INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
DONOVAN, Michael John, 1939-
MEASUREMENT OF EXTREME RESPONSE:
OPERATIONAL DEFINITIONS AND RELIABILITY.

The Ohio State University, Ph.D., 1976
Psychology, general

Xerox University Microfilms, Ann Arbor, Michigan 48106
MEASUREMENT OF EXTREME RESPONSE:

OPERATIONAL DEFINITIONS AND RELIABILITY

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Michael John Donovan, A.B., M.A.

The Ohio State University

1976

Reading Committee:

Frank M. Fletcher, Adviser
Samuel H. Osipow
W. Bruce Walsh

Approved By

Frank M. Fletcher
Adviser
Department of Psychology
In memoriam

Herbert A. Toops

He taught respect for the integrity of data.
ACKNOWLEDGEMENTS

I wish to express my deep gratitude to my advisor Dr. Frank M. Fletcher for his longstanding and most valued support and assistance and, above all, for the opportunity to benefit from the example of his commitment to inquiry.

To Dr. W. Bruce Walsh and Dr. Samuel H. Osipow, members of the committee, I also wish to express my gratitude for their longstanding support, encouragement and assistance.

To Dr. Naomi K. Keara a special appreciation is due, particularly for her assistance in the collection of the data. I am also grateful for the consultation on statistical points which was so freely given by Dr. Robert J. Wherry, Dr. B. J. Winer, and Dr. David A. Specht.

The indirect contributions of virtually all of the faculty and students with whom I have had the opportunity to interact at The Ohio State University must also be acknowledged.

The contributions to this thesis by my wife, Dr. Joyce J. Fitzpatrick, have ranged from keypunching to cooking, and far, far, inexpressibly far, beyond. It is with deep pleasure that I here publicly express my gratitude to her and to my mother, Mrs. John J. Donovan, for the gift of their unfailing trust.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>I. PROBLEM</td>
<td>1</td>
</tr>
<tr>
<td>Prior Research</td>
<td>1</td>
</tr>
<tr>
<td>Reliability and Generality</td>
<td>2</td>
</tr>
<tr>
<td>Sex Differences</td>
<td>3</td>
</tr>
<tr>
<td>Age</td>
<td>3</td>
</tr>
<tr>
<td>Intelligence, Education, Occupation, Culture</td>
<td>4</td>
</tr>
<tr>
<td>Adjustment</td>
<td>5</td>
</tr>
<tr>
<td>Specific Personality Attributes</td>
<td>5</td>
</tr>
<tr>
<td>Discussion</td>
<td>7</td>
</tr>
<tr>
<td>Effects of ER</td>
<td>8</td>
</tr>
<tr>
<td>Content</td>
<td>9</td>
</tr>
<tr>
<td>Meaning</td>
<td>10</td>
</tr>
<tr>
<td>Content and Meaning</td>
<td>12</td>
</tr>
<tr>
<td>Scoring</td>
<td>14</td>
</tr>
<tr>
<td>Overview</td>
<td>18</td>
</tr>
<tr>
<td>Formulation of Hypotheses</td>
<td>20</td>
</tr>
</tbody>
</table>
Hypothesis ONE.

Hypothesis Two.

Hypothesis Three.

Hypothesis Four.

Hypothesis Five.

Hypothesis Six.

II. METHODOLOGY.

Terminology.

Data Collection.

Sample.

Instruments.

Course Evaluation Form.

Study Behavior Form.

Instrumentation Rationale.

Midpoint Styles.

Deviation Hypothesis.

Reflection.

Scoring Methods.

Position Use Tally.

Deviations About Hypothetical Means.

Deviations About Empirical Means.

Deviations About Subject Means.

Indexes of Dispersion.

Score Deviations.

Combinations of Score Deviations.
III. RESULTS AND DISCUSSION

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computation and Reporting</td>
<td>37</td>
</tr>
<tr>
<td>Preliminary Evaluations</td>
<td>38</td>
</tr>
<tr>
<td>Characteristics of the Measures</td>
<td>39</td>
</tr>
<tr>
<td>Correlation Results</td>
<td>42</td>
</tr>
<tr>
<td>Hypothesis One</td>
<td>43</td>
</tr>
<tr>
<td>Hypothesis Two</td>
<td>49</td>
</tr>
<tr>
<td>Hypothesis Three</td>
<td>51</td>
</tr>
<tr>
<td>Other Statistical Results</td>
<td>53</td>
</tr>
<tr>
<td>Hypothesis Four</td>
<td>58</td>
</tr>
<tr>
<td>Hypothesis Five</td>
<td>63</td>
</tr>
<tr>
<td>Hypothesis Six</td>
<td>65</td>
</tr>
<tr>
<td>Normalization</td>
<td>68</td>
</tr>
<tr>
<td>Adjustment</td>
<td>72</td>
</tr>
<tr>
<td>Scoring Methods</td>
<td>73</td>
</tr>
<tr>
<td>Interrelation of Methods</td>
<td>75</td>
</tr>
<tr>
<td>Discussion</td>
<td>78</td>
</tr>
<tr>
<td>Limitations</td>
<td>83</td>
</tr>
<tr>
<td>Future Research</td>
<td>83</td>
</tr>
<tr>
<td>Summary</td>
<td>85</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>90</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>93</td>
</tr>
<tr>
<td>LIST OF REFERENCES</td>
<td>94</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table                                      Page
1. Scale Based Scoring Methods.             29
2. Score Based Scoring Methods.             30
3. Characteristics and Interrelationships of Scales and Measures. 40
4. Reliability of ER Scores Mean Coefficients and Their Significance Levels. 44
5. ER Scores and Content Scores Coefficients of Intercorrelations and Their Significance Levels. 45
6. Cumulative Percentages of Coefficients Attaining Specified Levels of Significance. 46
7. Differences Between Coefficients For All Possible Pairs of Scoring Methods Cumulative Percentages Attaining Specified Levels of Significance. 60
8. Tests of Statistical Significance Related to Hypotheses Four, Five, and Six. 62
9. Reliability of First Eleven Scoring Methods Coefficient Alpha Values For Each Measure and Across All Eighty Scales. 74
10. Coefficients of Intercorrelations of ER Raw Scores Summed Across Eight Measures For Selected Scoring Methods. 76

vii
CHAPTER I

PROBLEM

This research is an investigation of the existence and nature of a phenomenon which is called extreme response style (ER). This phenomenon may be described in general terms as the relative tendency to give extreme responses to multipoint scales. There exists a substantial body of research results which asserts that ER is a reliably observable differential characteristic of individuals. The very nature of this body of research results raises questions concerning the existence and nature of the phenomenon. The hypotheses and the methodology of this study are developed in an attempt to yield data analyses which seek results in answer to such questions about the existence and nature of ER.

Prior Research

Research on ER has developed within the general context of investigations of response styles. Concern for the accuracy of objective assessments has focused great attention on their reliability, validity, and generality. Unique and consistent patterns of response characteristic of individuals and independent of the purposes of these assessments have long been considered to be sources of error which should be excluded from these measurements (Cronbach, 1946, 1950; Horst, 1972). Others have urged that these characteristic response
patterns should be studied as personality attributes (Jackson & Messick, 1958; Lorge, 1937; Wiggins, 1962).

Three response styles have been identified and studied extensively. These are acquiescence (Bass, 1955; Couch & Kenniston, 1960), social desirability (Crowne & Marlowe, 1964; Edwards, 1953, 1957), and deviation (Berg, 1957, 1959). Research on these response styles has been summarized (Berg, 1967; Jackson & Messick, 1958; McGee, 1962b; Messick & Ross, 1962). A fourth response style, the style investigated here, ER, has also been extensively studied in the same context and the results of this research have also been reviewed (Hamilton, 1963; O'Donovan, 1965).

Reliability and Generality. Hamilton (1963) asserts that criticism of the lack of reliability of the other response styles (e.g., McGee, 1962a; Rorer, 1965) does not apply to ER. He marshalls in evidence numerous test-retest reliability coefficients, internal consistency reliability coefficients, and "generality" coefficients or correlations between different measures of ER.

Subsequent reports of a similar nature may also be cited. Merrens (1970) found such support on two new measures of ER. Also using new measures employing visual and auditory stimuli Merrens (1971) found substantial stability over time. Most impressive, however, are the findings of Goldsamt (1971) which showed that for 27 measures of ER reliability coefficients greater than .304 were found for 22.

Questions may be raised about the value of such results as evidence for the reliability of ER. Before these questions are addressed,
however, it is appropriate to consider the research which generated these results. This will serve to illustrate the importance of the questions concerning the reliability of ER and to provide a frame of reference within which questions about the value of the evidence may be discussed.

**Sex Differences.** Females have given more extreme responses than males in some samples: normal adults (Soueif, 1953), college students (Berg & Collier, 1953; Borgatta & Glass, 1961; Brown, 1964; Crandall, 1965; Hamilton, 1965; Peak, Monev, & Clay, 1960), prisoners (Borgatta & Glass, 1961), and abnormal adults (Brown, 1964; Parsonson, 1969a). Nonsignificant differences between males and females have been observed in other samples: normal adults (Brengelmann, 1960a; Zuckerman, Oppenheimer, & Gershowitz, 1965), normal children (Light, Zax, & Gardiner, 1965; Soueif, 1953), Negro college students (Berg & Collier, 1953), and abnormal adults (Borgatta & Glass, 1961; Brengelmann, 1960a, 1960b). In one study males were found to be significantly higher than females in ER (Brengelmann, 1959b).

**Age.** Substantial evidence demonstrates a curvilinear relationship of ER with age. Findings show ER to be high in children, to decrease to a minimum in young adults, and to increase with age after that period (Das & Dutta, 1969; Emmerich, 1971; Emmerich, Goldman, & Shore, 1971; Hesterly, 1963; Iwawaki, Zax, & Hitsuoka, 1969; Light, Zax, & Gardiner, 1965; Soueif, 1958; Zax, Gardiner, & Lowy, 1964). An interaction of age and sex may be observed in an earlier achievement of minimal ER for girls (Emmerich, 1971).
Intelligence, Education, Occupation, Culture. Without offering supportive evidence, Osgood (1941) concluded that ER was related to intelligence, education and occupation. Goldsamt (1971) failed to find a significant relationship between ER and Scholastic Aptitude Test -- Verbal scores. While Zuckerman and Norton (1961) found no relationship, other investigators, however, report a negative relationship between ER and intelligence: Brengelmann in his studies (1958, 1959a, 1960a, 1960b); Kerrick (1954) for high and low IQ groups; Light et al. (1965) for such groups of fourth and twelfth graders, but not for eighth graders; Wilkinson (1970) with Shipley-Hartford Vocabulary, Abstract and Total scores; and Das and Dutta (1969) report such results from a series of Indian studies.

Shulman (1973) studied 6,703 females in fourteen studies and found a negative relationship between ER and educational level.

Brengelmann (1959b) studied occupational status in a German sample and found higher ER for skilled and unskilled workers than for academic, business, and professional groups. He could not replicate these findings, however, on English samples of either normal or abnormal subjects (Brengelmann, 1960b). Souieif (1958) studied adolescent male Moslems and found less ER in the upper-middle class sample than in the lower-middle class sample. Actors and actresses have been shown to be higher on ER than high school teachers (Zuckerman et al., 1965) and novice nuns higher on ER than college females (Zax, Cowen, & Peter, 1963). The studies reviewed by Das and Dutta (1969) revealed higher degrees of ER in persons with positive attitudes toward religion.
Although no differences were found with younger subjects, twelfth grade Japanese were less extreme in their responses than Americans (Iwawaki et al., 1969). Two samples of German males made more extreme responses than did a sample of English males (Brengelmann, 1959b). Negro males have been found to make more extreme responses than white males, although this result was not found for females (Berg & Collier, 1953). In Egypt minority Christians give more extreme responses than do Moslems (Soueif, 1958).


Brengelmann conducted a series of five studies in which he used ten samples (1958, 1959a, 1959b, 1960a, 1960b). He evaluated correlations with dogmatism and low ambiguity tolerance or rigidity and found 20 of 23 to be significantly positive. Schutz and Foster (1963) found ER to be positively related to inflexibility.

Damarin and Messick (1965) related high ER to a need for certainty as did Brengelmann (1959a). Brim and Hoff (1957) showed high ER to be related to low tolerance for ambiguity. Goldsamt (1971) found five of nine correlations with need for certainty to be significantly positive. White and Harvey (1965) showed high ER to be related to concrete rather than abstract modes of cognition.

Three studies have found high anxiety subjects to be higher in ER (Berg & Collier, 1953; Lewis & Taylor, 1955; Norman, 1969), while two studies have failed to demonstrate a difference except in interaction with a third variable. Kerrick (1954) found high ER in high anxiety, high IQ subjects and Crandall (1965) in high anxiety females. This result was not found, however, in the Brengelmann studies (1958, 1959a, 1959b, 1960a, 1960b). In related research Schonpflug and Schonpflug (1972) found extreme response to be more frequent under intermediate stress.

The Brengelmann studies (1958, 1959a, 1959b, 1960a, 1960b) found no relationship of ER to introversion although Schutz and Foster (1963) did find a positive relationship. Iwawaki and Zax (1969) found the same relationship, although in interaction with neuroticism.
Several multivariable personality tests have been examined with regard to ER. Correlations with 16PF scales were examined by Jones and Rorer (1973) who deemed them not useful and by Borgatta and Glass (1961) for three samples which showed 6 of the 96 correlations to be significant, chiefly for mental patients and on scales relevant to self control and self determination. Borgatta and Glass (1961) also used the EPPS and found only 5 of 90 correlations to be significant. Also using the EPPS Zuckerman et al. (1953) found no significant correlations. With the MMPI scales Jones and Rorer (1973) found no useful correlations with ER nor did they find such correlations between ER and the scales of the CPI. Kerrens (1970, 1971) found two different sets of low correlations with the PRF. Hamilton (1965) found 2 of 40 correlations with Guilford-Zimmerman scales to be significant.

Discussion

While the findings which have been reviewed above are not entirely consistent, they do lend themselves to the formation of an impression of a person who makes extreme responses.

Such a person might be imagined to be a female of either very young or very old age, low intelligence, little education, in a low status or deviant occupation, from a minority religion and race and holding a positive attitude toward religion. She would be poorly adjusted and evidence psychopathology. She would be authoritarian, dogmatic, inflexible, intolerant of ambiguity and have a high need for certainty. She would be anxious and under stress, concretistic in her thinking, a neurotic introvert.
It may be suggested that if these terms and their contraries were to be evaluated by means of a semantic differential the responses to this impression would be demonstrated to be negative, in the extreme.

It may be further suggested that these results are an artifact of experimenter bias, that they serve to support a negative stereotype collectively held by the researchers, and that this accounts for two noteworthy phenomena: the remarkably small range of variables investigated and the relatively little attention given directly to the issue of extreme response.

**Effects of ER.** It is noteworthy that the research which has been presented includes no attempt to determine whether an extreme response set or style does undesirably influence or bias psychological measurements. One study with negative results has been recently reported. Jacobs (1971) administered some scales of the CPI in both 2 point and 6 point formats together with a measure of ER. He found that the 6 point format increased the reliability of the scales. The increase was completely attributable to female subjects since reliabilities were slightly higher for males with the 2 point format. The change in reliability was not attributable to ER.

After Cronbach (1946) first seriously raised the issue of extreme response style, when he warned that ER would falsely inflate reliability if multipoint rather than two point scales were used, fully twenty-five years passed before this first investigation of that subject by Jacobs (1971). Almost all of the research on personality reported above was conducted in that interval.
Content. Measures of ER may be distinguished by their content. Some instruments are measures of the very same personality traits with which ER has been associated: Rokeach's Dogmatism Scale (Biggs & Das, 1973; Gordon, 1971; Peabody, 1962), the California F Scale (Gordon, 1971; Klein, 1963; Peabody, 1962; Zuckerman et al., 1958), the Rationality Index and Wagner's Hand Test (Peiser, 1972), the Conservatism Scale and General Agreement Scale (Peabody, 1962), and the Parental Attitude Research Instrument (Zuckerman & Norton, 1961; Zuckerman et al., 1958).

Some investigators have apparently achieved supportive but independent findings by simple examination of their research instrument (Biggs & Das, 1973; Emmerich, 1971; Emmerich et al., 1971; Wilkinson, 1970).

Measures specifically created to evaluate response styles in an objective fashion through the use of ambiguous content have also been used: the Personal Friends Check List (Biggs & Das, 1973; Brengelmann, 1959a, 1959b, 1960a; Das & Dutta, 1969; Neillens, 1971; Souelf, 1958), the Perceptual Reaction Test (Berg, 1953; Berg & Collier, 1953; Forehand, 1962), the Word Reaction Test (Berg, 1953; Brengelmann, 1958), the Activities Checklist (Forehand, 1962; Rundquist, 1950), Attitude Statements (Peak et al., 1960), the Word Checklist (Rundquist, 1950), the Adjectives Checklist (Hamilton, 1965) and the Aphorisms Checklist (Hamilton, 1965; Schutz & Foster, 1963).

In his review Hamilton (1968) urged the use of content-free scales as being of primary importance. At least partially in consequence of this injunction novel measures were subsequently employed. Generally
these have taken what is described as the semantic differential format and employed various stimuli in place of verbal concepts: Rorschach Ink Blots (Iwawaki, Zax, & Mitsuoka, 1969; Iwawaki & Zax, 1969; Norman, 1969), Thematic Apperception Test pictures (Norman, 1969), tonal patterns (Merrens, 1971), and visual and auditory stimuli (Merrens, 1970). Goldsamt (1971) used rating of associative value, rather than semantic differential scales, to obtain responses to: CVC nonsense syllables, numbers, and random shapes.

Many investigators report having employed semantic differential scales (e.g., Arthur, 1966; Becker & Myers, 1970; Jones & Rorer, 1973; Mogar, 1960; Neuringer, 1961; Norman, 1969; Parsonson, 1969a, 1969b; Worthy, 1969). Often these semantic differential scales have been applied to concepts or stimuli that are related to the personality characteristics which are of interest to the investigators.

Meaning. The issue of the content of instruments has been discussed in the broader context of the issue of meaning in a lengthy review by O'Donovan (1965). He prefers the term polarization as being less emotionally loaded than the description of extreme response and summarizes the operational definition as the distance from the point of origin or neutral point on a rating scale. Nonetheless, with some additions he discusses the same body of research which has been reviewed here.

O'Donovan contrasts work linking ER to pathology with consideration of the linkage of ER to meaningful commitment and constructive behavior. Extreme response has been considered as an operational definition of decisiveness (Cronwell & Caldwell, 1962), saturation of

The relationship of ER to deviance may at first consideration seem simple and clear but upon further study it may become quite complex. For example a sample of psychologists has been found to endorse the same value pattern as the typical American but to do so more extremely (Johnston, 1964). Are such subjects deviant, nondeviant, deviantly nondeviant, or nondeviantly deviant? Further complexities may be introduced as evidenced by Zax, Cowen, and Peter (1963) who found novice nuns deviated from college females by more extreme endorsement of ratings of social desirability.

Meaningfulness is clearly related to ER as is indicated in the review by Underwood and Schulz (1960). Emotionality or affectivity ratings have been related to meaningfulness by Noble (1958). Osgood, Suci, and Tannenbaum (1957) have reported the relationship between ER and speed of response. Speed of response was linked to intensity of attitudes by Postman and Zimmerman (1945). Similar findings are reported in a line of studies dating to Henmon (1911).

Meaningfulness and ER are shown to be related in research on person perception or social psychophysics (Sherif & Hovland, 1961). Increases in ER are found with greater involvement of the subject (Hovland & Sherif, 1952; Janis, 1960; Pettigrew, Allport, & Barnett, 1958), with the introduction of high valence stimuli or positive
reinforcement in psychophysical judgments (Tajfel, 1959; Tajfel & Cawasjee, 1959) and with increased personal relevance of rating scales (Hastorf, Richardson, & Dornbusch, 1958).

The tendency of subjects to use extreme intervals in rating relatively vivid or meaningful affects was described by Block (1957). This leads to consideration of the findings presented linking ER to pathology. O'Donovan (1965) observes the inconsistency of those findings and additionally notes the failure of other work to support such conclusions (Barker, 1958; Paine, 1964). He cites literature to document the relationship of stereotypic behavior, or limited behavioral variability to pathology. He notes the flattened affect of schizophrenics among many such instances. Four studies have found less ER in schizophrenics (Bopp, 1955; Morris et al., 1960; O'Donovan, 1960; O'Donovan, Morris, & Eiduson, 1959).

Content and Meaning. Both the research findings which support the pathology hypothesis and the inconsistency of those findings may be partially attributable to the meaning of the content of the instruments used to assess ER. Merrens (1970, 1971), for example, using high content and low content measures, has observed somewhat divergent results.

Presentation of stimuli related to a personality attribute may be relatively meaningful to individuals to whom the attribute itself is relatively meaningful. In such a case the individuals would be expected to make relatively extreme responses. Whether the same individuals would continue to make relatively extreme responses to unrelated stimuli would remain problematic.
It may be argued that to a greater or lesser extent this is in fact what has occurred in much of the research on personality attributes associated with extreme response. In those instances where measures of the attributes under investigation or of related attributes have been used to assess ER this is clear.

Measures intended to be content-free or meaningless may operate in a similar fashion. It should be considered that such instruments have been created without attention to validity on a largely ad hoc basis by researchers concerned primarily with a relatively restricted domain of psychological phenomena. The paucity of research directly on the issue of ER itself may be advanced as an indication of this restriction and in suggestion of such a bias.

Hypothetical interpretations of various research results may be advanced. For example, it is possible that actors and actresses are more artistic than teachers and so find abstract designs more meaningful and respond to them more strongly (Zuckerman et al., 1965). Perhaps individuals with a high need for certainty are threatened by instructions to assign association values to nonsense syllables and so experience this request as meaningful and give extreme responses (Goldsamt, 1971). And the question remains: How do psychologists who find authoritarianism, dogmatism, inflexibility, intolerance of ambiguity, need for certainty and anxiety so meaningful as to be worthy of research go about constructing ad hoc, unvalidated, meaningless measures?

This criticism has not been extended to refute the findings which have been cited but rather to explain their inconsistency and highlight
their limitations. It is suggested that the research findings should be accepted as descriptive of the relationship between the personality attributes and extreme response to a more or less consistent set of related measures. The relationship of these findings to the generalized individual characteristics of extreme response style remains problematic, as does the very existence of such a characteristic.

Scoring. Scoring method may be an additional factor contributing to the inconsistency of findings in ER research. On this elemental level of operational definition unanimity of practice is not to be observed nor has extensive research on differences between methods been undertaken.

While in perhaps a majority of cases the method adopted has been to simply tally the total number of times that the respondent elects one of the single most extreme responses, it will be recalled that O'Donovan (1965) described the variable as deviation from the point of origin or neutral point on a rating scale. Other investigators have adopted similarly divergent practices.

When a 9, 10, or 11 point scale is used, for example, some investigators have tallied the two extreme categories at each end of the scales (e.g., Borgatta & Glass, 1961; White & Harvey, 1965). With a 9 point scale Peak et al. (1960) defined a bimodality index to be the mean number of responses on points 1, 2, 8, and 9 minus the average number of other responses. Some researchers have measured ER as the summed deviations from scale midpoints (e.g., Mogar, 1960; Neuringer, 1961; Peabody, 1962).
Although in general the impact of divergent operational definitions of ER has not been examined some research has been directed to the questions that arise. Goldsamt (1971) used 6, 12, and 18 point scales and scored them by two dichotomous and one intensive method. While he observed the scoring methods to load on the same factors and in general to be highly intercorrelated, only 5 of the 9 scores were significantly related to desire for certainty and he noted that the single extreme response method was slightly more reliable. Iwawaki and Zax (1969) related ER to psychopathology but could not replicate their findings when they scored for intermediate and neutral ratings. Worthy (1969), however, found use of the midpoint to be positively related to use of the extreme positions and asserted that it should, therefore, be considered an extreme response. Finally, Carl (1968) used 2, 3, 5, and 7 point scales with the MNPI items and tallied the frequency of response for each category. He identified five consistent factors, including the contrast between mild agreement and mild disagreement, use of the intermediate position, and use of the extremes on the seven point scales, but asserted that ER scores are scale specific except for the tendency to use the intermediate position.

Emmerich (1971) used 4 point scales with the terms "never" and "very often" used to describe the extreme points. He obtained a significant interaction for one term but not the other and suggested the difference was due to the interaction of two variables: extremeness and absoluteness.
Investigations have been undertaken by some researchers of the relationship of extreme positive responses to extreme negative responses (Brengelmann, 1959a, 1960a; Peabody, 1962). While the measures used are subject to the criticisms described above, it is noteworthy that the correlations reported are markedly lower than the reliability, test-retest, and between-test coefficients of the same measures. Brengelmann (1958, 1959a, 1959b, 1960a, 1960b) used extreme positive responses to relate ER to authoritarianism, dogmatism and rigidity for he found no such relationship with extreme negative responses. Curiously, Gordon (1971) used the same measures of dogmatism and authoritarianism as measures of ER and found that extreme negative responses contributed more variance to the ER scores. It would seem inappropriate to adduce results specific to extreme negative responses on measures of restricted content as evidence for the existence, much less the correlates, of the generalized personal characteristic described as extreme response style or set.

Such evidence, however, does raise questions concerning item characteristics relevant to ER. Nowakowska (1973) has observed item stability to be related to negative emotional reaction to content and to negative previous experience and item stability to be related to negative intellectual evaluation of both the question and the answer.

The most important determinant of the relative importance of two single inferences is their range and Warr (1971) has observed that the second determinant is evaluative direction with the positive being more important. Positive evaluative direction is defined as either the assertion of a positive characteristic or the denial of a negative
characteristic while negative direction is either the assertion of a negative or denial of a positive. Warr (1974) has further investigated the rules for means of an item pair in terms of their means and the range of the difference between their means to account for pushover phenomena.

Additional relevant work on item response would include the findings of Schonpflug and Schonpflug (1972) that instability of judgment frame is linearly related to stress and that under intermediate levels of stress responses are accentuated and variation of scale choices increases. Kurdock and Van Bruggen (1970) observed individuals who use fewer categories make fewer inclusions in their categories and that the receipt of relevant information reduced response extremities of both broad and narrow categories. Wheeler (1973) observed that individuals are veridical in judgments of odds for individual data without regard to range but conservative in inference when they revise over a sequence of data. This would seem relevant to Shulman's suggestion (1973) that more extreme responses are observed to a bipolar or semantic differential than to a Likert format because the former is more difficult to understand.

In all of the ER research which has been cited only one study permitted item characteristics to influence scoring (Emmerich et al., 1971). In that case intersubject standard deviations for items were summed to be used as ER scores, demonstrably discounting differences attributable to variation in item means.

In sum, scoring of ER has been relatively uninvestigated, inconsistent and without allowance for item characteristics. The
inconsistency and failure to take account of item characteristics in scoring may or may not have spuriously generated and/or restricted the findings of research on ER and relating it to other variables.

**Overview.** One may speculate that interest in ER begins in the following way. A psychologist asks some questions of his subjects. He reviews the answer sheets and notes that on the multipoint scales some subjects seem to have used the extreme response positions with relatively greater frequency. He then notes that these same subjects seem to have used the intermediate response positions with relatively less frequency.

At this point it might seem that no value judgment has been made and no negative inference has been drawn. Reasonable questions, however, arise. Why has the psychologist noted the behavior of these subjects who give many extreme and few intermediate responses? By definition there must be other subjects who have not caught his attention: those who with some consistency select relatively indefinite response positions and those who display relative inconsistency in their selection of response positions.

If it is assumed that the questions were not developed to measure a response style but rather some other variable, it may also be assumed that these extreme response positions were included because of a belief that they are valid to the measurement of the variable, that the measurement was intended to be reliable, and thus that consistent use of extreme response positions would be desirable from the point of view of the investigator as a contribution to both reliability and validity.
When a group of such investigators with virtual unanimity ungraciously infer that the tendency to use extreme response positions is a manifestation of pathology and attempt to demonstrate this within a relatively restricted domain, it is appropriate to speculate on this extreme response on the part of the investigators.

In view of the literature which has been cited above linking meaningfulness of stimuli to extremeness of response it may be inferred that the researchers were asserting that it is pathological to experience stimuli as meaningful or at least that it is pathological to respond to stimuli as if they are experienced as meaningful.

It may be noted that most of the relevant research was conducted in the interval between World War II and the Vietnam War. This body of research had its beginnings in the Eisenhower years, the era of the "man in the gray flannel suit," a time when stability, adjustment, and conformity were highly valued. Such research has almost ceased in the more recent period when greater value was accorded: consciousness, meaningfulness, awareness, sensitivity, expressiveness, and novelty or deviance in general.

It is notable that research on all four of the response styles: yea-saying, social desirability, deviant response, and extreme response, developed in the same context. Psychologists who were attempting to understand and measure undesirable deviance or pathology attempted to investigate response styles that might interfere with this research. They agreed in all four fields that the tendency to confound their measurements was itself an expression of undesirable deviance of pathology.
It may be suggested that much of this argument was circular, that it ended where it began. They looked at pathology; they saw pathology; they looked no further.

Formulation of Hypotheses

The research on ER which has been reviewed is extensive and it does support the proposition that certain personal attributes are related to ER on highly related measures. The research fails to demonstrate that the attributes are related to a generalized characteristic style of responding, however, because the question of the existence of such a style has never been carefully examined by evaluation of responses to meaningful but unrelated material.

A careful examination is necessary in order to determine whether a generalized characteristic style of response does exist which can reliably be established to differentiate individuals. This examination must be made through the use of material which is highly relevant to the experience of the subjects but which does not directly reflect personality attributes.

This examination must include examination of ER in the more broadly defined context of characteristic styles of responding to multipoint scales. A variety of such styles must be operationally defined with care. The examination should include consideration of the effects of the content of the stimuli, of the characteristics of both the items and the subjects.

Six specific hypothetical assertions require evaluation.
Hypothesis One. It is hypothesized that ER scores are reliable.

Hypothesis Two. It is hypothesized that ER scores are related to content.

Hypothesis Three. It is hypothesized that ER scores are reliable independent of their relationship to content.

Hypothesis Four. It is hypothesized that the reliability of ER scores is related to scoring methods.

Hypothesis Five. It is hypothesized that the relationship of ER scores to content is related to scoring methods.

Hypothesis Six. It is hypothesized that the reliability of ER scores independent of their relationship to content is related to scoring methods.
CHAPTER II

METHODOLOGY

Terminology

Initial to the discussion of methodology clarity requires explicit definition be given of the usage here of several terms which would otherwise be subject to various definitions.

The word scale is used to indicate an item in a multipoint format. The word measure is used to indicate a collection or set of scales which are intended to measure a single variable. The word instrument is used to indicate a collection or set of measures which are of common origin.

Data Collection

The data were collected at the University of Wisconsin at LaCrosse in May of 1969. The device which was used to collect the data included the two instruments which are described below and upon which the analyses were performed: the Course Evaluation Form and the Study Behavior Form. The device was also used to collect additional data which are not reported as part of the present study: the Marlowe-Crowne Social Desirability or Need for Approval Scale or Personal Reactions Inventory; the Locus of Control of Reinforcement or Internal-External Control or I-E Scale or Social Reactions Inventory, which was developed by Liverant and has been widely attributed to Rotter; the Role Scale or Anomie or
Personal Views Inventory; and personal, social, and academic descriptive information and opinions.

Independently collected were related data which also are not reported in the present study: High School Rank; ACT Scores; Number of Semester Hours Completed; Accumulated Grade-Point Average; Points Earned in Course; and Course Grade.

Sample

The sample consisted of undergraduates enrolled in introductory psychology courses, one course in two sections with two different instructors and a different course with a third instructor. There were a total of 121 males and 159 females or 230 subjects in all.

Instruments

Course Evaluation Form (CEF). Fletcher (1967, 1972) has developed an instrument on which students may report judgments about courses in which they have been enrolled. The format is similar to that of a semantic differential. Rather than functioning to differentiate among several semantic stimuli, however, the instrument serves to analyze a single stimulus, experience in a course. The instrument employs 7 point bipolar scales with verbal judgments used to anchor each pole. The instrument yields five measurements: Difficulty, Content, Instruction, Examinations, and General Evaluation. To the nine scales of each of these measures one supplemental scale was added, in the present study, in order to balance reflected scales with unreflected scales.

Study Behavior Form (SBF). The present writer prepared a set of sixteen items, also in a 7 point bipolar format, which are anchored by
verbal descriptions of study behaviors which Robinson (1961) has described as particularly effective or ineffective. The set of sixteen scales was presented to the student three times and on each presentation the student was instructed to respond in the context of his perception of the particular course in which the instrument was administered. The three presentations constitute three measures. The student was instructed to describe behavior in the course: by himself; by the average student; and by the ideal student. In the second presentation, in which the student was instructed to describe the average student, the verbal descriptions of study behavior which anchor the poles of the scales were reversed. That is, anchors which were on the right side in the first and third presentations were on the left side in the second presentation. The instrument then consists of three measures: Self, Average, and Ideal. Each measure contains sixteen scales. In order to simplify computation only ten of the sixteen scales are utilized for data analysis of the present study.

The ten seven-point bipolar scales which constitute each of the eight measures which are used in this study are listed in Appendix A.

Instrumentation Rationale

It was argued above that the existence of generalized styles or sets or biases in responses to multipoint scales cannot be demonstrated by the use of instruments which measure personality attributes for such evidence may simply demonstrate a characteristic of such measurements. Evidence gained from the use of the CEF and the SBF is not subject to such criticism.
On the basis of their content the CEF and the SBF do not appear to be measures of personality. The instruments do not ask questions about personality nor are they so vague or ambiguous that it could be argued that they are covert measures of personality. While it may be argued that all behavior is but the manifestation of personality, it seems clear that these instruments yield responses to transient extrasubjective phenomena and are not direct measures, either overt or covert, of basic personality characteristics.

Findings from research on the CEF (Fletcher 1967, 1972) both demonstrate and illustrate this point. Factor analytic techniques were employed in the development of the instrument through eight revisions. When administered to each of a number of groups of subjects, the five measures evidence remarkable factorial purity and stability. If the measures were measures of personality then their intercorrelations would be stable if based on groups of subjects with comparable personalities. Personality characteristic A, for example, would have the same association with personality characteristic B for two sections of introductory psychology students. An examination of the intercorrelations among CEF measures for 32 classes, 56 of which were introductory psychology sections, does not show such a stable relationship between measures. Indeed, the average range for the ten intercorrelations is fully .947. The factorial purity and stability of the measures is evidence that the variability of the intercorrelations is not attributable to unreliability in the measures. In fact the split-half reliability coefficients based on 31 sections of introductory
psychology are: Difficulty, .79; Content, .73; Instruction, .90; Examinations, .90; and General Evaluation, .86. Clearly the CEF yields highly reliable measures, not of personality, but of experience in a particular course.

The CEF is a sophisticated, well developed, highly stable instrument which may be expected to display relative homogeneity of item means and standard deviations. The 3BF is a crude, unanalyzed, ad hoc collection of items which it is expected subjects will perceive to be quite heterogeneous. The three measures of the 3BF are contrasting: Self, Average, and Ideal. For these reasons it was anticipated that, in contrast to the CEF measures, the 3BF measures would be of little internal consistency and would evidence relative heterogeneity of item means and standard deviations. If a response style is truly a generalized characteristic it should be demonstrable with both instruments.

**Multipoint styles**

This investigation is concerned with general styles of response to multipoint scales. No attempt is made to evaluate any styles of response which cannot be observed on all multipoint scales. For this reason this study does not attempt to evaluate tendencies to give responses which are acquiescent, socially desirable, or statistically infrequent. Multipoint scales may exist which cannot be scored for the presence of such tendencies.

**Deviation Hypothesis**

For the sake of clarity, it should be stated that the present study is not related to the Berg deviation hypothesis (Berg, 1957,
1961, 1967). The present investigation does consider individual difference variables and all such variables specify intersubject differences or deviation. The Berg hypothesis, however, is limited to consideration of behaviors or responses which differentiate individuals or groups from the general population at a statistical level of significance, or which differentiate two or more groups which have previously been defined as deviant. This study considers neither of those issues.

Reflection

On each of the eight measures employed in this study one half of the scales are reflected. This means, for example, that on Difficulty five scales are arranged so that the rightmost response position, response 7, is anchored by a verbal description of difficulty and the selection of response 7 adds 7 points to the raw score on Difficulty. The remaining five scales are anchored on the right side by a verbal indication of lack of difficulty and the selection of response 7 adds only 1 point to the raw score on Difficulty. The effect of this balance in reflection is to offset any impact on raw score of laterality, a tendency to mark response positions on a given side of all scales. Reflected scales are used in all analyses after reflection, that is as if they had been presented in reverse form.

Scoring Methods

The computational method by which an ER score is obtained constitutes an operational definition of ER. As has been noted above, several operational definitions of ER have been employed in the
research which has been reviewed. It has also been noted above that any single operational definition of ER constitutes but one of a variety of variables which may be considered under the more general ruberic of characteristic styles of response to multipoint scales. The operational definitions of ER or scoring methods considered in this study are listed in Table 1 and Table 2.

It is appropriate to examine the very phrase extreme response. Extreme is clearly the operative term. It is an adjective delimiting the noun response and finds its etymological base in a Latin superlative. In order to define a response as extreme or not extreme it is necessary to consider or evaluate the response within a context or frame of reference. The word extreme is an adjective which describes the relationship of a response to one or more other responses, either real or hypothetical.

With virtual unanimity previous researchers have defined extreme response in terms of deviation about a hypothetical mean response, the midpoint on a continuous scale, contiguously operationalized.

Position Use Tallies. The majority of researchers have defined extreme response as the maximum deviation from the hypothetical mean response. For them an ER score is the tally of the number of times the subject selects the single response position nearest to each end of a multipoint scale. Other researchers include in their definitions one or more adjacent response positions. The use of the hypothetical mean response position has itself been defined as an extreme response.

On a 7 point scale, for example, some would define ER as the tendency to use positions one and seven, others as one and two and six
Table 1

Scale Based Scoring Methods

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON</td>
<td>Content score; sum of the values of the responses; sum of $X$</td>
</tr>
<tr>
<td>SM 1</td>
<td>Sum of responses in positions one and seven.</td>
</tr>
<tr>
<td>SM 2</td>
<td>Sum of responses in positions two and six.</td>
</tr>
<tr>
<td>SM 3</td>
<td>Sum of responses in positions three and five.</td>
</tr>
<tr>
<td>SM 4</td>
<td>Sum of responses in position four.</td>
</tr>
<tr>
<td>SM 5</td>
<td>Sum of responses in positions one, seven, two, and six.</td>
</tr>
<tr>
<td>SM 6</td>
<td>Sum of responses in positions one, seven, three, and five.</td>
</tr>
<tr>
<td>SM 7</td>
<td>Sum of responses in positions one, seven, and four.</td>
</tr>
<tr>
<td>SM 8</td>
<td>Sum of the absolute values of deviations from the midpoint; sum of absolute values ($</td>
</tr>
<tr>
<td>SM 9</td>
<td>Sum of the absolute values of deviations from the empirical or observed group mean for the scale; sum of absolute values ($</td>
</tr>
<tr>
<td>SM 27</td>
<td>Sum of the absolute values of deviations from the empirical or observed group mean for the scale adjusted for differences between scales in variability; sum of absolute values ($</td>
</tr>
<tr>
<td>SM 10</td>
<td>Sum of the absolute values of deviations from the subject's own mean response on the measure; sum of absolute values ($</td>
</tr>
<tr>
<td>SM 11</td>
<td>Sum of the absolute values of deviations from the subject's own mean response on the measure standardized by scale; sum of absolute values ($X$ standardized by measure-Subject Mean standardized by measure).</td>
</tr>
<tr>
<td>SM 12</td>
<td>Standard deviation of subject on measure.</td>
</tr>
<tr>
<td>SM 13</td>
<td>Kurtosis of subject on measure.</td>
</tr>
</tbody>
</table>

Note. Scores are computed by each scoring method for each of the eight measures or collections of ten scales.
Table 2
Score Based Scoring Methods

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM 14</td>
<td>Absolute value of (CON - CON Mean for Measure).</td>
</tr>
<tr>
<td>SM 15</td>
<td>Absolute value of (SH 1 - SH 1 Mean for Measure).</td>
</tr>
<tr>
<td>SM 16</td>
<td>Absolute value of (SM 2 - SM 2 Mean for Measure).</td>
</tr>
<tr>
<td>SM 17</td>
<td>Absolute value of (SM 3 - SM 3 Mean for Measure).</td>
</tr>
<tr>
<td>SM 18</td>
<td>Absolute value of (SM 4 - SM 4 Mean for Measure).</td>
</tr>
<tr>
<td>SM 19</td>
<td>Absolute value of (SM 5 - SM 5 Mean for Measure).</td>
</tr>
<tr>
<td>SM 20</td>
<td>Absolute value of (SM 6 - SM 6 Mean for Measure).</td>
</tr>
<tr>
<td>SM 21</td>
<td>Absolute value of (SM 7 - SM 7 Mean for Measure).</td>
</tr>
<tr>
<td>SM 22</td>
<td>Absolute value of (SM 8 - SM 8 Mean for Measure).</td>
</tr>
<tr>
<td>SM 23</td>
<td>Absolute value of (SM 12 - SM 12 Mean for Measure).</td>
</tr>
<tr>
<td>SM 24</td>
<td>Absolute value of (SM 13 - SM 13 Mean for Measure).</td>
</tr>
</tbody>
</table>

Combinations of Score Deviations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM 25</td>
<td>Sum of (SM 15 + SM 16 + SM 17 + SM 18).</td>
</tr>
<tr>
<td>SM 26</td>
<td>Sum of (SM 15 Standardized + SM 16 Standardized + SM 17 Standardized + SM 18 Standardized).</td>
</tr>
</tbody>
</table>

**Note.** Scores are computed by each scoring method for each of the eight measures or collections of ten scales.
and seven, others as four, others as one and four and seven, etc.

It should be noted that the tendency to use one or more response positions is equivalent to the tendency to not use the remaining response positions and it is complementary to the tendency to use those remaining positions. These are alternative phrasings of the same operational definitions.

On 7 point scales, for example, the number of occasions on which position one or position seven is marked must be equal to the number of times the remaining positions are not marked and the number of times these remaining positions are selected must be equal to the number of scales less the number of times positions one and seven are marked.

Considered as tallies of the use of response positions relative to the hypothetical mean response, tendencies to respond to multipoint scales may be comprehensively subjected to operational definition for scales of any given dimension by a finite number of scoring methods (SM's). In the case of 7 point scales these are the following 7 SM's where the responses indicated are tallied:

<table>
<thead>
<tr>
<th>SM</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3 or 5</td>
</tr>
<tr>
<td>3</td>
<td>2 or 6</td>
</tr>
<tr>
<td>4</td>
<td>1 or 7</td>
</tr>
<tr>
<td>5</td>
<td>1 or 7 or 4</td>
</tr>
<tr>
<td>6</td>
<td>1 or 7 or 3 or 5</td>
</tr>
<tr>
<td>7</td>
<td>1 or 7 or 2 or 6</td>
</tr>
</tbody>
</table>
Deviations About Hypothetical Means. SM's 1 through 7 are absolute or discrete for scales. A scale response is simply tallied, that is scored as 0 or 1. At the level of scores, however, these SM's are continuous or relative. There would seem to be no compelling theoretical or empirical reason for this determination to limit SM's to those that operationally define ER as absolute at one level of behavior analysis but relative at another level. For this reason this study also investigates the tendency to make responses which are relatively extreme or deviant from the scale midpoint. Such a definition has been employed by several investigators. SM 8 is operationally defined as the sum across the scales of a measure of the number of response positions by which each response observed is removed from the scale midpoint or the sum of the absolute or unsigned values of the deviations about the hypothetical mean response.

For SM 8 on 7 point scales the values of the response positions are as follows:

<table>
<thead>
<tr>
<th>Value</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3 or 5</td>
</tr>
<tr>
<td>2</td>
<td>2 or 6</td>
</tr>
<tr>
<td>3</td>
<td>1 or 7</td>
</tr>
</tbody>
</table>

Deviations About Empirical Means. SM's 1 through 8 take no account of item characteristics. They make no allowance for the comparison of the behavior of a subject to the behavior of other subjects in response to a given scale. An illustrative item might
have the stem: "Children should be given enough food to sustain life," coupled with a 7 point scale anchored by the terms "agree" and "disagree." If 95% of some sample should select position 7, the position closest to the word agree, and the remainder be scattered in their selections then previous usage would define 7 as a more extreme response than 3. Consideration of such examples suggests that previous research has implicitly accepted the assumption of coincidence between scale means and scale midpoints, the identity of empirical and hypothetical means. Research and practice demonstrate that this does not occur. In the present study, therefore, examination is also made of $S_{II} 9$ which is operationally defined as the sum of absolute deviations about the empirical means of scales.

Similarly, previous practice does not admit the consideration of differences between items in their variability. If virtually all subjects elect position 4 on one scale and on another are about evenly divided between positions 1 and 7 then it may be argued that 7 is a more extreme response on the first item than it is on the second. $S_{II} 27$ is adjusted for differences in both item means and variability and is operationally defined as absolute deviation about empirical means of responses standardized by scale.

**Deviations About Subject Means.** Differences between subjects on the measures used are not considered in either the research which has been reviewed or the $S_{II}$'s which have been described. Scales ranging from like to dislike, for example, might be administered to evaluate several brands of chocolates. One subject might use response positions
have the stem: "Children should be given enough food to sustain life," coupled with a 7 point scale anchored by the terms "agree" and "disagree." If 95% of some sample should select position 7, the position closest to the word agree, and the remainder be scattered in their selections then previous usage would define 7 as a more extreme response than 3. Consideration of such examples suggests that previous research has implicitly accepted the assumption of coincidence between scale means and scale midpoints, the identity of empirical and hypothetical means. Research and practice demonstrate that this does not occur. In the present study, therefore, examination is also made of SM 9 which is operationally defined as the sum of absolute deviations about the empirical means of scales.

Similarly, previous practice does not admit the consideration of differences between items in their variability. If virtually all subjects elect position 4 on one scale and on another are about evenly divided between positions 1 and 7 then it may be argued that 7 is a more extreme response on the first item than it is on the second. SM 27 is adjusted for differences in both item means and variability and is operationally defined as absolute deviation about empirical means of responses standardized by scale.

Deviations About Subject Means. Differences between subjects on the measures used are not considered in either the research which has been reviewed or the SM's which have been described. Scales ranging from like to dislike, for example, might be administered to evaluate several brands of chocolates. One subject might use response positions
clustered about the midpoint. A second subject might give responses clustered near the term dislike. The difference between the ER scores obtained by the previous scoring methods would, perhaps, more accurately be described as a difference between the two subjects in liking for chocolates, the variable being measured, than as a difference in style of responding to multipoint scales. Such a between subjects difference is adjusted for by SM 10 which is operationally defined as absolute deviation of a response from the subject's own mean response.

Another SM to be examined in this study combines the consideration of both item and subject characteristics. SM 11 is operationally defined as the absolute value of the deviation from the subject's mean standardized response.

Adjustment for within subject variability is not to be considered for, while such adjustment would permit examination of between subject differences on a single scale, total scores for all subjects would be equal and thus not constitute a variable.

Indexes of Dispersion. Subject variability may also be assessed by the traditional indices: the standard deviation and an index of kurtosis. These indexes are SM 12 and SM 13. They are scoring methods in that the value of an index for a subject serves as the value of a score for that subject.

An index of skewness is not an index of variability but rather of directionality or laterality and therefore is not a scoring method appropriate to this study.
Score Deviations. Just as ER may be operationalized as absolute deviation from the empirical means of individual scales or items, it may also be examined by scoring methods which are operationally defined as the absolute deviation of a score on a measure from the empirical or observed group mean score on a measure. These scoring methods may be applied to the scores obtained by each of the scoring methods which have been previously defined. In this manner SM 15 through SM 22 scores are obtained from SM 1 through 8 while SM 23 and SM 24 scores are obtained from SM 12 and SM 13. Content scores serve as the basis for scores on SM 14.

A score deviation scoring method has not been applied to SM 9, deviation from empirical means, to SM 10, deviation from subject means, to SM 11, deviation from subject means standardized by scale, or to SM 27, deviation from empirical means standardized by scale. This omission is due to an oversight by the author in the midst of the complexities of computation.

It should be noted that these score deviation scores may also be interpreted as indexes of the hypothesized unitary deviation factor which is so often misattributed to Berg, but then so also might any measure of individual differences.

Combinations of Score Deviations. Combinations of a particular subset of these score deviation scoring methods scores may also be conceptualized as appropriate to the measurement of ER. This is accomplished by SM 25 which is defined as the combination by summation of the scores which indicate the extent to which a subject is extreme
in underutilization or overutilization of response positions: one and seven; two and six; three and five; and four. This scoring method is, of course, adjusted for empirical differences between means on the four scores to be combined, however, it is subject to bias by the differences between standard deviations. This latter fact may be adjusted for by standardization of the scores before summation, which is the operational definition of $\text{SI}_26$. 
CHAPTER III

RESULTS AND DISCUSSION

Computation and Reporting

The data were collected on Digitek answer sheets and transferred to card form through the facilities of the Orientation and Testing Center of The Ohio State University. Preliminary data analyses were performed through the facilities of the Academic and Research Computer Center of The Ohio State University and the Academic Computer Center of New York University by the utilization of a variety of IBM computers and a variety of programming resources.

The results which are presented here were computed on the CDC 6600 under both SCOPE and KP.CNOd at the Computing Center of the Courant Institute of Mathematical Sciences at New York University which is under the sponsorship of the Atomic Energy Commission. The bulk of the computation was performed by FORTRAN programs authored by the present writer which made incidental use of IBM Subroutines HINV, HPRD, and LOC. Additional computation was performed using the BMDX72 Factor Analysis program (Dixon, 1970) and the SPSS Subroutine Reliability (Specht, 1975).

Full presentation of the statistical considerations involved and results obtained in the computations is not feasible. The considerations alone would fill several volumes the size of the present one and full presentation of the results would require, at the very least, many
more volumes. The presentation and discussion here is, therefore, limited to those matters which are most meaningful and directly relevant to the evaluation of the hypotheses.

**Preliminary Evaluations**

The two methods for scoring by deviations about the empirical means of scales were found to be virtually identical. SM 9 was operationally defined as the summation across scales of the absolute values of the deviations of the responses of the subject about the empirical or observed mean for all subjects on each scale. SM 27 was defined identically with the exception that prior to summation the deviation of each individual response was divided by the standard deviation of the responses of all subjects for the particular scale. For all eight measures SM 9 correlated with SM 27 at magnitudes in excess of .99. Thus within the measures differences among scales in their pull for a dispersion of responses were not observed to be related to the variable defined by SM 9 and the effects of additional adjustments made under SM 27 cancelled each other out when summation was performed across scales. Due to this virtual correlational identity of the two scoring methods with these data, SM 27, the more complex formulation, was omitted from further study.

Similar note should be taken concerning the normalized adjusted scores. All scores, whether raw or computed by one of the scoring methods under study, were additionally evaluated after normalization by the method of unequal intervals (Ghiselli, 1964). All scores computed by the various scoring methods were evaluated after they were adjusted
for their relationship to the content of the measures by the computations of adjusted scores from which the effects of raw score were partialled out. These adjusted scores were also evaluated after normalization by the method of unequal intervals. The evaluations of these normalized adjusted scores were not observed to contribute in a meaningful way to the evaluation of the hypotheses and so they will not be reported.

Characteristics of the Measures

Characteristics of each of the eight measures are summarized in Table 3.

The most divergent scale means are observed on the SEP measures. This is in accord with the expectations advanced in the discussion of the measures. Mean dispersion of scale responses, however, is not observed either to differ markedly between measures or to be related to mean scale response.

No marked evidence of either skewness or kurtosis is observed but that which is observed may be seen to be truly reduced to inconsequential levels by the normalization procedure. The computational formulae for the distribution statistics are given in Appendix B.

Throughout this study a variety of reliability estimates have been computed (Specht, 1975) including the following: Cronbach's alpha, standardized item alpha, unbiased alpha under the parallel model, alpha and unbiased alpha under the strict parallel model, Guttman's remaining five lambdas, between forms coefficient, Spearman-Brown corrected split half coefficient, and Guttman's split half coefficient. Inasmuch as these indexes do not vary in an unanticipated fashion, alpha is the statistic reported on grounds that it is most basic (Nunnally, 1967).
Table 3
Characteristics and Interrelationships of Scales and Measures

<table>
<thead>
<tr>
<th>Measures</th>
<th>Mean of Scale</th>
<th>Mean of Scale</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Normalized Skewness</th>
<th>Normalized Kurtosis</th>
<th>Alpha 1</th>
<th>Intercorrelations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M's</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>CEF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Difficulty</td>
<td>4.40</td>
<td>1.50</td>
<td>10.67</td>
<td>-0.06</td>
<td>-0.13</td>
<td>-0.00</td>
<td>-0.06</td>
<td>0.89</td>
</tr>
<tr>
<td>2. Content</td>
<td>4.26</td>
<td>1.69</td>
<td>11.52</td>
<td>-0.06</td>
<td>-0.51</td>
<td>-0.01</td>
<td>-0.07</td>
<td>0.87</td>
</tr>
<tr>
<td>3. Instruction</td>
<td>4.39</td>
<td>1.76</td>
<td>14.46</td>
<td>-0.29</td>
<td>-0.67</td>
<td>0.00</td>
<td>-0.11</td>
<td>0.94</td>
</tr>
<tr>
<td>4. Examinations</td>
<td>4.57</td>
<td>1.57</td>
<td>11.76</td>
<td>-0.32</td>
<td>-0.11</td>
<td>0.00</td>
<td>-0.12</td>
<td>0.91</td>
</tr>
<tr>
<td>5. General Evaluation</td>
<td>4.83</td>
<td>1.63</td>
<td>11.95</td>
<td>-0.51</td>
<td>0.18</td>
<td>-0.02</td>
<td>-0.11</td>
<td>0.90</td>
</tr>
<tr>
<td>SBF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Self</td>
<td>3.77</td>
<td>1.63</td>
<td>6.69</td>
<td>-0.07</td>
<td>-0.42</td>
<td>-0.01</td>
<td>-0.07</td>
<td>0.44</td>
</tr>
<tr>
<td>7. Average</td>
<td>3.70</td>
<td>1.43</td>
<td>5.90</td>
<td>-0.18</td>
<td>0.05</td>
<td>-0.01</td>
<td>-0.11</td>
<td>0.41</td>
</tr>
<tr>
<td>8. Ideal</td>
<td>5.14</td>
<td>1.68</td>
<td>3.53</td>
<td>-0.79</td>
<td>0.06</td>
<td>-0.04</td>
<td>-0.14</td>
<td>0.68</td>
</tr>
</tbody>
</table>
It may be observed in Table 3 that the standard deviations and alphas of the measures covary in the expected fashion afforded by the measures as the range is attenuated or extended.

Similarly, the relatively greater reliability of Instruction in comparison to the other CEF scores and of Ideal in comparison to the other SEF scores may be attributed to their corresponding relatively greater standard deviations. Whether this relatively greater differentiation among students on these two measures is attributable to the characteristics of these particular students, courses, instruments, concepts or some interaction of these factors it is not possible to determine from these results.

The relatively reduced values of the standard deviation and of alpha of the SEF measures as contrasted to those of the CEF measures would appear to be directly attributable in part to the presence in the SEF measures of scales six and eight which are consistently and negatively correlated with the remaining scales on all three measures.

From the intercorrelations of the scores of these students in these courses on these instruments the following observations may be made about the measures: the CEF measures are substantially interrelated with the exceptions that Difficulty is but modestly related to Content and Instruction and not at all related to Examinations and General Evaluation. Difficulty and Examination are less related than the other CEF measures to the SEF measures. CEF measures are most related to Self, less to Average, and least to Ideal. Self is moderately related to Average and only slightly to Ideal. Average is not related to Ideal.
In sum, the characteristics and interrelationships of the scales and of the measures appear to satisfy, particularly in their heterogeneity, the criteria advanced for the selection of the instruments.

**Correlation Results**

All correlation coefficients reported are based upon observations of the 280 subjects. In all cases, therefore, there are 279 degrees of freedom associated with each correlation coefficient. All are Pearson product moment coefficients of correlation.

The coefficients differ in terms of whether they are based on raw, normalized, or adjusted scores. The coefficients also differ in terms of the type of scores upon which they are based: a score on a measure (10 scales) with a score on another measure (10 scales) or another score on the same measure (10 scales); a score on a measure (10 scales) with the same score summed across the remaining seven measures (70 scales); or a score summed across all eight measures (80 scales) with another score summed across all eight measures (80 scales).

Additionally, the coefficients differ in terms of whether they are simple coefficients or mean coefficients obtained by summing across a number of coefficients and dividing by the number of those coefficients. Taken strictly, such a computation of a mean coefficient entails a statistical error. This is because the correlation coefficient is not a normally distributed statistic. It is statistically appropriate, therefore, to first convert the distribution of coefficient values to a distribution of normal deviate values, summate the normal deviate
values, obtain their mean value, and then convert this mean normal
deviate value back to a correlation coefficient (McNemar, 1962).
The resultant coefficient may then be tested for significance in
exactly the same manner as is appropriate to the evaluation of any
one of the original coefficients. Both procedures were followed in
the data analysis. The exact procedure utilizing the normal deviate
produced mean coefficients of negligibly greater magnitude. The
mean coefficients obtained without recourse to the normal deviate
transformations are presented here, therefore, on grounds that they
will be more intelligible to the reader and result in a statistical
test which is only negligibly more conservative.

Since hypotheses concerning reliability imply positive inter-
correlation the relevant coefficients are subjected to one-tailed tests.
No directionality is implied, however, in the hypothesized intercor-
relation of content scores with ER scores and so the coefficients
relevant to those hypotheses are subjected to two-tailed tests.

Results of the correlational analyses are summarized in Table 4,
Table 5, and Table 6.

**Hypothesis One**

The first hypothesis asserts that ER scores are reliable. In
order to evaluate this assertion the data are examined at each of
four levels of analysis.

At the first level of analysis it is possible to evaluate the
correlations between two measures of the ER scores obtained by each
Table 4

Reliability of ER Scores
Mean Coefficients and Their Significance Levels

<table>
<thead>
<tr>
<th>Scoring Method</th>
<th>Raw</th>
<th>Normalized</th>
<th>Adjusted</th>
<th>Raw</th>
<th>Normalized</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM 1</td>
<td>.40</td>
<td>.0005</td>
<td>.41</td>
<td>.0005</td>
<td>.44</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 2</td>
<td>.24</td>
<td>.0005</td>
<td>.24</td>
<td>.0005</td>
<td>.23</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 3</td>
<td>.33</td>
<td>.0005</td>
<td>.34</td>
<td>.0005</td>
<td>.35</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 4</td>
<td>.26</td>
<td>.0005</td>
<td>.27</td>
<td>.0005</td>
<td>.23</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 5</td>
<td>.26</td>
<td>.0005</td>
<td>.27</td>
<td>.0005</td>
<td>.26</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 6</td>
<td>.20</td>
<td>.0005</td>
<td>.20</td>
<td>.0005</td>
<td>.21</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 7</td>
<td>.33</td>
<td>.0005</td>
<td>.34</td>
<td>.0005</td>
<td>.34</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 8</td>
<td>.36</td>
<td>.0005</td>
<td>.35</td>
<td>.0005</td>
<td>.35</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 9</td>
<td>.31</td>
<td>.0005</td>
<td>.30</td>
<td>.0005</td>
<td>.30</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 10</td>
<td>.25</td>
<td>.0005</td>
<td>.24</td>
<td>.0005</td>
<td>.26</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 11</td>
<td>.28</td>
<td>.0005</td>
<td>.24</td>
<td>.0005</td>
<td>.30</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 12</td>
<td>.26</td>
<td>.0005</td>
<td>.24</td>
<td>.0005</td>
<td>.27</td>
<td>.0005</td>
</tr>
<tr>
<td>SM 13</td>
<td>.05</td>
<td>.005</td>
<td>.05</td>
<td>.005</td>
<td>.05</td>
<td>.005</td>
</tr>
<tr>
<td>SM 14</td>
<td>.17</td>
<td>.01</td>
<td>.16</td>
<td>.01</td>
<td>.19</td>
<td>.005</td>
</tr>
<tr>
<td>SM 15</td>
<td>.21</td>
<td>.005</td>
<td>.13</td>
<td>.025</td>
<td>.12</td>
<td>.005</td>
</tr>
<tr>
<td>SM 16</td>
<td>.12</td>
<td>.05</td>
<td>.12</td>
<td>.05</td>
<td>.11</td>
<td>.005</td>
</tr>
<tr>
<td>SM 17</td>
<td>.15</td>
<td>.025</td>
<td>.12</td>
<td>.05</td>
<td>.15</td>
<td>.025</td>
</tr>
<tr>
<td>SM 18</td>
<td>.08</td>
<td>.04</td>
<td>.08</td>
<td>.04</td>
<td>.04</td>
<td>.04</td>
</tr>
<tr>
<td>SM 19</td>
<td>.08</td>
<td>.05</td>
<td>.11</td>
<td>.05</td>
<td>.11</td>
<td>.05</td>
</tr>
<tr>
<td>SM 20</td>
<td>.09</td>
<td>.09</td>
<td>.09</td>
<td>.09</td>
<td>.09</td>
<td>.09</td>
</tr>
<tr>
<td>SM 21</td>
<td>.13</td>
<td>.025</td>
<td>.12</td>
<td>.05</td>
<td>.13</td>
<td>.025</td>
</tr>
<tr>
<td>SM 22</td>
<td>.16</td>
<td>.01</td>
<td>.13</td>
<td>.025</td>
<td>.17</td>
<td>.01</td>
</tr>
<tr>
<td>SM 23</td>
<td>.13</td>
<td>.025</td>
<td>.13</td>
<td>.025</td>
<td>.13</td>
<td>.025</td>
</tr>
<tr>
<td>SM 24</td>
<td>.04</td>
<td>.03</td>
<td>.04</td>
<td>.03</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>SM 25</td>
<td>.16</td>
<td>.01</td>
<td>.15</td>
<td>.025</td>
<td>.16</td>
<td>.01</td>
</tr>
<tr>
<td>SM 26</td>
<td>.15</td>
<td>.025</td>
<td>.14</td>
<td>.025</td>
<td>.16</td>
<td>.01</td>
</tr>
</tbody>
</table>
Table 5
ER Scores and Content Scores

Coefficients of Intercorrelations and Their Significance Levels

<table>
<thead>
<tr>
<th>Column</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Coefficients Across 8 Measures</td>
<td>Coefficients for Sums of 8 Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raw</td>
<td>Normalized</td>
<td>Raw</td>
<td>Normalized</td>
</tr>
<tr>
<td>Scoring Method</td>
<td>Raw</td>
<td>Normalized</td>
<td>Raw</td>
<td>Normalized</td>
</tr>
<tr>
<td>SM 1</td>
<td>0.17</td>
<td>0.02</td>
<td>0.20</td>
<td>0.01</td>
</tr>
<tr>
<td>SM 2</td>
<td>0.12</td>
<td>0.08</td>
<td>0.27</td>
<td>0.001</td>
</tr>
<tr>
<td>SM 3</td>
<td>0.17</td>
<td>0.02</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>SM 4</td>
<td>-0.20</td>
<td>0.01</td>
<td>-0.22</td>
<td>0.01</td>
</tr>
<tr>
<td>SM 5</td>
<td>0.26</td>
<td>0.001</td>
<td>0.26</td>
<td>0.001</td>
</tr>
<tr>
<td>SM 6</td>
<td>0.05</td>
<td>-0.08</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>SM 7</td>
<td>0.03</td>
<td>0.06</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SM 8</td>
<td>0.25</td>
<td>0.001</td>
<td>0.26</td>
<td>0.001</td>
</tr>
<tr>
<td>SM 9</td>
<td>-0.13</td>
<td>0.05</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>SM 10</td>
<td>-0.08</td>
<td>0.09</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>SM 11</td>
<td>-0.16</td>
<td>0.02</td>
<td>-0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>SM 12</td>
<td>-0.05</td>
<td>0.07</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>SM 13</td>
<td>0.12</td>
<td>0.09</td>
<td>0.27</td>
<td>0.001</td>
</tr>
<tr>
<td>SM 14</td>
<td>-0.14</td>
<td>-0.05</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>SM 15</td>
<td>0.08</td>
<td>0.08</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>SM 16</td>
<td>0.16</td>
<td>0.04</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>SM 17</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.05</td>
<td></td>
</tr>
<tr>
<td>SM 18</td>
<td>-0.05</td>
<td>-0.01</td>
<td>-0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>SM 19</td>
<td>0.05</td>
<td>0.07</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>SM 20</td>
<td>0.06</td>
<td>0.06</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>SM 21</td>
<td>0.01</td>
<td>0.02</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>SM 22</td>
<td>0.04</td>
<td>0.06</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>SM 23</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>SM 24</td>
<td>0.05</td>
<td>0.01</td>
<td>0.18</td>
<td>0.02</td>
</tr>
<tr>
<td>SM 25</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>SM 26</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>
Table 6
Cumulative Percentages of Coefficients Attaining Specified Levels of Significance

<table>
<thead>
<tr>
<th>Description Line Score</th>
<th>Significance Levels: One Tailed Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.0005</td>
</tr>
<tr>
<td>ER Score</td>
<td></td>
</tr>
<tr>
<td>1 R</td>
<td>43.4</td>
</tr>
<tr>
<td>2 N</td>
<td>37.9</td>
</tr>
<tr>
<td>3 A</td>
<td>42.4</td>
</tr>
<tr>
<td>Mean Coefficients</td>
<td></td>
</tr>
<tr>
<td>4 R</td>
<td>42.3</td>
</tr>
<tr>
<td>5 N</td>
<td>42.3</td>
</tr>
<tr>
<td>6 A</td>
<td>42.3</td>
</tr>
<tr>
<td>Between Measures and Sums of Remaining Measures (728 r's)</td>
<td>7 R</td>
</tr>
<tr>
<td>8 N</td>
<td>99.0</td>
</tr>
<tr>
<td>9 A</td>
<td>98.1</td>
</tr>
<tr>
<td>Mean Coefficients for Measures (208 r's)</td>
<td>10 R</td>
</tr>
<tr>
<td>11 N</td>
<td>100.0</td>
</tr>
<tr>
<td>12 A</td>
<td>100.0</td>
</tr>
<tr>
<td>Between Measures and Sums of Remaining Measures (26 r's)</td>
<td>13 R</td>
</tr>
<tr>
<td>14 N</td>
<td>20.2</td>
</tr>
<tr>
<td>Mean Coefficients for Measures (26 r's)</td>
<td>15 R</td>
</tr>
<tr>
<td>16 N</td>
<td>7.7</td>
</tr>
<tr>
<td>For Sums Across Measures (26 r's)</td>
<td>17 R</td>
</tr>
<tr>
<td>18 N</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Significance Levels: Two Tailed Tests

|                        | .001 | .01  | .02  | .05  | ns  |
| ER Score Content Score |      |      |      |      |     |
| For Measures (208 r's) | 13 R | 20.2 | 28.8 | 34.6 | 41.3 | 58.7 |
| 14 N | 20.2 | 28.4 | 31.3 | 35.1 | 64.9 |      |
| Mean Coefficients for Measures (26 r's) | 15 R | 7.7 | 11.5 | 23.1 | 30.8 | 69.2 |
| 16 N | 7.7 | 19.2 | 23.1 | 23.1 | 76.9 |      |
| For Sums Across Measures (26 r's) | 17 R | 23.1 | 30.8 | 34.6 | 34.6 | 65.4 |
| 18 N | 11.5 | 19.2 | 19.2 | 23.1 | 76.9 |      |

R=Raw Score; N=Normalized Score; A=Adjusted Score
scoring method with scores obtained by the same scoring method for every possible pair of measures. This procedure yields a matrix containing 728 coefficients of reliability between pairs of ten scale measures (twenty-six scoring methods by the twenty-eight possible pairs of the eight measures taken two at a time and without replacement). The cumulative proportion of these coefficients which attain specified significance levels is reported in line 1 of Table 6. Of the 728 coefficients more than 71% attain the 5% level of significance. Line 2 of the same table reports the results for coefficients obtained in the same way but for which the scores were normalized before the correlations were computed. Of the 728 coefficients more than 66% attain the 5% level of significance.

At the second level of analysis it is possible to evaluate the mean coefficients of the correlations between two measures of the ER scores obtained by each scoring method with scores obtained by the same scoring method for every possible pair of measures. This procedure yields a vector containing 26 mean coefficients of reliability between pairs of ten scale measures averaged across the twenty-eight possible pairs of measures (twenty-six scoring methods). These mean coefficients are reported in column 1 of Table 4 and their significance levels are summarized in line 4 of Table 6. Of the 26 coefficients more than 34% attain the 5% level of significance. Similarly, the mean coefficients for coefficients obtained from normalized scores are presented in column 2 of Table 4 and their significance levels are summarized in line 5 of Table 6. Of the 26 coefficients more than 76% attain the 5% level of significance.
At the third level of analysis it is possible to evaluate the correlations of the $3R$ scores obtained by each scoring method on each measure with scores obtained by the same scoring method and summed across the remaining seven measures. This procedure yields a matrix containing 208 coefficients of reliability between a measure containing ten scales and a combination of seven measures each containing ten scales or a total of seventy scales (twenty-six scoring methods by eight measures). The cumulative proportion of these coefficients which attain specified significance levels is reported in line 7 of Table 6. Of the 208 coefficients fully 100\% attain the 1\% significance level. Line 8 of the same table reports the results for coefficients obtained in the same way but for which the scores were normalized before the correlations were computed. Of the 208 coefficients fully 100\% attain the 0.5\% level of significance.

At the fourth level of analysis it is possible to evaluate the mean coefficients of the correlations of the $3R$ scores obtained by each scoring method on each measure with scores obtained by the same scoring method and summed across the remaining seven measures. This procedure yields a vector containing 26 mean coefficients of reliability between measures containing ten scales and a combination of seven measures each containing ten scales or a total of seventy scales (twenty-six scoring methods). These mean coefficients are reported in column 4 of Table 4 and their significance levels are summarized in line 10 of Table 6. All of the 26 coefficients attain the 0.05\% level of significance. Similarly the mean coefficients for
coefficients obtained from normalized scores are presented in column 5 of Table 4 and their significance levels are summarized in line 11 of Table 6. All of the 26 coefficients attain the 0.05% level of significance.

It would appear from these results that hypothesis one, that ER scores are reliable, may as a general statement be accepted. Further partial support of the hypothesis is reported in Table 9 which will be discussed subsequent to the presentation of the results directly pertinent to the hypotheses.

Hypothesis Two

The second hypothesis asserts that ER scores are related to content. In order to evaluate this assertion the data are evaluated at each of three levels of analysis.

At the first level of analysis it is possible to evaluate the correlations between the ER scores obtained by each scoring method for each measure with the content score for that same measure. This procedure yields a matrix containing 203 coefficients of intercorrelation between content scores and ER scores on each ten scale measure (twenty-six scoring methods by eight measures). The cumulative proportion of these coefficients which attain specified significance levels is reported in line 13 of Table 6. Of the 203 coefficients more than 41% attain the 5% level of significance. Line 14 of the same table reports the results for coefficients obtained in the same way but for which the scores were normalized before the correlations were computed. Of the 203 coefficients more than 35% attain the
At the second level of analysis it is possible to evaluate the mean coefficients of the correlations between the ER scores obtained by each scoring method for each measure with the content score for that same measure. This procedure yields a vector containing 26 mean coefficients of intercorrelation between content scores and ER scores averaged across the eight measures containing ten scales each (twenty-six scoring methods). These mean coefficients are reported in column 1 of Table 5 and their significance levels are summarized in line 16 of Table 6. Of the 26 coefficients more than 30% attain the 5% level of significance. Similarly the mean coefficients for coefficients obtained from normalized scores are presented in column 2 of Table 5 and their significance levels are summarized in line 17 of Table 6. Of the 26 coefficients more than 23% attain the 5% level of significance.

At the third level of analysis it is possible to evaluate the correlations between the ER scores obtained by each scoring method and the content scores both summed across all eight measures. This procedure yields a vector of 26 coefficients of intercorrelation between content scores and ER scores both on the sum of the eight measures of ten scales or all eighty scales (twenty-six scoring methods). These coefficients are reported in column 3 of Table 5 and their significance levels are summarized in line 17 of Table 6. Of the 26 coefficients more than 34% attain the 5% level of significance. Similarly the coefficients obtained from normalized scores are presented in column 4 of Table 5 and their significance levels
are summarized in line 18 of Table 6. Of the 26 coefficients more than 23\% attain the 5\% level of significance.

It would appear from these results that hypothesis two, that ER scores are related to content, may as a general statement be accepted. The evidence in support of hypothesis two is not nearly as strong as the evidence in support of hypothesis one but nonetheless it does compel assent.

Hypothesis Three

The third hypothesis asserts that ER scores are reliable independent of content. In order to evaluate this assertion ER adjusted scores have been generated which are independent of content scores. This was accomplished by partialling out the content component of each ER score. In order to evaluate hypothesis three the data are examined at each of four levels of analysis.

At the first level of analysis it is possible to evaluate the correlation between two measures of the ER adjusted scores obtained by each scoring method with adjusted scores obtained by the same scoring method for every possible pair of measures. This procedure yields a matrix containing 728 coefficients of reliability between pairs of ten scale measures (twenty-six scoring methods by the twenty-eight possible pairs of the eight measures taken two at a time and without replacement). The cumulative proportion of these coefficients which attain specified significance levels is reported in line 3 of Table 6. Of the 728 coefficients more than 73\% attain the 5\% level of significance.
At the second level of analysis it is possible to evaluate the mean coefficients of the correlations between two measures of the ER adjusted scores obtained by each scoring method with adjusted scores obtained by the same scoring method for every possible pair of measures. This procedure yields a vector containing 26 mean coefficients of reliability between pairs of ten scale measures averaged across the twenty-eight possible pairs of measures (twenty-six scoring methods). These mean coefficients are reported in column 3 of Table 4 and their significance levels are summarized in line 6 of Table 6. Of the 26 coefficients more than 84% attain the 5% level of significance.

At the third level of analysis it is possible to evaluate the correlations of the ER adjusted scores obtained by each scoring method on each measure with adjusted scores obtained by the same scoring method and summed across the remaining seven measures. This procedure yields a matrix containing 203 coefficients of reliability between a measure containing ten scales and a combination of seven measures each containing ten scales or a total of seventy scales (twenty-six scoring methods by eight measures). The cumulative proportion of these coefficients which attain specified significance levels is reported in line 9 of Table 6. Of the 208 coefficients fully 100% attain the 1% significance level.

At the fourth level of analysis it is possible to evaluate the mean coefficients of the correlations of the ER adjusted scores obtained by each scoring method on each measure with adjusted scores
obtained by the same scoring method and summed across the remaining seven measures. This procedure yields a vector containing 26 mean coefficients of reliability between measures containing ten scales and a combination of seven measures each containing ten scales or a total of seventy scales (twenty-six scoring methods). These mean coefficients are reported in column 6 of Table 4 and their significance levels are summarized in line 12 of Table 6. All of the 26 coefficients attain the 0.05% level of significance.

It would appear from these results that hypothesis three, that ER scores are reliable independent of content, may as a general statement be accepted. The evidence in support of hypothesis three is equally as strong as the evidence in support of hypothesis one.

Other Statistical Results

Hypotheses four, five, and six assert a relationship between scoring methods and the reliability of ER scores, the relation of ER scores to content, and the reliability of ER scores independent of their relationship to content. Clearly, the best method by which to evaluate these hypotheses is the direct examination of the same results which were examined in the evaluation of hypotheses one, two, and three.

In addition, however, it is instructive to consider the results of other statistical analyses based upon these results. The results of these additional analyses also shed light upon the questions of the qualities of and interrelationships among scoring methods and the interrelationships between scoring methods and measures. These later
matters are discussed subsequent to the report of the results related to the hypotheses.

The assertion of the final three hypotheses that the scoring methods are related to the coefficients resulting from the use of the scoring methods may also be phrased in the null form as an assertion that within each group of twenty-six the coefficients do not differ from one another to a greater extent than would be expected by chance. The experienced eye may simply look at the coefficients in order to evaluate this assertion. The most direct statistical test of the assertion would be to evaluate the differences between all possible pairs of coefficients within each group. The normal deviate procedure could be employed (McNemar, 1962). For the coefficients summarized in Table 6 this would amount to 43,550 tests of significance. This would, perhaps, be but minimally more informative than simple visual inspection.

In addition, it is possible to consider whether there exist specific relationships between the scoring methods and the coefficients. Support for the existence of any specific relationship requires the rejection of the general denial of relationships and therefore the acceptance of the hypotheses.

One specific possible relationship which will be considered is that of a nonrandom relationship between the scoring methods and the average value of the coefficients which they yield. This may be tested by the analysis of variance.
The analysis of variance requires an assumption of equality of variances. This assumption may be evaluated either by the $F_{\text{max}}$ procedure developed by Hartley or by Cochran's $C$. The appropriate tables, however, do not extend to the degrees of freedom involved in the present analysis (Winer, 1962). The values of these statistics will, however, be reported in terms of values which might be expected in extended tables. Evidence of heterogeneity of variance between groups of coefficients is support of another specific relationship between the scoring methods and the coefficients. It is evidence of a relationship between the scoring methods and the variability of the coefficients that they yield.

Because it is believed that a relationship also exists between the coefficients and the measures or pairs of measures upon which they are based the technique which is employed is the repeated measures analysis of variance. In this analysis the effect of a simple relationship between the coefficients and the measures or pairs of measures upon which they are based or main effect of measures or pairs of measures upon coefficients is removed from the analysis because it is unrelated to the hypotheses. This increases the accuracy of the analysis (Winer, 1962).

An assumption of such repeated measures analysis of variance is that the effects of scoring methods and the effects of measures or pairs of measures upon the coefficients are additive. This means that the effects are simple or independent rather than interactive. In order to determine whether this assumption is met, a test of nonadditivity has been developed (Tukey, 1949). Strictly taken, this test
is applicable only to the two factor analysis of variance. It is often found, however, to be of considerable value in the repeated measures analysis of variance. It is a test for the presence of a linear by linear interaction effect. It does not evaluate, for example, linear by quadratic or cubic by cubic effects. If nonadditivity is determined to be present the procedure provides for an improved evaluation of the main effect by removal of the spurious inflation of the error term due to the interaction (Scheffe, 1960). The observation of a significant level of additivity in the present analyses is evidence of a specific relationship between the scoring methods and the coefficients. It is evidence that the scoring methods and the measures or pairs of measures linearly interact in their relationships with the coefficients.

In addition to the assumption of homogeneity of variance the repeated measures analysis of variance requires the assumption of homogeneity of covariance. In this analysis this means an assumption that across measures or pairs of measures all scoring methods are uniformly interrelated in the coefficients which they yield. The combined assumptions of homogeneity of variance and covariance may be evaluated by a Chi Square procedure (Box, 1950). Such evidence of heterogeneity is support of another specific relationship between the scoring methods and the coefficients. It is evidence that the scoring methods and the measures or pairs of measures differentially interact in their relationships with the correlations.
When the Chi Square procedure indicates the presence of heterogeneity in the variance-covariance matrix a conservative test of the F value for main effect may be conducted by the use of degrees of freedom appropriate when such heterogeneity is at a maximum (Box, 1950). The T Square statistic, which assumes multinormal distribution, is an exact test of the same proposition because its quality of invariance renders it immune to the effects of such heterogeneity (Hotelling, 1931).

Subsequent to the repeated measures analyses of variance a posteriori multiple comparisons of differences between means are not conducted by any of the methods available (Winer, 1962) because the necessary tables of the Studentized Range statistic are not sufficiently extensive (Pearson & Hartley, 1956).

The correlation coefficient, as it was previously noted, does not distribute normally and the transformation of coefficients to normal deviate values may be appropriate (McNemar, 1962). All tests related to the final three hypotheses were conducted both on the coefficients and on their normal deviate values.

In summary, hypotheses four, five, and six are evaluated by observation of the relationships of the scoring methods to the coefficients which were evaluated under hypotheses one, two, and three. First, inspection will be made to determine differences in coefficients produced by different scoring methods. Second, differences within certain sets of coefficients will be tested by the normal deviate procedure. Additionally, where possible, tests will be made for differences in the average magnitude and in the variability of the
coefficients produced by the different scoring methods and for both linear by linear and differential interactions between the scoring methods and the measures or pairs of measures in their effect upon the correlations which they yield.

**Hypothesis Four**

The fourth hypothesis asserts that the reliability of ER scores is related to scoring methods.

In the evaluation of the first hypothesis examination was made of the reliability coefficients or correlation between ER scores obtained by each scoring method on every possible pair of measures, a matrix of 728 coefficients (twenty-six scoring methods by the twenty-eight possible pairs of the eight measures taken two at a time and without replacement). Inspection of this matrix and of the corresponding matrix of coefficients based upon normalized scores and of the matricies of normal deviate values of these coefficients supports the hypothesis that the reliability of ER scores is related to scoring methods.

As has been noted, the mean coefficients of reliability of the ER scores obtained by each of the twenty-six scoring methods averaged across the coefficients obtained between the twenty-eight possible pairs of measures are reported in column 1 of Table 4. The normal deviate procedure was used to test for significant differences between each of the 325 possible pairs of mean coefficients (twenty-six coefficients taken two at a time and without replacement).
The results of these tests are reported in Table 7. For the 325 possible pairs of mean coefficients more than 28% of the differences attain the 5% level of significance. Similar tests were made for differences between the mean coefficients for coefficients obtained from normalized scores which are reported in column 2 of Table 4. The results of these tests are also reported in Table 7. For these 325 possible pairs of mean coefficients more than 32% of the differences attain significance at the 5% level.

Also as has been noted, the mean coefficients of the reliability of ER scores obtained by each of the twenty-six scoring methods summed across the coefficients obtained between each of the eight measures and the sum of the ER scores for the remaining seven measures are reported in column 4 of Table 4. The same tests for significant differences between pairs of these mean coefficients were made. The results of these tests are reported in Table 7. For the 325 possible pairs of mean coefficients more than 21% of the differences attain the 5% level of significance. Similar tests were made for differences between the mean coefficients for coefficients obtained from normalized scores which are reported in column 5 of Table 4. Results of these tests are also reported in Table 7. For these 325 possible pairs of mean coefficients more than 24% of the differences attain the 5% level of significance.

The matrix of 728 reliability coefficients for ER scores was subjected to the repeated measures analysis of variance and other procedures which have been discussed. Similar analyses were also performed upon the matrix of coefficients based on normalized scores.
Table 7
Differences Between Coefficients For All Possible Pairs of Scoring Methods
Cumulative Percentages Attaining Specified Levels of Significance

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
<th>Significance Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>.001</td>
</tr>
<tr>
<td><strong>Mean Coefficients ER Scores</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Pairs of Measures</td>
<td>R</td>
<td>5.8</td>
</tr>
<tr>
<td>(325 differences)</td>
<td>N</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>5.8</td>
</tr>
<tr>
<td>Between Measures and Sums of Remaining 7 Measures</td>
<td>R</td>
<td>2.8</td>
</tr>
<tr>
<td>(325 differences)</td>
<td>N</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>2.5</td>
</tr>
<tr>
<td><strong>Mean Coefficients ER Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>For Measures</td>
<td>R</td>
<td>8.9</td>
</tr>
<tr>
<td>(325 differences)</td>
<td>N</td>
<td>10.5</td>
</tr>
<tr>
<td>For Sums Across Measures</td>
<td>R</td>
<td>28.6</td>
</tr>
<tr>
<td>(325 differences)</td>
<td>N</td>
<td>18.8</td>
</tr>
</tbody>
</table>

1R=Raw Score; N=Normalized Score; A=Adjusted Score
The same analyses were also performed upon the matrices of normal deviate values for these matrices of coefficients. Results of these analyses are reported in Table 3.

The $F$ test for main effect, the $F$ test for main effect adjusted for nonadditivity, the $F$ test for main effect in the presence of maximum heterogeneity, and the $F$ test of $T$ Square all show differences to exist in the average values of the coefficients yielded by different scoring methods.

The observed values of the $F_{\text{max}}$ statistic and of Cochran's $C$ cannot be directly evaluated because the requisite tables are insufficiently extensive. They do, however, seem to indicate that differences exist in the variability of the coefficients yielded by different scoring methods.

The $F$ test for nonadditivity indicates the presence of a linear by linear interaction between scoring methods and pairs of measures in the determination of the coefficients. The Chi Squares indicate the presence of differential interaction.

These results are observed both for coefficients based upon raw scores and for coefficients based upon normalized scores whether they are evaluated directly or by their normal deviate values.

The evidence is incompatible with the null proposition of lack of relationship. The fourth hypothesis which asserts that the reliability of ER scores is related to scoring methods is therefore accepted.
Table 8

Tests of Statistical Significance Related to Hypotheses Four, Five, and Six

<table>
<thead>
<tr>
<th>Description</th>
<th>df</th>
<th>1 R</th>
<th>1 P</th>
<th>N</th>
<th>1 P</th>
<th>A</th>
<th>1 P</th>
<th>df</th>
<th>1 R</th>
<th>1 P</th>
<th>N</th>
<th>1 P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>25, r</td>
<td>73.336</td>
<td>.001</td>
<td>94.617 .001</td>
<td>77.762 .001</td>
<td>25, r</td>
<td>3.618 .0001</td>
<td>3.944 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>675 z</td>
<td>74.824</td>
<td>.001</td>
<td>95.216 .001</td>
<td>79.338 .001</td>
<td>175 z</td>
<td>3.587 .0001</td>
<td>3.938 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Corrected for Maximum Heterogeneity</td>
<td>1, r</td>
<td>73.336</td>
<td>.001</td>
<td>94.617 .001</td>
<td>77.762 .001</td>
<td>1.7 r</td>
<td>3.618 .10</td>
<td>3.944 .10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27 z</td>
<td>74.324</td>
<td>.001</td>
<td>95.216 .001</td>
<td>79.338 .001</td>
<td>z</td>
<td>3.587 .10</td>
<td>3.938 .10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Corrected for Additivity</td>
<td>25, r</td>
<td>74.117</td>
<td>.001</td>
<td>96.547 .001</td>
<td>78.335 .001</td>
<td>25, r</td>
<td>3.757 .0001</td>
<td>4.762 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>675 z</td>
<td>76.769</td>
<td>.001</td>
<td>99.513 .001</td>
<td>81.115 .001</td>
<td>175 z</td>
<td>3.639 .0001</td>
<td>4.838 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F of F Square</td>
<td>25, r</td>
<td>81.233</td>
<td>.0019</td>
<td>107.197 .001</td>
<td>63.456 .0028</td>
<td>175 r</td>
<td>3.639 .0001</td>
<td>4.838 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 z</td>
<td>50.093</td>
<td>.0039</td>
<td>31.600 .0019</td>
<td>30.986 .0019</td>
<td>175 r</td>
<td>3.639 .0001</td>
<td>4.838 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variability and Interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F max</td>
<td>26, r</td>
<td>4.565 na 3</td>
<td>4.166 na</td>
<td>4.896 na</td>
<td>26, r</td>
<td>37.169 na</td>
<td>161.402 na</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27 z</td>
<td>4.830 na</td>
<td>4.822 na</td>
<td>5.136 na</td>
<td>7 z</td>
<td>46.090 na</td>
<td>196.020 na</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cochran's C</td>
<td>26, r</td>
<td>.0920 na</td>
<td>.0770 na</td>
<td>.0975 na</td>
<td>26, r</td>
<td>.1074 na</td>
<td>.1172 na</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27 z</td>
<td>.0920 na</td>
<td>.0901 na</td>
<td>.0909 na</td>
<td>7 z</td>
<td>.1164 na</td>
<td>.1260 na</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Nonadditivity</td>
<td>1, r</td>
<td>7.832 .0001</td>
<td>15.034 .0001</td>
<td>7.031 .0001</td>
<td>1, r</td>
<td>7.636 .0062</td>
<td>37.290 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>674 z</td>
<td>18.187 .0001</td>
<td>30.339 .0001</td>
<td>15.396 .0001</td>
<td>174 z</td>
<td>5.959 .0156</td>
<td>41.014 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi Square</td>
<td>349 r</td>
<td>1026 .0001</td>
<td>794 .0001</td>
<td>947 .0001</td>
<td>174 z</td>
<td>5.959 .0156</td>
<td>41.014 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>349 z</td>
<td>1030 .0001</td>
<td>795 .0001</td>
<td>953 .0001</td>
<td>174 z</td>
<td>5.959 .0156</td>
<td>41.014 .0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 R=Raw; N=Normalized; A=Adjusted 2 r=Coefficient values; z=Normal deviate values of coefficients 3 na=Not available
Hypothesis Five

The fifth hypothesis asserts that the relationship of ER scores to content is related to scoring methods.

In the evaluation of the second hypothesis examination was made of the coefficients of intercorrelation between the content score for each measure and the ER scores for that measure obtained by each of the scoring methods, a matrix of 208 coefficients (eight measures by twenty-six scoring methods). Inspection of this matrix and of the corresponding matrix of coefficients based upon normalized scores and of the matrices of normal deviate values of these coefficients supports the hypothesis that the relationship of ER scores to content is related to scoring methods.

As has been noted, the mean coefficients of intercorrelation between the content scores and each of the ER scores obtained by each of the twenty-six scoring methods averaged across the coefficients obtained for each measure are reported in column 1 of Table 5. The normal deviate procedure was used to test for significant differences between each of the 325 possible pairs of mean coefficients (twenty-six coefficients taken two at a time and without replacement). These results are reported in Table 7. For the 325 possible pairs of mean coefficients more than 38% of the differences attain significance at the 5% level. Similar tests were made for differences among the mean coefficients obtained from normalized scores which are reported in column 2 of Table 5. The results of these tests are also reported in Table 7. For the 325 possible pairs of mean coefficients more than 33% of the differences attain significance at the 5% level.
Also as has been noted, the coefficients of intercorrelation between content and the ER scores obtained by each of the twenty-six scoring methods summed across the eight ten scale measures or all eighty scales are reported in column 3 of Table 5. The same tests for significant differences between pairs of these coefficients were made. The results of these tests are reported in Table 7. For the 325 possible pairs of coefficients more than 43% of the differences attain the 5% level of significance. Similar tests were made for differences between coefficients obtained from normalized scores which are reported in column 4 of Table 5. Results of these tests are also reported in Table 7. For these 325 possible pairs of coefficients more than 34% of the differences attain the 5% level of significance.

The matrix of 208 intercorrelation coefficients between ER scores and content scores was subjected to the repeated measures analysis of variance and other procedures which have been discussed. Due to the dimensions of the matrix the Chi Square and T square statistics could not be computed. Analyses similar to those performed upon this matrix of coefficients were also performed upon the matrix of coefficients based on normalized scores. The same analyses were also performed upon the matrices of normal deviate values for these matrices of coefficients. Results of these analyses are reported in Table 8.

The F test for main effect, the F test for main effect adjusted for nonadditivity, and the F test for main effect in the presence of maximum heterogeneity show differences to exist in the average values of the coefficients yielded by different scoring methods. While the
The F test corrected for the presence of heterogeneity only approaches the 5% level, the test is so conservative that the result may be accepted as evidence in support of the relationship.

The observed values of the $F_{\text{max}}$ statistic and of Cochran's $C$ cannot be directly evaluated because the requisite tables are insufficiently extensive. They do, however, seem to indicate that differences exist in the variability of the coefficients yielded by different scoring methods.

The $F$ test for nonadditivity indicates the presence of a linear by linear interaction between scoring methods and measures in the determination of the coefficients.

These results are observed both for coefficients based upon raw scores and for coefficients based upon normalized scores whether evaluated directly or by their normal deviate values.

The evidence is incompatible with the null proposition of no relationship. The fifth hypothesis which asserts that the relationship of ER scores to content is related to scoring methods is therefore accepted.

Hypothesis Six

The sixth hypothesis asserts that the reliability of ER scores independent of their relationship to content is related to scoring method.

In the evaluation of the third hypothesis examination was made of the reliability coefficients or correlation between ER adjusted scores obtained by each scoring method on every possible pair of measures,
a matrix of 728 coefficients (twenty-six scoring methods by the twenty-eight possible pairs of the eight measures taken two at a time and without replacement). Inspection of this matrix and of the matrices of normal deviate values of these coefficients supports the hypothesis that the reliability of ER scores independent of their relationship to content is related to scoring methods.

As has been noted, the mean coefficients of reliability of the ER adjusted scores obtained by each of the twenty-six scoring methods averaged across the coefficients obtained between the twenty-eight possible pairs of measures are reported in column 3 of Table 4. The normal deviate procedure was used to test for significant differences between each of the 325 possible pairs of mean coefficients (twenty-six coefficients taken two at a time and without replacement). The results of these tests are reported in Table 7. For the 325 possible pairs of mean coefficients more than 29% of the differences attain the 5% level of significance.

Also as has been noted, the mean coefficients of the reliability of the ER adjusted scores obtained by each of the twenty-six scoring methods summed across the coefficients obtained between each of the eight measures and the sum of the ER adjusted scores for the remaining seven measures are reported in column 6 of Table 4. The same tests for significant differences between pairs of these mean coefficients were made. The results of these tests are reported in Table 7. For the 325 possible pairs of mean coefficients more than 21% of the differences attain the 5% level of significance.
The matrix of 728 reliability coefficients for ER adjusted scores was subjected to the repeated measures analysis of variance and other procedures which have been discussed. The same analyses were also performed upon the matrices of normal deviate values for these matrices of coefficients. Results of these analyses are reported in Table 8.

The F test for main effect, the F test for main effect adjusted for nonadditivity, the F test for main effect in the presence of maximum heterogeneity, and the F test of T Square all show differences to exist in the average values of the coefficients yielded by different scoring methods.

The observed values of the F max statistic and of Cochran's C cannot be directly evaluated because the requisite tables are insufficiently extensive. They do, however, seem to indicate that differences exist in the variability of the coefficients yielded by different scoring methods.

The F test for nonadditivity indicates the presence of a linear by linear interaction between scoring methods and pairs of measures in the determination of the coefficients. The Chi Squares indicate the presence of differential interaction.

These results are observed for coefficients based upon ER adjusted scores whether they are evaluated directly or by their normal deviate values.

The evidence is incompatible with the null proposition of lack of relationship. The sixth hypothesis which asserts that the reliability
of ER scores independent of their relationship to content is related to scoring methods is therefore accepted.

Normalization

As has been noted, content scores and ER scores were analyzed both in their raw form and after normalization. Through comparison of the results of these analyses it is possible to determine the effect of the shapes of distributions upon the findings presented here.

While perfectly normal distributions are seldom observed in practice, it remains common practice among psychologists to assume that the theoretical normal distribution is an accurate representation of much of reality and that departures from this theory are evidence of erroneous practice or, in other words, artifacts of measurement (Ghiselli, 1964). Measurements which yield distributions that approximate the normal do facilitate communication. For example, the magnitude, and therefore the meaning, of the correlation coefficient is related to the distribution characteristics of the data which it describes. While in general departures from normalcy tend to reduce the value of the coefficient, if the departures are similar for both variables they will inflate it (McNemar, 1962).

This latter point is particularly pertinent to this study for two reasons. First, it is generally true that the fewer scales a measure contains, the more unlikely it is that scores on the measure will approximate normalcy in their distribution and ten is a fairly small number of scales. Second, it is possible unusual scoring methods, particularly those which employ folded distributions such as
through 3M 26, will yield unusually nonnormal distributions. For these two reasons it is proper to examine the results of this study in order to determine the extent to which they are attributable to such failure to approximate the normal distribution.

The results reported in this study are based upon substantially more than 1,000 distributions. A conventional approach here might be to examine indexes of skewness and kurtosis for each in an effort to determine its normalcy. While such indexes have been computed, they are not to be considered here because they do not directly address the question of the influence of departures from normalcy upon the results of this investigation.

If a normally distributed variable is subjected to the normalization procedure described previously, the original and the normalized distributions should yield a coefficient of intercorrelation that, within rounding error, approximates a value of 1.0. The raw scores and the normalized scores described above have been intercorrelated. A mean value of .97 is observed for the resulting 216 coefficients (eight measures for content and each of the twenty-six scoring methods). Only 42 coefficients have a value of less than .96, 14 of less than .90, with the lowest being .79. The latter is for ER scores on Ideal computed according to 3M 13. Coefficients of less than .96 are evenly divided among the eight measures except that they are about twice as frequent on the Ideal measure. No values of less than .96 are observed for fifteen of the scoring methods. Of the 42 coefficients with values of less than .96 two are from content scores (Difficulty and Ideal are both .94), four from 3M 1, eight from 3M 13, six from 3M 15, four from
SM 18, three from SM 23, and six from SM 24.

Given the assumption that the normalized scores are more accurate reflections of reality, Table 6 indicates that the use of raw scores artifactually inflates the magnitude of ER score reliability coefficients and also inflates the intercorrelation coefficients observed between ER scores and content scores. Inspection of Table 4 indicates that raw score based estimates of reliability for ER scores are inflated for the less reliable scoring methods but veridical for the more reliable scoring methods. Similarly, inspection of Table 5 suggests that raw score based estimates tend to overstate the relationship between content and ER scores based on the less reliable scoring methods and, perhaps, to display a mixed influence on the assessments of relationships to ER scores based on the more reliable scoring methods. Discrepancies between results obtained from raw scores and normalized scores are particularly noteworthy for SM 13, SM 15, SM 23, and SM 24, as would be anticipated from the observation of lower values for the intercorrelation of their raw and normalized scores.

Inspection of Table 7 shows that coefficients of the reliability of ER scores demonstrate greater differentiation among scoring methods when they are based upon normalized rather than upon raw scores. Initial inspection indicates the contrary result for coefficients of intercorrelation between ER scores and content scores, however, at more stringent levels of significance and for scores based on ten scales greater differentiation of the coefficients is observed when they are based upon normalized scores.
Inspection of Table 3 shows that differences among scoring methods in the values of the average coefficients of reliability for ER scores and of intercorrelation between ER scores and content scores are more clearly observable among coefficients based upon normalized scores than among those based upon raw scores. This appears to be a consequence of the fact which was just noted that normalized scores tend to produce lower coefficients for the less reliable scoring methods. A negligible reduction in the variability of the variances of the reliability coefficients and a similar increase for the intercorrelation coefficients may be attributable to the same fact. The apparently much larger $F_{\text{max}}$ values for the intercorrelation coefficients based on normalized scores is a spurious consequence of a meaninglessly closer approach to a value of zero for the smaller variance. The increase in clarity in the observation of the linear by linear interactions and minor reduction of heterogeneity in the variance-covariance matrices may also be attributable to the more veridical coefficients obtained from the normalized scores.

In sum, the normalization of scores apparently increases the accuracy of the results chiefly by reducing the magnitude of coefficients associated with scoring methods of lower reliability. These more accurate results yield slightly less support for hypotheses one, two, and three and slightly more and clearer support for hypotheses four, five, and six. In no case, however, is the evaluation of a hypothesis substantially related to whether analysis is based upon raw scores or normalized scores.
Adjustment

The employment of ER adjusted scores from which the content score components have been partialled out permits the evaluation of the obverse of a proposition advanced by Cronbach (1946). He asserted that when measures consisting of multipoint scales are scored an extreme response score component of these scores will spuriously enhance or decrease reliability and validity depending upon the relationship between the extreme response component and the true score or the criterion. In this investigation the examination of differences between results based upon raw scores and those based upon adjusted scores permits an assessment of whether content score components spuriously enhance or decrease the observed reliability of extreme response scores.

A matrix of 208 intercorrelations between raw scores and adjusted scores (eight measures by twenty-six scoring methods) yield a mean coefficient approaching .99, when coefficients based on Ideal which have a mean of .92 are excluded. Only 22 values of less than .96 are observed: three for Difficulty, three for Examinations, five for General Evaluation, and eleven, the remaining one-half, for Ideal. These lower coefficients are observed primarily for SK 5, SK 8, and SK 13 and for Ideal on SH 1, SH 3, SH 7, SH 9 and SH 11. The relatively reduced values observed for Ideal may be attributable to the fact that this measure which yields the most skewed content scores also yields the apparently greatest differences in both skewness and kurtosis between raw scores and adjusted scores produced by the various scoring methods. Examination of Table 4 suggests that a content score component may spuriously very slightly enhance the reliability
of SM 4 and SM 5 and spuriously very slightly decrease the reliability of SM 9 and SM 11. No other results give any meaningful indication of the possible influence of the presence of a content score bias component in ER scores.

Scoring Methods

From the results presented in Table 4 and in Table 5 and from the alpha coefficients of reliability presented in Table 9, it is possible to gain a general impression of the relative reliabilities of the twenty-six scoring methods.

Least reliable are SM 13 and SM 24: kurtosis and the folded distribution of kurtosis scores. Slightly more reliable are SM 18 and SM 20: the folded distributions of the tallies of 4's and of 1, 7, 3, and 5's. Slightly more reliable are SM 23, SM 16, SM 19, and SM 21: the folded distributions of the standard deviation scores and of the tallies of 2 and 6's, of 1, 7, 2, and 6's, and of 1, 7, and 4's.

Somewhat more reliable are SM 25, SM 26, SM 14, SM 22, and SM 17: the two combinations of folded distribution scores, and the folded distributions of content scores, of deviations about the hypothetical mean, and of the tallies of 3 and 5's. Somewhat more reliable are SM 6 and SM 15: the tallies of 1, 7, 3, and 5's and the folded distribution of the tallies of 1 and 7's.

More reliable are SM 2, SM 4, SM 10, and SM 12: the tallies of 2 and 6's, and of 4's, and the deviations about the subject means, and the standard deviation scores. More reliable are SM 7, SM 3, SM 11, and SM 9: the tallies of 1, 7, and 4's, and of 3 and 5's, the deviations
Table 9
Reliability of First Eleven Scoring Methods
Coefficient Alpha Values for Each Measure and Across All Eighty Scales

<table>
<thead>
<tr>
<th>Measures</th>
<th>CON</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>.39</td>
<td>.35</td>
<td>.61</td>
<td>.59</td>
<td>.72</td>
<td>.31</td>
<td>.54</td>
<td>.67</td>
<td>.85</td>
<td>.84</td>
<td>.74</td>
<td>.84</td>
</tr>
<tr>
<td>Content</td>
<td>.87</td>
<td>.82</td>
<td>.67</td>
<td>.67</td>
<td>.61</td>
<td>.79</td>
<td>.55</td>
<td>.70</td>
<td>.81</td>
<td>.80</td>
<td>.75</td>
<td>.80</td>
</tr>
<tr>
<td>Instruction</td>
<td>.94</td>
<td>.90</td>
<td>.72</td>
<td>.76</td>
<td>.83</td>
<td>.88</td>
<td>.47</td>
<td>.81</td>
<td>.91</td>
<td>.91</td>
<td>.75</td>
<td>.91</td>
</tr>
<tr>
<td>Examinations</td>
<td>.91</td>
<td>.86</td>
<td>.75</td>
<td>.74</td>
<td>.79</td>
<td>.86</td>
<td>.68</td>
<td>.77</td>
<td>.83</td>
<td>.87</td>
<td>.72</td>
<td>.87</td>
</tr>
<tr>
<td>General Evaluation</td>
<td>.90</td>
<td>.35</td>
<td>.65</td>
<td>.74</td>
<td>.70</td>
<td>.35</td>
<td>.57</td>
<td>.73</td>
<td>.86</td>
<td>.85</td>
<td>.72</td>
<td>.85</td>
</tr>
<tr>
<td>Self</td>
<td>.44</td>
<td>.71</td>
<td>.56</td>
<td>.60</td>
<td>.53</td>
<td>.68</td>
<td>.39</td>
<td>.63</td>
<td>.71</td>
<td>.64</td>
<td>.67</td>
<td>.65</td>
</tr>
<tr>
<td>Average</td>
<td>.41</td>
<td>.75</td>
<td>.65</td>
<td>.70</td>
<td>.64</td>
<td>.77</td>
<td>.58</td>
<td>.67</td>
<td>.77</td>
<td>.75</td>
<td>.72</td>
<td>.75</td>
</tr>
<tr>
<td>Ideal</td>
<td>.68</td>
<td>.91</td>
<td>.81</td>
<td>.92</td>
<td>.66</td>
<td>.36</td>
<td>.31</td>
<td>.88</td>
<td>.90</td>
<td>.71</td>
<td>.70</td>
<td>.72</td>
</tr>
<tr>
<td>Total</td>
<td>.93</td>
<td>.95</td>
<td>.86</td>
<td>.90</td>
<td>.83</td>
<td>.94</td>
<td>.33</td>
<td>.91</td>
<td>.95</td>
<td>.93</td>
<td>.88</td>
<td>.93</td>
</tr>
</tbody>
</table>
about the subject means standardized by scale, and the deviations about
the empirical means for scales.

Most reliable are SM 5, SM 8, and SM 1: tallies of 1, 7, 2, and 6's, deviations about the hypothetical mean, and tallies of 1 and 7's.

In sum: score based methods are less reliable than item based meth­ods and the methods most frequently employed in previous research are the most reliable.

**Interrelation of Methods**

A variety of factor analytic techniques were applied in the analy­ses of various sets of scores. The results of these analyses do not shed much light upon the interrelationships among the various scoring methods. This appears to be attributable to the presence of the folded distribution scores, SM 15 through SM 26. While these scores are relatively moderate in the degree of their intercorrelation, in part arti­factually interrelated in SM 25 and SM 26 which are composites of SM 15 through SM 18, they are so numerous that they consistently emerge as a first factor. Given this fact and an awareness of the relatively lower reliability of these score based scores and of the index of kurtosis it is instructive to directly examine the coefficients of intercorrelation among the remaining and more reliable scoring methods.

The intercorrelation coefficients for the sums of scores, summed across the eight measures, are reported in Table 10 for these more reliable scoring methods. Examination of these coefficients and of those computed for each measure and for each type of scores generally indicates the existence of four additional groupings of scoring methods.
Table 10
Coefficients of Intercorrelations of ER Raw Scores
Summed Across Eight Measures
For Selected Scoring Methods

<table>
<thead>
<tr>
<th>SM</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.19</td>
<td>-.69</td>
<td>-.35</td>
<td>.75</td>
<td>.50</td>
<td>.77</td>
<td>.86</td>
<td>.84</td>
<td>.61</td>
<td>.52</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-.30</td>
<td>-.46</td>
<td>.51</td>
<td>-.60</td>
<td>-.50</td>
<td>.29</td>
<td>.12</td>
<td>.08</td>
<td>.03</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.02</td>
<td>-.80</td>
<td>.28</td>
<td>-.68</td>
<td>-.67</td>
<td>-.66</td>
<td>-.49</td>
<td>-.40</td>
<td>-.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.61</td>
<td>-.43</td>
<td>.33</td>
<td>-.71</td>
<td>-.49</td>
<td>-.33</td>
<td>-.26</td>
<td>-.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.04</td>
<td>.34</td>
<td>.95</td>
<td>.81</td>
<td>.59</td>
<td>.48</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.21</td>
<td>.34</td>
<td>.32</td>
<td>.22</td>
<td>.21</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.39</td>
<td>.51</td>
<td>.38</td>
<td>.34</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.66</td>
<td>.62</td>
<td>.51</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.67</td>
<td>.69</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.94</td>
<td>.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>.70</td>
<td>.10</td>
<td>-.57</td>
<td>-.39</td>
<td>.68</td>
<td>.24</td>
<td>.44</td>
<td>.71</td>
<td>.81</td>
<td>.18</td>
<td>.17</td>
<td>.22</td>
</tr>
<tr>
<td>15</td>
<td>.76</td>
<td>-.33</td>
<td>-.40</td>
<td>-.22</td>
<td>.44</td>
<td>.53</td>
<td>.62</td>
<td>.58</td>
<td>.64</td>
<td>.34</td>
<td>.34</td>
<td>.34</td>
</tr>
</tbody>
</table>
The first and clearly most generally potent of the groupings consists of what may be termed the customary methods, SM 1, SM 5, SM 8, SM 9: tallies of 1 and 7's, of 1, 7, 2, and 6's, deviations about the hypothetical mean, and deviations about the empirical mean. It is noteworthy that the method based upon empirical means most clearly enters this group when the coefficients are based upon adjusted scores. In addition, SM 14, the folded distribution of the content scores, approaches entry to the group and it is most related to SM 9, deviations about the scale means. Also, but more weakly, characteristics of the group are observable in SM 15, the folded distribution of the tallies of 1 and 7's.

The second less potent but much more internally consistent group consists of what might be termed the subject methods of scoring, SM 10, SM 11, and SM 12: deviations about subject means, deviations about subject means standardized by scale, and standard deviation scores.

A third grouping, which does not constitute a consistent factor, is observable in the three remaining methods of simple tallies, SM 2, SM 3, and SM 4: tallies of 2 and 6's, of 3 and 5's, and of 4's.

The last group to be identified consists of the remaining two methods of compound tallies, SM 6 and SM 7: tallies of 1, 7, 3, and 5's, and of 1, 7, and 4's. These methods relate to others in a primarily artifactual manner.

It is noteworthy that the two potent groups of methods, the customary methods and the subject methods, embrace most of the more reliable methods and a substantial degree of intercorrelation is observed between the groups.
Discussion

A large body of previous research has found ER to be a reliable phenomenon. With somewhat less consistency, the same body of research has found ER to be related to negative or pathological personal characteristics. A number of criticisms have been made and questions raised about this body of previous research. Those criticisms and questions led to this study.

A basic criticism of the previous research concerned the content of the measures from which ER scores have been obtained. The bulk of the previous research has obtained ER scores from measures of psychopathology or from measures the nature of which was so unknown or so ambiguous that it could be argued that they were measures of psychopathology or of other personality characteristics. It could, therefore, be argued that the ER scores were nothing more than a consequence of the content of the measures, a consequence of the personal characteristics. If this were true then the reliability and correlates of ER scores which have been found in the previous research might be the reliability and correlates of the measures and worthless as evidence for the existence of any extreme response phenomenon, much less of its reliability and correlates.

In the previous discussion of the instruments it was pointed out that it is not reasonable to argue that the CET measures are measures of personal characteristics. For this reason the findings in this study in support of hypothesis one, that ER is reliable, are evidence that ER is a reliable phenomenon and not merely an artifact of the measurement of personal characteristics.
Even with this evidence that ER may be reliably observed on measures other than measures of personal characteristics, the possibility would remain that ER scores are consequences of the content of the measures upon which they are based, of a content component which ER scores may contain. The findings of this investigation in support of hypothesis two, that ER is related to content, mean that ER scores do contain a content component. This result means that it is possible that the negative or pathological characteristics which previous research has found to be correlates of ER may in fact have simply been correlates of the content components of the ER scores, or, in other words, that they were attributable to the content of the measures employed. The inconsistency in the findings of previous research concerning the correlates of ER scores, therefore, might be due to the fact that the ER scores were based upon measures of differing content which would yield different content components in the ER scores.

Given that ER scores do contain a content component, it is necessary to question whether ER is a valid or merely an artifactual phenomenon. The findings of this study in support of hypothesis three, that ER is reliable independent of content, demonstrate that ER is a valid and reliable phenomenon and not merely an artifact of a content component. In general, the reliabilities of ER scores are virtually identical to the reliabilities of the ER adjusted scores from which the content component has been removed. Future research, therefore, could employ such ER adjusted scores in order to investigate possible correlates of ER which are not artifacts of a content component.
An additional possible criticism of the previous research concerned a more technical point. It was conceivable that the meaning of the content and the form of its presentation might have spuriously combined to generate evidence for the reliability and correlates of ER. It was possible that these results could be attributable to the nonnormality of the distribution characteristics of the scores. The results of the present investigation do show that the reliability of ER is to some extent spuriously magnified by such departures from normality, that the reliability coefficients based upon normalized scores are somewhat lower in value than are those based upon raw scores. Similarly, the results of this study show that departures from normal distribution magnify the correlations between ER scores and content scores and therefore, increase the content component of ER scores.

The results of the investigation do show ER to be reliable even when based upon normalized scores and thus lend support to the previous findings regarding the reliability of ER scores. The fact that in this study departures from normality increased the content component of raw scores suggests, however, that previous findings concerning correlates of ER might be attributable not to ER but to a content component of ER scores which is magnified by this departure from normality.

Another basic criticism of previous research concerns the issue of scoring methods. As has been noted above, the body of previous research on the reliability and correlates of ER has not been consistent in the operational definitions or scoring methods employed
nor has the question of possible differences between scoring methods been investigated with care.

The findings in the present investigation in support of hypotheses four, five, and six demonstrate that in a variety of ways scoring methods do not all produce equal coefficients of reliability, of correlation with content, or of reliability independent of content. In addition, the coefficients of intercorrelation between ER scores obtained by the different methods of scoring are operational definitions of different variables. While these different variables do overlap a great deal, they overlap differentially. It is possible, therefore, that some of the inconsistency in the findings of previous research regarding the correlates of ER may be attributable to the fact that different variables were observed because different scoring methods were employed.

Lastly, it may be noted that the magnitudes of the values of coefficient alpha observed in the present study may be pertinent to the question of the relationship of the meaningfulness of content to ER. On the basis of the values of coefficient alpha which are observed in this study, the values of the coefficients of reliability evaluated in this study would be expected to be greater than the values of the coefficients which have been observed. While it is possible that this discrepancy may be attributable to the brevity of the measures or to points in psychometric theory, it is also possible that the discrepancy may be attributable to the meaningfulness of content. For example, under S:1 for the ten scale measures the average value of coefficient alpha is .83. This means that under
The ER scores for ten scale measures should yield coefficients of intercorrelation with an average value of .83. The observed average value is .40.

It is possible that the ER score on a measure contains two components. In the present case, for example, all scales are related to experience in a course. The ER scores for a measure may contain one component which reflects the general meaningfulness of experience in the course and another component which reflects the meaningfulness of the particular aspect of experience in the course which is tapped by the measure. Thus an ER score on Difficulty would reflect two components: the general meaningfulness of experience in the course and the specific meaningfulness of the experience of the difficulty of the course. Under such a hierarchical explanation the average value of coefficient alpha would be an estimate of the reliability of the combined general and specific meaningfulness components, and the average coefficient observed would represent the reliability of general meaningfulness.

Contextual support for the interpretation of meaningfulness as a differential determinant of ER may be found in the results of investigations by Warr and Coffman (1970) and by Warr and Rogers (1974) and in the results of a series of fifteen studies based in personal construct theory which are reported by Bonarius (Bonarius, H., in press; Bonarius, J., 1966, 1967, 1963, 1971; Bonarius, ..., 1968; Heertens, 1967; Overtoom, 1963; Schaap, 1968; Van der Werff, 1965; Van Harten, 1970; Van Os, 1963; Visser, 1968; Wissena, 1968).
In sum, the findings of the present investigation suggest that the following points may be true with regard to criticisms of and questions about the findings of previous research concerning the reliability and correlates of \( R \). The reliability of \( R \) is demonstrable, although at somewhat reduced levels, with measures other than measures of personal characteristics, with normalized scores, and with adjusted scores from which the content component has been removed and this reliability is differentially determined by scoring methods. The correlates of \( R \) which have been observed in previous research may be attributable to the content of the measures employed, the nonnormality of the scores, and/or the particular scoring methods employed.

**Limitations**

Several limitations of the present investigation may be noted. The subjects were American college students. The measures all addressed their immediate involvement in a particular experience, enrollment in psychology courses. The scales were 7-point bipolar scales with contrasting verbal anchors at their end points. The data were collected by the instructors at the termination of the courses. Twenty-seven specific scoring methods were employed. The question of whether the findings of this investigation may be generalized to other conditions can only be answered by further investigations, involving other populations, settings and measures.
**Future Research**

Some suggestions may be made about the possibilities of related research in the future. Such research should investigate the possible relationship between meaningfulness of content, taken as an individual differences variable, and ER. It should also consider whether established differences in the meaningfulness of content between measures, taken as stimulus variables, generate corresponding differences in the reliability of the ER scores which they generate. Additionally, such research should attempt to explore the possible interrelationship between conditional or stimulus variables and individual difference or respondent variables as determinants of ER. Research should be undertaken to determine whether measures of hierarchical specificity in their content are similarly hierarchical in their meaningfulness, and, if so, the consequences in terms of the ER scores which they yield.

Attempts should be made to develop additional scoring methods and to employ measures containing a larger number of scales. Measures should be employed which display a greater diversity of content reliability and distribution characteristics, and which contain scales of more diverse distribution characteristics. The differential impact on ER scores of such diverse measures in interaction with the various scoring methods should be carefully explored. The interrelationships among scoring methods should be further elucidated, with particular attention being given to those cases where such interrelationships are calculated for scores obtained from different measures.
The search for individual difference correlates of ER should continue. The very factors which the results of the present investigation suggest may have spuriously generated evidence of such correlates might have served in an equally spurious manner to suppress evidence of such correlates. A more broadly defined range of possible correlates should be considered, including not only affective but also cognitive and behavioral variables.

Finally, it is suggested that investigations of ER should evaluate results obtained for raw, normalized, and adjusted scores so that their differential implications may be observed.

Summary

The tendency to employ relatively extreme response positions on multipoint scales was placed in the context of response styles or biases by Cronbach (1946). Research related to this extreme response style was reviewed by O'Donovan (1965), who formulated the issue in terms of two competing explanatory hypotheses: the pathology hypothesis and the meaningfulness hypothesis. A subsequent review of related research by Hamilton (1968) noted O'Donovan's formulation but focused almost exclusively upon marshalling evidence in support of the reliability and correlates of ER in support of the pathology hypothesis. Such research found ER to be characteristic of those who are female, of either very young or very old age, low intelligence, little education, in low status or deviant occupations, from minority religions or races or cultures, and of those holding positive attitudes toward religion. This research also found ER to be associated
with a number of characteristics of psychopathology, especially dogmatism, authoritarianism, ethnocentrism, and the desire for certainty. Subsequent to Hamilton's review such findings in support of the pathology hypothesis continued to appear, albeit with decreasing frequency. As Bonarius (in press) has noted, until recent years relatively few investigations were undertaken in attempts to support the meaningfulness hypothesis.

Criticism of the prior research on ER in support of the pathology hypothesis may be based upon the findings in support of the meaningfulness hypothesis and directed to certain possibly crucial limitations of the previous research on ER. The most basic criticism concerns the content of the measures which have been employed to obtain ER scores. Most studies have employed measures of pathological personal characteristics or measures of such an unknown or ambiguous nature that it could be argued that they are measures of pathological personal characteristics. If ER scores are an index of the meaningfulness of content then in such studies evidence for the reliability and correlates of ER might be nothing more than a spurious consequence of the content of the measures employed. Another fundamental criticism of previous research is to be found in the inconsistency of the operational definitions of ER or scoring methods which have been employed. Most commonly, ER has been conceptualized and operationalized in terms of deviations from a hypothetical mean response position, the midpoint on a multipoint scale. Some studies have simply tallied utilization of the single most extreme positions or of the two most extreme positions while others have included or separately considered the midpoint position. Some
studies have defined ER as the degree of deviation from the hypotheti-
cal mean response, or from the mean response of the group, or from the
mean response of the individual on the scale. Previous practice has
been inconsistent in the adoption of operational definitions or scoring
methods and has failed to attend to the consequences of the selection
of particular methods. Lastly previous research has not attended to
the question of the extent to which findings may have been a conse-
quence of departures from normalcy in ER score distributions.

In an attempt to address these issues six hypotheses were evalu-
ated in the present investigation. It was hypothesized that ER scores,
even when they are obtained from measures which are not measures of
personal characteristics are: reliable, related to content scores, and
reliable independent of their relationship to content scores. It was
also hypothesized that the scoring methods by which ER scores are
obtained differentially determine: reliability, relationship to content,
and reliability independent of content.

Data were obtained from 280 American college students enrolled in
psychology courses, 121 males and 159 females. The students responded
to five Course Evaluation Form measures (Fletcher 1967, 1972) and to
three Study Behavior Form measures developed by the present writer.
Each measure contained ten 7-point bipolar scales with contrasting
verbal anchors at their end points and balanced for reflection. ER
scores obtained by each of twenty-seven scoring methods were computed
for each measure. Seven methods exhausted the possible combinations
of position use tallies relative to the midpoint. Other methods
evaluated deviations about hypothetical means, empirical means, empirical means standardized by scale, subject means, and subject means standardized by scale. Subject standard deviations and indexes of kurtosis were also employed. In addition folded distributions of most of these scores and combinations thereof were employed. For the data in this study, two methods, deviations from empirical means and deviations from empirical means standardized by scale, were found to yield virtually identical results and, therefore, the latter method was eliminated from further evaluation.

Evaluation of the results of this study led to the acceptance of all six hypotheses. In addition, normalization of score distributions was found to somewhat reduce reliabilities and relationships to content. In general, folded distributions, methods of compound tallies, and the index of kurtosis were found to be less reliable and less related to content scores. The most potent grouping of methods proved to consist of the conventional methods: tallies of 1 and 7's, of 1, 7, 2, and 6's, deviations about the hypothetical means, and deviations about the empirical means. A somewhat less potent but much more internally consistent grouping consisted of the subject methods: deviations about subject means, deviations about subject means standardized by scale, and standard deviation scores.

The following conclusions were based upon the findings of this investigation. The reliability of ER is not merely an artifact either of the measurement of personal characteristics or of a content component of ER scores. Content components of ER scores may, however,
have accounted for prior evidence of correlates of ER and such relationships may have been spuriously magnified by departures from normality. The inconsistency in the findings of previous research concerning correlates of ER may have been attributable to inconsistency in either the content of the measures or of the scoring methods employed. It is suggested that future research should attempt to further elucidate the consequences of content components, of the meaningfulness of content, of departures from normalcy of distribution, and of scoring methods upon the reliability and correlates of ER.
APPENDIX A

LIST OF SCALES

Course Evaluation Form

Difficulty (Reflected)

1. My easiest course -------------- My toughest course
2. Easier than most courses ------ Harder than most courses
3. Minimal studying --------------- Time consuming
4. Little work demanded ----------- Much work demanded
5. Easy A or B course -------------- Difficult C course

Difficulty (Not Reflected)

6. Little effort required ------- Extra effort required
7. Leisurely course ------------- Strenuous course
8. Less studying than usual ------ More studying than usual
9. Easier than expected --------- Harder than expected
10. Minimal mental demands ------- Exhausting mental demands

Content (Reflected)

1. Stress on memorizing ----------- Stress on understanding
2. Exams emphasize memory -------- Exams emphasize thinking
3. Ordinary retention ------------ Ordinary reasoning
4. Stereotyped ----------------- Scholarly
5. Retention of facts ------------- Mastery of ideas

Content (Not Reflected)

6. Routine thinking -------------- Critical thinking
7. Tedium ------------------------ Challenging
8. Rote memory ------------------ Reflective consideration
9. Memory emphasized ------------ Reasoning required
10. Memorize ideas --------------- Interpret ideas
Instruction (Reflected)

1. Unimpressive instructor ------- Stimulating instructor
2. Poor lectures --------------- Good lectures
3. Instructor below average ------ Instructor above average
4. Boring instructor ----------- Interesting instructor
5. No lectures effective ------- All lectures effective

Instruction (Not Reflected)

6. Lectures poorly presented ----- Lectures well presented
7. Unimaginative instructor ------ Creative instructor
8. Low quality lectures -------- High quality lectures
9. Routine instruction --------- Challenging instruction
10. Poor instructor ----------- Excellent instructor

Examinations (Reflected)

1. Exams below average --------- Exams better than most
2. Exams detracted from course --- Exams enhanced course
3. Low quality exams ---------- High quality exams
4. Exams distracting --------- Exams helpful
5. Exams confusing ----------- Exams reasonable

Examinations (Not Reflected)

6. Poor exams ---------------- Good exams
7. Poor coverage in exams ------ Good coverage in exams
8. Inadequate exams ---------- Well constructed exams
9. Unfair exams ---------------- Fair exams
10. Inappropriate exams -------- Appropriate exams

General Evaluation (Reflected)

1. Course lacks purpose ------- Meaningful course
2. Little application --------- Pertinent to everyday life
3. Will not recommend -------- Recommend to others
4. Have had enough ----------- Eager for more
5. Vague course objectives ---- Clear course objectives

General Evaluation (Not Reflected)

6. Unimportant ---------------- Significant
7. Sorry I took course --------- Glad I took course
8. Worthless ------------------- Valuable
9. Waste of time --------------- Worth taking
10. Pointless course ---------- Productive course
Study Behavior Form

Self (Not Reflected)

1. Study in bursts ------------ Study methodically
2. Never use library ----------- Extensive use of library
3. Can't concentrate on subject - Give undivided attention
4. Often miss class ----------- Attend every class
5. Review for test only ------- Continual review

Self (Reflected)

6. Depend on self ------------- Seek help from others
7. Dread coming to class ------- Eager to attend class
8. Study everything ----------- Study important points only
9. Very inefficient study ------ Highly effective study
10. Insufficient study ------- Fully adequate study

Average (Reflected)

1. Study in bursts ------------ Study methodically
2. Never use library ----------- Extensive use of library
3. Can't concentrate on subject - Give undivided attention
4. Often miss class ----------- Attend every class
5. Review for test only ------- Continual review

Average (Not Reflected)

6. Depend on self ------------- Seek help from others
7. Dread coming to class ------- Eager to attend class
8. Study everything ----------- Study important points only
9. Very inefficient study ------ Highly effective study
10. Insufficient study ------- Fully adequate study

Ideal (Not Reflected)

1. Study in bursts ------------ Study methodically
2. Never use library ----------- Extensive use of library
3. Can't concentrate on subject - Give undivided attention
4. Often miss class ----------- Attend every class
5. Review for test only ------- Continual review

Ideal (Reflected)

6. Depend on self ------------- Seek help from others
7. Dread coming to class ------- Eager to attend class
8. Study everything ----------- Study important points only
9. Very inefficient study ------ Highly effective study
10. Insufficient study ------- Fully adequate study
APPENDIX B

FORTRAN COMPUTATIONAL FORMULAE FOR DISTRIBUTION STATISTICS

SUBROUTINE DIST(N,X,Y)
DIMENSION X(280),Y(4)

N IS THE NUMBER OF VALUES IN THE DISTRIBUTION.
X IS THE VECTOR OF VALUES OF WHICH THE FIRST N ARE THE
DISTRIBUTION.
Y IS THE VECTOR OF THE CHARACTERISTICS OF THE DISTRIBUTION.
Y(1) IS THE ARITHMETIC MEAN.
Y(2) IS THE STANDARD DEVIATION.
Y(3) IS THE COEFFICIENT OF SKINNESS, G1.
Y(4) IS THE COEFFICIENT OF EXCESS (KURTOSIS), G2.
Y(2), Y(3), AND Y(4) ARE SAMPLE STATISTICS OR UNBIASED
ESTIMATES OF POPULATION PARAMETERS.

S=N
A=0.
B=0.
C=0.
D=0.
DO 1 I=1,N
A=A+X(I)
B=B+(X(I)**2)
C=C+(X(I)**3)
D=D+(X(I)**4)
1 CONTINUE
Y(1)=A/S
Y(2)=((S*B)-(A**2))/(S*(S-1.))
Y(3)=(((S**2)*C)-(3.*S**2*A)+(2.*S**2*A**2))/((S*(S-1.))*(S-2.))
Y(4)=(((S**3)+(S**2))*D)-(4.*((S**2)+S)*C*A)-(3.*((S**2)+S))
1*B**3*12.*S**2*A**2*6.*A**4)/((S*(S-1.))*(S-2.))*(S-3.))
IF(Y(2).LE.0.)GO TO 2
Y(4)=(Y(4))/(Y(2)**2))
Y(3)=Y(3)/(Y(2)**2)*SQR(Y(2)))
Y(2)=SQR(Y(2))
2 CONTINUE
RETURN
END
APPENDIX B

FORTRAN COMPUTATIONAL FORMULAE FOR DISTRIBUTION STATISTICS

SUBROUTINE DIST(N,X,Y)
DIMENSION X(280),Y(4)

N IS THE NUMBER OF VALUES IN THE DISTRIBUTION.
X IS THE VECTOR OF VALUES OF WHICH THE FIRST N ARE THE DISTRIBUTION.
Y IS THE VECTOR OF THE CHARACTERISTICS OF THE DISTRIBUTION.
Y(1) IS THE ARITHMETIC MEAN.
Y(2) IS THE STANDARD DEVIATION.
Y(3) IS THE COEFFICIENT OF SKININESS, G1.
Y(4) IS THE COEFFICIENT OF EXCESS (KURTOSIS), G2.
Y(2), Y(3), AND Y(4) ARE SAMPLE STATISTICS OR UNBIASED ESTIMATES OF POPULATION PARAMETERS.

N=S
A=0.
B=0.
C=0.
D=0.
DO 1 I=1,N
A=A+X(I)
B=B+(X(I)**2)
C=C+(X(I)**3)
D=D+(X(I)**4)
1 CONTINUE
Y(1)=A/S
Y(2)=((S**3)-(A**2))/(S*(S-1.))
Y(3)=((S**2)*C)-(3.*S**2*A)+2.*(A**3))/(S*(S-1.)*(S-2.))
Y(4)=(((S**3)+(S**2))*D)-(4.*((S**2)+S)*C*A)-(3.*((S**2)-S))
Y(3)=((S**3)+12.*S**2*A)-(6.*A**4))/((S**2)-1.)*(S-2.)*(S-3.))
1 CONTINUE
IF(Y(2).LE.0.)GO TO 2
Y(4)=(Y(4)/(Y(2)**2))
Y(3)=Y(3)/(Y(2)**QRT(Y(2)))
Y(2)=QRT(Y(2))
2 CONTINUE
RETURN
END
LIST OF REFERENCES


Brengelmann, J. C. Abnormal and personality correlates of certainty. Journal of Mental Science, 1959a, 105, 142-162.


Edwards, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, 1953, 37, 90-93.


Fletcher, F. M. Student assessment of an educational experience. Unpublished manuscript, The Ohio State University, 1972.


Henmon, V. A. The relation of the time of a judgment to its accuracy. *Psychological Review*, 1911, 18, 186-201.
Hesterly, S. O. Deviant response patterns as a function of chrono-

Horst, P. Factor scores independent of item traits. Educational and

Hotelling, H. The generalization of Student's ratio. Annals of

Hovland, C. I., & Sherif, M. Judgmental phenomena and scales of
attitude measurement: Item displacement in Thurstone scales.

Isaacson, G. A comparative study of the meaningfulness of person and
cultural constructs. Unpublished master's thesis, University of
Missouri, 1962.

Iwawaki, S., & Zax, M. Personality dimensions and extreme response
tendency. Psychological Reports, 1969, 25, 31-34.

Iwawaki, S., Zax, M., & Mitsuoka, S. Extremity of response among
79, 257-259.

Jackson, D. N., & Messick, S. Content and style in personality

Jacobs, D. R. Effects of number of response choices on scale validity
(Doctoral dissertation, Wayne State University, 1971). Dissertation

Johnston, H. T. An empirical study of the value orientation of

Jones, A. Distributions of traits in current Q-sort methodology.

Jones, R. R., & Rorer, L. G. Response biases and trait descriptive

Kanungo, R., & Lambert, W. L. Effects of variations in amount of
verbal repetition on meaning and paired-associate learning.

Kerrick, J. S. The effects of intelligence and manifest anxiety on
attitude change through communications (Doctoral dissertation,
University of Illinois, 1954). Dissertation Abstracts International,
1954, 54, 2661.


Noble, C. E. Emotionality (e) and meaningfulness (m). *Psychological Reports, 1953, 4*, 16.


