

MANIPULATING RESPONSE SET  
IN THE TRUE-FALSE TEST

Thesis for the Degree of Ph. D.  
MICHIGAN STATE UNIVERSITY  
SARAH S. KNIGHT  
1972

THESIS



This is to certify that the  
thesis entitled

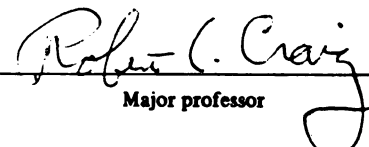
MANIPULATING RESPONSE SET IN THE TRUE-FALSE TEST

presented by

Sarah S. Knight

has been accepted towards fulfillment  
of the requirements for

Ph.D. degree in Department of Counseling,  
Personnel Services, and  
Educational Psychology

  
Major professor

Date 1972/July 31

O-7639



## ABSTRACT

### MANIPULATING RESPONSE SET IN THE TRUE-FALSE TEST

By

Sarah S. Knight

True-false test items are potentially effective and efficient measures of academic achievement. Numerous criticisms of this item type have been made, however. Most of the criticisms can be dealt with by careful attention to item construction, by the weight of logical argument and by research evidence. Despite these efforts, the susceptibility of true-false tests to the effects of response set remains as a limitation on their academic measurement potential. This research study was designed to assess various methods of manipulating response set in the true-false test.

Response set, in the form of a set to say true, has generally been considered to be a kind of response style which appeared consistently across tests for a given subject. Response set might instead be characterized as test specific and temporary, appearing across tests for a given subject because the various tests had certain influential characteristics in common. This was the thesis of this research study. It was tested experimentally by manipulating the test characteristics of item format (true-false, two-response multiple-choice), response option order (true-false, false-true), and test instructions (emphasis on true, emphasis on false).

The experimental treatments were five combinations of the test characteristics, constituting five levels of emphasis on the true and false response options. One treatment involved two-response multiple-choice items and the other four treatments involved the same set of true-false items, which were systematically derived from the two-response multiple-choice items.

The experiment yielded answers to three questions: 1) Can response set be manipulated by alterations in emphasis in T-F tests? 2) Does alteration of response option order affect response set? 3) Does the two-response multiple-choice item format yield response set?

First, response set was altered as a result of some of the manipulations. Decreasing emphasis on "true" yielded a decreasing set to respond "true." When the emphasis was on "false," the set to respond true increased as the emphasis on "false" increased. It was concluded that response option order reversal might have acted to suppress the subjects' tendency to respond "true," and that emphasis on either response prompted the subject to attend more closely to the entire response continuum, thereby enhancing a response style consisting of response acquiescence.

Second, response option order had no significant effect on response set in the T-F tests.

Third, the two-response multiple-choice items showed only slight evidence of generating any response set. On the basis of mean response bias scores, the data indicated that there was a significant tendency for subjects to prefer response option two.

Response set is generally not considered to be present in a test if the subject's bias score is not reliable. On this basis, the two-response multiple-choice items elicited little or no response bias, because the bias scores were of extremely low reliability.

In general, the two-response multiple-choice item format appeared to be the most effective manipulation in dealing with response set. When the response category labels were eliminated, response set decreased to a near-zero level.

MANIPULATING RESPONSE SET

IN THE TRUE-FALSE TEST

By

Sarah S. Knight

A THESIS

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Personnel  
Services, and Educational Psychology

1972

#### ACKNOWLEDGMENTS

The patient counsel and continued aid of Dr. Robert C. Craig, chairman of my guidance committee, are most gratefully acknowledged. Sincere thanks are also extended to my committee members, Dr. Robert L. Ebel, Dr. Donald M. Johnson and Dr. Byron H. VanRoekel for their valuable suggestions and critiques of this research study.

## TABLE OF CONTENTS

	Page
LIST OF TABLES . . . . .	v
 Chapter	
I. PROBLEM AND RELATED RESEARCH . . . . .	1
Introduction . . . . .	1
Response set and true-false tests . . . . .	2
Manipulating response set . . . . .	6
Other criticisms of the T-F test . . . . .	9
The problem . . . . .	12
Hypotheses . . . . .	13
Definition of terms . . . . .	14
Overview . . . . .	14
II. METHOD . . . . .	15
Research design and analysis . . . . .	15
Design . . . . .	15
Analysis . . . . .	17
Test development . . . . .	19
Item conversion . . . . .	20
Item pretesting . . . . .	23
Final test assembly . . . . .	25
Subjects . . . . .	26
Test administration . . . . .	28
Summary . . . . .	29



Chapter	Page
III. RESULTS . . . . .	31
Results concerning response set . . . . .	31
Hypothesis 1 . . . . .	31
Hypothesis 2 . . . . .	33
Hypothesis 3 . . . . .	34
Bias score reliability . . . . .	34
Test analysis results . . . . .	35
Summary . . . . .	37
IV. DISCUSSION . . . . .	40
T-F tests . . . . .	42
Multiple-choice test . . . . .	46
Suggestions for future research . . . . .	47
Summary . . . . .	49
REFERENCES . . . . .	51
APPENDIX A: Written and oral test instructions . . . . .	55

## LIST OF TABLES

Table	Page
1. Description of experimental treatment conditions . .	15
2. Number of true and false items at each difficulty level . . . . .	26
3. Allocation of subjects in the experimental design . .	27
4. Analysis of variance of bias scores . . . . .	32
5. Means and variances of bias scores for each treatment . . . . .	32
6. Summary statistics for each of the five test forms .	36
7. Distribution of item difficulty indices for each of the five test forms . . . . .	36
8. Distribution of item discrimination indices for each test form . . . . .	38

## Chapter I

### Problem and Related Research

#### Introduction

Classroom tests using a true-false item format, while not currently popular, do have proponents (Ebel, 1965, 1970, 1971). True-false items are attractive tools for achievement testing because they tend to be relatively easy to construct and because students can respond quickly to them. If criticisms of true-false items can be obviated, these items can become an effective as well as an efficient achievement testing technique.

Most criticisms of the true-false item can be countered on the basis of logic and experimental evidence, but there remains the stubborn and pervasive problem of the item's susceptibility to response set. Numerous studies attest that students display a definite tendency to use the response "true" when they are in doubt about the answer to a test item. This pattern of behavior is generally considered to be a kind of response style, a response tendency which the student brings to any true-false test, regardless of the specifics of that test.

There is, however, another way in which response set can be characterized. Instead of being a person's general style of responding to any true-false test, it could be a response tendency

which is temporary and test-specific. The set to respond "true" may show up consistently across true-false tests because the tests happen to have certain characteristics in common. That is the thesis of this research study. It was tested experimentally by manipulating the test characteristics of item format, response option order and test instruction.

Experimental treatment conditions consisted of various combinations of item format, response option order and test instructions. So combined, these test characteristics constituted several levels of emphasis on the true and the false response options. It was hypothesized that the direction and amount of response set that students displayed on a test would shift according to the emphasis placed on the response options. Successful manipulation of response set from a set to respond "false," through a set to respond "true," would favor the hypothesis that response set is test-specific. Further, such evidence should indicate methods for reducing or eliminating response set as a limiting factor in the utility of true-false tests.

#### Response set and true-false tests

True-false (T-F) tests are said to yield responses which are influenced by response set. Unlike other criticisms made of the T-F test, there have been no substantial rebuttals of this point. Ebel mentions it (1965), but does not accord it any extensive consideration. Therefore, of all the criticisms made of the T-F test, response set assumes the central position.

Studies by Cronbach (1941, 1942, 1946, 1950) are the major sources of information on response sets in tests in general, and in

T-F tests in particular. Response set is defined as, "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in different form" (Cronbach, 1946, p. 476). The specific response set with respect to T-F items is considered to be response acquiescence, defined as the subject's responding "T" more often on the average than "F," and/or the subject's tendency to respond "T" rather than "F" when in doubt (Cronbach, 1941, 1946).

Response acquiescence (RA) has been shown to exist in the responses to T-F tests by many studies prior to those of Cronbach (Arnold, 1927; Fritz, 1927; Granich, 1931; Krueger, 1932). In 1941, Cronbach compared the performance of students taking multiple-choice tests and T-F tests. His results indicated little significant difference between the two tests. Items keyed "F" had considerably higher reliability and validity than the items keyed "T." The number correct on "T" items was greater than the number correct on "F" ones. Theoretical considerations in the same paper led to the prediction that: response acquiescence would restrict the range of test scores; response acquiescence would have greatest effect on difficult items; the acquiescent subject would achieve a low score on the test if less than 50% of its items were keyed "T," and a high score if more than 50% of the items were keyed "T." Acquiescence was measured by tabulating the total number of "T" responses given by each student.

Cronbach offered the following practical example of the effects of RA on test scores. Suppose a 10-item test has 5 T and 5 F items. An examinee who simply guesses and marks as many F as T could get a

score from 10 to 0. An acquiescent examinee, guessing and marking 7 items T, could get a score ranging from 8 to 2. Now suppose the test has 7 T and 3 F items. The same examinees, following the same response patterns, could achieve scores from 8 to 2 when simply guessing and from 10 to 4 when they acquiesced. High scores would be likely to go to the acquiescent examinee if a test contains more than 50% T items.

It also follows that RA could be expected to inflate scores on the portion of the test containing just T items and similarly deflate the scores on the F items. When this obtains, low reliability and validity for T items and high reliability and validity for F items follows. Item statistics become uninterpretable. It becomes impossible to tell how much of the determined item difficulty and discrimination is due to the effects of RA, and how much to knowledge.

In 1942, Cronbach focused on the questions: 1) Are T items in general less reliable and valid than F items? 2) Is RA a consistent individual difference? 3) Can new test instructions obviate acquiescent behavior? The first two questions were answered affirmatively. Cronbach showed, however, that his new instructions did not eliminate RA effects. The latter result was attributed to the instructions being ill-conceived originally. As predicted in 1941, there were indications that RA did in fact limit score range as evidenced by test variances.

A review of the existence and extent of response sets was undertaken subsequently (Cronbach, 1946). Evidence was found for several types of response sets in several test types, among them T-F tests. The effect of response sets appeared to increase in ambiguous,

unstructured, or difficult situations. It was noted also that response sets could be compared to constant errors in psychophysics. In light of the survey's findings, Cronbach recommended a number of techniques for decreasing response set effects and thus increasing test validity, such as increasing test structure, use of the multiple-choice format, and avoidance of unreasonably difficult items.

Using the general technique of deriving a "bias" (response set) score and measuring its internal consistency to "prove" the set's existence, Cronbach (1950) studied response set with respect to test design. Response set was found to be consistent across as well as within test administrations, leading to the conclusion that it was a stable, personality-like trait. Relative to two-choice test formats, it was noted that altered test instructions yielded altered test biases on the same test (Rubin, 1940). This was not related to T-F tests, however. Apparently it was felt that the demonstration of parallel form reliability for response set was sufficient to characterize RA as a stable factor.

Following Cronbach's lead, Miklich (1965) investigated the relationship between RA and item importance and ambiguity, using personality/aptitude test items. The results confirmed that ambiguous items did elicit RA. An important interaction was found between importance and ambiguity, however. Important but ambiguous items tended to elicit agreement (RA), while unimportant ambiguous ones tended to yield disagreement.

Miklich (1968) designed a study to demonstrate that the response set in T-F items was not acquiescence, but rather test-taking carefulness (TTC). A set of maximally difficult (realistic nonsense) items

was generated. One half of the items contained specific determiners usually associated with true statements (the "pseudo-true" items), and the remaining items contained specific determiners usually found in false statements (the "pseudo-false" items). Difficult items should tend to yield RA. Therefore, if RA was operative, it should appear as an excess of T responses over all items, regardless of specific determiners. If TTC was operative, a negative correlation between the number of T responses given to pseudo-true and pseudo-false items should obtain. Evidence favored TTC, the more T responses subjects tended to give to pseudo-true items, the fewer they tended to give to pseudo-false ones. It should be noted, however, that this result does not eliminate RA as an explanation for response set in T-F tests because 1) the experimental situation was so extreme as to be highly unrealistic, and 2) careful item construction should eliminate TTC while probably leaving RA to a discriminable degree.

#### Manipulating response set

Regardless of the exact source of response set in T-F tests, there is abundant evidence of its existence. If T-F items are considered as special cases of a two-choice judgment task, there also exists evidence that response set can be manipulated. Cronbach has noted that alteration of instructions could affect response set, and that response set was related to the concept of constant error in psychophysical judgments, but he failed to develop either notion with respect to T-F tests.

In 1930, Fernberger showed that in a weight lifting experiment, different instructions yielded different category widths for a three category scale of heavier, equal or lighter. When the instructions



emphasized finding a difference between the standard and variable weights, very few weights were judged equal to the standard. When the instructions allowed free use of the equal category, many weights were judged equal to the standard.

Rubin (1940) altered the instructions on the Seashore Pitch Test (a High, Low, two-choice test), to emphasize the second response option, L. He got fewer (56.8%) errors of "L marked H" as compared with the number of such errors (60%) when the option H was emphasized. Altered instructions reduced set. Similarly, Rubin obtained a corresponding significant result when subjects were to imagine a coin toss and record its results. He found more "heads" responses when "heads" was emphasized in the instructions, and the same effect for "tails" responses.

In psychophysical judgments, Goodfellow (1940) and Gault and Goodfellow (1940) reported parallel findings. When instructional emphasis was placed on the "yes" response (stimulus present) the usual response set, reporting the stimulus to be absent, was reversed.

Holland (1961) demonstrated that specific alterations in task instructions led to differences in flicker-fusion thresholds as well as to over-all differences in what should normally be very similar sets of results. The task was a two-choice one, flicker was judged to be either present or absent. Task instructions emphasized the presence of flicker (report as soon as any unsteadiness appears in the light source) or alternatively, the absence of the steady light (report only when you are certain that flicker is present, i. e., the steady light is definitely absent). Instructional emphasis on flicker presence depressed the threshold, while emphasis on steady

light absence raised the threshold.

LeFurgy (1966) manipulated response set with a conditioning technique. He trained subjects to associate size of circle with the response categories positive, negative and neutral. One group of subjects learned that large circles were positive, medium ones were neutral and small ones were negative. Another group learned the associations in the reverse order. When tested on a second set of circles, subjects who learned that large circles were positive used that response category more frequently, using it for a wide range of larger circles. For subjects who learned that small circles were positive, that response category was used most frequently, including a wide range of smaller circles.

Using T-F tests, Bugelski and Herson (Bugelski and Herson, 1966; Herson, 1966, 1967) demonstrated that a response set could be conditioned on "ambiguous" items. In effect, they emphasized either the T or F response to ambiguous items by using a training session prior to testing, in which subjects answered test items and then were verbally informed of the correct response by the experimenter. When ambiguous items were called "T" by the experimenter, subjects in subsequent testing continued tending to respond "T" to such items, and vice versa when "F" was conditioned.

Mathews (1927) studied the effects of response option position in two-response items. Subjects were children in grades five and six. He found that alternating the horizontal ordering of a pair of responses definitely affected response selection. The response in the left position was consistently favored. When vertical ordering of the response pairs was altered, the upper position was consistently

avored. Further, response position influence was greatest where guessing was greatest.

There are a number of studies which indicate that an item's position in a series might affect the response given to it. Good-fellow (1940) reported that subjects' responses followed some definite patterns in two-choice situations. The subjects tended to avoid symmetric series of responses (ABABA) while they also tended to alternate responses. George (1953) found that subjects could be induced to respond in a predictable way to a paired-comparison task. He constructed a series of comparisons so that the subject gave the same response (heavier or lighter) to the first two sets, then, when the subject encountered a third, much more difficult comparison, he displayed a definite tendency to repeat the previous response.

#### Other criticisms of the T-F test

It remains to establish that the T-F test is a viable technique for measuring academic achievement. Therefore, the following major criticisms of the T-F test will be considered: 1) T-F tests are very susceptible to chance error introduced by guessing. 2) Such tests lack reliability. 3) The test content concerns trivialities. 4) It encourages rote learning. 5) "Absolute" truth or falsity in items is difficult or impossible to attain. 6) T-F tests lack validity.

First, it has been charged that excessive error can be introduced by guessing. Ebel (1965, 1970) responds that blind guessing is the major concern. Informed guesses provide valid indications of achievement. Blind guessing has been shown to occur relatively rarely, 3% to 8% of total test responses (Ebel, 1968), and its

effects vary inversely with test length. Further, it has been shown that reliable T-F tests can be constructed (Ebel, 1968; Burmester and Olson, 1966). This could not obtain if blind guessing seriously affected the test scores.

Burmester and Olson (1966) showed that unreliability is not a necessary characteristic of the T-F test. Two-choice tests do yield somewhat lower reliabilities than three-choice tests of comparable length, composition, respondents, etc. (Ebel, 1969; Williams and Ebel, 1957), while three-choice items tend to be more ideal than four-choice items (Costin, 1970). Reliability can be raised with increased test length, however, and given that two-choice items can be composed and responded to in much less time than three- or four-choice items, two-choice items become preferable.

The T-F items are not limited to potentially trivial specifics, Ebel maintains (1970). It follows that if T-F items are not so limited, they are not likely to encourage rote learning or to measure only very low levels of knowledge. He demonstrated this point by generating a series of good T-F items measuring understanding of event or process, principle applications, knowledge of functional relationships and problem solving. As Ebel says, "Surely there is nothing intrinsically trivial about a statement whose truth is open to question" (Ebel, 1970, p. 382).

The careful, skillful item writer can generate T-F items which are neither trivial nor ambiguous. Emphasis is on the application of care and skill. Items thus generated will, of course, be affected by ambiguities present in the nature of the language, but can be otherwise free of them. Another source of ambiguity is the basic

assumption that items must be absolutely true or false, raising the problem of defining truth. Ebel proposes two steps to deal with the definitional problem: students should be instructed to respond T if the item contains more truth than falsity, and the reverse for F responses. Second, the definite truth or falsity should obtain in the opinion of qualified experts (Ebel, 1971). And again, the occurrence of high test reliabilities argues that unambiguous test items have in fact been produced in considerable quantity.

Critics say T-F tests lack validity. What they refer to in part is the proposition that T-F tests cannot measure educational achievement due to its previously-discussed faults. That is, apparent concern lies with a kind of logical validity attained by T-F tests. Assuming the following four statements are correct and follow each other logically, then it is clear that T-F tests can attain logical validity.

- "1. The essence of educational achievement is the command of useful verbal knowledge.
2. All verbal knowledge can be expressed in propositions.
3. A proposition is simply a sentence that can be said to be true or false.
4. The extent of a student's command of a particular area of knowledge is indicated by his success in judging the truth or falsity of propositions related to it (Ebel, 1970, pp. 373-374).

Whether T-F tests can attain a respectable level of concurrent or predictive validity remains open to question. Frisbie (1971) compared parallel forms of T-F and multiple-choice items and found that two of eight sets of comparisons were of significantly less than perfect correlation at  $\alpha > .10$ . These results suggest, with a relatively large chance of error, that T-F tests can be constructed so

that they achieve tolerable levels of concurrent validity, using scores on a parallel multiple-choice test form as the criterion variable. Response set operating in the T-F test forms remains a plausible explanation for the two rather low correlations between T-F test scores and multiple-choice test scores which Frisbie found.

Cronbach (1942) calculated the predictive validity of several T-F tests. The criterion was a summation of scores on other, non T-F, tests taken in the same classes. The resulting correlations were low, ranging from 0.30 to 0.67. Again, the plausible explanation is response set, as Cronbach pointed out.

Apart from the criticisms, the T-F test has some intrinsic assets. Students can respond to considerably more T-F (or two-choice) items per unit of time than they can to three-, four-, or five-option multiple-choice items. When four-response multiple-choice items were compared with T-F items, it was found that about three T-F items could be responded to for every two multiple-choice items (Frisbie, 1971). True-false items are also often judged faster and occasionally easier to write than the latter.

### The problem

Objective achievement tests, T-F tests included, have certain characteristics like instructions, response option order and item format, which can and frequently do vary across tests. That is, the nature of such characteristics tends to be specific to each test. If one closely observes a T-F test, it is clear that the instructions emphasize "T," at least to the extent that it is always mentioned first. It is further emphasized by being listed first or assigned the number "1" on machine score answer sheets, and whenever such

two-choice tests are discussed, they are labeled true-false tests.

This being the case, is response set a student's response style with respect to T-F tests, or is it a temporary phenomenon resulting from consistent but test-specific emphasis on "T" across tests? Note, even the "standard" and altered test instructions used by Cronbach (1942) inadvertantly emphasized T by placing it first and mentioning it first, e. g., "This is a T-F test, circle T before a statement if it is always true. Circle F if the statement is false in any way" (p. 409).

Cronbach has mentioned a number of methods aimed at compensating for response set in T-F tests. The basic assumption underlying these methods is that response set is stable and depends on each person's style of responding. If it can be shown that altered emphasis produces altered response set, then it would argue that the response set is momentary, with the test itself as the source, thus casting doubt on Cronbach's recommendations.

Therefore, before methods of compensating for response set can be applied, it must be clear whether the phenomenon is intrinsic to the test or is the result of students' response styles. This research study is intended to clarify the point.

### Hypotheses

Three hypotheses were tested in this research study.

1. The response set found with T-F tests is a temporary phenomenon, whose source is intrinsic to the test, rather than a response style whose source is the person taking the test and unrelated to specific test characteristics.

The corresponding research hypothesis: Systematic variation

in the degree and direction of emphasis of response options will result in a corresponding variation in response set found with T-F tests.

2. The order in which the response options are presented will affect response set displayed on T-F tests.
3. Two-response multiple-choice items, which correspond to T-F items, will elicit no response set.

#### Definition of terms

Specific test characteristics refer to the test's instructions, the form of its items and the order in which its response options are presented. Emphasis is a variable made up of various combinations of test characteristics.

Response set is the tendency to use one response option over any others when there is doubt about a test item's keyed response.

#### Overview

The design and analysis of the research, test development and administration, and the research subjects are discussed in Chapter II. Research results are presented in Chapter III. The final chapter contains a summary of the study, a discussion of the results and recommendations for further research.



## Chapter II

### Method

#### Research design and analysis

Design. The experiment involved one treatment dimension which consisted of five levels of emphasis (E) on either T or F. The degree of emphasis for the test in each treatment condition was determined by various combinations of test instructions, item format and response option order. A description of the content and level of emphasis for each treatment condition is presented in Table 1.

Table 1. Description of experimental treatment conditions

Treatment	Emphasis	Item format	Instructions	Response option order
E <sub>1</sub>	high T	T-F	stress on T	T = 1, F = 2
E <sub>2</sub>	low T	T-F	minimum stress on T and F	T = 1, F = 2
E <sub>3</sub>	none	2RMC	best answer	
E <sub>4</sub>	low F	T-F	minimum stress on T and F	F = 1, T = 2
E <sub>5</sub>	high F	T-F	stress on F	F = 1, T = 2

The five levels of emphasis constituted the independent variable in this experiment. The dependent variable was response set, operationally defined as the number of responses in position one on an answer sheet, minus the number of responses occurring in position two on the same answer sheet (for conditions  $E_4$  and  $E_5$  the definition was the reverse, the number of responses marked two minus the number of responses marked one). Thus derived, this number was called the subject's bias score.

Both the amount and direction of emphasis varied across the five treatment levels yielding treatment conditions of high emphasis on T ( $E_1$ ), slight T emphasis ( $E_2$ ), no emphasis on either T or F ( $E_3$ ), slight F emphasis ( $E_4$ ), and high emphasis on F ( $E_5$ ). Conditions  $E_1$  and  $E_5$  obtained maximal emphasis by combining response option order and test instructions so that they emphasized the same response. A condition of no emphasis was achieved in condition  $E_3$  by using two-response multiple-choice (2RMC) instead of T-F items. Low emphasis occurred in conditions  $E_2$  and  $E_4$  through a combination of instructions that placed a minimum of stress on either T or F, and a response option order in which first place went to the response to be emphasized.

The 2RMC item format was chosen for two reasons. First, since it is a "two-choice" item form and hence related to T-F items, its use in a test which directly parallels the T-F test yields a condition in which emphasis is at an irreducible minimum. Neither its instructions nor its responses listed on the answer sheet involve the notions of true or false. Second, Cronbach (1946, 1950) and others (Wevrick, 1962) found multiple-choice items to be virtually free of response set.

Whether this attribute would extend to the 2RMC form was of experimental interest.

All levels of the treatment were administered in each of two classrooms. Within each of the classes, subjects were randomly assigned to treatments, and each subject received only one treatment. The experimental design was what is known as a randomized block design, with five levels of the treatment factor (emphasis) and two levels of the block factor (classes).

Factors concerned with internal and external validity of the study were controlled. In Campbell and Stanley's (1963; Bracht & Glass, 1968) terms, the design can be considered as an elaboration of the true experiment "posttest-only control group design." Threats to internal validity were controlled by random assignment of subjects to treatments, with each subject receiving only one treatment, and all treatments being administered to each class of subjects. Treatment administration and the measurement of its effects were carried out at the same time.

Research generalizability was limited by the nature of the subject population, the differences between actual classroom testing and the administration of treatments, the specialized type of T-F item which was used, and the subject matter included in the tests.

Analysis. The analysis assumed five levels of treatment, with subjects randomly assigned to treatments. Each subject received only one treatment, and all levels of treatment occurred in each of two classes.

Hypothesis one, concerning the effects of varying emphasis

on response set was tested using an analysis of variance for a randomized-block design. Given a significant treatment effect, post hoc multiple comparisons were made using Scheffé's technique.

Hypothesis two, concerning the effect of response option order on response set was tested with a post hoc analysis. Mean bias scores for treatment conditions  $E_2$  and  $E_4$  were compared with Scheffé's method for multiple comparisons.

The question of the amount of response set elicited by the 2RMC treatment condition, hypothesis three, was answered with a one-sample t-test. The obtained mean bias score for treatment condition  $E_3$  was tested against the score which was equivalent to zero bias.

Theoretically, the bias score as defined earlier could range between +74 and -74. In order to avoid the possibility of negative numbers in the analysis, 74 was added to each bias score. Thus the bias score which was analyzed was:

$$x - y + 74,$$

where  $x$  = number of 1 responses marked (2, for  $E_4$  and  $E_5$ ),

$y$  = number of 2 responses marked (1, for  $E_4$  and  $E_5$ ).

No bias was represented by a score of 74, high negative bias by a score of 0, and high positive bias by a score of 148.

Complete item analyses were performed on the test responses in each treatment condition. The item statistics which were computed were difficulty (proportion correct) and discrimination ( $r$  point-biserial). The  $r$  point-biserial was selected as the discrimination index instead of the more traditional  $r$  biserial because it was most plausible to assume only two distinct positions, right and wrong on the item continuum, given items which were cast in a true-false

format. The discrimination indices which were obtained from the tests in the experimental treatments were therefore somewhat lower than they would be had  $r$  biserials been computed.

Summary statistics for the entire test were: mean, variance, reliability ( $KR_{20}$ ), and standard error of measurement. This detailed information was used for a close examination of the functioning and comparability of the T-F and 2RMC item types.

In order to estimate the reliability of the bias scores, an odd-even split-half reliability coefficient was computed for each of the treatment conditions. The Spearman-Brown prophecy formula was applied to these reliability coefficients in order to estimate their magnitudes on tests twice as long.

### Test development

Social psychology was selected as the test's content domain. The items which formed the initial test item pool were drawn from a collection of four-response multiple-choice social psychology items for which item statistics were available. All of these items were constructed by subject-matter experts, and were designed for use with students similar to those in the research sample. The item statistics were based on several hundred responses per item.

Each of the introductory psychology classes in this study involved a unit on social psychology, thus a test involving this content area achieved considerable face validity in terms of being an integral part of the course. The original items were also selected because they were likely to be sufficiently difficult for the research sample. Item difficulty is an important factor in response set

and this level of difficulty was expected because the items were written for students who had had a more extensive unit on social psychology than the research subjects.

The following three items were typical of those in the original multiple-choice item pool:

1. Role determinants of personality rest mainly upon
  1. the socialization process.
  2. idiosyncratic experiences.
  3. biological inheritance.
  4. accidental events which occur in our lives.
2. Victims of violence within lower-class Black ghettos are usually
  1. close acquaintances of the offender.
  2. white merchants.
  3. strangers to the offender.
  4. local political leaders.
3. Which of the following contributes most to the present problem of poverty in the United States?
  1. The decline of the Protestant ethic
  2. Technological innovation
  3. Lack of motivation to work among persons in the lower socioeconomic groups
  4. The progressive increase in the cost of living

Item conversion. From each of the original four-response multiple-choice items, two new items were developed. One item of the pair was a two-response multiple-choice (2RMC) item, and the other was a special type of T-F item. It was intended that the 2RMC and T-F forms should be as comparable as possible; therefore each T-F item was a statement of comparison between two alternatives. The truth

or falsity of these statements depended on the order in which the alternatives were compared, thus making the criterion of truth internal to each item.

The item conversion methods, which were unique to this research, were kept as uniform as possible across items. First, the multiple-choice item was reduced from four to two responses, one of which was the one keyed correct. The second response was the distractor which drew the most responses from students who scored low on the test, i.e., the distractor which was the most discriminating. Occasionally there was no most discriminating distractor, or if there was it was not of the same character as the correct response. In these special instances, either the second most discriminating distractor became the second response or one of the three distractors was suitably altered to become the second response.

True-false items were developed from the 2RMC items. Their general form was the 2RMC stem combined with the two 2RMC response options so that a statement of comparison between the 2RMC response options resulted. Thus the T-F item was the 2RMC item recast as a single statement.

There were some instances where two statements were required for clarity of communication. However, the general item conversion rule-of-thumb was to form the T-F item using a single statement, sacrificing as little of the item's original wording as possible.

Whether a T-F item was formed as a true or a false statement was decided on a random basis within each of several specified levels of item difficulty. This was done to ensure a balance of items keyed

T and F within as well as across levels of item difficulty.

The following items demonstrate the item conversion process. The pairs of items were developed from the three items presented above as representative of the original multiple-choice item pool. The keyed response for each 2RMC item is indicated by an asterisk beside that response. The keyed response for each T-F item follows that item in parentheses.

The first example item became the following when only the keyed answer and the most discriminating distractor were retained:

2RMC<sub>1</sub>    Role determinants of personality rest mainly upon

1. idiosyncratic experiences.

\*2. the socialization process.

From this new item the T-F mate was then formed:

T-F<sub>1</sub>    Role determinants of personality rest more on idiosyncratic experiences than on the socialization process. (F)

The 2RMC/T-F pair corresponding to the second example item became:

2RMC<sub>2</sub>    Victims of violence within lower-class Black ghettos are usually

\*1. close acquaintances of the offender.

2. strangers to the offender.

T-F<sub>2</sub>    Victims of violence within lower-class Black ghettos are more often close acquaintances than strangers to the offender. (T)

The third pair became:

2RMC<sub>3</sub>    Which of the following contributes most to the present problem of poverty in the United States?

\*1. Technological innovation

2. The progressive increase in the cost of living



- T-F<sub>3</sub>     The progressive increase in the cost of living contributes more to the current United States poverty problem than does technological innovation. (F)

Based on pretest data, the T-F<sub>3</sub> item was changed from a false to a true statement. The following T-F<sub>3</sub> item which was included in the final test form demonstrates the determination of truth or falsity on the basis of alternative order in the comparative statement.

- T-F<sub>3</sub>,     Technological innovation contributes more to the current United States poverty problem than does the progressive increase in the cost of living. (T)

Item T-F<sub>2</sub> could similarly be changed to form a false statement:

- T-F<sub>2</sub>,     Victims of violence within lower-class Black ghettos are more often strangers than close acquaintances of the offender. (F)

Item pretesting. There were 120 pairs of T-F/2RMC items available for pretesting. For the 120 T-F items, an attempt was made to balance the number of items keyed T and F within each of the quartiles of the distribution of item difficulties. Over all, there were 60 items keyed T and 60 items keyed F in the item pool.

For pretesting, the 120 pairs of items were first split into two sets of 60 items, again with an attempt to balance the number of items keyed T and F within and across the item difficulty distribution. The 60 pairs of items were then separated to form a 60-item 2RMC test and a 60-item T-F test. Thus, pretesting was accomplished by breaking the item pool into four separate tests, two 2RMC and two T-F.

Two introductory psychology classes participated in the pre-testing. Each class was randomly divided in half, with one half taking a 2RMC test and the other half taking the matching T-F test. This technique yielded an average of 24 responses per item and allowed a comparison between the T-F/2RMC pairs.

Since an unusual form of T-F item was used, response rates on the 2RMC and T-F items were compared. About five minutes into the testing period, the subjects were asked to circle the number of the item on which they were currently working. Five minutes later they were asked to do this again. The resulting data indicated that the rate of responding was similar for the two item types: 12.5 T-F compared to 11.6 2RMC items in the first class; 9.3 T-F items compared to 10.7 2RMC items in the second class.

The preliminary test forms were introduced to the subjects as pretests for their unit on social psychology. The class instructors presented the tests, and stressed that while the subjects' course grade would be unaffected by the test, they were very interested in how much their students knew of social psychology prior to formal instruction in the subject. There was no mention of the presence of two different test forms. The students marked their answers on machine score sheets which had the item response option lettered, with T and F response options also indicated.

The item responses were machine scored and analyzed. The resultant item statistics were item difficulty, in terms of the proportion of respondents answering the item correctly ( $p$ ), and item discrimination in terms of  $r$  biserial coefficients. Summary statistics for the test included an internal consistency reliability estimate,  $KR_{20}$ . Of these statistics, item difficulty indices were of especial interest, while the others served as rough indicators of the way the final forms of the test could be expected to function. Item difficulty indices were used to determine which 2RMC/T-F pairs were to be retained for use in the final test forms, and to assure that the T-F test would have

items keyed T and F balanced within and across item difficulty levels.

Final test assembly. Based on pretest data, 74 item pairs were selected for use in the final test form. Items which were too hard ( $p \leq .20$ ) or too easy ( $p \geq .90$ ) were excluded, as well as item pairs whose T-F and 2RMC halves were not of reasonably comparable difficulty. Item pairs were rejected if the difference between the item difficulties of the T-F and 2RMC halves was greater than .20. Several item pairs were rejected because of awkward, uncorrectable T-F statements.

The total number of items on the final test was set at 74 for two reasons. First, the pretest response rate data indicated that everyone could reasonably be expected to complete a test of this length within the available 60 minutes of testing time. Second, this even number of items could conveniently be split into an equal number of items keyed T and F so that the subjects' bias scores could be derived.

At some difficulty levels the pretest data indicated that the number of true and false items were out of balance. To reestablish that balance, the excess items within each difficulty level were randomly selected and rewritten so that their keyed response was altered.

The distribution of true and false items across item difficulty levels for the final test can be seen in Table 2. The entries in Table 2 were derived from the pretest data.

The items in the final test form were arranged according to increasing item difficulty (or decreasing p values). The T-F and 2RMC test items were arranged in the same order, the difficulty

levels of the T-F items being used to determine the order.

Table 2. Number of true and false items at each difficulty level

Difficulty (p)	T	F
.30-.39	5	5
.40-.59	13	13
.60-.79	18	17
.80-.89	1	2

### Subjects

The subjects for this study were members of four introductory psychology classes at a Michigan university. Both the classes and their members participated in the research on a volunteer basis. The classes used were those in which the instructors agreed to incorporate the experimental treatments in their usual teaching procedures. The subjects were not required to respond to the experimental treatment, although they were urged to do so.

About 75% of the subjects were college freshmen, 20% sophomores, and 5% juniors. Approximately 45% of the subjects were male, 55% were female. They represented a broad range of both academic achievement and academic majors. Due to the specific academic requirements of this university, the students who enroll in introductory psychology can be reasonably expected to be representative of the freshman-sophomore student body.

The subjects who participated in the test development phase of this research were members of the two smallest of the four classes,

with enrollments of about 60 students each. The subjects in the final experimental phase of the research were members of the two largest classes, with enrollments between 200 and 300. A total of 99 students took part in test development; 315 students were in the final experiment. There were at least 60 students from the latter group in each treatment condition.

The number of subjects allocated to each treatment X class combination in the research design is presented in Table 3. The number in parentheses within each cell indicates the number of subjects whose bias scores were analyzed in the randomized blocks analysis of variance. In cells where some subjects were eliminated, the elimination was done at random. When the number of subjects was reduced as specified, the design became an orthogogonal one with proportional cell frequencies.

Table 3. Allocation of subjects in the experimental design

Classes	Levels of emphasis				
	$E_1$	$E_2$	$E_3$	$E_4$	$E_5$
1	40 (40)	43 (40)	44 (40)	45 (40)	42 (40)
2	20 (19)	23 (19)	19 (19)	20 (19)	19 (19)

Test administration

Five sets of test instructions were developed for the five levels of emphasis. The crucial paragraphs from the instructions for each level of emphasis were as follows. The test booklet cover sheets with the complete test instructions for each treatment condition are in Appendix A.

- E<sub>1</sub> The test consists of statements which are either true or false. You might think a statement is more true than false, then you should mark the number on your answer sheet which corresponds with true.

If you think a statement is more true than false, mark 1 on the answer sheet for that statement. If you think it is more false than true, mark 2 for that statement. Remember,

1 = true

2 = false

- E<sub>2</sub> The test consists of statements which are either true or false. If you think a statement is more true than false, mark 1 on the answer sheet for that statement. If you think it is more false than true, mark 2 for that statement. Remember,

1 = true

2 = false

- E<sub>3</sub> The test consists of multiple choice items. For each item there are 2 choices. Select the one best answer for each item, and mark its number in the appropriate space on the answer sheet.

- E<sub>4</sub> The test consists of statements which are either true or false. If you think a statement is more false than true, mark 1 on the answer sheet for that statement. If you think it is more true than false, mark 2 for that statement. Remember,

1 = false

2 = true

- E<sub>5</sub> The test consists of statements which are either false or true. You might think a statement is more false than true, then you should mark the number on your answer sheet which corresponds with false.

If you think a statement is more false than true, mark 1 on the answer sheet for that statement. If you think it is more true than false, mark 2 for that statement. Remember,

1 = false

2 = true

One of the classes used in the experimental phase of this research had just completed their unit on social psychology while the other class was just beginning their unit at the time the experimental treatments were administered. This necessitated slightly different verbal instructions for the treatments in these classes. The scripts which were used by the two instructors as a basis for introducing the treatments are in Appendix A.

One major intent of the instructors' verbal introduction of the treatments was to establish the experimental tests as being of especial interest to him. That is, it was intended to establish the experimental tests as quasi classroom tests, even though course grades would be unaffected by the results. Verbal instruction was also intended to ensure that the subjects' attention would be focused on the instructions written on the cover sheet of the test booklet.

### Summary

Experimental treatments were administered within the framework of a randomized blocks design, which had five levels of the treatment factor, emphasis, and two levels of the blocking factor, classes. The independent variable was emphasis and the dependent variable was response set, represented by bias score.

An analysis of variance for randomized-blocks was computed for the bias scores of 295 subjects, followed by post hoc comparisons using Scheffé's method. A one-sample t-test was computed for the 2RMC mean bias score.

Two sets of 74 items, 2RMC and T-F, were systematically developed from a common set of four-response multiple-choice items. These items formed five treatment condition tests, which were administered to 315 subjects. Each subject received either the 2RMC or the T-F items. Subjects responding to the T-F items did so under one of four instructional/response-option-order variations.

The experimental subjects were representative of the freshman student body of a Michigan university, and they received the experimental treatments under reasonably standardized conditions. The testing conditions were structured so that they appeared to be a natural part of the subject's class.



## Chapter III

### Results

The results are presented in two sections. Results concerned with response set are presented in the first section. The second section is descriptive, containing the test analysis data from the five experimental test forms.

#### Results concerning response set

Hypothesis 1. The major research hypothesis was that systematic variation of the degree and direction of emphasis would result in a corresponding variation in response set. It was tested with an analysis of variance for randomized blocks. The hypothesis that the five treatment mean bias scores were not different was tested against the alternative hypothesis that at least two mean bias scores were different.

The dependent variable of bias was calculated as follows:

$$x - y + 74 = \text{bias}$$

where  $x$  = number of 1 responses (2, for  $E_4$  and  $E_5$ )

$y$  = number of 2 responses (1, for  $E_4$  and  $E_5$ )

The resultant analysis of variance appears in Table 4. As the levels of significance for the F statistic show, there is clear evidence in favor of the alternative hypothesis: at least two of the treatment means were different. Further, there is no

evidence to indicate that the two different classes performed differently either over-all or at any particular treatment X class combination.

Table 4. Analysis of variance of bias scores

Source	df	MS	F	$\alpha$
Replications	285	184.93		
Classes	1	45.22	.240	<.6214
Emphasis	4	1884.70	10.188	<.0001
Emphasis X Classes	4	97.24	.526	<.7170

Given that treatment differences existed in the data, it was important to locate the source of the differences. The mean bias values on which the analysis of variance was computed, and their variances, appear in Table 5.

Table 5. Means and variances of bias scores for each treatment

Classes	Levels of emphasis				
	$E_1$	$E_2$	$E_3$	$E_4$	$E_5$
Class 1	85.20	83.10	71.20	79.45	79.45
Class 2	86.00	83.79	70.21	73.89	80.32
Pooled means	85.46	83.32	70.88	77.66	79.73
Pooled variances	191.75	175.28	67.48	206.12	150.63

Inspection of the means suggested that the bias scores obtained by the subjects in the  $E_3$  treatment condition were a source of the highly significant treatment effect found in the analysis of variance. To test this, the following contrasts were tested with the Scheffé method of multiple comparisons:

$$(1) E_1 - E_3 \pm S \sqrt{\text{Var } \hat{L}}$$

$$(2) E_3 - E_5 \pm S \sqrt{\text{Var } \hat{L}}$$

$$(3) E_2 - E_3 \pm S \sqrt{\text{Var } \hat{L}}$$

$$(4) E_3 - E_4 \pm S \sqrt{\text{Var } \hat{L}}$$

$$(5) E_1 - E_4 \pm S \sqrt{\text{Var } \hat{L}}$$

$$(6) E_1 - E_5 \pm S \sqrt{\text{Var } \hat{L}}$$

where  $E_1 \dots 5$  are pooled treatment means

$\text{Var } \hat{L}$  = estimated contrast variance

$$S = \sqrt{4 F_{.05, 4, 285}}$$

Contrasts 1, 2, 3 and 5 were significant at  $\alpha = .05$ . As observation of the obtained cell means suggested, condition  $E_3$  was responsible for three of the significant treatment differences pointed up by the analysis of variance.

In summary, the evidence partially supported hypothesis one. Systematic variations in emphasis did result in differences in response set, although the differences did not directly correspond with the direction of variation of emphasis. The no-emphasis condition ( $E_3$ ) yielded bias scores which were, as intended, significantly less than those from the two true-emphasis conditions. However, condition  $E_3$  also yielded bias scores which were significantly less than those of  $E_5$ , where the reverse was intended. Finally, there was no difference between conditions  $E_1$  and  $E_5$  bias scores, where differences were anticipated.

Hypothesis 2. Response option order will affect the response

set found in true-false tests. The test of this hypothesis was a post hoc Scheffé's multiple comparison between treatment conditions  $E_2$  and  $E_4$ , as follows:

$$E_2 - E_4 \pm S \sqrt{\text{Var } \hat{L}}$$

where  $E_2$  and  $E_4$  are pooled treatment means,

$\text{Var } \hat{L}$  = estimated contrast variance

$$S = \sqrt{4 F_{.05, 4, 285}}$$

The contrast failed to achieve significance with  $\alpha = .05$ . Therefore, hypothesis two was not supported by the data. There was no evidence that response option order affected the subjects' bias scores.

Hypothesis 3. Multiple-choice items with two response options will yield no response set. Reference to Table 5 indicates that treatment condition  $E_3$  did yield the lowest mean bias scores. Further, the post hoc multiple comparisons above showed that the  $E_3$  means were significantly different than the means of conditions  $E_1$ ,  $E_2$ , and  $E_5$ . To ascertain whether the  $E_3$  pooled mean was significantly different than the value indicating no bias, a one-sample t-test was performed.

For the t test, the hypothesis of no difference between the  $E_3$  mean and a value of 74 (0 bias) was tested against the alternative that the  $E_3$  mean was different than 74. Based on the data from all 63 subjects in  $E_3$ , the statistic t was significant ( $\alpha \leq .002$ ,  $df = 62$ ). It was concluded that the  $E_3$  pooled mean represented a bias score which was significantly less than 74, and thus subjects in  $E_3$  displayed a significant tendency to prefer response option two. Hypothesis three was therefore not supported by the evidence.

Bias score reliability. Cronbach maintained that response set

could be said to exist if it could be shown to be reliable. Reliability coefficients were therefore computed for the bias scores in each of the treatment conditions. The odd-even reliability coefficients, corrected with the Spearman Brown prophecy formula were .68, .72, -.15, .79, and .46 for treatment conditions one through five respectively.

Treatment condition five gave bias scores of relatively low reliability. Condition three, the 2RMC condition, yielded a coefficient so low that it is doubtful whether its bias scores did in fact represent the existence of any response set.

#### Test analysis results

The validity of the experimental results with respect to response bias depends on the character of the test from which the results were derived. These tests should contain comparable and sufficiently difficult items. They should be generally comparable in reliability and item discrimination to each other and to the classroom test in order to enhance the generalizability of the results.

In order to obtain good descriptions of the subjects' responses to the five test forms, the data from the two classes participating in the research were pooled. The test analyses were therefore based on at least 60 responses per item per test form.

Summary statistics for each of the five tests are presented in Table 6. It can be seen that, with the exception of condition  $E_3$ , the mean test scores appear similar across conditions. Condition  $E_3$  had the highest mean test score. Test forms  $E_4$  and  $E_5$  displayed the highest reliability coefficients, as well as the greatest variance in test scores.

Table 6. Summary statistics for each of the five test forms

Statistic	Test form				
	E <sub>1</sub>	E <sub>2</sub>	E <sub>3</sub>	E <sub>4</sub>	E <sub>5</sub>
No. examinees	60	66	63	65	61
Mean no. correct	43.1	42.4	45.4	40.9	41.4
SD	6.1	5.6	5.9	8.0	6.8
Reliability (KR <sub>20</sub> )	.55	.47	.55	.73	.63
SEM	4.05	4.06	3.94	4.12	4.11

Table 7. Distribution of item difficulty indices for each of the five test forms

Difficulty *	Test form				
	E <sub>1</sub>	E <sub>2</sub>	E <sub>3</sub>	E <sub>4</sub>	E <sub>5</sub>
91 - 100		1	1		
81 - 90	3	1	8		1
71 - 80	14	13	15	3	4
61 - 70	14	15	17	23	27
51 - 60	20	18	12	30	14
41 - 50	14	16	13	10	20
31 - 40	8	8	7	7	8
21 - 30	1	2	1	1	
Mdn. p	57.5	56.6	62.9	56.8	53.6

\* p = proportion of subjects correctly responding to an item  
(decimals deleted)

The distributions of item difficulties for each of the test forms are shown in Table 7. Test forms  $E_4$  and  $E_5$  had a slightly restricted range of difficulty levels relative to the other three forms. With the exception of form  $E_3$ , the median item difficulty was similar for all test forms. The multiple-choice items in the  $E_3$  condition appear to have been somewhat easier for the research subjects than were the comparable true-false items.

The over-all level and distribution of item difficulties for all of the test forms suggests that the items were appropriate for studying response set. As Cronbach has indicated, difficult items increase the likelihood of the appearance of response set. The present items can be considered to have been very difficult for the subjects, when the obtained difficulty levels are compared with the value shown by Lord (1952) to be ideal for two-choice items,  $p = .85$ , assuming there is no correction for guessing and assuming that all those who do not know the answer to an item respond randomly.

The obtained item discrimination indices are shown in Table 8. Because the indices are  $r$  point-biserial coefficients, they are somewhat lower than the more traditional  $r$  biserial coefficients generally would have been.

### Summary

Hypothesis one, that systematic variation in emphasis would result in a corresponding variation in response set was partially supported. Conditions  $E_1$ ,  $E_2$ , and  $E_3$  yielded bias scores as predicted, with  $E_1$  having the highest scores and  $E_3$  the lowest. Conditions  $E_4$  and  $E_5$ , however, had associated bias scores which were the reverse of the predicted negative bias scores.

Table 8. Distribution of item discrimination indices for each test form

Discrimination *	Test form				
	E <sub>1</sub>	E <sub>2</sub>	E <sub>3</sub>	E <sub>4</sub>	E <sub>5</sub>
51 - 60			2	4	
41 - 50	1	1	2	8	3
31 - 40	14	5	11	11	17
21 - 30	16	23	16	21	13
11 - 20	19	21	15	13	20
00 - 10	18	17	20	7	13
-01 - -10	5	5	5	7	5
-11 - -20	1	1	2	2	3
-21 - -30		1	1	1	

\* r point-biserial (decimals deleted)



Hypotheses two and three were not supported by evidence. Differences in response set did not obtain when response option order was reversed. Moreover, although the 2RMC condition ( $E_3$ ) did yield the lowest bias scores, these scores were significantly less than a score corresponding to zero bias.

Reliability coefficients altered with the Spearman Brown prophecy formula showed that a very low coefficient,  $-.15$ , came from the  $E_3$  treatment condition bias scores.

Detailed test analyses of the tests used in each of the five treatment conditions indicated that they were appropriate for research on response set in two-choice classroom tests.

## Chapter IV

### Discussion

As Cronbach (1941, 1942, 1946, 1950) and his predecessors noted, response set, in the form of RA, tended to appear when students responded to T-F tests. The present results essentially reaffirmed this point. When two-choice items were cast in a form requiring a value-laden response of either true or false, then subjects tended to say T when in doubt. When the two-choice item was cast instead in a 2RMC format whose responses involved neither the concepts true or false, then only very weak evidence for the presence of any response set appeared. Item format thus appeared to be the most potent and effective factor in manipulating response set.

Is RA a general response style? The research results point to the conclusion that it might be when the subject responds to items whose response options include the words true and false. These responses have connotations which extend beyond the test items and may well shade all such responses to all T-F items.

Cronbach (1942, 1946) had suggested that appropriate test instructions could aid in reducing RA in T-F tests, although he had not successfully used them for that purpose. As with Cronbach, the instructional manipulation used in this experiment was not effective in altering response set. The instructions were not effective in the sense

that they appeared to increase RA where an increase in the tendency to respond F was predicted.

Individual differences in response set are important considerations. Cronbach (1942) concluded that because there was a variation in response between subjects which remained relatively stable across tests, RA scores represented real individual differences. The present experiment clearly elicited varying amounts of response set for subjects within each treatment condition. The central question, however, was whether or not different treatment conditions yielded different degrees of variation in response set. A Cochran's test for homogeneity of variance (C) was computed, and found nonsignificant at  $\alpha \leq .05$ . Thus there was no evidence to conclude that individual variation in response set was different in any of the treatment conditions, although the bias scores in  $E_3$  had notably less variation than those in the other conditions.

Another indication of individual differences in response set is reliability of the bias scores within and across treatment conditions. Within treatment conditions, except for  $E_3$ , subjects in general tended to be moderately consistent in displaying response set throughout the test. Across treatment conditions subjects clearly responded differently. Assuming that the bias score reliability estimates are reasonable Pearson product-moment coefficient estimates, the five coefficients were compared with a variant of the Fischer  $r$  to  $Z$  transformation. The resulting statistic was significant ( $\alpha \leq .01$ ,  $df = 4$ ). Subjects displayed response set more consistently in some conditions than in others. Examination of the data suggested that the source of the difference was condition  $E_3$ .

The 2RMC items appeared to elicit a much less consistent response set from the subjects.

Frisbie (1971) found that four-response multiple-choice items were either as reliable or more reliable than their T-F counterparts. Although subjects in the different treatment conditions of this experiment were equated on the basis of random assignment of subjects to treatments, a comparison similar to Frisbie's can be made between the total number correct on the 2RMC and T-F tests. If the Kuder-Richardson formula 20 reliability coefficients are considered to be reasonable estimates of Pearson product-moment coefficients, than the coefficients for the five test forms can be compared in the same manner as the bias score reliability coefficients, with a variant of the Fischer  $r$  to  $Z$  transformation. The resultant statistic was not significant ( $\alpha > .05$ ,  $df = 4$ ). There were no significant differences among the five reliabilities; the 2RMC test appeared no more, or less, reliable than the T-F counterparts.

The conclusion that each T-F test was generally as reliable as each other T-F test and as reliable as the 2RMC test must be tempered somewhat because its basis is weak. Each reliability coefficient was based on only about 60 cases, and the scores for the tests were quite positively skewed.

#### T-F tests

When ordered on the basis of decreasing mean bias scores, the treatments assumed the order:  $E_1$ ,  $E_2$ ,  $E_5$ ,  $E_4$ ,  $E_3$ . Only one significant difference occurred between the T-F treatment mean bias scores,  $E_1$  was significantly larger than  $E_4$ . The size of the obtained treatment means are suggestive, however. If response set was completely unaffected by manipulations of response option order and instructions,

the  $E_1 - E_4$  difference should not have occurred, and further, there would be no reason to anticipate any pattern in the results. Logically, all mean bias scores should have been close to those of  $E_2$ . This is because the  $E_2$  treatment could be considered to be an example of the typical T-F test, with both instructions and response option order conforming to the typical case.

The results appear anomalous in light of these considerations. In fact, high emphasis on either response option appeared to boost bias scores above those of their treatment counterparts which had the same response option order but different instructions. With the  $E_5$  condition, the increased tendency to mark T was enough that while  $E_1$  and  $E_4$  had significantly different bias scores,  $E_1$  and  $E_5$  did not.

Based in part on the non-significant trends in the experimental results, the following might be hypothesized. The altered response option order depressed the tendency to mark T in  $E_4$ , and added stress on the response option F partially counteracted these effects. That is, an unusual response option order decreased the subjects' response set (true RA in Cronbach's terms), and any attention otherwise specifically drawn to the responses functioned only to alert the subject to the importance of the entire response continuum, not to just one portion of that continuum.

To investigate this hypothesis, an experiment could be conducted involving four treatment groups: 1) high T emphasis, response option order T-F; 2) high T, F-T; 3) high F, T-F; 4) high F, F-T. If response option order functioned to depress response set, then conditions 1 and 3 should yield high, similar bias scores while conditions 2 and 4

should have similar, lower bias scores. Given these results it could also be concluded that emphasis on one response option draws attention to both options. If it was the case that response option order did lower RA and that the effect was one which lasted across tests, then it would become one method for coping with response set in T-F tests.

Other indications of the source of the experimental results might come from an examination of subjects' item-by-item test responses. Did subjects in the different conditions, especially  $E_1$  compared with  $E_5$ , tend to answer correctly or guess on different items? Did either true or false items contain unanticipated specific determiners? For each test, did the mean number correct vary across tests in a pattern which paralleled the experimental results?

Unfortunately, neither close examination of each test item nor of the tests' over-all performance gave much indication of why such experimental results occurred. When the item difficulty indices, which gave the proportion of subjects answering each item correctly, were correlated for all pairs of tests, all coefficients were positive and reasonably high (.62 to .84) for the T-F tests. In particular, the correlation between  $E_1$  and  $E_5$  yielded a coefficient of .72. Thus it appeared that for a large proportion of the test items, subjects responded similarly in both groups.

When the T and the F items were correlated separately across tests, generally high positive coefficients again resulted. There were two exceptions. The correlation coefficient for T items between  $E_1$  and  $E_4$  was positive but rather low (.42) and for T items between  $E_4$  and  $E_5$  it was moderate and positive (.53). The T items in  $E_4$

appeared to function differently than they did in  $E_1$  and  $E_5$ . Response order reversal without extra stress on the response appeared to yield T items which had a tendency to be more difficult. If the T items tended to have higher difficulty indices, then fewer subjects than in  $E_1$  and  $E_5$  were answering the item as it was keyed, and hence there was apparently less "guessing T when in doubt." This might be taken as evidence that response option reversal does in fact have the potential for depressing the tendency to respond T.

It is improbable that an entire class of items tended to contain specific determiners in the usual sense. The type of T-F item used argues against the presence of specific determiners. Truth or falsity depended only on the order of the comparison between alternatives, not on wording changes. Many of the T-F items did use positive comparisons (more than, greater than) rather than negative ones (less than). These items were about equally distributed between items keyed T and F. While this type of specific determiner might help to account for the response set favoring T that occurred across tests, research evidence suggests that it is unlikely. Whipple (1957) studied the effects of positive and negative phrasing in T-F items, and found only a very slight tendency for subjects to say T to positively stated items.

The summary statistics for each test as a whole revealed little. The mean number correct within each treatment condition was similar. However, condition  $E_4$  yielded the highest reliability estimate and its scores had the largest variance. The higher reliability in the  $E_4$  test might have had two sources. First, the subjects' test scores were

most variable in this condition, due possibly to the unfamiliarity of the response option order. Second, the T items apparently functioned better, eliciting fewer guesses, thus removing their generally depressive effects on the over-all test reliability (Cronbach, 1941, 1942). When these points are considered apart from questions of test validity, they add weight to the case for the potential usefulness of response option order reversal in dealing with RA.

No test, no matter how reliable, is of much value as an achievement test if it is not also valid. The validity of T-F and 2RMC tests under the present experimental manipulations remains to be shown. Assuming that appropriate criteria were available, the role of RA in them would have to be assessed. Logically, the behavior of responding T when in doubt would play little or no role in strictly academic achievement. When this is the case, as Cronbach suggested (1946), the presence of any response set would reduce test validity. Thus while  $E_4$  might point the way to dealing with RA, it would have to reduce RA to a near-zero level while maintaining a reasonably high reliability to achieve the necessary validity to put it into contention with the 2RMC test.

#### Multiple-choice test

Did, or did not 2RMC items elicit response set? Yes, they yielded a significant tendency to use response option two, if the conclusion is based solely on the test of whether or not the mean bias score was different from zero. No, if the conclusion is based on whether or not the bias scores were reliable. Cronbach maintained that they must achieve reliability in order to be considered to exist,



and the reliability estimate for the 2RMC bias scores was negative and very low (-.15).

Unless and until the results can be replicated, and they have a reasonably high reliability estimate, it seems most conservative to conclude that the experimental data gave only weak evidence that the 2RMC test yielded response set, a conclusion which is in agreement with those of Cronbach (1946, 1950) and Wevrick (1962) concerning tests with four- and five-response multiple-choice items.

Comparing the T-F and 2RMC item pairs, the 2RMC items tended to be easier across all test pairs. This could be attributed to the higher apparent verbal difficulty of the T-F items. The two item forms compare like the outline of a paragraph (2RMC) and the fully written paragraph (T-F).

A test with 2RMC items would be recommended as the best method of dealing with response set appearing in a two-choice format, if the test could achieve adequate reliability, in addition to remaining unaffected by response set. Such a test would be most likely to achieve adequate validity as well. The 2RMC test would be recommended over the T-F test form  $E_4$  on the assumption that improved reliability is likely to be achieved sooner and with a more lasting effect than is a reduction in RA through response option order reversal.

#### Suggestions for future research

Two-response multiple-choice items may prove to be an excellent compromise between T-F and four- and five-response multiple-choice items. They should require a minimum of composition time to obtain a maximum of clarity. Students have been shown to be able to respond to them nearly as quickly as they can to T-F items (Ruch &

Stoddard, 1925). Therefore the following suggestions for research with 2RMC items are made.

1. Replicate this experiment with more, and different subjects to find out if response set does reliably appear on 2RMC tests.
2. If response set does appear, find out if it can be manipulated with instructions.
3. Investigate the optimum number of items for the desired level of reliability.
4. Investigate the concurrent and/or predictive validity of an achievement test with 2RMC items.
5. Compare the validity of 2RMC tests and matching T-F tests, where the T-F test is administered with reversed response option order.

Concerning T-F tests, there still remain unanswered questions generated by this experiment. The following studies are suggested as methods of answering some of the questions.

1. Compare the standard T-F item with the comparative type used in this study to see if the latter yields more or less response set.
2. As suggested in the discussion, investigate whether response option order reversal (F-T, instead of T-F) depresses subjects' bias scores.
3. If response option order reversal does depress subjects' bias scores, find out if this is a stable phenomenon or if it is the result of novelty.

### Summary

This research study was an investigation of the effects of manipulating test instructions, item format and response option order on response set in T-F tests. Subjects were each given one of five tests with various combinations of these variables.

The questions to be answered by the experiment were: 1) Can response set be manipulated by alterations in emphasis in T-F tests? 2) Does alteration of response option order affect response set? 3) Does the 2RMC item format yield response set?

Response set was altered in some instances, corresponding to the degree of emphasis placed on one (T) response. However, response set did not parallel the direction of emphasis used in each treatment. The first three treatments with decreasing emphasis on T showed a decreasing set to respond T, but when the emphasis was on F, set to respond T increased again with increasing F emphasis. It was concluded that response option order reversal might have acted to suppress the subjects' tendency to respond T, and that emphasis on either response prompted the subject to attend more closely to the entire response continuum, thereby enhancing a response style consisting of response acquiescence.

Response option order had no significant effect on response set in the T-F tests.

The 2RMC items showed only slight evidence of generating any response set. On the basis of mean response bias scores, the data indicated that there was a significant tendency for subjects to prefer response option two. Response set is generally not considered to be present in a test if the subject's bias score is not reliable.

On this basis, the 2RMC items elicited little or no response bias, because the bias scores were of extremely low reliability.

In general, the 2RMC item format appeared to be the most effective manipulation in dealing with response set. When the response category labels were eliminated, response set decreased to a near-zero level.

## REFERENCES

## References

- Arnold, H. L. An analysis of discrepancies between true-false and simple recall examinations. Journal of Educational Psychology, 1927, 18, 414-420.
- Bracht, G. H. & Glass, G. V. The external validity of experiments. American Educational Research Journal, 1968, 5, 437-474.
- Bugelski, B. R. & Hersen, M. Conditioning acceptance or rejection of information. Journal of Experimental Psychology, 1966, 71, 619-623.
- Burmester, M. A. & Olson, L. A. Comparison of item statistics for items in multiple-choice and alternative response form. Science Education, 1966, 50, 467-470.
- Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.
- Costin, F. The optimal number of alternatives in multiple-choice achievement tests: some empirical evidence for a mathematical proof. Educational and Psychological Measurement, 1970, 30, 353-358.
- Cronbach, L. J. An experimental comparison of the multiple true-false and multiple multiple-choice tests. Journal of Educational Psychology, 1941, 32, 533-543.
- Cronbach, L. J. Studies of acquiescence as a factor in the true-false test. Journal of Educational Psychology, 1942, 33, 401-416.
- Cronbach, L. J. Response sets and test validity. Educational and Psychological Measurement, 1946, 6, 475-494.
- Cronbach, L. J. Further evidence on response sets and test design. Educational and Psychological Measurement, 1950, 10, 3-30.
- Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N. J.: Prentice-Hall, 1965.

- Ebel, R. L. Blind guessing on objective tests. Journal of Educational Measurement, 1968, 5, 321-325.
- Ebel, R. L. Expected reliability as a function of choices per item. Educational and Psychological Measurement, 1969, 29, 565-570.
- Ebel, R. L. The case for true-false test items. School Review, 1970, 78, 373-389.
- Ebel, R. L. How to write true-false items. Educational and Psychological Measurement, 1971, 31, 417-426.
- Fernberger, S. W. The use of equality judgments in psychophysical procedures. Psychological Review, 1930, 37, 107-112.
- Frisbie, D. A. Comparative reliabilities and validities of true-false and multiple-choice tests. Unpublished Ph.D. dissertation, Michigan State University, 1971.
- Fritz, M. F. Guessing in a true-false test. Educational Research Bulletin, 1942, 21, 9-12.
- Gault, R. H. & Goodfellow, L. D. Sources of error in psychophysical measurements. Journal of General Psychology, 1940, 23, 197-200.
- George, F. H. 'Either-or' questions in series. British Journal of Psychology, 1953, 44, 243-247.
- Goodfellow, L. D. The human element in probability. Journal of General Psychology, 1940, 23, 201-205.
- Granich, L. A technique for experimentation on guessing in objective tests. Journal of Educational Psychology, 1931, 23, 81-91.
- Hersen, M. Generalization of positive and negative response biases. Journal of Experimental Psychology, 1966, 72, 834-840.
- Hersen, M. Experimentally induced response biases as a function of positive and negative wording. Journal of Experimental Psychology, 1967, 74, 588-590.
- Holland, H. C. Judgments and the effects of instructions. Acta Psychologica, 1961, 18, 445-457.
- Kreuger, W. C. F. An experimental study of certain phases of a true-false test. Journal of Educational Psychology, 1932, 23, 81-91.
- LeFurgy, W. G. The induction of anchoring effects in absolute judgments through differential reinforcement. Journal of Psychology, 1966, 63, 73-81.
- Lord, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulty. Psychometrika, 1952, 17, 181-193.

- Mathews, C. O. The effect of position of printed response words upon children's answers to questions in two-response types of tests. Journal of Educational Psychology, 1927, 18, 445-457.
- Miklich, D. R. Item characteristics and agreement-disagreement response set. (Doctoral dissertation, University of Colorado) Ann Arbor, Mich.: University Microfilms, 1966. No. 66-3259
- Miklich, D. R. & Gordon, G. P. Test-taking carefulness vs response set on true-false examinations. Educational and Psychological Measurement, 1968, 28, 545-548.
- Rubin, H. K. A constant error in the Seashore Test of Pitch Discrimination. Unpublished master's thesis, University of Wisconsin, 1940.
- Ruch, G. M. & Stoddard, G. D. The comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 1925, 16, 89-103.
- Smith, K. An investigation of the use of "double-choice" items in testing achievement. Journal of Educational Research, 1958, 51, 387-389.
- Wesman, A. G. Writing the test item. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1971.
- Wevrick, L. Response set in a multiple-choice test. Educational and Psychological Measurement, 1962, 22, 533-538.
- Whipple, J. W. A study of the extent to which positive or negative phrasing affects answers in a true-false test. Journal of Educational Research, 1957, 51, 56-63.
- Williams, B. J. & Ebel, R. L. The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. Fourteenth yearbook of the National Council on Measurement in Education, 1957, 63-65.

#### General References

- Glass, G. V. & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N. J.: Prentice-Hall, 1970.
- Hays, W. L. Statistics. New York: Holt, Rinehart and Winston, 1963.



- Henryssen, S. Gathering, analyzing, and using data on test items.  
In R. L. Thorndike (Ed.), Educational measurement. Washington,  
D. C.: American Council on Education, 1971.
- Johnson, D. M. Systematic introduction to the psychology of thinking.  
New York: Harper and Row, 1972.
- Rorer, L. G. The great response-style myth. Psychological Bulletin,  
1965, 63, 129-154.

## **APPENDIX**

## APPENDIX A

### Written and oral test instructions

Written instructions for treatment condition E<sub>1</sub>

#### INSTRUCTIONS:

1. Put your name and student number on the answer sheet.
2. NOTE: on this answer sheet, the answer spaces go ACROSS the WIDTH of the page.
3. The test consists of statements which are either true or false. You might think a statement is more true than false, then you should mark the number on your answer sheet which corresponds with true.

If you think a statement is more true than false, mark 1 on the answer sheet for that statement. If you think it is more false than true, mark 2 for that statement. Remember,

1 = true

2 = false

4. There is no penalty for guessing, so respond to EVERY statement.

Written instructions for treatment condition E<sub>2</sub>

## INSTRUCTIONS:

1. Put your name and student number on the answer sheet.
2. NOTE: on this answer sheet, the answer spaces go ACROSS the WIDTH of the page.
3. The test consists of statements which are either true or false.

If you think a statement is more true than false, mark 1 on the answer sheet for that statement. If you think it is more false than true, mark 2 for that statement. Remember,

1 = true

2 = false

4. There is no penalty for guessing, so respond to EVERY statement.

Written instructions for treatment condition E<sub>3</sub>

## INSTRUCTIONS:

1. Put your name and student number on the answer sheet.
2. NOTE: on this answer sheet, the answer spaces go ACROSS the WIDTH of the page.
3. The test consists of multiple choice items. For each item there are 2 choices. Select the one best answer for each item, and mark its number in the appropriate space on the answer sheet.
4. There is no penalty for guessing, so respond to EVERY item.

Written instructions for treatment condition E<sub>4</sub>

## INSTRUCTIONS:

1. Put your name and student number on the answer sheet.
2. NOTE: on this answer sheet, the answer spaces go ACROSS the WIDTH of the page.
3. The test consists of statements which are either true or false.

If you think a statement is more false than true, mark 1 on the answer sheet for that statement. If you think it is more true than false, mark 2 for that statement. Remember,

1 = false

2 = true

4. There is no penalty for guessing, so respond to EVERY statement.

Written instructions for treatment condition E<sub>5</sub>

INSTRUCTIONS:

1. Put your name and student number on the answer sheet.
2. NOTE: on this answer sheet, the answer spaces go ACROSS the WIDTH of the page.
3. The test consists of statements which are either false or true. You might think a statement is more false than true, then you should mark the number on your answer sheet which corresponds with false.

If you think a statement is more false than true, mark 1 on the answer sheet for that statement. If you think it is more true than false, mark 2 for that statement. Remember,

1 = false

2 = true

4. There is no penalty for guessing, so respond to EVERY statement.

## Oral instructions for Class 1

I'm interested in how aware you are of some elements of social psychology, since you've nearly finished introductory psychology, and have had some exposure to social psychology. This test will help me to get this information. Its results will help me with some future course planning. The results won't affect your grade in this course.

This test is different than the kind that you are used to. It is essential for you to read the instructions very carefully.

Notice that the answer sheets are slightly different than the ones you usually use. (Hold up the answer sheet and demonstrate the following, making sure that the ENTIRE class sees it.) The answer spaces go in order across the width of the answer sheet, starting with 1 here . . .

Answer all of the test items. You might find some of them a little difficult, because they were originally written for people who had had an entire course in social psychology, but still, do your best to give me an indication of what you know about social psychology and the social issues in this test.

Remember, the test is different than the ones you're used to, so you must read the instructions very carefully.

(go ahead... When you're through, put the answer sheet back in the test booklet and drop it in the box at the door when you leave.)



## Oral instructions for Class 2

I'm interested in how aware you are of some elements of social psychology. Everyone comes into psychology with some information about social psychology, and you know about many issues. I would like to know how much you do know about the subject before you have any formal instruction in it. The test results will help me with some future course planning. They won't affect your grade in this course.

This test is different than the kind that you are used to. It is essential for you to read the instructions very carefully.

Notice that the answer sheets are slightly different than the ones you usually use. (Hold up the answer sheet and demonstrate the following, making sure that the ENTIRE class sees it.) The answer spaces go in order across the width of the answer sheet, starting with 1 here . . .

Answer all of the test items. You might find some of them a little difficult, because they were originally written for people who had had an entire course in social psychology, but still, do your best to give me an indication of what you know about social psychology and the social issues in this test.

Remember, the test is different than the ones you're use to, so you must read the instructions very carefully.

(go ahead... When you're through, put the answer sheet back in the test booklet and drop it in the box at the door when you leave.)



