a reluctant attitude on the part of either administrators or teachers was considered to be grounds for the non-selection of particular schools. Volunteers were sought who would accept the necessary curriculum and schedule modifications which would result from use of the specified learner activities and student testing. With regard to geographic location, the anticipated need for frequent visits to the participating schools by PARTNERS staff members inhibited the consideration of schools more than one and one-half hours driving time from Arlington. Other considerations involved school size and the ethnic composition of the student body in grades eight and eleven. Because of the noted restrictions to school selection the inclusion of a proportionate number of students from each ethnic group was not possible. However, the desirability of obtaining responses from each of the three major ethnic groups — anglo, black and Mexican-American — was recognized and was a consideration in school selection. School size was also important in that small classrooms would have required the participation of an unacceptably large number of teachers to assure that a minimum number of students responded to each test item. Following consultation with WLC/MRC personnel, this minimum was determined to be 50 students.

By applying the foregoing general criteria 33 schools in sixteen school districts were identified. No difficulties were experienced in obtaining the approval of administrators in any district or school contacted. One hundred thirty-eight teachers in the 33 schools volunteered to participate. The expressed desire to become involved in this aspect of the PARTNERS program by all of the administrators and a large majority of the teachers contacted was particularly gratifying.

Experimental and Comparison Groups

The study design required students in each of the classrooms participating to function in a dual capacity, as members of both experimental and control groups. For example, an experimental class was pretested, instructed and posttested utilizing appropriate test items. The same class also functioned as a control for another experimental group by being pre- and posttested utilizing test items unrelated to the instructional material to which the class had been exposed. This methodology was feasible because the association between items written for different learner outcomes is weak to non-existent. In addition, the total number of students and classrooms required for the study was reduced by approximately 50% by utilizing this particular technique.

Statistical Procedures

In criterion-referenced testing strong emphasis is placed on the effectiveness of test items to discriminate between those students who have profited from instruction and those students who have not. Three types of indices were used in this study to determine "sensitivity-to-instruction."

- The Internal Sensitivity Index (ISI) measures item quality from the perspective of the total test's discriminating power.
- The External Sensitivity Index (ESI) and the Roudabush "S" measures an individual item's ability to reflect learning (independent of the test).
- The Objective Sensitivity Index (OSI) measures the total test's ability to discriminate between learners and nonlearners.

This study utilized experimental and comparison groups for each test with both groups receiving the pre- and posttests and the experimental group receiving instruction. A Z-test was utilized to detect statistically significant differences between the indices reported for the experimental and the comparison groups. (See Appendix I.)

The Internal Sensitivity Index (ISI) is computed as follows:

$$\text{ISI} = \frac{n_2 - n_1,}{n}$$

where n_1 is the observed frequency of students who answered item i correctly on the posttest but failed the pre- and posttest, n_2 is the observed frequency of students who answered item i correctly on the posttest but failed the pretest and passed the posttest, and n is the total number of respondents who correctly answered item i.

The External Sensitivity Index (ESI) is computed as follows:

$$\text{ESI} = \frac{m_2 - m_1,}{m}$$

where m_1 is the observed frequency of respondents who missed item i on the pretest and posttest, m_2 is the observed frequency of respondents who missed item i on the pretest but responded correctly on the posttest, and m is the total number of respondents.

The Objective Sensitivity Indes (OSI) is computed as follows:

$$\text{OSI} = \frac{N_2 - N_1,}{N}$$

where N_1 is the number of respondents who failed the pretest and the posttest, N_2 is the number of respondents who failed the pretest but passed the posttest, and N is the total number of respondents.

The Roudabush "S" is an index of the degree to which examinees are selecting the correct response to the item as a function of the instruction received between pre- and posttest, that is, a sensitivity index. This index is simply the proportion of cases that missed the item on the pretest and then answered it correctly on the posttest after a correction for guessing had been applied.

The values for each index range from -1 to +1. A score of -1 would occur when no one learned. Such a result suggests that either instruction failed to benefit any of the students, or, more realistically, that the item fails to discriminate among learners. A score of +1 is obtained when all students miss an item on the pretest and correctly answer it on the posttest. This is the ideal situation, the item shows maximum change in the direction of learning. Any scores on the pass-fail and pass-pass cells will lower the absolute values of the indices.

The difference in the proportion of gainers (those passing the posttest and failing the pretest) out of the total number of potential gainers (those who failed the pretest) for the experimental and comparison groups was also computed. A Z-test of significance was conducted if the number who failed the pretest was large (greater than 20). For small samples, Fisher's "exact test" was used (cf., Snedecor and Cochran, 1967). Similiar tests were conducted on the proportion of gainers (experimental vs. control) *for each item.* Specifically Z or Fisher's tests were conducted using the difference in the proportions of (1) those passing the posttest among those failing the pretest and correctly answering the item, and (2) those correctly answering the items on the posttest among those missing the item on the pretest.

A sample page of printout from the sensitivity-to-instruction analysis conducted by WLC/MRC statisticians is given in the following table.

TABLE WLC/MRC Sensitivity to Instruction Sample Printout

ANALYSIS FOR > = B

FORM NUMBER 02

ITEM	GROUP	N	N1	N2	ISI	P2	Z1	SIGL	Z2	SIGL	N3
1	EXP	39	20	12	-0.21	0.38	2.15	.016	2.47	.007	1
1	CON	27	17	1	-0.59	0.06					4
2	EXP	25	10	11	0.04	0.52	1.17		1.31	.096	0
2	CON	11	4	1	-0.27	0.20					1
3	EXP	33	14	12	-0.06	0.46	3.10	.001	3.07	.001	1
3	CON	28	19	1	-0.64	0.05					3
4	EXP	10	5	2	-0.30	0.29	0.96				0
4	CON	12	7	0	-0.58	0.00					2
5	EXP	26	17	12	-0.14	0.41	1.78	.038	1.82	.035	1
5	CON	16	9	1	-0.53	0.10					2

GROUP	CAP N	CAP N1	CAP N2	OSI	CAPN3
EXP	46	27	12	-0.33	1
CON	37	26	1	-0.68	5

Z1 = 2.26 SIG. LEV = .012 Z2 = 2.72 SIG. LEV. = .003

M	M1	M2	ESI	P2	Z1	SIGL	Z2	SIGL	S	ESI*	M3
46	5	15	0.22	0.75	2.26	.012	1.87	.031	0.84	0.18	2
37	5	3	-0.05	0.38					-0.19	-0.65	5
46	19	17	-0.04	0.47	2.25	.012	2.43	.098	0.55	-0.14	2
37	20	4	-0.43	0.17					-0.07	-.098	6
46	11	16	0.11	0.59	0.74		0.54		0.69	0.03	2
37	6	6	0.00	0.50					0.30	-0.24	3
46	35	5	-0.65	0.13	-1.14		0.00		0.10	-0.72	1
37	21	3	-0.49	0.13					-0.03	-0.84	4
46	8	14	0.13	0.64	2.49	.006	2.47	.007	0.72	0.05	2
37	11	3	-0.22	0.21					-0.30	-1.62	11

The significance tests (for differences in ISI, ESI, and proportion of gainers) were very useful since the indices are quite new and little is known about what constitutes a "good" value (of ISI, ESI, etc.). For example, if an index is high, say greater than 0.75, and significantly *higher* than for the comparison group, it can be inferred that the item is really sensitive to instruction. Contrariwise, if there is no significant difference and both values of the index are either high, medium, or low, it cannot be inferred that the item is sensitive to instruction. It may be sensitive to some other (apparently common) factor, which is not related to instruction. In the case where the indices are low or negative, and the difference is significant, the interpretation is questionable. The interpretation of the test for difference in proportion of gainers was more straight-forward since a significant difference in these tests indicated that the item manifested a real difference in gain between those who had been given instruction and those who had not. The mastery level established for passing or failing was varied for this study. These indices were computed by WLC/MRC for the 50%, 70%, 80%, and 90% levels.

The participating teachers were asked to rate their students as follows. students who generally earned grades equal to or greater than B, and students who generally earn grades less than B. All indicators of sensitivity to instruction were computed for the total sample and for these two sub-groups. (The analysis given in the sample printout is for the "B or above" group.)

Results of the Study

Sufficient data for analysis purposes were received on items addressed to 51 of the 52 objectives selected for the study. The 51 separate tests were composed of 111 items, many of which required multiple responses. The computer treated each of the separate responses as individual items. This resulted in a total item count of 215.

The Internal Sensitivity Index (ISI) and Objective Sensitivity Index (OSI) are both dependent upon established mastery levels to determine the number and percent of students passing and/or failing the tests. Testing results were analyzed at various mastery levels from 50% to 90%. These levels represent the percent of the total number of items written for a single objective which a student must answer correctly to achieve mastery of the objective. As the mastery level criteria was lowered, the values of both the ISI and the OSI tended to increase. However, a lower mastery level — say 50% as opposed to 80% — resulted in more students passing the pretest. This caused the indices to reflect learning for a smaller percent of the sample. With the mastery level established at 80% a higher percentage of the students failed the pretest thereby increasing the number who might profit from instruction and providing a more reliable indicator of sensitivity.

The Internal Sensitivity Index (ISI) measures item quality from the perspective of the total test's ability to discriminate between mastery and non-mastery of the objectives. One hundred two items were found to have a positive ISI score at the 80% mastery level. The Z-test for ISI yielded questionable results, since many of the statistically significant ISIs were negative or quite low. Using the test for difference in proportion of gainers, it was found that at the 80% mastery level, twelve items were found to be significantly different at the .10 level, eight at the .05 level; and sixteen .01 level.

The External Sensitivity Index (ESI) measures an individual item's ability to reflect learning. One hundred one items were found to have positive ESI scores. Using the test for difference in proportion of gainers (on the items), four items showed a statistically significant difference between the experimental and the comparison groups at the .10 level; and fourteen items were significant at the .01 level.

The Roudabush S'' is a measure of an item's sensitivity and includes a correction for guessing. Roudabush found that at least 50 cases are needed to establish a reliable index; i.e., at least 50 students who fail the pretest should be instructed and subsequently posttested. Nineteen items in this study met this criteria and thirteen of these items had a positive index.

The Objective Sensitivity Index (OSI) measures the total test's (for an objective) ability to discriminate between learners and non-learners. Nine objectives had a positive OSI score at the 80% mastery level. Using the test for difference in proportion of gainers (on the tests), two objectives were found to be significant at the .10 level, five objectives were significant at the .05 level, and six were significant at the .01 level.

When the total sample was divided into two groups by grade average , A or B students and C or poorer students according to teacher ratings, analysis of comparative data yielded the predictable results. The students rated B or above yielded higher sensitivity indices than those rated below B. For example, at the 80% mastery level eighteen objectives show a positive OSI score for A or B rated students and nine objectives showed a positive OSI score for those rated below B.

Limitations of the Sensitivity to Instruction Study

Several factors combine to severely limit the usefulness of the data collected for this study. First among these is the item/objective/outcome relationship which existed when the sensitivity to instruction study was initiated. Final review and acceptance of the objectives and related test items prepared by WLC/MRC had not been completed by TEA or by the PARTNERS project prior to printing of the mini-tests to be used in the study. Subsequent joint review of the objectives and items by the parties concerned (WLC/MRC, TEA, and PARTNERS) resulted in the elimination of approximately 25% of the items developed by WLC/MRC to that time. In addition, major revisions and format changes were made to more than half of the remaining items. These revisions or changes were based upon the professional judgment of the three parties participating in the review. The items and objectives eliminated or revised did not adequately address the elements of the basic learner outcomes for which they had been prepared. The ultimate result of the changes made was to reduce by approximately 65% the number of viable items used in this study.

A second factor limiting the usefulness of the data relates to the quality of the test items available. Prior to this study the test items utilized had not been pilot-tested or tried out in any fashion with students. There was therefore no information available with regard to the readability, understandability or appropriateness of the test items in a testing environment. (The test items had been reviewed by educators, and by students at the junior and senior levels as single items but not in a test context format.) Erratic student responses, characterized by unsymmetrical foil distribution patterns for many items, in both the control and experimental groups, are believed to be directly related to this factor. In addition, a large number of items were correctly answered on the pretest by very high percentages of the students. For example, 84 items were correctly answered on both the pre- and posttests by 80% or more of the students participating. An adequate tryout or a pilot testing program would have identified many of the test items as being too easy for eighth and/or eleventh-grade students.

A third factor, which is attributable in part to the fact that the items had not been previously tried out, also tends to limit the value of the data collected. This involves the number of students included in the study. Roudabush found that for a reliable sensitivity index to be computed (Roudabush "S") the number of students failing the pretest (and therefore requiring instruction) should be at least 50. In many instances, fewer than a dozen of the students participating in this study failed to pass the pretest. In fact, only nineteen of the 215 items utilized met the criteria established by Roudabush. This situation could be avoided in the future by conducting an adequate tryout or pilot-test to eliminate inappropriate items prior to a study of this type.

A fourth factor, was the question of instruction to objectives. The extent to which the quality and effectiveness of instruction varied across objectives directly influences the sensitivity indices. The variability of instruction presents a confounding variable which disturbs the comparability of the sensitivity indices across objectives.

CHAPTER VI

SYSTEMS FOR REPORTING FIELD TEST RESULTS TO TEACHERS

Introduction

The ultimate success or failure of the measurement system will depend largely upon the usefulness of the information that the tests generate. Thus, it is essential that test data reported to the potential users of the information be written so that it can be easily understood. The systems used for reporting the results of the March field tests to students and school personnel were of a developmental nature, and criticism from those receiving the test results was encouraged.

The purpose for reporting the test results is to provide students and school personnel diagnostic information about student performance in terms of the behaviors described by the learner outcomes. Two types of reports were used. (1) a modified version of the SCORE (WLC/MRC) student report and (2) a TEA-devised report.

WLC/MRC Format

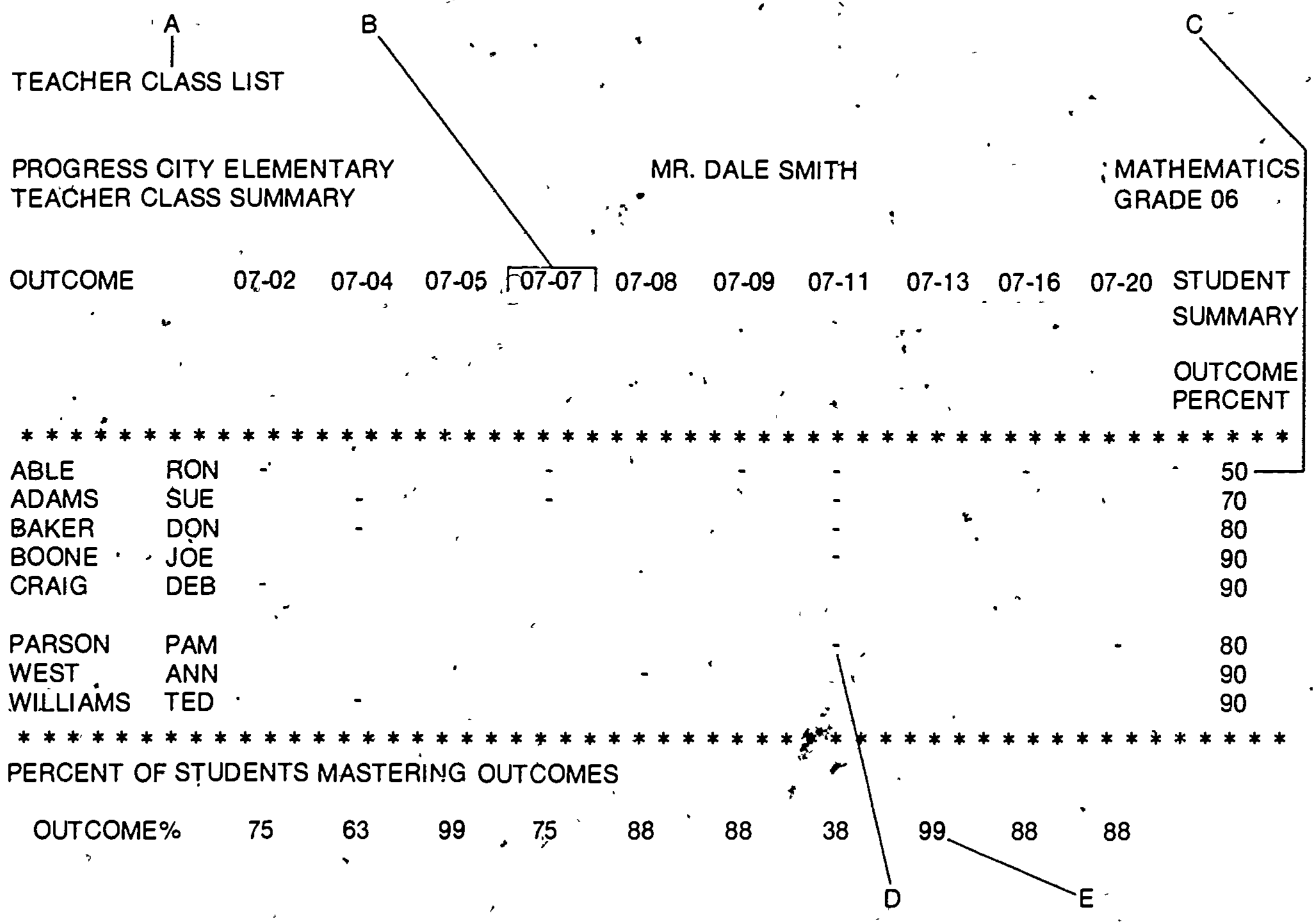
The modified SCORE report contains information on (1) whether each student mastered each outcome, (2) the percent of *outcomes* mastered by each student, and (3) the percent of *students* mastering each outcome. A 50% mastery level was used, i.e., a student must have correctly answered at least *half* of the items measuring an outcome to be classified as having "mastered" the outcome. The 50% level was used, rather than a higher, more stringent level, since no instruction toward the learner outcomes was assumed. A sample report (Westinghouse Learning Corporation SCORE Class List) is provided in Table 1. The outcome "legend;" i.e., the numerical outcome codes with the corresponding outcome descriptions, is provided (for test booklet 11) in Table 2.

TEA Format

The TEA report format contains concise statements reflecting the degree of outcome mastery rather than the mastery/nonmastery format used in the SCORE system report. An individual report is provided for each member of the class which indicates his or her performance on the test. A copy of a TEA style report (for test booklet 52) is given in Table 3.

The teachers were asked to evaluate the two types of systems. The SCORE format was favored, although the response to the questionnarie was spotty due to the fact that it was sent out rather late in the school year.

SCORE REPORT — TABLE 1



TEACHER CLASS LIST

PROGRESS CITY ELEMENTARY
TEACHER CLASS SUMMARY

MR. DALE SMITH

MATHEMATICS
GRADE 06

OUTCOME		07-02	07-04	07-05	07-07	07-08	07-09	07-11	07-13	07-16	07-20	STUDENT SUMMARY OUTCOME PERCENT
ABLE	RON	-			-		-	-		-		50
ADAMS	SUE		-		-			-				70
BAKER	DON		-					-				80
BOONE	JOE							-				90
CRAIG	DEB	-										90
PARSON	PAM							-			-	80
WEST	ANN					-						90
WILLIAMS	TED		-									90

PERCENT OF STUDENTS MASTERING OUTCOMES

OUTCOME%	75	63	99	75	88	88	38	99	88	88

A The class list is a performance record for each student in a teacher's class for each outcome tested.

B Numeric representation of the outcomes as listed in the teacher's outcome legend.

C Percent of outcomes mastered by each student.

D Interpretation of the outcome mastery is as follows. if a minus appears under an outcome, the student has not mastered that outcome. A blank designates mastery of the outcome.

E The percent of the class mastering each outcome is also summarized.

Modified SCORE Report
Table 1.

OUTCOME LEGEND - TEST 11 - TABLE 2

01-03:

The student should understand the necessity for having a satisfying job when setting his career goal.

01-04:

The student should understand that he will work better when he accurately matches his personal goals with his career choice.

01-05:

The student should be able to identify career directions which are available to him.

01-07:

The student should be able to use his/her own resourcefulness to solve personal problems such as. He wants to go to college, but there is not enough money for tuition. He could look for a job, put in a request for financial aid, or apply for a loan.

TEA REPORT - Table 3 -

STUDENT'S NAME ______________________ TEACHER'S NAME ______________________

TEST BOOKLET 52

CATEGORY V: Skills in Human Relationships

STUDENT OUTCOME: The student should be able to understand that some means of communication work more effectively in some situations than others.

A. HAS DEMONSTRATED THE ABILITY TO SELECT THE MOST EFFECTIVE MEANS OF COMMUNICATION	HAS DEMONSTRATED THE ABILITY TO SELECT SOME EFFECTIVE MEANS OF COMMUNICATION	DOES NOT APPEAR TO HAVE HAD EXPERIENCES IN SELECTING EFFECTIVE MEANS OF COMMUNICATION
XXX	______	______

STUDENT OUTCOME: The student should be able to understand that there will be many instances in his life when he will have to to make compromises.

A. RECOGNIZES THAT THERE ARE TIMES WHEN COMPROMISES ARE NECESSARY	RECOGNIZES SOME INSTANCES WHEN COMPROMISE IS NECESSARY	DOES NOT RECOGNIZE INSTANCES WHERE NEED FOR COMPROMISE IS NECESSARY
XXX	______	______

STUDENT OUTCOME: The student should be able to give examples of the advantages and disadvantages of being a leader and/or follower.

A. IS ABLE TO GIVE EXAMPLES OF THE ADVANTAGES OF BEING A LEADER AND/OR A FOLLOWER	IS ABLE TO GIVE SOME EXAMPLES OF THE ADVANTAGES OF BEING A LEADER OR FOLLOWER	DOES NOT RECOGNIZE THE ADVANTAGES OF BEING A LEADER OR FOLLOWER
XXX	______	______

B. THE STUDENT MADE 26 POSITIVE RESPONSES OUT OF 33 POSSIBLE RESPONSES TO STATEMENTS RELATING TO A DISPLAY OF RESPECT FOR PEOPLE OF DIFFERENT RACES OR ETHNIC ORIGINS

C. HAS INDICATED AN INSTANCE IN WHICH PEOPLE HAVE BEEN TREATED UNFAIRLY BECAUSE OF THEIR RACE	HAS NOT INDICATED AN INSTANCE IN WHICH PEOPLE HAVE BEEN TREATED UNFAIRLY BECAUSE OF THEIR RACE
XXX	

STUDENT OUTCOME: The student should understand that there are other individuals with whom he as a worker must interact.

A. HAS IDENTIFIED SITUATIONS WHEN IT IS NECESSARY TO INTERACT WITH OTHERS	HAS IDENTIFIED SOME SITUATIONS IN WHICH IT IS NECESSARY TO INTERACT WITH OTHERS	DOES NOT APPEAR TO HAVE HAD SUFFICIENT EXPERIENCES TO IDENTIFY THE NEED FOR INTERACTION WITH OTHERS
XXX		

STUDENT OUTCOME: The student should understand the benefits of and necessity for being sensitive to others.

A. ABLE TO SELECT SITUATIONS THAT SHOW THE BENEFITS AND NECESSITY FOR BEING SENSITIVE TO OTHERS	ABLE TO SELECT SOME SITUATIONS THAT SHOW THE BENEFITS AND NECESSITY FOR BEING SENSITIVE TO OTHERS	DID NOT INDICATE THE ABILITY TO SELECT SITUATIONS THAT SHOW THE BENEFITS AND NECESSITY FOR BEING SENSITIVE TO OTHERS
XXX		

TEA REPORT - Table 3 -

STUDENT OUTCOME: The student should be able to understand how attitudes based upon prejudice affect behavior of other individuals, for example: if he feels all blacks are inferior, blacks may sense this and become hostile toward him.

A. DID SELECT SITUATIONS THAT SHOW HOW ATTITUDES BASED UPON PREJUDICE AFFECT THE BEHAVIOR OF OTHERS	DID SELECT SOME SITUATIONS THAT SHOW HOW ATTITUDES BASED UPON PREJUDICE AFFECT THE BEHAVIOR OF OTHERS	DID NOT SELECT SITUATIONS THAT SHOW HOW PREJUDICES AFFECT THE BEHAVIOR OF OTHERS
______	XXX	______

CHAPTER VII

STATISTICAL PROCEDURES FOR DEVELOPMENT OF THE SURVEY INSTRUMENT

A survey or diagnostic instrument comprising about 45 items was developed to be used at the eighth-grade level. The purpose of this test is to diagnose further measurement of student performance with one or more of the sixteen category tests. The category tests would then prescribe instructional strategies.

A stepwise regression procedure (cf., Draper and Smith, 1966) was employed to select one or (at most) two items which correlate highly with the "outcome" scores. The dependent variable in this framework is the outcome score, and the independent variables comprise (1) the "scores" on each item within the outcome (0 = wrong, 1 = correct) and (2) a control variable to indicate, and thus control for, the grade tested (upper or lower). The data may be fitted to the following regression equation:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + Bp\, Xp + e,$$

where B_0 is the Y-intercept, B_1 is the regression coefficient for the 'control' variable, $B_i + 1$ is the coefficient corresponding to the ith item "score" ($i = 1, 2 \cdots, p$) and,

$$X_1 = \begin{cases} 0 \text{ if the student is in lower grade} \\ 1 \text{ if the student is in upper grade} \end{cases}$$

$$X_i + 1 = \begin{cases} 0 \text{ the student answers item i incorrectly} \\ 1 \text{ if the student answers item i correctly} \end{cases}$$

The variable X_1 was always included. The other variables (no more than 2) were selected in a stepwise manner as follows:

1. The variable with highest partial correlation with Y (holding X_1 fixed) is selected.
2. The variable with highest partial correlation, holding the item selected in step 1 fixed, is selected.

Tests of statistical significance for each item entered were conducted. They were all highly significant due to the large number of subjects.

The decision was made to use two other criteria. (1) addition to R^2, the multiple correlation coefficient, and (2) "beta weight" times the corresponding zero-order correlation or point biserial (cf., Draper and Smith, 1966). These procedures yielded more or less the same results.

The outcomes were grouped (using subjective judgement) into subcategories or "clusters". If performance on outcomes can be predicted with reasonably high R^2 (say .3 and above) then one would expect that summing over outcomes within a "part" would give even better predictability on the sub-categories. Due to the practical constraints regarding test length, a few sub-categories are estimated by only one item.

CHAPTER VIII

IMPLICATIONS

Introduction

From the beginning, the Texas Career Education Measurement Series project has been visualized and conducted as a developmental effort. The building of objective-based measure for this project has included many types of procedures that either have been developed by others in the recent past or have been designed for this project. Some of the steps taken have followed the precedents for test development while other procedures have not followed the traditional mode. The purpose of this chapter is to assist those who are either contemplating or conducting test development efforts similar to this one by discussing some of the implications for test development.

Implementation of the Study

Basing a measurement system on learner outcomes that have been developed from the perceptions of students, educators, and those outside of the field of education brings credibility to the development of an objective-based test. There is evidence of less difficulty in obtaining assistance from schools. Early planning with schools is still necessary to assure timely field tests and item tryouts.

Development of test instruments should not be undertaken in an objective-based system until the objectives are organized and written in appropriate form. Specification of the behavior domains to be measured are a prerequisite to selection/development of items to measure those behaviors.

Planning of the system for reporting results from the measurement instruments should begin with the initial development procedures. The reporting of results should become an important guide to the types of items developed. If the "how to report" frame of reference is ignored, one result can be that after items are written, it is discovered that the results cannot be reported in a useful manner.

Item Development

The development of items for an area such as career education which does not have an organized group of professionals who represent that discipline requires special attention in the item development phase. For example,

- Item writing is particularly difficult — even for professional item writers.
- If local school personnel are to be involved in item development, sufficient preparation for the task must be provided.

Contributions from local school personnel can be obtained more effectively if item writing is conducted away from their regular duties. Time should be set aside for them to work without conflict with their daily routine.

In writing items for objective-based instruments, there should be a large number of items written in order to have sufficient coverage of objectives in the final instruments. Although item attrition for objective-based measures may occur for different reasons than for norm-referenced tests, one should expect to reject 30% to 50% of the items during the development and review processes.

Sensitivity-to-instruction is an important concept for objective-based measures. A study of this type should be conducted after items have been validated for a given set of objectives in order to avoid interaction of two dependent variables — quality of instruction versus item validity.

Item Review and Revision

Student review of items is very productive. If the students perceive that their input is important and will be used, they will furnish useful information about items. Items should be discussed with a small group of students (3-5) of the appropriate age. A student sampling plan should be devised to ensure that each item will be reviewed by students of each ethnic, sex, geographical, etc. sub-population.

Continued revision of bad items soon becomes inefficient. If an item is unacceptable after two revisions, that item should be discarded and a new one developed for the objective.

Analysis of Data

Significant advances have been made in the kinds of statistical analyses that are available for item and test construction in an objective-based measurement system. Further testing of these procedures will provide evidence of their usefulness for other test developers. The procedures presented in Chapter IV are primarily useful for items of a multiple choice format and do provide additional information for decision-making about items. However, as the amount and types of information about items increase, additional attention must be given to the "decision model for item acceptance". The relative weight to be given the results from two or more statistical procedures requires additional investigation.

General Implications

There is evidence from this project that a state department of education, a regionally-based project, and a commercial contractor can function together to develop new measurement instruments. Although special attention must be given to communication between the three organizations, serendipities of the following type can result:

- A cadre of people at the regional level and the state level can obtain experience in test development. This expertise is beneficial for future development and revision of objective-based measures.
- A commercial contractor can gain in knowledge of local, regional, and state educational policies and procedures. In addition, a large number of students and teachers can be involved in item tryout and revision procedures at a reduced cost to the contractor.
- An increased level of awareness is developed throughout the schools and regions from participating in the development process.
- Positive results are obtained from the involvement of personnel from several areas of specialization — special education, vocational education, curriculum, guidance, measurement, etc.

When procedures are designed for local school participation in a developmental project, management plans must take into consideration the local school calendar in order to provide sufficient time to schedule project activities around school holidays.

The procedures used to develop this measurement system imply that the career education tests are now in their "first version". Objective-based measurement must be in a continuous state of refinement to retain relevancy to priority objectives. Future administrations of the 16 category tests and the survey test will provide additional student data upon which the system will be tested and refined.

REFERENCES

Campbell, D., and Stanley, J., *Experimental and Quasi-Experimental Designs for Research,* Rand McNally & Co., Chicago, 1963.

Cochran, W., *Sampling Techniques,* John Wiley & Sons, Inc., New York, 1963.

Draper, N., and Smith, H., *Applied Regression Analysis*, John Wiley & Sons, Inc., New York, 1966.

Goodman, L., and Kruskal, W., "Measures of Association for Cross Classification," *Journal of the American Statistical Association*, Vol. 49, pp. 732-764, 1954.

Kosecoff, J., and Klein, S., "Instructional Sensitivity Statistics Appropriate for Objective-Based Test Items," paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April, 1974.

Lord, F., and Novick, M., *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, Mass., 1968.

Magnusson, D., *Test Theory*, Addison-Wesley, Reading, Mass., 1967.

Roudabush, "Item Selection for Criterion-Referenced Tests," a paper presented at the American Educational Research Association meetings in New Orleans, February, 1973.

Snedecor, G., and Cochran, W., *Statistical Methods* (6th Edition), Iowa State University Press, Ames, Iowa, 1967.

Veale, J., and Foreman, D., "Cultural Validity of Items and Tests. A New Approach," SCORE technical report, Westinghouse Learning Corporation, Iowa City, Iowa, 1975.

Wilks, S., *Mathematical Statistics*, John Wiley & Sons, Inc., 1962.

APPENDIX A

Student Review of
Career Education Items

Recorder ______________ Date ______________

Item # ______________________________

Campus ______________________________

District ______________________________

	B	M	A
M			
F			

Sections I and III are to be completed for all items in the package. In Section II, complete only that portion appropriate to the format (multiple-choice, open-ended, etc.) of each particular item of the package.

SECTION I

1. Relationship to objective. *Does the item get at the objective? ________ Could the relationship be made more direct? ________ If yes, how?

2. Credibility. Is the response to the item likely to reflect what the student considers to be the truth — or would the item lead the student toward giving an "expected" or "socially acceptable" response?

3. Bias/Offensiveness: *Is there anything offensive about the item? ________ If yes, what?

   *Might the item be unfair to students of a particular race or sex? ________ How?

4. Understandability: *Was there any trouble understanding the item or the directions? Yes ________ No ________ If yes, what caused the trouble?

5. Appropriateness. Underline the phrase which best describes how the students felt about the content of the item. too Mickey Mouse, too advanced, unrelated to student interests, dated, interesting and appropriate, other (specify).

SECTION II

(complete only the entry compatible with the item being reviewed)

1. Multiple-choice item with one or more choices designated as "correct":
   Do you agree that the "correct" response(s) is (are) indeed correct? ______ If no, why?

   Are some of the other responses defensible as being correct? ______ If yes, which ones?

   Do you think any smart student could, regardless of whether he had mastered the objective, be able to eliminate some of the response choices? ______ If yes, which ones?

2. Multiple-choice item with no response designated as "correct":
   Are there enough response choices that each student could express his feeling? ______ If no, which choices should be added?

3. Matching item:
   Do you feel that some of the matching pairs would fail to give any information as to whether the student had mastered the objective? ______ If so, which pairs?

4. Checklist:
   Would the person who is supposed to complete the checklist be able to do so without an excessive amount of effort?

5. Open-ended item:
   Is the scoring guide clear? ______ Are the responses to the item likely to provide the information sought? ______

6. Individually administered items:
   Do you see any way that 2 or 3 group administered items (such as multiple-choice or matching) could get the same information? ______

SECTION III

Comments concerning item attributes not mentioned above:

General evaluation of item potential - Excellent _____ Good _____ Fair _____ Poor _____
Comments:

Suggestions for revisions: (where possible, enter onto item)

Does this item need additional review? _____ Why?

APPENDIX B

Teacher/Counselor Review
of Career Education Items

Recorder ____________ Region ____________

Reviewers: Specialty:

____________ ____________

____________ ____________

____________ ____________

Item # ____________ Date ____________

Sections I and III are to be completed for all items in the package. In Section II, complete only that portion appropriate to the format (multiple-choice, open-ended, etc.) of each particular item of the package.

SECTION I

1. Relationship to objective. *Is the relationship between the objective and what is measured by the item acceptably close? YES ______ NO ______ How could the item be changed so as to bring about a closer relationship between the objective and the item?

2. Credibility. Is the item likely to obtain a true picture of the student's knowledge, feelings, or plans (as distinguished from an "expected" or "socially acceptable" response)? ________ If no, why?

3. Bias/offensiveness. *Is the item biased against or likely to be offensive to students of a particular race, sex, geographic location, size and/or type of community, or socio-economic status? YES ________ NO ________ If yes, indicate the nature of the difficulty and, if possible, how the bias or offensiveness might be reduced.

4. Understandability. Which words, if any, would be likely to cause difficulty among students at the sixth grade reading level?

   Is the sentence structure easy to follow?

   *Would the item and its directions be understandable by 90% of 8th grade Texas students? YES ______ NO ______ How could the item or its directions be improved?

5. Appropriateness: Is the item appropriate for grade level 8? ________ 11? ________ If no, why?

6. Usefulness. Does the item provide information useful in identifying the students instructional needs? ________ If no, could the item be changed to do so? ________ How?

SECTION II

(identify and complete only the entry appropriate for the item being reviewed)

1. Multiple-choice item with one or more choices designated as "correct":
Is there any quarrel that the response choice(s) designated as "correct" are more correct or desirable than the response choices not designated as "correct"? ______ If yes, explain.

Are any of the response choices so weak that a student who lacks the knowledge (or the desired attitude), but is "test-wise" enough to use the process of elimination, can guess the correct response at an above chance level? ______ If yes, how could the "weak" response choices be strengthened?

2. Multiple-choice item with no response designated as "correct":
Do the response choices provide wide enough coverage to enable the student to give a reasonably accurate expression of his attitude or plan? ______ If no, what should be added or changed?

3. Matching item:
Are any of the matching pairs "weak", i.e., fail to provide information as to the student's master of the objective? ______ If so, which pairs?

4. Checklist:
*Would teachers (or students, as appropriate) find the checklist feasible of
- completion? YES ______ NO ______.
- scoring? YES ______ NO ______

5. Open-ended item:
Is the scoring guide clear? ______

*Would scoring of the item by teachers be feasible? YES ______ NO ______ If no, why not?

Will responses to the item provide the information sought? ______ If no, why not?

Could another type of item be used to gain similar information? ______ If so, how?

6. Individually administered items:
*Would scoring of the item by teachers be feasible? YES ______ NO ______ Could 2 or 3 group administered items (such as multiple choice or matching) get the same information? ______ If yes, how?

SECTION III

Comments concerning item attributes not mentioned above:

General evaluation of item potential - Excellent ________ Good ________ Fair ________ Poor ________
Comments:

Suggestions for revisions: (where possible, enter onto item)

Does this item need additional review? ________ Why?

APPENDIX C
School Districts Which Participated

School Districts	Student Reviews	Item Tryouts (Phase I)	Item Tryouts (Phase II)	Sensitivity-To-Instruction	Field Administration
Abilene ISD					X
Alamo Heights ISD		X			X
Aldine ISD	X	X			X
Aledo ISD		X	X		
Alief ISD	X	X			
Amarillo ISD					X
Anthony ISD	X				
Apple Springs ISD					X
Arlington ISD		X	X	X	
Austin ISD		X			X
Beaumont ISD					X
Boerne County Line ISD		X			
Brazosport ISD					X
Breckenridge ISD					X
Bryan ISD					X
Burleson ISD		X	X		
Calhoun County ISD					X
Carroll ISD		X	X		
Carrollton-Farmers Branch ISD	X			X	
Carrizo Springs ISD		X			
Castleberry ISD				X	X
Chapel Hill ISD					X

APPENDIX C
School Districts Which Participated

School Districts	Student Reviews	Item Tryouts (Phase I)	Item Tryouts (Phase II)	Sensitivity-To-Instruction	Field Administration
Cleburne ISD				X	
Clyde ISD					X
Collinsville ISD					X
Corpus Christi ISD					X
Cotulla ISD		X			
Crockett County Cons. ISD					X
Crystal City ISD		X			
Cypress-Fairbanks ISD	X				
Dallas ISD	X	X	X	X	X
Dayton ISD		X			
Denison ISD		X			X
Denton ISD				X	X
Dimmit ISD					X
Donna ISD					X
Duncanville ISD					X
Eagle Pass ISD		X			
Eanes ISD	X				
Ector ISD	X				
Ector County ISD					X
Edgewood ISD					X
Edinburg ISD			X		
Edna ISD					X

APPENDIX C
School Districts Which Participated

School Districts	Student Reviews	Item Tryouts (Phase I)	Item Tryouts (Phase II)	Sensitivity-To-Instruction	Field Administration
El Paso ISD					X
Everman ISD					X
Flatonia ISD		X			
Forney ISD				X	
Fort Stockton ISD					X
Fort Worth ISD	X	X	X	X	X
Fredericksburg ISD		X			
Galena Park ISD					X
Galveston ISD					X
Garland ISD			X		X
Gatesville ISD					X
Giddings ISD	X				
Goose Creek ISD	X	X			X
Granbury ISD				X	
Greenville ISD					X
Gregory-Portland ISD					X
Groesbeck ISD					X
Hamshire-Fannett ISD					X
Harlandale ISD					X
Hearne ISD					X
Hidalgo ISD					X

APPENDIX C
School Districts Which Participated

School Districts	Student Reviews	Item Tryouts (Phase I)	Item Tryouts (Phase II)	Sensitivity-To-Instruction	Field Administration
Highland ISD	X				
Houston ISD	X	X			X
Hughes Springs ISD					X
Hurst-Euless-Bedford ISD		X	X	X	
Irving ISD		X	X	X	X
Joshua ISD		X	X		
Kerrville ISD					X
Kendale ISD		X	X		
Kilgore ISD					X
Killeen ISD					X
Kingsville ISD					X
Klein ISD					X
Lackland ISD		X			
Lake Dallas ISD		X	X		
Lampasas ISD					X
La Porte ISD					X
Lewisville ISD		X	X	X	X
Little Cypress-Mauriceville ISD	X				
Lockney ISD					X
Lorenzo ISD					X

APPENDIX C
School Districts Which Participated

School Districts	Student Reviews	Item Tryouts (Phase I)	Item Tryouts (Phase II)	Sensitivity-To-Instruction	Field Administration
Lubbock ISD					X
McAllen ISD			X		
Mesquite ISD	X			X	
Midland ISD					X
Mineral Wells ISD					X
Mission ISD			X		
Moody ISD					X
Nederland ISD					X
New Boston ISD					X
North East ISD	X				X
North Forest ISD		X			X
Northside ISD					X
Palmer ISD		X	X		
Pearsall ISD		X			
Pflugerville ISD	X	X			
Pharr-San Juan-Alamo ISD			X		X
Plano ISD				X	X
Post ISD					X
Pottsboro RISD		X			
Quinlan ISD		X	X		

APPENDIX C
School Districts Which Participated

School Districts	Student Reviews	Item Tryouts (Phase I)	Item Tryouts (Phase II)	Sensitivity-To-Instruction	Field Administration
Red Oak ISD		X	X		
Richardson ISD					X
Rio Hondo ISD					X
Robinson ISD					X
San Antonio ISD	X	X			
Santa Rosa ISD					X
Sherman ISD					X
South San Antonio ISD					X
Spring Branch ISD		X			X
Stamford ISD					X
Taylor ISD					X
Temple ISD					X
Terrell ISD				X	
Texas City ISD					X
Tyler ISD					X
United ISD					X
Van Alstyne ISD		X			
Victoria ISD					X
Waco ISD					X

APPENDIX C
School Districts Which Participated

School Districts	Student Reviews	Item Tryouts (Phase I)	Item Tryouts (Phase II)	Sensitivity-To-Instruction	Field Administration
Water Valley ISD					X
Weatherford ISD				X	
West Orange-Cove Cons. ISD	X				X
Wharton ISD					X
Whitewright ISD		X			
Wichita Falls ISD					X
Willis ISD					X
Wilmer-Hutchins ISD		X	X		X
Wylie ISD		X	X		
Ysleta ISD					X

APPENDIX D

THE WLC/MRC INSTRUMENT ANALYSIS PROGRAM PACKAGE:

INTERPRETIVE GUIDE

The WLC/MRC instrument analysis package is a "generalized" computer program for analyzing items and instruments. It goes beyond the standard item analysis, applying statistical tests of significance to determine whether or not (i) students are, as a group, guessing at the item, (ii) the foils are attracting uniformly, and (iii) an item or instrument is culturally biased. In addition, traditional statistics are computed such as "p-values," foil distributions, point biserials, and KR_{20} reliability coefficients. The package comprises two components. (1) an "item analysis" which includes the cultural validity analysis, and (2) an "objective analysis," which includes mastery/non-mastery statistics as well as KR_{20}'s.

Description of the "Item Analysis" Printout

The printout for the "item analysis" includes the following statistics:

1. **P-values**
   The percent correctly answering each item is presented.
2. **Foil distribution**
   The percent answering each wrong response as well as omits, double marks, and "invalids" is presented.
3. **Z-test for "chance level of functioning (guessing)"**
   The hypothesis H_0. $p = 1/r$ is tested, where r = number of responses, against the one-sided alternatives H_1. $p < 1/r$ and H_2: $p > 1/r$, respectively, using a large sample (approximate) test. The hypotheses H_0, H_1, and H_2 correspond to "guessing," "below chance," and "above chance," respectively. If instruction has been given to the objectives tested, acceptance of H_2 means that there is evidence that the item is appropriate for the grade level tested. (If instruction has not been given, this test is still informative, but acceptance of H_2 should *not* be considered a requirement for inclusion of the item in the instrument.)
4. **Chi-square test for uniform foil response distribution**
   A chi-square test of the hypothesis that the foils (incorrect responses) are uniformly attractive is conducted and relevant statistics are printed.
5. **Internal consistency**
   A point biserial yields information about the internal consistency of the test, i.e., the extent to which "the items measure the same thing." Moreover, the "maximum" point biserial (corresponding to the case $p = 1/2$) is calculated. This statistic indicates the extent of the influence of the p-value on the point biserial.
6. **Breakdown by "cultural" groups**
   The above statistics are computed for each cultural group, for each cultural variable (e.g., ethnic background, sex, SES, etc.).
7. **Cultural validity analysis**
   Statistics for testing and measuring the cultural validity of items, objectives, and the total test are computed. The approach is that described in the SCORE technical report "Cultural Validity of Items and Tests. A New Approach" by James R. Veale and Dale I. Foreman. The conditional "foil" response distributions are investigated using chi-square and other procedures for measuring the degree of heterogeneity of these distributions across cultural groups.

In all of the above procedures which involve significance tests, significance levels (i.e., the probability of "more extreme" values under the null hypothesis) are computed if they are less than 0.10. This enables the user to specify his own "critical" level of significance (e.g., .01, .05, or .10). A sample item analysis is given in Table 1 on page 4.

Decision Model for "Item Analysis"

A "global" analysis of the printout data is suggested for determining the viability of the items. A "decision model" (Table 2) is presented to indicate one possible set of criteria. (Notation. X^2 = chi-square statistic, V = Cramér's statistic which measures degree of association or heterogeneity, T_{95} = lower 95 percent confidence limit for the Goodman-Kruskal T statistic, L^*_{95} = lower 95 percent confidence limit for Goodman-Kruskal L* statistic, "PT-BISER" = point biserial correlation coefficient, "Max PT-BISER" = "maximum" point biserial, Z = statistic for testing chance level of functioning (guessing).

The specific numerical cutoffs for the "rejection," "questionable," and "acceptance" levels are only rough guidelines for analysis. We do not favor a "weighting" system for evaluating items (e.g., assigning weights to the four types of analyses and numerical ratings to the three levels), since this would imply a further "abstraction" of the observed data beyond the statistical analysis. Moreover, it involves a high degree of arbitrariness. However, such a system may be of use in special situations.

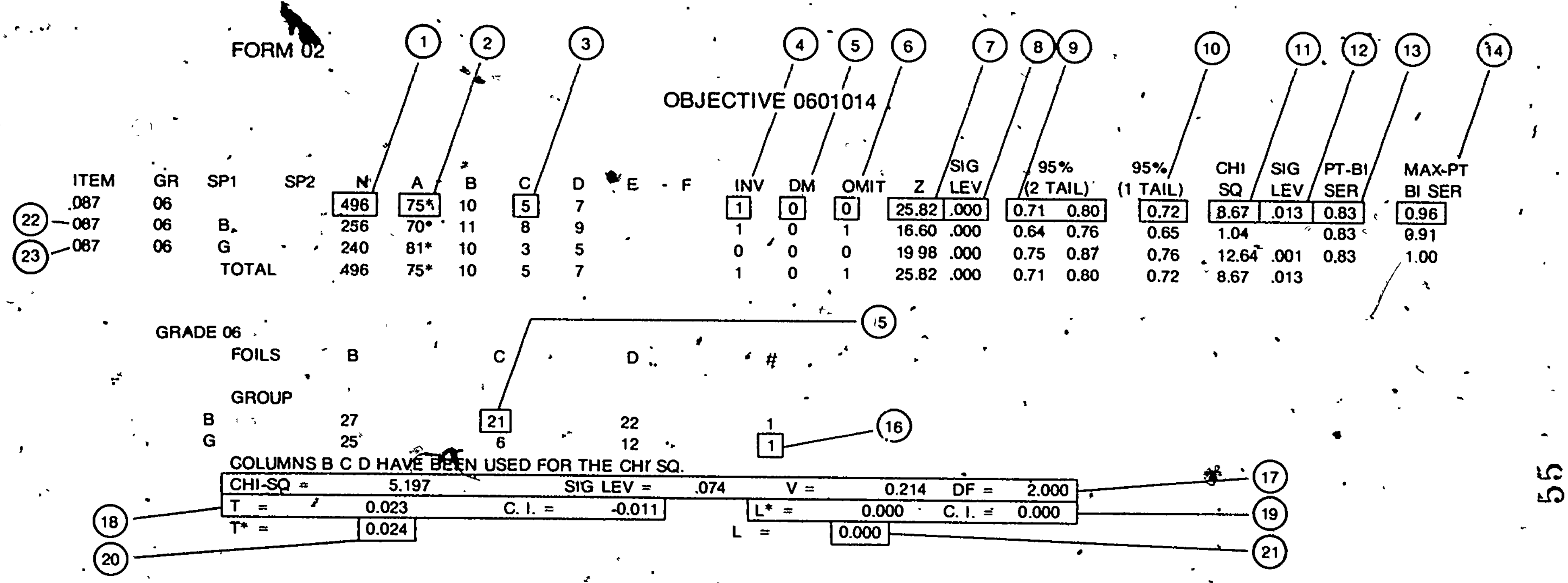
Objective Analysis

The printout for the "objective analysis" includes the following:

1. **Percent mastering objectives**
   The percent of respondents "mastering" each objective is printed. This is computed by determining the number of respondents who correctly answered a sufficiently high number of items in each objective. (For example, if there are five items and a 70 percent mastery level is used, a student must answer at least four items to be classified as a "master.")
2. **Upper confidence limits for percent mastering**
   Upper 95 and 99 percent confidence limits are computed using standard statistical procedures. This yields the *largest* probable values of the percent mastering objectives for the *population* based on the sample data.
3. **KR_{20} reliability coefficients**
   A KR_{20} reliability coefficient is computed for each instrument.

The cultural validity analysis may be conducted at the (i) item, (ii) objective, and (iii) total test levels. Similarly, the point biserial, percent mastering, and KR_{20} statistic may be computed with respect to objectives (two hierarchical levels) and total test.

TABLE 1. WLC/MRC Item Analysis Printouf



```
FORM 02
                                   OBJECTIVE 0601014
                                                                   SIG     95%         95%     CHI   SIG   PT-BI  MAX-PT
ITEM  GR  SP1    SP2    N    A   B  C  D  E - F   INV  DM  OMIT  Z      LEV   (2 TAIL)   (1 TAIL)  SQ    LEV   SER    BI SER
.087  06                496  75* 10  5  7          1    0   0    25.82  .000  0.71 0.80   0.72     8.67  .013  0.83   0.96
-087  06  B.            256  70* 11  8  9          1    0   1    16.60  .000  0.64 0.76   0.65     1.04        0.83   0.91
-087  06  G             240  81* 10  3  5          0    0   0    19 98  .000  0.75 0.87   0.76     12.64 .001  0.83   1.00
          TOTAL         496  75* 10  5  7          1    0   1    25.82  .000  0.71 0.80   0.72     8.67  .013

GRADE 06
     FOILS      B          C          D          #

     GROUP
  B             27         21         22         1
  G             25         6          12         1
     COLUMNS B C D HAVE BEEN USED FOR THE CHI SQ.
     CHI-SQ =        5.197        SIG LEV =      .074     V =         0.214   DF =      2.000
     T    =          0.023        C. I. =       -0.011    L* =        0.000   C. I. =   0.000
     T*   =          0.024                                L  =        0.000
```

KEY

1 N- counts (number of respondents)

2 P- values (percent correct)

3 Foil response percentage (foil 'C')

4 Percentage of invalid responses

5 Percentage of double marks

6 Percentage of omits

7 Z- statistic for testing 'grade level appropriateness' of item (deviation from random guessing)

8 Significance level of Z- statistic

9 95% confidence interval for (true) p-value

10 95% lower confidence limit for (true) p-value

11 Chi-square statistic for testing for uniform foil response distribution

12 Significance level for Chi-square

13 Point biserial correlation

14 "Maximum" point biserial

15 Frequency of foil response for cultural group ('B')

16 Frequency of double marking for cultural group ('G')

17 Chi-square cultural validity analysis (χ^2 statistic, significance level, Cramér's V, degrees of freedom)

18 Goodman-Kruskal 'T' statistic and lower 95% confidence limit

19 Goodman-Kruskal 'L*' statistic and lower 95% confidence limit

20 Goodman-Kruskal 'T*' statistic

21 Goodman-Kruskal 'L' statistic

22 Statistics 1-14 for subpopulation 'B' (boys)

23 Statistics 1-14 for subpopulation 'G' (girls)

TABLE 2. Decision Model For Item Analysis

	CULTURAL VALIDITY		INTERNAL CONSISTENCY	CHANCE LEVEL OF FUNCTIONING	UNIFORM FOIL DISTRIBUTIONS
	Large Samples	Small Samples			
Acceptance level (item O.K.)	χ^2 not significant at .10 level **and** $V < 0.3$	$T_{95} \leq 0.03$ **and** $L^*_{95} \leq 0.03$	$r_{pb} \geq 0.4$ **and** max $r_{pb} \geq 0.6$	$Z > 0$ **and** Z significant at the .05 level	χ^2 not significant at .05 level
Questionable level revise or eliminate item	χ^2 significant at .10 level **or** $V \geq 0.3$ (but condition for "rejection level" given below **not** satisfied)	$T_{95} \geq 0.03$ **or** $L^*_{95} \geq 0.03$ (but condition for "rejection level" given below not satisfied)	$0 \leq r_{pb} \leq 0.4$ **and** max $r_{pb} \geq 0.6$ **or** $0.4 \leq r_{pb} \leq 0.6$ **and** $0.4 \leq$ max $r_{pb} \leq 0.6$	Z not significant at 0.05 level	χ^2 significant at .05 but not .01 level
Rejection level probable elimination of item	χ^2 significant at .05 level **and** $V \geq 0.5$	$T_{95} \geq 0.25$ **or** $L^*_{95} \geq 0.25$	$r_{pb} \leq 0.4$ **and** max $r_{pb} \leq 0.6$	$Z < 0$ **and** Z significant at .05 level	χ^2 significant at .01 level

APPENDIX E

STUDENT INFORMATION SHEET

On your Answer Sheet there is a section labeled "STUDENT INFORMATION," columns 3 through 9. Each column of numbered ovals corresponds to a question on this page. read each question, 3 through 9, and darken the oval that matches the number of your response in the appropriate column on your Answer Sheet

3. To which group do you belong?

   1. Mexican-American
   2. Black
   3. Anglo
   4. American Indian
   5. Oriental
   6. Other

4. Which language is spoken in your home?

   1. Spanish
   2. German
   3. Czech
   4. French
   5. Chinese
   6. Italian
   7. Polish
   8. English
   9. Other

5. Outside of school, how long do you usually watch TV on a school day?

   1. None
   2. 1 or 2 hours
   3. 3 or 4 hours
   4. 5 or 6 hours
   5. More than 6 hours

6. How many books do you have in your home?

   1. Few
   2. Many

7. Do you have encyclopedias in your home?

   1. Yes
   2. No

8. Does your family receive a daily newspaper?

   1. Yes
   2. No

9. Does your family receive magazines through the mail?

   1. Yes
   2. No

CULTURAL VALIDITY OF ITEMS AND TESTS: A NEW APPROACH

James R. Veale and Dale I. Foreman

Technical Report No. 1

Abstract*

The question of cultural bias in test instruments is a critical one for test development. Most of the procedures for detecting cultural bias which have been heretofore advanced assume that either (i) an unbiased external criterion for ability is available, or (ii) the total score on the test is a reasonably good approximation of the student's ability.

The approach taken in this paper is based on the variation among conditional foil response distributions for the various cultural groups in the population tested. It does not involve measures of ability and thus does not require either of the above assumptions. Both large sample and small sample procedures are presented.

*Reprints are available from WLC/MRC, Iowa City, Iowa

APPENDIX G

GUIDE TO THE STATISTICS USED IN THE CULTURAL VALIDITY ANALYSIS

This appendix includes a brief discussion of the statistics used in the cultural validity analysis.

1. **Chi-square statistics.**
   A chi-square statistic is computed for each item to test the statistical significance of cultural heterogeneity of foil responses, i.e., to test the hypothesis that cultural groups and foil response are independent. The usual formula was applied to the contingency table consisting of foil responses (column) for the various cultural groups (rows). Significance levels were computed and (when they were less than 0.10) printed.
2. **Cramér's V statistic.**
   Cramér's V is a measure of the degree of cultural variation in foil responses, defined as follows:

$$V = \sqrt{\frac{X^2}{N \min\{g-1, f-1\}}}$$

   where X^2 is the aforementioned chi-square statistic, N is the number of incorrect (foil) responses, g is the number of cultural groups, and f is the number of foils (plus "double marks," if any). The V statistic ranges from zero to unity, with zero corresponding to *no* cultural variation and unity corresponding to *extreme* cultural variation.
3. **The Goodmen-Kruskal measures of heterogeneity.**
   Goodman and Kruskal (1954) developed several measures of association which have a probabilistic interpretation. Two if these statistics, denoted T and L, are defined as follows:

$$T = \frac{N \sum_a \sum_b O^2_{ab}/O_{a.} - \sum_b O^2_{.b}}{N^2 - \sum_b O^2_{.b}}$$

$$L = \frac{\sum_a O_{am} - O_{.m}}{N - O_{.m}}$$

   where O_{ab} is the observed number of responses to foil b in cultural group a, $O_{a.}$ is the total number of foil responses in cultural group a, $O_{.b}$ is the total number of responses to foil b. O_{am} is the *maximum* number of foil responses in cultural group a, $O_{.m}$ is the maximum total number of foil responses (after summing over cultural groups), and N is the total number of foil responses.

   The above statistics, and the slightly modified statistics denoted T* and L*, have operational meaning whatever the sample size (N), unlike the chi-square (which requires large samples). They measure the proportion of errors in predicting the foil responses of randomly chosen individuals that can be eliminated by incorporating knowledge of the individual's cultural group. They all range from zero to unity, with zero corresponding to *no* gain in predictive utility with knowledge of cultural groups (no cultural variation) and unity corresponding to *perfect* predictive utility with knowledge of cultural group (extreme cultural variation).
4. **Lower 95 percent confidence limits for T and L*.**
   Lower 95 percent confidence limits for (the true values of) T and L* were also computed. This takes into account the sampling error, which is important since we are sampling approximately 600 students (per instrument), rather than testing the entire population of Texas students.

5. **Degree of cultural variation.**

Professional judgment was employed to rate the degree of cultural variability exhibited by the item data, using *all* of the statistics discussed above. The rating scale was:
1 = very high variability,
2 = high variability, and
3 = moderate variability.

For a more detailed discussion of the approach and techniques used for measuring cultural variation, see Veale and Foreman (1975).

APPENDIX H

ANALYSIS FOR ITEMS EXHIBITING CULTURAL VARIATION

This appendix includes a content analysis of items manifesting some degree of cultural variation according to the statistics described in Chapter 4 and in Appendix 6. Tables listing the items having cultural variations and probable cause(s) for the variation(s) are displayed. (For example, an item may be culturally biased as reflected by the variation in foil responses across groups due to a factor inherent in the respondent's cultural background which results in a distortion of the p values for the groups.) It should be understood that these analyses consist of data based hypotheses of one test development specialist.

BOOKLET 11

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0103/3 *	07			X (A,D)		
0103/3	10			X (A,D)		
0104/11	07		X			
0104/11	10		X			
0104/12 *	07	X (E)			X	
0104/13	10					X (Easy)
0104/15	07		X			
0104/16	07	X (S,E)				
0105/17A	07	X (E)			X	
0105/17A	10	X (E)			X	
0105/17C	07	X (E)			X	
0105/17E *	07				X	
0105/17D	07	X (E)			X	
0105/17B	07	X (E)			X	
0107/7 *	10	X (E)		X (B)		

Key for the above table and other tables in this appendix.

* = item is included in the content analysis

E = ethnic variable

S = sex variable

BAD FOIL

Booklet 11, Item 3:

Which *ONE* of the following is the *BEST* reason why people need to be satisfied with their jobs?

(A) If they make the effort, people can learn to get along on a job.
(B) Satisfied people do better work and are happier.
(C) Satisfied people do not have to try very hard to better themselves.
(D) People should seek job satisfaction from their family and friends.

Foils "(A)" and "(D)" do not relate to the question that was asked. Any answer to a question should certainly answer the question (only wrongly if it is a foil.) Both "(A)" and "(D)" need to be revised to answer the question "Why do people need to be satisfied?" or replaced with other foils.

BAD FOIL + CULTURAL BIAS

Booklet 11, Item 7:

As a pharmacist working for a chain drugstore, your goal is to operate your own business. You realize that a new pharmacy probably would be successful if opened in a recently developed area of town. You would like to quit your job and establish your own business, but you do not have enough money to do so. so.

Which of the following actions would *BEST* solve your problem and help you reach your goal?

(A) forget about operating your own business
(B) sell your home and car to raise the money
(C) go into partnership with someone with money to invest
(D) read all the latest magazines on drugstore operation

Foil "(B)" is not attractive to any of the students. It is logical that nearly everyone is sufficiently security oriented (conservative) to resist giving up anything that they already possess in order to engage in a speculative venture. That is exactly what is suggested in foil "(B)", "sell your house and car to raise the money."

Another problem with the foil in relation to the item is that no where in the item does it say "you" own a house and car. Most kids would not consider foil "(B)" since they cannot relate to such ownership.

Mexican-Americans are overly attracted to "(D)". It is possible that through their background (poor reading) and their view of the background of those who are successful (and can read) they believe reading proficiency will yield success.

CULTURAL BIAS

Booklet 11, Item 12:

Grace wants a job where she does not have to deal with many strangers.

Which career do you feel would *BEST* match Grace's goal?

(A) receptionist
(B) bookkeeper
(C) public librarian
(D) salesperson

In this item, there is a Mexican-American/black "interaction" at grade 7 with foils "(C)" and "(D)". (Mexican-Americans were more attracted to "(C)", while blacks were more attracted to "(D)". Unless the students were specifically taught the duties of these jobs, it is likely that the responses would be highly influenced by either lack of experience or by some key word association. For example, the most difficult word, "receptionist," is chosen very frequently. This very often happens when the students have little knowledge of concept. Moreover, it is interesting to note that among the foil responses, "(C)" and "(D)" are proportionately more attractive with minorities than with "others" (primarily anglos). With specific education to these occupations, the variation may be eliminated.

BAD FORMAT + DIAGNOSTIC

Booklet 11, Item 17:

At this time, which of the following do you think is your career direction? Darken (A) on your Answer Sheet for the *ONE* direction which you have chosen. Darken (B) for the others.

(A) (B) a. enter a trade or technical school
(A) (B) b. prepare for immediate employment
(A) (B) c. enter college
(A) (B) d. do not work at a job
(A) (B) e. some other direction

This item has no correct answer. It is asking a student to select a career direction. The data can then be used as census data to help plan for counseling, etc.

Unfortunately, the "Yes/No" format was confusing to the Mexican-Americans and blacks. The data show that many minority students marked "Yes" to several of the career directions. They did not understand that only one "Yes" should be marked. These data in their present form are of little use.

A better format would be to eliminate foil "(E)" and make this item a four choice multiple choice asking the student to "Mark the *ONE* career direction you choose."

BOOKLET 12

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0102/5 *	08			X(All Correct)		
0102/24	08		X			
0102/24	11		~~X~~			
0112/8 *	11			X (A,B)		
0112/10	08	X (E)		X (B)		
0112/10	11		X	X (B)		
0112/11	08		X	X (C)		
0112/11	11	X (S)		X (C)		
0202/16	11					X
0202/18 *	08	X (E)			X	
0202/21	08	X (E)				
0202/22	08		X			

BAD FORMAT

Booklet 12, Item 5:

> Read the following paragraph and answer questions 5 and 6 on your Answer Sheet.
>
> Carol, who is a volunteer worker at General Hospital, is graduating from high school. She hopes to make nursing her career. The hospital has offered her a job as a nurse's aide. Carol is trying to decide whether to take the job or to enter her local community college to become a Licensed Vocational Nurse.
>
> 5. If Carol decides to take the job, which *ONE* of the following might be a result of that decision?
>
> (A) She might get to be a doctor.
> (B) She may never become a nurse.
> (C) She will always work as a nurse's aide.
> (D) She would still be able to to school.

The stem of this item is stated in such a way that all answers are correct. The question asks "which *ONE* of the following might be . . . ". Any of the answers *might* be a result of the decision. It should be restated in such a way that the student will select the *most likely* result of the decision and then make sure there is only *ONE* most likely decision among the answers.

BAD FOIL

Booklet 12, Item 8:

> Joe never did well in school. Five years ago, he dropped out and began doing odd jobs around the neighborhood. He lived with his folks and paid part of the living expenses with his earnings.
>
> A year ago, Joe and LaWanda married. Now they and their baby live with his folks, but they would like very much to be able to move to a place of their own. Joe worries a lot about taking care of his family. Joe keeps trying to get a steady job. He wants to get training. He needs a high school diploma. His friends tell him that he is crazy to think that things will ever get better.

Given the factors influencing Joe's life-style, which *ONE* of the following statements *BEST* describes Joe's chances of meeting his needs and wants.

(A) Because of Joe's educational level, he will not have difficulty meeting his needs and wants.
(B) Because of Joe's marriage, he will meet all his needs and wants.
(C) Because of Joe's educational level and family responsibilities, he will have a difficult time meeting his needs and wants.

Foils "(A)" and "(B)" are too easily eliminated. One problem is that because two foils are parallel (negatives), "(A)" and "(C)", the student can automatically eliminate "(B)". This is a common problem in test construction.

Secondly, it is obvious that Joe's low education level is going to limit his success in meeting his needs and wants. This leaves "(C)" as the only choice.

CULTURAL BIAS + BAD FORMAT

Booklet 12, Item 18:

Which *ONE* of the following would *NOT* be a good way to learn about the supply and demand of a particular occupation?

(A) going to the local employment office
(B) talking to personnel directors
(C) talking to those currently employed in the field
(D) determining the number of workers in your local community.

There is evidence that the blacks and Mexican-Americans are *NOT* reading the negative stem as a negative. Both are going to foils, each different, that would be, in their mind, *BEST* places to learn about supply and demand of an occupation. The concept of supply and demand may be too difficult for eighth graders.

BOOKLET 21

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0201/18 *	10		X			
0205/13	10		X	X (B)		
0205/16	07		X			
0205/16 *	10		X			
0207/10	07			X (D)		
0207/17	07		X			
0210/5	07	X (E)				
0210/6	07		X			
0210/7	07		X			

DIAGNOSTIC

Booklet 21, Item 16:

Which of the sources below would give you the *BEST* information (job description, location within the United Sates, salary, requirements) about all types of employment?

(A) a local employment agency
(B) "Help Wanted" section of newspaper
(C) *Occupational Outlook Handbook*
(D) state employment office

The incorrect responses to this question should lead into instructional strategies which will clarify the typical types of information that can be obtained from each source. One potential source of problems at present may be the lack of knowledge of many about the existence of the *Occupational Outlook Handbook*. Also, most people are aware of the "Help Wanted" section of the newspaper and state employment offices. This could cause differential attraction to "(B)" and "(D)" due to their common occurrence.

DIAGNOSTIC

Booklet 21, Item 18:

Part 1
On the line below, write the occupation title you chose from the Occupation List on Page 3.

____

Think about the occupation you chose. Have you ever talked to someone who works in that field in order to get more information about the field? If so, darken (Yes), otherwise darken (No).

Part II
If you answered "Yes" what did he/she tell you about his/her job that might be useful to you?

____

____

____

Scoring Key:

Mark "(A)" if the student indicates his/her career of interest, "Yes" for Part I, and at least one piece of useful information that the person told him/her about his/her job in Part II.
Mark "(B)" if the student indicates his/her career of interest and "Yes" for Part I only.
Mark "(C)" otherwise.

Examples of Useful Information:

- types of skills and knowledge areas required
- job outlook for the future
- types of job characteristics relevant to the job
- salary expectations
- types of employee benefits that probably exist
- chances for advancement in the chosen career

This is an open-ended item with a scoring guide. The differential response patterns for this type of item mean either the scoring guide is inappropriate, incomplete, or otherwise disfunctional or that the information is diagnostic of different population deficiencies. In this case, the scoring guide is appropriate. The strong Mexican-American affinity to "C" implies that fewer of the group have talked to someone who works in a field of their interest.

BOOKLET 32a

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0301/2 *	11		X	X (D)		
0301/3	08			X (C,D)		
0302/11 *	08	X (E)		X (Stem,A)		
0302/12B	08		X			
0302/12B	11		X			

DIAGNOSTIC AND BAD FOIL

Booklet 32a, Item 2:

Which *ONE* of the following would you probably be required to write in on a job application form?

(A) names and addresses of references
(B) names of stores where you have charge accounts
(C) names of your teachers
(D) names of foreign countries in which you have traveled

Only one application would include the question "What foreign countries have you traveled in?" That is a security clearance for a government job. The foil "(D)" is very out of line with the other responses making it unattractive or unreasonable. Something like "names and addresses of all your schools" would be better.

The other foils give diagnostic information, such as an indication of where you would have to list your charge accounts. Each of these wrong responses could be used by the teacher to teach the student where their use would be appropriate.

BAD FOIL + CULTURAL BIAS

Booklet 32a, Item 11:

John is 16 years old and will be interviewed for a part-time position as a machinist. The personal quality his prospective employer will think *MOST* important is

(A) his previous years of work experience.
(B) his high school grade average.
(C) his appearance.
(D) his attitude.

There are several problems with this item. First, the question asks for a personal quality and the keyed answer "(B)" ("His previous year of work experience") is not a personal quality. Further, foil "(C)" is not selected. This seems logical since it is also not a personal quality but a physical quality. Blacks selected foil "(D)" heavily.

BOOKLET 32b

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0307/1	11		X	X (B)		
0307/4 *	08		X	X (D)		

BAD FOIL + DIAGNOSTIC

Booklet 32b, Item 4:

Which *ONE* of the following situations indicates job success?

(A) You work for a company that has signed a new labor contract and has given all employees an eight percent raise.
(B) You are asked to work overtime on Friday afternoons for the next two months.
(C) During a conference, you are asked for your advice on changing employee work routines.
(D) You are asked to proofread your letters before they are mailed out.

Foil "(B)" seems on the surface to be a good foil. In other words, being asked to work overtime means the boss likes your work and therefore you are successful. This foil is not attractive to the students.

Otherwise, all the foils provide diagnostic information for the teacher and students who select them Foil "(A)", for example, identifies the student who is unable to discriminate between a general increase and a personal raise for a good job done.

BOOKLET 32 c

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0301/31 *	08		X			

DIAGNOSTIC

Booklet 32c, Item 31:

Scoring Key:
This is a summary score which ties together items 1-30.

Mark "(A)" if all 30 categories are scored (A).
Mark "(B)" if all "(*)" categories are scored (A), and one or more of the other categories are scored (B).
Mark "(C)" if five to seven of the "(*)" categories are scored (A), and one or more of the other categories are scored (B).
Mark "(D)" if less than five of the "(*)" categories are scored (A), and one or more of the other categories are scored (B).

This item is an application blank that is scored according to degree of correctness. For example, those who scored in category "(B)" have completed the necessary (i.e. critical) parts of the application. Their response is sufficient to be able to obtain a job.

The item is scored according to written criteria and is, therefore, diagnostic. Students missing parts of the item can be instructed to improve their subsequent responses.

APPLICATION FOR EMPLOYMENT

SOCIAL SECURITY NUMBER: ______________________

PERSONAL INFORMATION DATE: ______________________

NAME PREFIX
MR.______ MRS.______ MISS ______
DR.______ MS.______

NAME ____
LAST FIRST MIDDLE

PRESENT ADDRESS ____
STREET CITY STATE ZIP

PERMANENT ADDRESS ____
STREET CITY STATE ZIP

PHONE NO.

IF RELATED TO ANYONE IN OUR EMPLOY,
STATE NAME AND DEPARTMENT REFERRED BY

EMPLOYMENT DESIRED

POSITION DATE YOU CAN START SALARY DESIRED

ARE YOU EMPLOYED NOW? IF SO MAY WE INQUIRE OF YOUR PRESENT EMPLOYER?

EVER APPLIED TO THIS COMPANY BEFORE? PLACE DATE

EDUCATION	NAME AND LOCATION OF SCHOOL	YEARS ATTENDED	DATE GRADUATED	MAJOR COURSE OF STU
ELEMENTARY SCHOOL			▨	▨
JUNIOR HIGH OR MIDDLE SCHOOL			▨	▨
HIGH SCHOOL				
COLLEGE				
TRADE, BUSINESS OR CORRESPONDENCE SCHOOL				

WHAT FOREIGN LANGUAGES DO YOU SPEAK FLUENTLY? READ WRITE

ACTIVITIES (CLUBS, HOBBIES, INTERESTS, ETC.)

SIDE ONE

FORMER EMPLOYERS (LIST BELOW LAST FOUR EMPLOYERS, STARTING WITH LAST ONE FIRST.)

DATE MONTH AND YEAR	NAME AND ADDRESS OF EMPLOYER	SALARY	POSITION	REASON FOR LEAVING
FROM				
TO				
FROM				
TO				
FROM				
TO				
FROM				
TO				

REFERENCES. GIVE BELOW THE NAMES OF THREE PERSONS NOT RELATED TO YOU, WHOM YOU HAVE KNOWN AT LEAST ONE YEAR.

NAME	ADDRESS	BUSINESS	YEARS ACQUAINTED
1			
2			
3			

IN CASE OF EMERGENCY NOTIFY

NAME ADDRESS PHONE NO.

I AUTHORIZE INVESTIGATION OF ALL STATEMENTS CONTAINED IN THIS APPLICATION. I UNDERSTAND THAT MISREPRESENTATION OR OMISSION OF FACTS CALLED FOR IS CAUSE FOR DISMISSAL FURTHER, I UNDERSTAND AND AGREE THAT MY EMPLOYMENT IS FOR NO DEFINITE PERIOD AND MAY. REGARDLESS OF THE DATE OF PAYMENT OF MY WAGES AND SALARY, BE TERMINATED AT ANY TIME WITHOUT ANY PREVIOUS NOTICE

DATE SIGNATURE

DO NOT WRITE BELOW THIS LINE

INTERVIEWED BY DATE

REMARKS:

HIRED DEPT. ASSIGNMENT POSITION REPORTING DATE SALARY WAGES

APPROVED: 1. 2. 3.

EMPLOYMENT MANAGER DEPT. HEAD- GENERAL MANAGER

SIDE TWO

BOOKLET 41

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0401/1	07					X
0401/1	10					X
0401/3 *	07	X (E)				
0407/10	07		X			
0408/13 *	07		X			
0401/4A-D	07,10				X	
0407/9A-D	07,10				X	
0407/12A-F	07,10		X		X	

NO CLEAR EVIDENCE

Booklet 41, Item 3:

Lem works in a supermarket as produce manager. He supervises the stock boys and sets a good example in his work. His work is always outstanding. Lem sometimes uncovers pricing errors which would cost the store a lot of money. The food in his department is always fresh. Lem is careful to insure that his customers are well satisfied.

How would Lem's work likely affect his status in the store?

(A) Lem would probably be offered a job by another store.
(B) Lem would be looked up to by his fellow employees as a good worker.
(C) Lem would feel that he is better than everyone else.
(D) Lem's boss might think that Lem is out to get his job.

In this item, blacks tend to respond more to "(D)". It could be interpreted that anyone who puts out extra effort is out to get someone else's job. This could result in the selection of foil "(D)" by those who have that outlook.

DIAGNOSTIC

Booklet 41, Item 13:

Juan, a social worker, has completed a case which required a great deal of time and effort Select the *ONE* statement which indicates a behavior that shows Juan takes pride in his successful accomplishment.

(A) Juan told a fellow worker how good he felt about the job.
(B) Juan left work early because the task was completed.
(C) Juan decided to apply for a new job that would pay more money and would not demand so much time.
(D) Juan talked with Helen about a case on which she was working.

Each of the incorrect responses indicate different results which might stem from an incorrect interpretation of the meaning of taking pride in one's accomplishments. For example, an individual may think that leaving early was an indication of pride.

Each incorrect response indicates a mind set that the student has which could be corrected with different instructional approaches. This offers an ideal diagnostic setup which can help combine testing results with instruction for things such as grouping students for instruction and definition of the instructional program.

BOOKLET 42a

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0403/8	11		X			
0403/10A	08		X			
0403/10C	08		X			
0403/10C	11		X			
0403/10H *	11		X			
0403/10D	11		X			
0403/10J	11		X			
0403/10K	08		X			
0403/10K	11		X			
0403/11F	08		X			
0403/11D	08		X			
0403/11D	11		X			
0403/11I	08		X			
0403/11I	11		X			
0403/11G	11		X			
0403/11A	08		X			
0403/11A	11		X			
0403/11C	11		X			
0405/2 *	11	X (E)		X (B)		
0405/5B	08	X (E)				
0405/5D	11	X (E)				
0405/5C	08		X			
0405/6C	08		X			
0405/6C	11		X			
0405/6B	08		X			
0405/6B *	11		X			
0405/6A	08		X			
0405/6E	11		X			
0414/13A	08		X			
0414/13F	08		X			
0414/14E	11		X			
0414/14D *	08		X			
0414/14G *	08		X			
0414/14A	08		X			
0414/14A	11		X			
0414/14C	11		X			
0414/14F	11		X			
0414/15A	08		X			
0414/15C	11		X			
0414/15D	08		X			

BAD FOIL + CULTURAL BIAS

Booklet 42a, Item 2:

> Imagine that you began managing a local volunteer project for development of a park in your neighborhood. The project involved much time and planning for getting jobs done by other people, including. earth movers, planters, tree trimmers, electricians, etc. You find that more and more of your time is taken up with this project. Problems arise and it is difficult to get cooperation from others. You feel discouraged and would like to drop the project.
>
> Which *ONE* of these statements shows a *BENEFIT* you might gain by staying with the project?
>
> (A) You will make a lot of money if you stay with the project.
> (B) You will make a lot of new friends if you stay with the project.
> (C) You will be asked to serve as chairperson of other volunteer projects.
> (D) You will gain some personal fulfillment if you achieve your goal.

This is a very easy item. Foil "(B)" is not drawing anyone and should be replaced. The cultural variation is primarily due to an avoidance by Mexican-Americans of the idea that they might be asked to be chairperson of anything. This whole concept seems to be less than concrete. As a result it is difficult to measure with any degree of success.

DIAGNOSTIC

Booklet 42a, Item 6:

6. On your Answer sheet, indicate whether you strongly agree, agree, are undecided, disagree or strongly disagree with each statement below by darkening the letter as follows:

STRONGLY AGREE	AGREE	UNDECIDED	DISAGREE	STRONGLY DISAGREE
(A)	(B)	(C)	(D)	(E)

(A) (B) (C) (D) (E) a. A person should practice disciplining himself/herself to complete tasks which should be done but are unpleasant.

(A) (B) (C) (D) (E) b. A person should stay with a task which is boring but must be done.

(A) (B) (C) (D) (E) c. If a person has a lot of work to do, he/she should not complete all the work.

(A) (B) (C) (D) (E) d. A person should not put off work until the last minute.

(A) (B) (C) (D) (E) e. When a person has a job to do but also wants to do something for fun, he/she should finish the job first and get it out of the way.

(A) (B) (C) (D) (E) f. There are some tasks which have to be done, even though a person does not want to do them.

(A) (B) (C) (D) (E) g. A person probably feels good about himself/herself if he/she sticks with a task until it is completed.

(A) (B) (C) (D) (E) h. A person should not attempt to complete a task which he/she does not think he/she would like.

The information in this item is useful in identifying the beliefs of students and the strengths of their beliefs. Black students tended to disagree with part a.

These data should be useful in planning instructional strategies.

DIAGNOSTIC

Booklet 42a, Item 10:

Read the following list of words and phrases. On your Answer Sheet, darken (A) for each item which may have influenced your attitude toward work. Darken (B) for each item which probably did not.

(A) (B) a. reading
(A) (B) b. mathematics
(A) (B) c. athletics
(A) (B) d. sex
(A) (B) e. age
(A) (B) f. family
(A) (B) g. socio-economic background
(A) (B) h. education
(A) (B) i. work experience
(A) (B) j. culture
(A) (B) k. peers (friends)
(A) (B) l. media; television, motion pictures, newspapers, magazines, etc.

There are no correct answers for these items. They are survey questions which naturally have different response patterns for different sexes and ethnic groups since each person's attitude toward is influenced by different things.

BOOKLET 42b

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0418/8 *	08	X (E)		X (C)		
0418/12C	08	X (E)	X			
0418/12E *	11		X			
0418/12D *	08		X			
0418/12B	08		X			
0418/12B	11		X			
0418/12A	08		X			
0421/4	08			X (A)		

BAD FOIL + CULTURAL BIAS

Booklet 42 b, Item 8:

Which statement reflects a *POSITIVE ATTITUDE* toward lawyers?

(A) Lawyers take advantage of people in trouble.
(B) Lawyers help people deal fairly with each other.
(C) Lawyers will not help people who do not have money to pay legal fees.
(D) Lawyers help people to cheat on their income tax.

There are two problems with foil "(D)". (1) it is not a positive attitude and (2) most people are aware of the fact that lawyers are strictly accountable to staying within the law and that to help you cheat on income tax is outside the law. As a result, no one chose this foil. Foil "(C)" is not attractive to Mexican-Americans.

DIAGNOSTIC

Booklet 42b, Item 12:

> On your Answer Sheet, indicate whether you strongly agree, agree, are undecided, disagree, or strongly disagree with each of the following statements by darkening the letter which you feel is appropriate as follows:
>
> | STRONGLY AGREE (A) | AGREE (B) | UNDECIDED (C) | DISAGREE (D) | STRONGLY DISAGREE (E) |
> |---|---|---|---|---|
>
> (A) (B) (C) (D) (E) a. Being a lawyer is a more useful occupation in society than being a mail carrier.
>
> (A) (B) (C) (D) (E) b. Artists perform useful tasks in our society.
>
> (A) (B) (C) (D) (E) c. Auto mechanics have less dignity than teachers.
>
> (A) (B) (C) (D) (E) d. The dignity of a job depends on the amount of education required.
>
> (A) (B) (C) (D) (E) e. The dignity of a job depends on the salary involved.
>
> (A) (B) (C) (D) (E) f. The dignity of a job depends on the quality of performance of the people involved.

Item 12 is a difficult format for many students. Some of the statements are about things that are tradition bound. Such as part (a). Minorities agreed with (a), showing that they have grown up with the idea that high class white collar jobs are more useful.

BOOKLET 51

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0504/8 *	07	X (E)	X			
0504/8	10	X (E)	X			
0504/10A *	07			X (Wrong Key)		
0504/11D *	07		X			
0504/11D	10		X			
0504/11A	07		X			
0504/11A	10		X			
0504/11F	07		X			
0504/11B	07	X (S)	X			
0504/11C	07		X			
0504/11C	10		X			
0504/11E	07	X (E)				
0504/11E	10	X (E)				
0506/17 *	07			X (B)		
0506/19	07			X (B)		
0506/22 *	07			X (A)		
0506/24	07			X (D)		

CULTURAL BIAS

Booklet 51, Item 8:

A team of people was chosen to discuss school bus routes and solve problems with time schedules. The team had a hard time arranging a plan of action. Everyone talked at once, argued back and forth, and did not listen to the chairperson. Each member voiced his/her ideas to one or two other members rather than directing his/her comments to the entire group. At the end of the project, the committee still had not agreed upon a clear-cut set of suggestions.

On your Answer Sheet, darken the letter which shows to what degree this group of people worked with each other as a team.

(A) They had a very effective system of procedure as a team.
(B) They had the makings of a good team, but one or two people spoiled it.
(C) They were not effective as a team.
(D) They would have been effective as a team had they had more time to work.

Blacks tended to select foil "(A)" which is the opposite of the situation that is true. This could be symptomatic of a gang approach to decision making where only one or two key people are involved in making decisions. This would cause members to only speak to those one or two key people who are in control.

Also, in many city gangs, there is arguing among members during times of decision making with no clear purpose being defined by the group. This would, then, seem to be an effective procedure to those with inner city experience.

BAD FOIL + INAPPROPRIATE KEY

Booklet 51, Item 10:

> Suppose you are part of a team assigned to recommend special units of study for the drafting of a building construction project. Frank, the chairperson of the group, seems to be losing interest in the project.
>
> For each of the three questions below, darken (A) on your Answer Sheet if your answer to the question is "Probably so." Darken (B) if your answer is "Probably not." Darken (C) if you do not know what you would do.
>
> a. Will you ask Frank to let the person whom you thought could do a better job be the chairperson?
>
> (A) Probably so
> (B) Probably not
> (C) I don't know what I would do.
>
> b. If you think a suggested unit is not a good one, will you volunteer your opinion?
>
> (A) Probably so
> (B) Probably not
> (C) I don't know what I would do.
>
> c. Will you agree to recommend only units that the majority wants?
>
> (A) Probably so
> (B) Probably not
> (C) I don't know what I would do.

Part a is keyed "(B)". Many students chose "(A)", which may indeed be a more appropriate response. It may be argued that if a person loses interest in a project that he/she is in charge of, it is appropriate to suggest that that person be replaced with someone else who has a much greater interest in the project.

DIAGNOSTIC

Booklet 51, Item 11:

> As a part of a social studies project, a class divided into groups of seven to write answers to social situations. Your group was given three questions to discuss. As a member of this group, how would you probably have worked in the group?

On your Answer Sheet, darken (A) for each of the statements that you think is true about yourself. Darken (B) for the ones that you do not think are true about yourself.

(A) (B) a. I would be a leader in the group.
(A) (B) b. I would not be a leader, but I would be active in expressing my feelings.
(A) (B) c. I would go along with whatever the leaders decided.
(A) (B) d If the group couldn't decide on the answer, I would take a vote and write the answer favored by the majority.
(A) (B) e. I would prefer not to participate actively, but I would be willing to write the answers.
(A) (B) f. I'm not sure how I would work in the group.

In this item there are no correct answers and the main purpose of data gathered in this item is detecting differences in values across cultural groups. In part "e" Mexican-Americans respond with a higher proportion of "Yes" responses.

Many girls marked "Yes" to part "b" indicating the lack of interest in being the leader of a group. This is consistent with traditional sex roles. The "interaction" that exists here could be changed with the new roles emerging due to the women's rights movement.

BAD FOIL

Booklet 51, Item 17:

Juan and his boss are walking to the front door of the building where they both work. His boss opens the door for Juan and motions for him to go ahead into the building. What would be the best thing for Juan to do?

(A) say "No, thank you," and wait until his boss goes in
(B) go in and apologize to his boss for not opening the door for him
(C) go in and say "Thank you"
(D) go in and say nothing but watch for a chance to open a door for his boss

Although the probable intent of foil "(B)" was to identify those students who would demean themselves in front of the boss, this is a highly unlikely occurance in these days of equal rights. It is particularly notable that the girls were the ones least likely to select the foil.

BAD FOIL

Booklet 51, Item 22:

The film ran a little late during third period, so the students left without putting the chairs back in place. This was

(A) a polite thing to do because the teacher would not mind their leaving the chairs out.
(B) not a polite thing to do because those leaving or coming into the room could stumble over the clutter of chairs.
(C) not a polite thing to do because they could have left before the film was over.
(D) a polite thing to do because they knew that the students in the next period would have to move the chairs anyway.

The logic in foil "(A)" is not sound. Very few students, even at grade 7, are going to equate "politeness" and "not minding" on the part of another. Politeness is an action that results in appreciation not passive acceptance of inconsiderate action.

BOOKLET 52

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0509/16 *	08					X
0510/29	11	X (E)	X			
0510/30 *	08			X (D)		
0510/30	11			X (D)		
0510/31 *	08		X			
0515/21	08		X			
0518/5	08	X (E)	X			
0518/5	11		X			
0522/2 *	08			X (B)		
0522/2	11	X (S)		X (B)		

BAD FOIL

Booklet 52, Item 2:

> An office staff of about fifty people was planning to have a Christmas party. Which *ONE* of the following means of communication would be the *MOST* effective way to ensure that everyone in the office would know about the party?
>
> (A) posting a bulletin board announcement
> (B) telling a few workers to pass the word to the others
> (C) passing around a written notice
> (D) making telephone calls to employees' homes

Word of mouth is commonly known to be a poor way of sending information. It is often inefficient, inaccurate, and incomplete. Most people probably knew this, and so foil "(B)" is an unlikely response.

NO CLEAR EVIDENCE

Booklet 52, Item 16:

> Read the following descriptions of people interacting in work situations. Which description do you think is the *BEST* example of *RESPECTFUL* behavior between people of different races?
>
> (A) Mike, a black, and Charles, an anglo, work together on a government research project. When Mike and Charles disagree, Mike goes directly to the supervisor to complain.
> (B) Mr. Green, an anglo, and Mr. Swartz, a black, have worked next to each other on the same job for ten years. Mr. Green and Mr. Swartz have seldom talked to each other.
> (C) Fred Bear has worked in a factory close to the Indian reservation for five years. He has been a faithful and hard working employee. Mr. Bear wants to take Thursday off from work to attend a tribal celebration. The boss has threatened to fire him if he takes that day off.
> (D) Mei Lee lives and works in Chinatown. Sally Sands, a college student, has been hired as a summer employee at the plant where Mei Lee works. Mei Lee introduced Sally to other workers on the job.

This is a very easy item. Sometimes this results in chance patterns of cultural variation. There doesn't seem to be any clear evidence of bias in the item. It is not clear why girls would be drawn to "(B)" and boys drawn to "(C)".

BAD FOIL

Booklet 52, Item 30:

Listed below are attitudes or beliefs expressed by some people. Select the belief/attitude which you think *MOST* indicates prejudice.

(A) People should be judged by their performance.
(B) People with long hair are generally lazy.
(C) It is difficult to know what people are really like when you first meet them.
(D) Most women have good eyesight.

Foil "(D)" is an inappropriate foil that seems to have been thrown in because something better could not be thought of. It is better to reduce the number of foils instead of including one that is ineffective. A better foil might be "It is easy to judge people after you first talk to them."

DIAGNOSTIC

Booklet 52, Item 31:

Which *ONE* of the following statements describes what might happen if the people of one race are *PREJUDICED* against the people of a different race?

(A) Communication will increase between the people of different races.
(B) People of different races will like each other better.
(C) Clashes between the people of different races will decrease.
(D) Understanding between people of different races will be hard to achieve.

This item has foils that are diagnostic of a clear understanding of what is meant by "PREJUDICED". A response to a wrong foil shows that there is confusion on the part of the student about the term and indicates the direction of the confusion.

BOOKLET 61

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0601/1 *	10	X (E)		X (D)		
0601/2	10	X (E)	X			
0601/3A	07		X			
0601/3A	10		X			
0601/3B	07		X			
0601/3C *	07		X			
0601/3C	10		X			
0601/3D	07		X			
0605/7	07	X (E)	X			
0605/7	10		X			
0605/9	07		X			

DIAGNOSTIC

Booklet 61, Item 3:

If you have problems or need advice, people with professional training can often help you. Darken (A) below if there is someone on the school staff to whom you would feel free to go if you had a problem concerning the following. Darken (B) for the others.

(A) (B) a. your schoolwork
(A) (B) b. your home life
(A) (B) c. a career choice
(A) (B) d. your personal life

This item has no correct answers. Reports based upon the data collected are useful in diagnosing the willingness of a student to utilize school staff for various types of personal problems. It is certain that willingness to use school staff is going to differ among all types of students. Data on this item could be used to identify students who need to be made aware of the types of help that a staff member can give as well as their willingness to give the help.

CULTURAL BIAS + BAD FOIL

Booklet 61, Item 1:

> Jeanie has been worried about her relationship with her boyfriend. Her parents don't like him and this adds to the problems that already exist. Jeanie cannot concentrate in school and her teacher is worried about her work. The only person she sees and talks to almost every day is her young aunt who happens to be a counselor at her school. She feels that she must talk to someone about her problems.
>
> Of the following, select the *ONE* person with whom Jeanie would probably first discuss the problem.
>
> (A) her parents
> (b) her boyfriend
> (C) her aunt, the school counselor
> (D) her teacher

This item is very highly tied to cultural background. The person a student is most likely to go to first to discuss a problem differs according to the background (ethnic and otherwise) of the child. For example, blacks were more likely to discuss the problem with parents [foil "(A)"], Mexican-Americans were about evenly divided between parents "(A)" and boyfriend "(B)" while "others" (mostly anglos) were heavily attracted to foil "(B)". Moreover, foil "(D)" is a very unlikely choice.

BOOKLET 62

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0602/1A	08		X			
0602/1A	11		X			
0602/1B	08		X			
0602/1B	11		X			
0602/1C	08		X			
0602/1C	11		X			
0602/1E	08		X			
0602/1E	11		X			
0602/1F	08		X			
0602/1G	08		X			
0602/1G	11		X			
0602/1I	08		X			
0602/1I	11		X			
0602/1J	08		X			
0602/1J	11		X			
0602/1K	11		X			
0602/1M	08		X			
0602/1N	08		X			
0602/1O	08		X			
0602/1P	08		X			
0602/1R	08		X			
0602/1S	08		X			
0602/1S	11		X			
0602/1U	08		X			
0602/1U	11		X			
0602/1V	08		X			
0603/7D	08		X			
0603/7D	11		X			

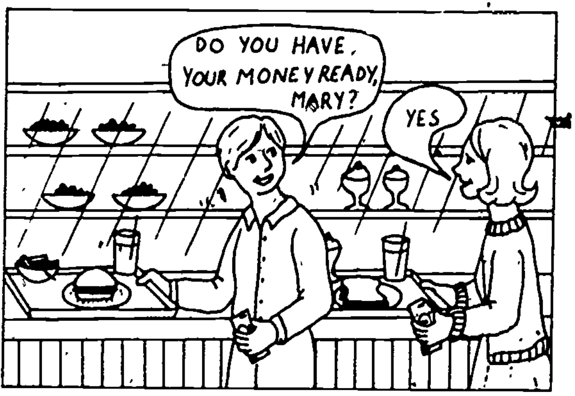
BOOKLET 11

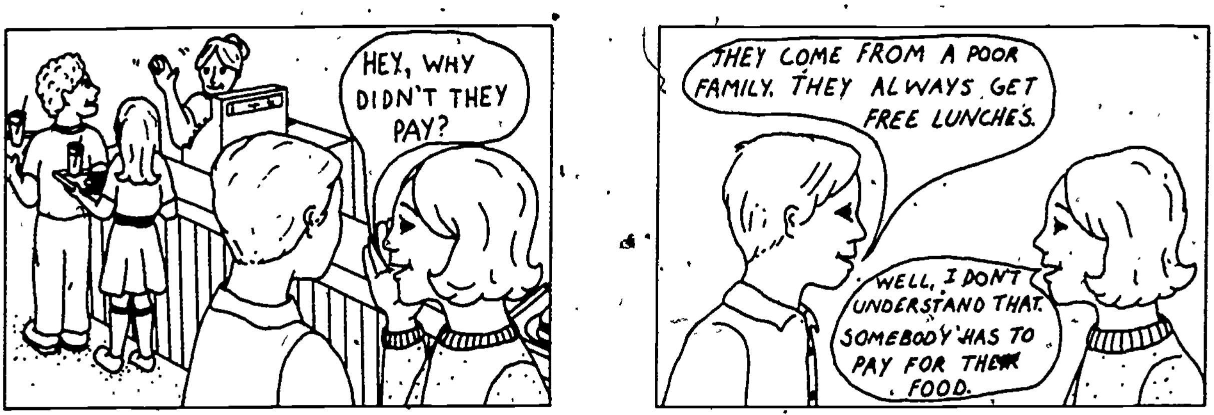
Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0603/7M	08		X			
0603/7M	11		X			
0603/7N	11		X			
0603/7G	08		X			
0603/7G	11		X			
0603/7A	08		X			
0603/7A	11		X			
0603/7O	08		X			
0603/7O	11		X			
0603/7H	11		X			
0603/7F	11		X			
0603/7Q	08		X			
0603/7Q	11		X			
0603/7I	08		X			
0603/7I	11		X			
0603/7R	11		X			
0603/7E	08		X			
0603/7E	11		X			
0603/7J	08		X			
0603/7J	11		X			
0603/7C	08		X			
0603/7C	11		X			
0603/7K	08		X			
0603/7F	08		X			
0603/7L	08		X			
0603/7L	11		X			
0604/3	08	X (E)		X (B)		
0604/4	11			X (C,D)		
0707/8 *	08	X (E)		X(A,B,D)		
0707/9	08			X (B,C)		
0707/12 *	11		X	X (C)		
0709/10 *	11	X (E)		X (A)		
0711/13	08	X (E)		X (B)		

CULTURAL BIAS + BAD FOIL

Booklet 62, Item 8:

YOUR COUSIN MARY HAS COME TO VISIT YOU FOR A WEEK. SHE LIVES ON A RANCH AND GOES TO A SMALL COUNTRY SCHOOL. IT IS EXCITING FOR HER TO SEE YOUR CITY AND YOUR SCHOOL.....




HEY, WHY DIDN'T THEY PAY?
THEY COME FROM A POOR FAMILY. THEY ALWAYS GET FREE LUNCHES.
WELL, I DON'T UNDERSTAND THAT. SOMEBODY HAS TO PAY FOR THEIR FOOD.

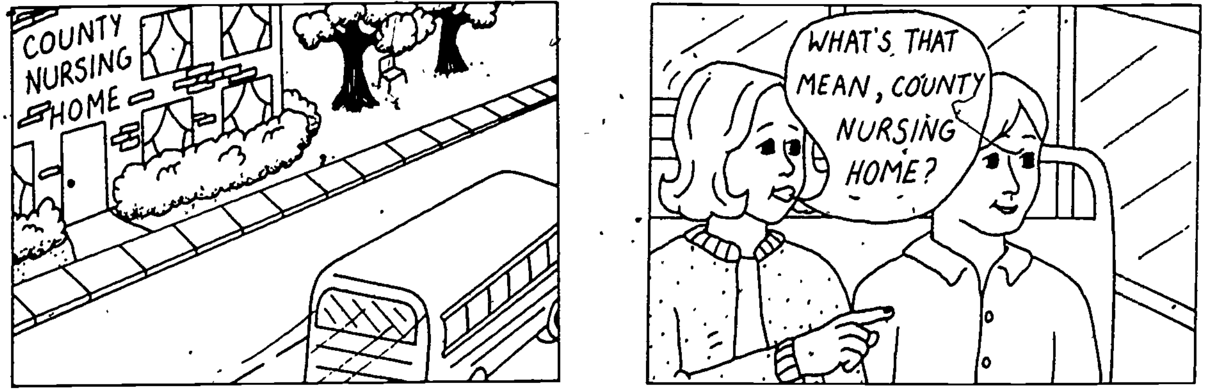
How were the two lunches probably paid for?

(A) The cashier liked them and paid.
(B) The teachers would pay for the lunches.
(C) The school kept a special lunch fund.
(D) Other students joined in and paid for the lunches.

This is a inappropriate item. None of the foils are viable choices. Moreover, blacks were attracted to Foil A.

BAD FOIL + DIAGNOSTIC

Booklet 62, Item 12:



The nursing home is used for what or whom?

(A) older people who cannot take care of themselves
(B) babies who cannot take care of themselves
(C) young plants that people buy for homes and businesses
(D) people who are training to be nurses

Although foil "(C)" could be a common word confusion (nursing home for nursery) no one is attracted to it. The foil should be replaced with some other idea.

The other two foils are good diagnostic statements which would help identify student problems.

BOOKLET 71a

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0702/15	07			X (B)		X
0702/16 *	10	X (E)		X (D)		
0702/20	10	X (E)		X (A)		
0704/10 *	07	X (E)		X (A)		
0704/11 } *	07		X			
0704/11	10		X			
0704/13	10			X (A)		
0705/1 *	07		X	X (D)		
0705/3	07		X	X (C)		
0705/4	07		X			
0705/6	10			X (D)		
0705/7 } *	07			X (C)		
0705/7	10			X (C)		
0705/8 *	10					X

BAD FOIL + DIAGNOSTIC

Booklet 71a, Item 1:

> In the United States, we all have "freedom of speech." This means that we have
>
> (A) freedom to say anything we want to, anytime, and about anyone.
> (B) freedom to speak our thoughts, but not to put them into printed form.
> (C) freedom to say or print what we want, as long as it is not false information.
> (D) freedom to appear on radio or television whenever we want.

Foil "(D)" is a bad foil which attracts no one. It is clear that nearly everyone knows that it is difficult to appear on television.

The other foils are diagnostic in that they are very common misconceptions, or misinterpretations. They can be specifically taught by the teacher.

BAD FOILS

Booklet 71a, Item 7:

> The United States Constitution guarantees its citizens many rights and freedoms. However, citizens can only have these rights as long as they
>
> (A) remain registered voters in the U.S.
> (B) do not infringe upon the rights of others.
> (C) are either fully employed or are in school.
> (D) have never been arrested for a major offense.

Foil "(C)" is very unattractive to students of both grades. Why not use something like "are living in their home town when the election is held" or "are living in the United States?"

NO CLEAR EVIDENCE

Booklet 71a, Item 8:

> You have a number of rights that are granted by the government. Which *ONE* of the following is *NOT* one of those rights?
>
> (A) the right to free speech
> (B) the right to print money
> (C) the right to an education
> (D) the right to a trial by jury

~~There is no clear evidence in the item or the response pattern about the strong evidence of cultural variability.~~ Part of this may be due to the fact that the item is very easy, thus leaving only scant numbers to respond hapazardly to the foils.

BAD FOIL + CULTURAL BIAS + DIAGNOSTIC

Booklet 71a, Item 10:

> Harold has a good job working as a delivery man for a parcel firm. Gina works as a checker in a supermarket. They have been married for four years. Last year they borrowed money from the bank to buy a small home. Gina is thinking about quitting her job. The *MOST* probable result of Gina's not working would be that
>
> (A) Gina and Harold would probably concentrate on furnishing their new home more quickly.
> (B) the bank would probably repossess their house.
> (C) Gina and Harold would probably use their charge accounts more.
> (D) Gina and Harold would probably buy fewer luxury-type items.

There is a large minority response to foil "(C)" which encourages the increased use of charge accounts. Although this foil may indicate an ethnic variation, it could be considered diagnostic of a need to educate certain groups to the need to utilize charge accounts with care.

Foil "(A)" is totally ineffective. A stronger foil should be written to replace it.

DIAGNOSTIC

Booklet 71a, Item 11:

> Some people do not have work to do. Which *ONE* of the following is the *MOST LIKELY* effect of not working?
>
> People who do not work
>
> (A) will probably not experience the personal satisfaction they might have experienced doing a job.
> (B) will be totally unhappy and very poor.
> (C) will want to begin working at any kind of job right away.
> (D) will feel that no working is so great that they will encourage everyone they know not to work.

The foils in item 11 direct the teacher to information that would help the student realize the value of the "right" job. It also will help direct teaching to orient the students toward understanding 1) What are the effects of not working?, 2) What are the effects of having a job you don't like?, or 3) Why would you not like being out of work?

BOOKLET 71b

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0709/1	10					X
0709/15	07		X			
0711/11	07		X	X (D)		
0716/4 *	10			X (C)		
0716/5	07		X			
0716/8 }*	07	X (S)				
0716/8	10	X (S)				

BAD FOIL

Booklet 71b, Item 4:

For many years, thousands of women have worked in the lower-paying jobs in business. Regardless of their experience or education, it has been difficult for them to advance to, or be hired for, management-level jobs. The women s liberation movement has done much to expose this waste of human resources and to make such discriminatory practices unlawful.

Which *ONE* of the following is an improvement in our economic system that should result from the efforts of the women's liberation movement?

(A) People will be hired according to qualifications.
(B) More men will choose to do manual labor.
(C) Secretaries will be paid lower wages.
(D) More men will be hired as business managers.

Foil (C) is a direct contradiction to the information in the paragraph. This has resulted in an extremely low response to the foil. A different incorrect response such as only women will be hired into the higher paying jobs" would be more appropriate than the current foil.

CULTURAL BIAS

Booklet 71b, Item 8:

People have become more aware of the frequent inequality in wages of paid male and female employees doing the same type of work. What effect has this increased awareness had on our economic system?

(A) upward adjustment of some women's wages
(B) reduction of the number of employed women.
(C) downward adjustment of the gross national product

This item seems to be overly negative toward the effects of the women's movement. There seems to be little explanation that can clarify the pattern of incorrect responses, however.

BOOKLET 72a

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0707/9	11					X
0708/1 *	11		X	X (D)		
0708/6 *	08		X			
0713/12 *	08	X (E)	X			

DIAGNOSTIC

Booklet 72a, Item 1:

Which *ONE* of the following is paid for by state taxes?

(A) postal service
(B) national defense
(C) highway maintenance
(D) telephone service

This is an excellent example of a diagnostic foil item. It identifies a misunderstanding of the source of financing for various public agencies. Any student not understanding taxes would likely be drawn off by the foils, giving teachers information to be used to correct the deficiencies.

DIAGNOSTIC + CULTURAL BIAS

Booklet 72a, Item 12:

> Which *ONE* of the following quotations reflects an individual's positive attitude toward participation in the economic system of the United States?
>
> (A) "Big businesses cheat on their taxes, so I do too."
> (B) "Irish wool is of better quality than local wool."
> (C) "I've invested my savings in a local corporation."
> (D) "I think that I should be able to get money any way I can."

Blacks are drawn to "(A)" and Mexican-Americans to "(D)". If minority individuals have experienced discriminating practices, this might explain the aforementioned variation in foil responses. However, all of the foils are diagnostic and can be used as instructional guide lines. An improved correct answer could reduce cultural variation.

DIAGNOSTIC

Booklet 72a, Item 6:

> Which *ONE* of the following levies taxes?
>
> (A) counties
> (B) churches
> (C) banks
> (D) stores

This is a very simple item which requires knowledge of two things. 1) What are taxes? and 2) What charges are legitimate for counties, churches, banks, and stores to make for services rendered? Each one collects money but only one levies taxes, the counties.

BOOKLET 72b

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0720/14 *	11			X (A,C,D)	X (Item)	
0722/8 *	08	X (S)	X			
0722/18A	11		X			
0722/18H	08		X			
0722/18H	11		X			
0722/18D	11		X			
0722/18I	08		X			
0722/18I	11		X			
0722/18B	08		X			
0722/18B	11		X			
0722/18J	08		X			
0722/18J	11		X			
0722/18F	08		X			
0722/18F	11		X			
0722/18K	08		X			
0722/18K	11		X			
0722/18E	08		X			
0722/18L	11		X			
0722/18M	08		X			
0722/18G	08		X			
0722/18G	11		X			
0722/18N	11		X			
0723/10	11	X (E)		X (B,C,D)		

CULTURAL VARIATION

Booklet 72b, Item 8:

> Which *ONE* of the following is an example of soil conservation?
>
> (A) oilwell drilling
> (B) contour farming
> (C) open pit mining
> (D) lumbering

This item is oriented toward jobs that have been traditionally male-dominated. As a result, the content makes it difficult for girls to know enough to respond to any answer. This variation could be corrected by improving the instructional program.

BAD FOIL

Booklet 72b, Item 14:

> Which *ONE* of the following would be a good safety practice for employees to observe when working in a factory?
>
> (A) Employees should keep bathroom doors locked at all times.
> (B) Employees should be able to follow fire drill procedures quickly.
> (C) Employees should bring chairs from home if those supplied by the company are uncomfortable.
> (D) Employees should organize and demand higher wages.

The item components are not related in a meaningful way. None of the foils relate to anything tied to a safety practice.

BOOKLET 81a

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0802/3	07		X			
0802/4	10		X	X (D)		
0802/6 *	07	X (E)				
0808/9	07		X			
0808/10	07		X			
0808/12	07		X			
0810/13 *	07		X	X (A)		
0810/18	07		X			
0816/21 *	10		X			
0816/22	10		X			

DIAGNOSTIC

Booklet 81a, Item 6:

> With new machines and computers changing routine jobs, some assembly line and office workers may be fearful of
>
> (A) overproduction of goods.
> (B) losing their jobs.
> (C) increase in cost of goods.
> (D) longer working hours.

The Mexican-Americans were attracted to foil "(D)", "longer working hours." This is probably due to the lack of knowledge of computers and a lack of experience. For many, it could mean that longer working hours would be required to learn how to use the machinery. This would be a common misconception for someone who has not been oriented toward mechanization.

DIAGNOSTIC - BAD FOIL

Booklet 81a, Item 13:

> Which *ONE* of the following is the *BEST* reason why many companies choose to pay salesmen on the basis of how much of the company's product they sell?
>
> (A) Such pay will not show on the company's records.
> (B) The salesmen will make more money if they are paid that way.
> (C) The salesmen will sell more of the product if they are paid that way.
> (D) The salesmen will not have to be paid any fringe benefits.

Foil "(A)" is not attractive to the students. It seems that all students are aware that all payroll that is "regular" is recorded both in company records as well as in tax records. Foils need to be likely choices.

The other two foils provide likely reasons for paying a commission. A student who choses one of these foils has a definite lack of understanding that can be rectified with some instruction.

DIAGNOSTIC

Booklet 81a, Item 21:

> Resources become goods when they are made ready for human use. Water may be considered "goods" rather than a resource when
>
> (A) it is flowing in a river.
> (B) a dam is built to stop flooding.
> (C) it is piped into your home.
> (D) it is polluted by chemicals from factories.

Each of the foils represents a logical misunderstanding. Blacks selected foil "(D)" frequently. If the student did not know the term "goods", he/she may be drawn off by the term "pollution".

BOOKLET 81b

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0817/2 *	10		X			
0817/5	07		X			
0817/5	10		X			
0817/7	07		X			
0820/20	07		X			
0826/10	07		X	X (D)		
0826/11	07	X (E)				
0827/15 *	07		X	X (C)		
0827/15	10		X	X (C)		
0827/16	07		X			
0827/18	07		X			
0827/18	10		X			

DIAGNOSTIC

Booklet 81b, Item 2:

> Eight big logging companies raised the price of raw lumber by a large amount. The housing industries felt they had no choice but to raise the price of the homes they offered for sale. What effect did the price change probably have on the demand for the houses?
>
> (A) The demand was probably greater.
> (B) The demand was probably less.
> (C) The demand was probably the same.

The supply and demand items are all diagnostic if they have only the three possible results, go up, stay the same, go down. This is a good example of an item which diagnoses a problem which instruction can supposedly rectify. A wrong response gives a clear direction for instruction.

BAD FOIL + DIAGNOSTIC

Booklet 81b, Item 15:

> A summer heat wave in New York caused people much discomfort. More people began to drink lemonade during the days of the heat wave. What effect did this action of consumers have on production?
>
> (A) More lemons were grown.
> (B) More lemon trees were planted immediately.
> (C) More lemons probably were used to make scented wax.
> (D) More lemons probably were used to make frozen juice.

Foil "(C)" is inappropriate. If there aren't enough lemons for lemonade, surely they won't have enough to increase the production of lemon scented wax. This foil should be replaced.

The other two foils are diagnostic of a lack of understanding that it takes a very long time to grow lemon trees which will bear fruit. These misunderstandings are important keys to additional instruction.

BOOKLET 82

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0804/19	08		X			
0804/20	08		X			
0804/20	11		X			
0804/21	08		X			
0804/21	11		X			
0811/7 *	08		X			
0811/7	11		X			
0811/8	08		X			
0811/11	08		X			
0821/3 *	11	X (E)	X			
0831/15	08		X			

DIAGNOSTIC + CULTURAL BIAS

Booklet 82, Item 3:

> The Peterson family had to make an important decision concerning their budget. They had not expected to have to spend their savings for repairs on their storm-damaged house. The family had to decide whether to borrow money for their planned trip to the Rocky Mountains or to spend their time at home and save for next summer's trip. The family realized that they would need to work more in order to repay any money borrowed. After discussing the problem, the family voted to take the trip they had planned.
>
> Which *ONE* of the following was the major factor affecting the decision of the Petersons'?
>
> (A) They placed a high value on saving money.
> (B) They placed a high value on vacation travel.
> (C) They placed a low value on home repairs.
> (D) They placed a low value on work.

Foil "(C)" is very attractive to blacks, this may be due to cultural background factors. Moreover poorer readers probably didn't realize that the repairs were already done. If the repairs were not done, foil "(C)" could be considered correct also.

Booklet 82, Item 7:

When Andy was 22, he bagan working for the Pioneer Motor Freight Company as a dock worker. His job was to load and unload trucks. Andy's gross income per week was $160.00 His take-home pay after deductions was $110.37.

After working at Pioneer for two years, Andy no longer worked on the docks. He drove the company trucks regularly. Then his gross income per week was $200.00.

As a result of making more money, Andy had

(A) more money deducted from his paycheck than before.
(B) less money deducted from his paycheck than before.
(C) the same amount of money deducted as before.

The cultural variability indices are very high for this item (especially at grade 11 where V = 0.61). There is a possibility that this was due to cultural background factors. A more likely explanation, however, is that the item is diagnostic of the students' understanding of the relationships of changes in income to the amount of deductions.

BOOKLET 91

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0904/3 *	10			X (D)		

BAD FOIL

Booklet 91, Item 3:

Which *ONE* of the following courses would probably be *MOST* helpful to you if you wished to be a bank clerk after graduating from high school?

(A) chemistry
(B) bookkeeping
(C) home economics
(D) American history

This is a very easy item. This may be due to the inappropriate foils used or because all tenth graders know the requirements for being a bank clerk. In particular, foil "(D)" attracted no students.

Outcome/ Item Number	Grade	Bias(Type)	Diagnostic	Bad Foil(#)	Bad Format	No Clear Evidence
0901/1 *	11			X (A,B)		X
0902/15	08		X			
0902/15	11		X			
0903/9C	08		X			
0903/9G	08		X			
0903/9G	11		X			
0903/9A	08		X			
0903/9H	08		X			
0903/9D	08		X			
0903/9I	08	X (S)	X			
0903/9I	11	X (S)	X			
0903/9J	08		X			
0903/9K	11		X			
0903/9M	11		X			
0903/9N	08		X			
0903/9N	11		X			
0903/9O	08				X	
0915/12 *	08			X (D)		

DIAGNOSTIC

Booklet 92, Item 1:

To be a legal secretary, one need *NOT* be able to

(A) spell correctly.
(B) type accurately.
(C) have a good command of the English language.
(D) debate a legal case.

This item has two foils ["(A)" and "(B)"] are not operating. For example, foil "(B)" is too easy because there is a tendency to associate typing with secretarial positions, whether it be legal secretary or not.

DIAGNOSTIC

Booklet 92, Item 9:

On your Answer Sheet, darken (A) for all those learning experiences *OUTSIDE* of school which you feel are important to you in making a decision about your career. Darken (B) for the others.

(A) (B) a. talking to parents about their jobs
(A) (B) b. talking to friends about their present or future jobs
(A) (B) c. seeing examples of jobs on televison
(A) (B) d. seeing people you don't know working on various jobs
(A) (B) e. reading books or magazines about people with various jobs
(A) (B) f. having had experiences with jobs after school or during the summer
(A) (B) g. talking with relatives about their jobs
(A) (B) h. belonging to a club or group
(A) (B) i. participating in a sport
(A) (B) j. traveling or moving to another city
(A) (B) k. taking lessons in painting, piano, guitar, dancing, etc.
(A) (B) l. visiting a job location
(A) (B) m. working at home on a hobby or project
(A) (B) n. doing volunteer work (such as Candy Striper) in the community
(A) (B) o. having had no outside school experience which has been important

This is a multipart item with no correct answers. The results are primarily used for survey purposes to identify various learning experiences *OUTSIDE* of school that students have had. For this reason, the responses are diagnostic of student experience.

BAD FOIL

Booklet 92, Item 12:

> Reading the editorial sections of newspapers will give you
>
> (A) individuals, views on various issues.
> (B) factual information only.
> (C) the best information available on various issues.
> (D) information concerning television scheduling.

Few students chose foil "(D)". This may be because the content of the foil is very different from that of the other three. One alternative would be to use something like "a summary of the most important events of the week" or "current book review information". The second suggestion would identify those who confused book editor with editorial.

APPENDIX I

SIGNIFICANCE TESTS FOR ISI AND ESI AND OSI

One may test the statistical significance of the difference in the ISI (ESI, OSI) for the experimental and comparison groups, by the following statistic:

$$Z = \frac{(P_2 - P_1) - (P_2' - P_1')}{\sqrt{\hat{\sigma}^2_{P_2 - P_1} + \hat{\sigma}^2_{P_2' - P_1'}}}$$

where

$$\hat{\sigma}^2_{P_2 - P_1} = \frac{P_1(1-P)}{n-1} + \frac{P_2(1-P_2)}{n-1} + \frac{2P_1P_2}{n},$$

$$\hat{\sigma}^2_{P_2' - P_1'} = \frac{P_1'(1-P_1')}{n-1} + \frac{P_2'(1-P_2')}{n'-1} + \frac{2P_1'P_2'}{n},$$

$$P_i = n_i/n, \quad i = 1, 2,$$

$$P_i' = n_i'/n', \quad i = 1, 2,$$

and (n, n_1, n_2) refers to the experimental (instruction) group, while (n', n_1', n_2') refers to the comparison (no instruction) group.

The above Z statistic is approximately (for "large" n, e.g., n greater than 30) distributed according to a standard normal, under the null hypothesis (true index for experimental group is equal to true index for the comparison group). The formulation utilizes well-known properties of the multinomial distribution, the formula for the variance of a linear combination of correlated variables, and the central limit theorem (cf., e g , Wilks, 1962.). A single-tailed test may be conducted and the upper-tail significance probability calculated A significant Z (e.g., at .05 or .01 level) indicates (i) the ISI of the experimental group is significantly higher than that of the comparison group and, (ii) the magnitude and statistical significance of the ISI may be attributed to the item's sensitivity to *instruction*, and not some extraneous factor, such as maturation (cf., Campbell and Stanley, 1963.).

An analogous Z test may be conducted for the ESI and OSI (just substitute m and N respectively for the n in the above formulation).

AN ALTERNATIVE PROCEDURE FOR DEVELOPING THE SURVEY TEST

The following heuristic procedure is an alternative to the stepwise regression analysis:

1. Compute the mean outcome score for (i) students in upper grade (10 and 11) and (ii) students in lower grade (7 or 8) for each outcome. Denote these by $\overline{Y_1}$ and $\overline{Y_2}$, respectively. For each item (within a given outcome) and each student, perform the following analysis:

   1. If the student's score is *less* than the mean outcome score, (either $\overline{Y_1}$ or $\overline{Y_2}$ depending on which level the student is at), and he answers the item *correctly*, assign the rating 0.
   2. If the student's score is *less than* the mean outcome score and he answers the item *incorrectly*, assign the rating 1.
   3. If the student's score is *greater than* the mean outcome scores and he answers the item *correctly*, assign the rating 1.
   4. If the student's score is *greater than* the mean outcome score and he answers the item *incorrectly*, assign the rating 0.
   5. If the student's score is exactly equal to the mean score assign the rating 1/2.
   6. Sum the ratings for each item, over all students and over both grade levels.
   7. Select the item with the highest rating, item M_1, say, (Denote this rating R_{M1}).
   8. Subtract the score for item M_1 ($X_{M1} = 0$ if the response is incorrect, $X_{M1} = 1$ if the response is correct) from the students outcome scores and recompute the ("adjusted") mean outcome scores, for the 2 grade levels.
   9. Repeat steps 1 through 7 with the adjusted means, i.e., perform steps 1 through 7 *after eliminating item M_1*.
   10. Select the item (M_2) with the highest "adjusted" rating, using the "adjusted" means. Denote this rating R_{M2}.
   11. Compute the ratio R_{M2}/R_{M1}.

Select the items M_1 and M_2 only if R_M and the ratio R_{M2}/R_M are sufficiently high.

THEORETICAL BASIS FOR THE USE OF 'SELF-WEIGHTING' ESTIMATORS IN THE FIELD TEST DATA

The determination of the number of schools to be selected in each stratum is by "proportional allocation" with respect to the number of students in each strata. This may be stated formally as follows:

$$n_h = \left(\frac{M_{ho}}{M_o}\right) n ,$$

where
n = number of schools in the sample (for a given instrument at a given grade level)
M_{ho} = number of students in all schools in stratum h (at a given grade level)
M_o = total number of students in the state (at a given grade level)
n_h = number of schools to be selected from stratum h (for a given instrument at a given grade level)

The selection of schools within strata is with probability proportional to size of school (p.p.s.). This may be stated mathematically as follows:

$$(2) \qquad Z_{hi} = \frac{M_{hi}}{M_{ho}}$$

$i = 1, 2, \ldots, N_h$, $h = 1, 2, 3$, where Z_{hi} is the probability of selecting the ith school in stratum h, and M_{hi} is the number of students in the ith stratum h, an N_h is the number of schools in stratum h.

The classical unbiased estimator of the (true) p-value (proportion getting an item correct) is.

$$(3) \qquad \hat{p} = \sum_h \frac{M_{ho}}{M_0} \left\{ \frac{1}{n_h \; M_{ho}} \sum_i \frac{M_{hi} \; P_{hi}}{Z_{hi}} \right\}$$

(cf., Cochran, 1963.) Substituting (1) and (2) into (3), one obtains

$$(4) \qquad \hat{p} = \sum_{hi} p_{hi}/n ;$$

Thus, p is the (unweighted) average of the p-values for each school. When m_{hi} = m, i.e., the number of students selected from each school is the same, (4) reduces to:

$$(5) \quad \hat{p} = \sum_{hi} x_{hi} \Big/ \sum_{hi} m_{hi} \text{ ,}$$

where X_{hi} = number of students answering item correctly in the ith school in stratum h. The estimator $\hat{p}$ in (5) is the simple proportion, the number answering correctly divided by the total number of students taking the test. We selected one classroom per campus. Although classroom size will vary somewhat across schools, it was the judgment of WLC/MRC statisticians that this would not markedly affect the estimates obtained using (5). The usual estimates of point biserials and KR-20 reliability coefficients were employed. These are considered as measures describing various characteristics of the tests, rather than estimates of population parameters. Thus weighting factors were not considered for these measures.

APPENDIX L
CRITERIA FOR ITEM ACCEPTABILITY

*Review Strategy Being Utilized

CRITERIA AREA	RATIONALE/DESCRIPTION	*	CRITERIA
I. Except for technical terms, each item must not be above the sixth-grade reading level.	The most important standard that must be used for judging items is — does it communicate to students. Judgements about the readability of items will be made throughout the development of the instrument. Reading level should be at the sixth grade.	A.	1. Any technical term used in items will contain an explanation of the term within the item.
		B.	2. At least 4 out of each 5 students will indicate they had no trouble understanding the item.
		C.	3. Responses from educators must indicate 90% agreement that the item is readable by 90% of 8th grade Texas students.
		D.	4. Not more that 10% of the students will indicate difficulty in understanding the item.
		E.	5. Not more than 5% of students will indicate difficulty in understanding the item.
II. Each item must be a direct measure of some knowledge, skill, or attitude called for in the objectives.	A purpose for developing the measurement instruments is to determine the status of students in relationship to the Basic Learner Outcomes. There must be a clear and direct linkage among learner outcomes, objectives, and items. If an item is a linked component of a direct measure of an outcome, it may be necessary to validate responses to the direct measure.	A.	1. Each item will be checked for association with the objective and outcome. Twenty-five percent of the groups questioning item-objective relationship will be cause for revision. Seventy- five percent of this group questioning this relationship will be cause for item elimination.

*Review Strategy Being Utilized

CRITERIA AREA	RATIONALE/DESCRIPTION	*	CRITERIA
		C.	2. Questions directed toward item-objective relationships will be included in materials developed for regional review meetings. Items may not receive more than 20% of the responses rejecting the item-objective relationship. We want to consider the revision approach using objectors comment.
		E.	3. Correlation of performance on each item with performance on the group of items measuring the same objective will be calculated. Point biserial correlations below 0.3 will be investigated further.
III. Each item should be inoffensive to reviewers representing students and educators.	The measurement instruments developed from the items should be useful to students and teachers throughout Texas. The major subcultures to consider in the development of items are those of the blacks, Mexican-Americans, both sexes and various SES groups. It may be impossible to develop instruments that do not reflect any culture bias, however, items should be reviewed to determine and eliminate as much as possible, any wording, examples, etc. that might not be understood by or be offensive to persons that are members of the above-mentioned subcultures.	A.	1. Each item will be screened for cultural aspects.
		B.	2. Students selected for student review will include: ● at least 2 ethnic minority group students, ● at least 2 persons of each sex Not more than 1 student may object to an item.
		C.	3. No more than 10% of the educators may judge an item objectionable.

*Review Strategy-Being Utilized

CRITERIA AREA	RATIONALE/DESCRIPTION	*		CRITERIA
		D.	4.	No more than 10% of the responses may indicate offensiveness or bias.
		E.	5.	No more than 5% may indicate offensiveness.
IV. Each item must be constructed so that the format and directions are simple and clear.	Meeting this objective will require a mindset similar to that expressed in objective 1 — Does it communicate with students? The items are meant to measure certain skills, attitudes, and knowledge called for in the learner outcomes and are not meant to measure the student's ability to decipher items.	A.	1.	Items will be reviewed by PACE and ACE for format simplicity and clearness.
		B.	2.	At least 4 out of 5 students who participate in student review must indicate they had no problems understanding item directions and/or format.
		C.	3.	Response from educators at the regional meetings must indicate 90% agreement that the eighth grade Texas students could understand the item direction and format.
		D.	4.	Not more than 10% of the students in item tryouts will indicate difficulty with understanding item directions as determined by interview.
		D.	5.	A record of student responses to items will be kept. Any response pattern questioned will be investigated to determine if a possible cause is related to difficulty in understanding directions.

*Review Strategy Being Utilized

CRITERIA AREA	RATIONALE/DESCRIPTION	*	CRITERIA
		E.	6. Not more than 10% of students participating in field testing will indicate difficulty with understanding item directions.
		E.	7. Any questionable response pattern will be investigated further to determine if a possible cause is related to difficulty in understanding directions.
V. Each item must be clearly scorable by an accompanying key, guide, or scheme for constructing a guide. In particular, any teacher ought to be able to score the responses or observe the behaviors.	It is planned that the measurement instruments developed from the items will be used by teachers of varying expertise and backrounds. Plans for scoring items, whether they be checking student responses or observing behavior should be as clear and as simple as possible. Any scoring schemes should keep any special training to a minimum. The scoring of items should require no more of the teachers time than is necessary for checking responses, tallying responses, and recording the tally.	A.	1. Items will be checked by PACE and ACE for simplicity of scoring and for feasibility and/or simplicity of any observations/checklists that teachers would be asked to do.
		B.	2. Any problems with scoring will be reported.
		C.	3. Responses from educators must indicate that 90% agree that the scoring of checklist used with the item is feasible for teachers to score.
		D.	4. No more than 5% of the teachers should express any difficulties on applying the scoring grade.

*Review Strategy Being Utilized

CRITERIA AREA	RATIONALE/DESCRIPTION	*	CRITERIA
VI. Each item must be accompanied by clear instructions for administration.	The measurement instruments that result from the item development will be used by a wide variety of teachers, therefore items must be	A.	1. Items will be reviewed for ease of administration and clearness of instruction by MRC, PACE, and ACE.
	● easy to adminster with a minimun amount of training necessary for the teacher	C.	2. Responses from educators must indicate a 90% agreement that the items are easy to administer and the item instructions are clear.
	● contain clear instructions for administration	D.	3. Questions will be asked of educator-administrators about ease of administration and clearness of directions. No more than 15% of the responses should indicate any difficulty.
		E.	4. 50 teachers will be selected at random to respond to questions about administration of items (instruments) and clearness of directions. If more than 5 teachers indicate problems, revisions will be made.

APPENDIX M

Materials Used in Conducting Spring Field Test

Texas Education Agency

- STATE BOARD OF EDUCATION
- STATE COMMISSIONER OF EDUCATION
- STATE DEPARTMENT OF EDUCATION

201 East Eleventh Street
Austin, Texas
78701

Letter sent to all Executive Directors of the twenty education service centers

The State Board of Education has identified Career Education as one of the top priorities for development. As a part of this priority, a set of important student outcomes in Career Education has been identified. Based on these student outcomes, we are now building a measurement system for Career Education that is described in the attached summary. We believe this system will provide information useful to you and your staff in the counseling and instruction of your students.

The measurement system is in the developmental stage. Test items have been written and grouped into trial instruments at two levels of student development. In order to insure that these instruments are of the highest possible quality, it is essential that they be pilot tested with a sample of Texas students in grades seven through eleven starting in mid-March.

We have drawn a random sample of school campuses that represent different types of Texas students. One or more campuses in your school district are included in the sample. Would you be willing to cooperate with us in this effort by allowing some of your students to take one of these instruments? It would require less than one hour of class testing time (an ordinary class period) for each participating student and an additional one and one-half hours time for each teacher to prepare for the administration of the test.

Attached is a list of campus(es) and number of classrooms requested to participate in your school district. I would appreciate it if you would return the enclosed form to let us know whether or not you can assist us. If you have additional questions or would like further information, please contact Keith Cruse or Bill Fischer of the Division of Program Planning and Needs Assessmen (512/475-2066).

I hope that you will feel that your school district can work with us on this important effort to strengthen the opportunity for all students in Texas to achieve the essential outcomes in Career Education.

Very truly yours,

M. L. Brockette
Commissioner of Education

"An Equal Opportunity Employer"

Texas Education Agency

201 East Eleventh Street
Austin, Texas
78701

- STATE BOARD OF EDUCATION
- STATE COMMISSIONER OF EDUCATION
- STATE DEPARTMENT OF EDUCATION

Letter sent to all Executive Directors of the twenty education service centers

As we described to you earlier in Texas Elementary and Secondary School Planning Council meetings, one of the Agency activities for the priority area of Career Education is the development of a measurement system for the "Basic Learner Outcomes for Career Education." Plans for the March administration of this measurement system have been revised to increase the usefulness of these tests. Rather than a statewide administration of the instruments, we are preparing to pilot test 22 developmental instruments which measure a set of outcomes from each of the nine categories of the basic learner outcomes.

A small random sample of 84 school districts has been drawn for pilot testing these instruments. Attached is a letter that we mailed to the superintendents of the schools in the sample. Additional information provided to these superintendents is also enclosed, along with a list of sample schools in your region.

If you or your staff members have specific interest in this activity, we welcome your inquiries and participation as we proceed with the next phase of this project. Keith Cruse, Division of Program Planning and Needs Assessment (512/475-2066) will be available to respond to your questions and provide additional information. Further details will be provided to guidance and career education coordinators in future statewide meetings.

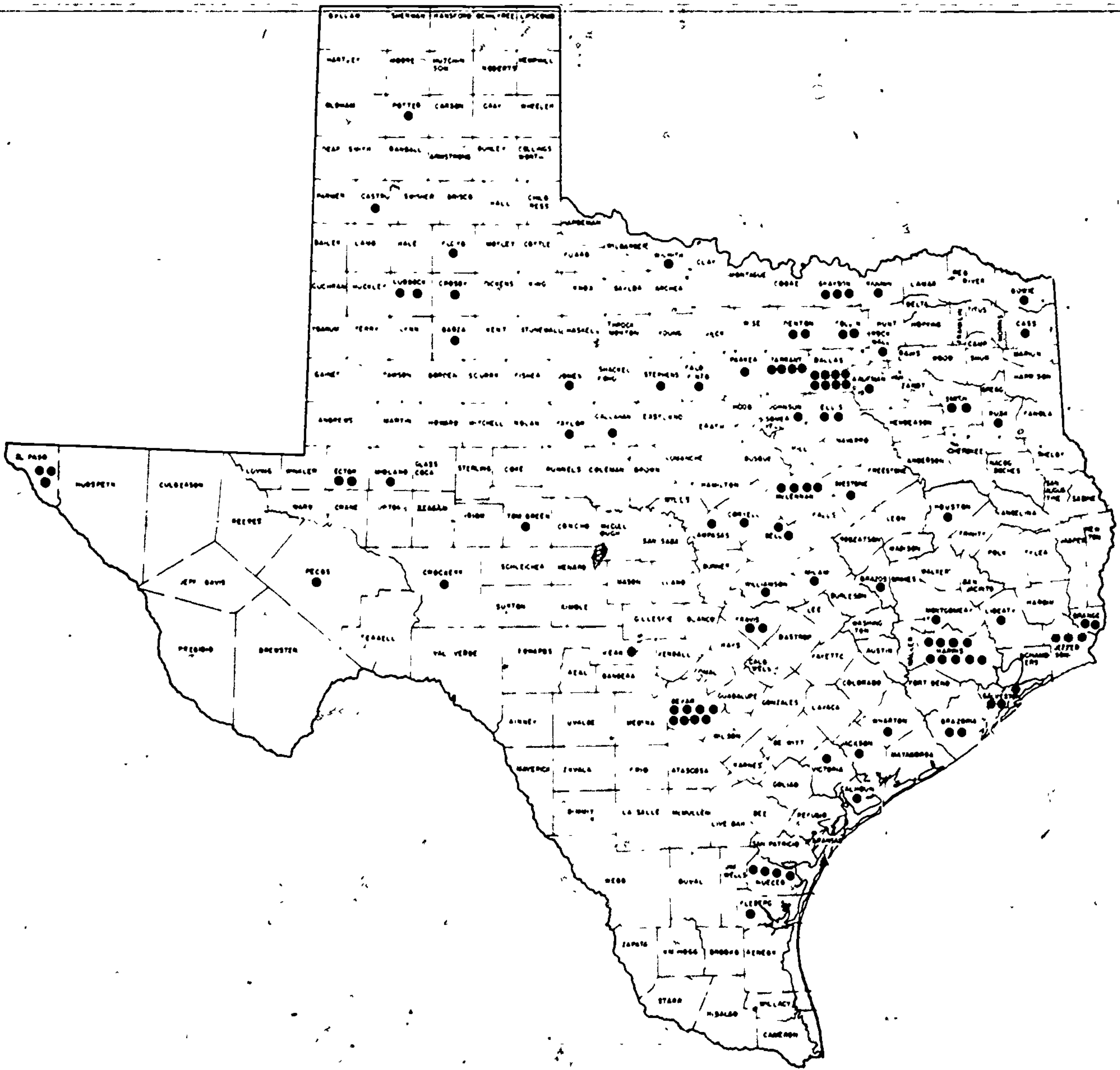
Yours truly,

Charles W. Nix
Associate Commissioner
for Planning and Evaluation

CWN:jr

Attachments

"An Equal Opportunity Employer"



POTTER
CASTRO
FLOYD
LUBBOCK
CROSBY
GARZA
WILBARGER
GRAYSON
FANNIN
BOWIE
DENTON
COLLIN
CASS
TARRANT
DALLAS
KAUFMAN
JONES
STEPHENS
PALO PINTO
PARKER
SMITH
ELLIS
JOHNSON
TAYLOR
CALLAHAN
EL PASO
ECTOR
MIDLAND
McLENNAN
LIMESTONE
HOUSTON
TOM GREEN
CORYELL
BELL
PECOS
CROCKETT
WILLIAMSON
MILAM
BRAZOS
WALKER
MONTGOMERY
LIBERTY
ORANGE
JEFFERSON
TRAVIS
HARRIS
KERR
BEXAR
GALVESTON
WHARTON
BRAZORIA
VICTORIA
JACKSON
CALHOUN
JIM WELLS
NUECES
KLEBERG
CAMERON

A MEASUREMENT SYSTEM FOR CAREER EDUCATION

SELECTING CLASS GROUPS FOR INSTRUMENT ADMINISTRATION

An important purpose for the pilot testing of the Career Education instruments is to get an accurate assessment of how all types of students at specific grade levels react to the instruments in general, and, more specifically, to the kinds of questions that are asked. The information provided by students from your school and from other schools will be grouped together and used to project how the instruments will be used when administered to students all over the state. As you can see, if the instruments are tried only with one type of student, such as the top students in each school, the information will give a distorted impression of how students perform. You are being requested to use the following guidelines when you select class(es) for participation in the pilot testing. These guidelines are for the purpose of helping you select the kinds of classes to provide the types of students that are needed. In no way is the overall performance of your school being evaluated.

Guidelines for Selection of Classes:

The following points should be considered when selecting a class(es) for participation in the pilot testing. The class(es) should:

- be representative of the ethnic make-up of the school.
- contain students with a mixture of abilities, not "honors" classes, that would lack an overall representation of student abilities.
- have a high majority of students at the grade level requested. It is realized that in high school it might be difficult to select a class that contains just one grade level of students.
- have from 20 to 35 students.

INSTRUCTIONS FOR COMPLETING THE

ASSESSMENT IN CAREER EDUCATION TEST EVALUATION FORM

Enclosed is the Assessment in Career Education (ACE) test evaluation form. This form asks questions about your perceptions and those of your students concerning the organization and sequencing of directions, instructions, and items contained in the Career Assessment Instrument. For this reason, it will be necessary for you to become thoroughly familiar with the questions on the form before you have administered the test instrument.

If any additional comments or space is needed to further elaborate on any of the questions on the evaluation form, please feel free to use the remainder of this instruction sheet. In addition, you will find an attached mailer so that you may return the evaluation form upon completion.

COMMENTS:

Form No. ________________________

Regional ESC No. ________________________

Campus Name ________________________

ACE TEST EVALUATION FORM

I. Presentation of Orientation Session

Yes No

______ ______ 1. I understand my role in and the purpose of the Career Education Assessment Instruments.

______ ______ 2. The orientation session was useful in providing answers to all questions that arose concerning the test and its administration.

______ ______ 3. I clearly understood the instructions which outlined the tasks I was to perform as the test administrator.

II. Instrument Design

______ ______ 4. The items on the test were in a logical sequence and well organized.

______ ______ 5. After the students received the *instructions* for the test instrument, did they understand what they were supposed to do? If they did *not*, what seemed to be the problem?

____

____

______ ______ 6. Were there *directions* within the test questions that at least three students did not seem to understand? If there were, please record the *number* of the item(s) and give a short comment about the problem with the item direction.

Item No. ____ ____

Item No. ____ ____

______ ______ 7. Were there *words* used in the test questions that at least three students did not know? If there were, please record the number of the item(s) and the word(s).

Item No. ____ ____

Item No. ____ ____

______ ______ 8. Were there any items that *offended* any students? If there were, please record the number of the item(s) and comment.

Item No. ____ ____

Item No. ____ ____

______ ______ 9. Did the students have any problems answering the *Student Information Form* questions found on the back of the test instrument? If they did, please identify which question and identify the problem.

Ques. # ____ ____

Ques. # ____ ____

III. Test Management

______ ______ 10. Indicate the size of group in which the test instrument was administered.

____

______ ______ 11. Were there any problems with the format of the answer sheet that caused students trouble? If there were, please identify the trouble.

____

Student Information ______________________________

______ ______ 12. Did you have any problems in scoring the open-ended item(s)? If you did, please record the item number and comment about the problem. (if applicable)

Item No ____ ______________________________

____

Item No. ____ ______________________________

____

______ ______ 13. Do you think the information received from this kind of item has enough value in relationship to the time it takes to score the item? (if applicable)

Comments ______________________________

____

______ ______ 14. Approximately what number of students finished the test in:

20 min.	40 min.	55 min.	Did Not Finish In One Testing Session
______	______	______	______

______ ______ 15. What subject do you teach? (main assignment)

____