

MULTI-TRAIT MULTI-METHOD STUDY OF ADULT
TEMPERAMENT

Thesis for the Degree of M. A.

MICHIGAN STATE UNIVERSITY

HOWARD MARK ERMAN

1977

6102001

ABSTRACT

MULTI-TRAIT MULTI-METHOD STUDY OF ADULT TEMPERAMENT

By

Howard Mark Erman

Though child temperament has been repeatedly studied over the last 15 years, there have been few studies of adult temperament. Studies demonstrating temperament influences on adult behavior are necessary for any general theory of temperament. However, before such evidence can be developed, two interrelated problems must first be met: (1) there must be agreement on what traits can be called temperaments; and (2) there is a need for an adult temperament scale firmly grounded in a theory of temperament. An adult temperament scale is an essential research tool. This study addresses these two interrelated problems.

Temperament as defined by Allport (1961) and elucidated by Buss and Plomin (1975) leads to five theoretical guidelines for an adult temperament scale: (1) temperament categories should be compatible with child development research; (2) the categories should be based on empirical evidence of long-term stability in adults; (3) the categories should reflect style rather than content; (4) the categories should include broad personality dispositions rather than

specific acts; and (5) the categories of temperament should be independent of one another.

Early attempts at studying temperament, particularly Sheldon (1942), were riddled with methodological errors which later invalidated many of the conclusions. The more recent studies, particularly Thomas et al. (1963, 1968), have been successful because they established careful methodological controls. Many of the precautions in the successful studies are part of normal test construction precautions, such as running careful reliability checks. However, some precautions were unique to the problem of measuring temperament. By contrasting the failed studies of temperament with the successful studies, two empirical guidelines for an adult temperament scale can be derived: (6) the unit of temperament analysis, namely a measure of the presence or absence of stable behavioral tendencies, should differ from the units of the questionnaire data; and (7) the best units for questionnaire data are neutral and concrete descriptions of common, everyday behaviors. In the following two examples, the second meets these empirical guidelines while the first does not:

Example 1. I tend not to be impulsive.

Example 2. I do not run out of toothpaste at home
because I keep a spare tube on hand.

Among currently available adult temperament measures, the only test that meets the five theoretical criteria--the EASI-III Temperament Survey developed by Buss and Plomin (1975)--fails to meet the empirical guidelines. Its title is an acronym for its four categories of temperament: Emotionality, Activity, Sociability, and Impulsivity.

A new temperament scale, the Temperament Scale-Erman or TS-E, was specifically developed to meet all seven temperament scale guidelines. Its four categories of temperament are the same as the EASI-III but the definitions of these categories are modified to reflect the work of Bronson (1969, 1971). Existing scales and the descriptive work of Buss and Plomin (1975) provided some items; however, most items were created in small group meetings where individuals were given a definition of a temperament category, asked to name people they knew who might be at one of the extremes of the temperament continuum, and then asked to cite the specific behaviors that led them to categorize the people they had named. Sexually biased or socially desirable items were eliminated. The resulting questionnaire used 20 items per scale, all forced-choice, true-false items. Special instructions were developed to control for cases of the impossible forced choice.

Subjects and Procedure

A multi-trait multi-method study of the four temperament traits was conducted; however, the entire multi-trait multi-method matrix was not generated because it would have placed excessive time demands on subjects. Multi-trait refers here to the four temperament traits studied.

The first method was a reliability study of the Temperament Scale-Erman. This study was primarily designed to see if the four scales were independent; independence would demonstrate discriminant validation. In addition the internal reliability of each scale was measured. The reliability study of the 80-item TS-E used 71 male

and 111 female students enrolled in an undergraduate Abnormal Psychology class at Michigan State University.

The second method was a peer-nomination study. This was used to test criterion validity of the EASI-III and of an expanded (25 items per scale) TS-E. A group of advanced undergraduate students, labeled "seekers," nominated two "subjects" whom they considered as high and low, respectively, on a given dimension of temperament. If seeker perceptions were matched by subject test scores, this would be a form of convergent validity. Different seekers were used for each temperament scale: 25 seekers for Activity; 7 for Emotionality; 11 for Sociability; and 12 for Impulsivity. Each scale was analysed by a matched t-test. In addition, a two-way within subject ANOVA was conducted; it was based on two levels of subjects (high, low) and two tests (TS-E, EASI-III) for each seeker.

Results or Findings

The reliability study showed low correlations between individual TS-E scales, ranging from $-.13$ to $.08$. This is a form of discriminant validation: these low correlations indicate that the four scales of the TS-E measure independent personality characteristics. Internal reliability, as measured by coefficient alpha, ranged from $.44$ to $.60$ for the TS-E scales. These reliabilities are high considering the relatively short length of the TS-E scales (20 items), the true-false (0,1) nature of scale items, the use of behavioral rather than attitudinal items, and above all the fact that this was the first time the TS-E was used.

For the criterion validity study, the following findings were significant ($p < .05$).

1. Criterion validity, in the form of differentiating high nominated subjects from low subjects, was established for all four scales of the Temperament Scale-Erman.
2. Criterion validity was also established for the following EASI-III scales: Impulsivity, Sociability, Activity.
3. When a TS-E scale is used in conjunction with an EASI-III scale, all four temperament scales were able to differentiate high subjects from lows.
4. The TS-E Impulsivity scale was better than the EASI-III at discriminating high nominated subjects from lows.
5. Scores were higher on the TS-E Emotionality scale than on the EASI-III Emotionality scale.

Implications and Conclusions

This study demonstrated the existence of four independent temperament traits. It also established criterion validation for two paper and pencil tests which meet the theoretical guidelines for temperament. The two tests thus provide a way of solving the two inter-related problems of (1) defining and (2) measuring adult temperament.

The TS-E, which meets the empirical guidelines, is already better than the EASI-III in differentiating high nominated subjects from low subjects. The psychometric properties of the TS-E can be improved by lengthening each scale and weeding out poor items. Ways

to improve the TS-E are outlined and important future validity tests of the TS-E are presented. If an improved TS-E passes these validity tests, then a vast range of adult temperament issues becomes open to simple and efficient study.

MULTI-TRAIT MULTI-METHOD STUDY OF ADULT TEMPERAMENT

By

Howard Mark Erman

A THESIS

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

MASTER OF ARTS

Department of Psychology

1977

Copyright by

HOWARD MARK ERMAN

1977

I would like to dedicate this thesis to my wife, Mary Corcoran. Her understanding and support, but above all her love, were central to this much as they are central to so many other things in my life.

ACKNOWLEDGMENTS

The completion of this project depended on help from a large number of people. Professor Gary Stollak, the chairman of my committee, first suggested to me that there was a need for a new adult temperament scale. He remained throughout the project a source of encouragement and sound advice. My original interest in temperament can be traced back to one of the best classes I have had in graduate school, a developmental seminar run by Professor Ellen Strommen. I remembered again why I thought of Ellen as a superb teacher when I read the useful comments she wrote about this project.

Professor Lawrence Messe suggested the method for analysing the validation study. He helped me interpret the data, steered me back to the central issues of my research when I wandered off into the kinds of by-ways where novices are apt to stray, and he provided useful comments on the final draft. But above all I want to thank Larry for his faith in me and in the value of this project; at times when I despaired of ever succeeding or completing anything, he provided a personal interest which carried me through. This concern was invaluable.

Professor Charles Hanley, though the last member to join my masters committee, was certainly first in terms of contributions to

this research. When I started this project, I knew very little about test construction; Professor Hanley was an expert, but more importantly for me, a patient expert. He convinced me to use true-false (0,1) items; he alerted me to dangers such as social desirability factors in the wording of items; he helped me to interpret the reliability study results; and, most importantly, he suggested the validation method used in the study. Professor Hanley never let me lose sight of the major issues of the study or of the theoretical meaning of its findings. He also suggested the title of the thesis. All the members of my committee were accessible for my questions, but as this list of contributions suggests, it was to Professor Hanley that I turned most frequently. I found him stimulating, supportive, and, though he might be among the last to acknowledge this, warm and friendly.

My wife Mary handled all the computer work in this thesis. She also handled me, patiently listening when I provided long and tedious explanation of what I was doing, and providing me with both technical and personal assistance.

I am particularly thankful to four undergraduates who provided invaluable time and assistance at each step of the project. Gayle Baugh, Gil Simon, and Fred Simon joined at the very beginning of the project and worked all winter quarter. They made critical contributions in deciding what the four categories would be, they helped pull useful items from old scales, and they helped run the small groups devoted to writing new items. We then met as a group to review every item to see if the item was appropriate for a given

temperament, and to eliminate any item which seemed sex-biased or biased by a social desirability factor. They also helped run the reliability study. In the spring quarter, Fred Simon was replaced by Geri Sims. The validation study took all quarter to run and Geri, Gil, and Gayle gave countless hours to insuring its success. In terms of their intelligence and their dedication to research, these four undergraduates could match most graduate students.

Finally, I want to thank the large number of personal friends who attended small group meetings devoted to writing new items for the scales. These included Linda Cohen, David and Peggy Hayes, Andrew and Paula McNitt, Professor Brian and Sally Silver, Meredith Taylor, and once again, my wife Mary. Since writing items is rarely an exciting prospect, I assume they helped out of personal friendship; I am grateful.

TABLE OF CONTENTS

	Page
LIST OF TABLES	viii
 Chapter	
I. STATEMENT OF THE PROBLEM AND REVIEW OF THE LITERATURE .	1
Theoretical Guidelines for Adult Temperament	
Questionnaire	2
Category Determination in Traditional Measures	
of Adult Temperament	5
Sixteen Personality Factor (16 PF) of Cattell	
and Eber (1962)	5
California Psychological Inventory (CPI)	
(Gough, 1964)	6
The Guilford-Zimmerman Temperament Survey	
(Guilford & Zimmerman, 1949)	7
Thorndike Dimensions of Temperament (Thorndike,	
1963)	10
Thurstone Temperament Schedule (Thurstone, 1953) .	10
Johnson Temperament Analysis (Johnson, 1944) . . .	11
Empirical Guidelines for an Adult Temperament	
Questionnaire	12
EASI-III (1975): A Temperament Scale That Meets the	
Theoretical Guidelines	26
II. DEVELOPMENT OF THE TEMPERAMENT SCALE-ERMAN (TS-E) . . .	31
Choosing the Categories	31
Writing the Items	35
Format of the Questionnaire	38
III. A MULTI-TRAIT MULTI-METHOD STUDY OF TEMPERAMENT	43
The Reliability Study	44
Method	44
Results	44
The Validity Study	46
Method	46
Results	51
Summary of Results	56

Chapter	Page
IV. DISCUSSION	57
TS-E or EASI-III?	58
Further Development of the TS-E	61
Behavior Validation of the TS-E	63
APPENDICES	
A. TEMPERAMENT SCALE-ERMAN (TS-E)	66
B. TEMPERAMENT SCALE-ERMAN (TS-E): BY TEMPERAMENT CATEGORY	73
LIST OF REFERENCES	80

LIST OF TABLES

Table	Page
1. Definitions of the Four Categories of Temperament in the Temperament Scale-Erman (TS-E)	36
2. Reliability Study: Inter-Scale Correlations for Male, Female, and Pooled Subjects on TS-E	45
3. Reliability Study: Means, Standard Deviations, and Internal Reliability for TS-E	46
4. Means and Standard Deviations of Reliability and Validity Studies	52
5. Results of Matched T-Tests for Individual Scales, Validity Study	53
6. Results of Two-Way Analysis of Variance, Validity Study .	55

CHAPTER I

STATEMENT OF THE PROBLEM AND REVIEW OF THE LITERATURE

While the last 15 years has seen an increasing study of infant temperament (Birns, Barten, & Bridger, 1969; Carey, 1970; Freedman & Keller, 1963; Graham, Rutter, & George, 1973; Thomas, Chess, Birch, Hertzig, & Korn, 1963; Thomas, Chess, & Birch, 1968; and Wilson & Lewis, 1974), there have been few studies of adult temperament. The study of adult temperament could provide critical evidence for any general theory of temperament and its influence on behavior. Should evidence for consistency of specifiable characteristics of temperament emerge, a broad range of issues would then merit further investigation including the effects of temperament on marital interaction and satisfaction, on job choice, on child caregiving behavior, and on both treatment and therapist of choice in clinical setting.

In part the lack of adult temperament studies is due to the lack of agreement on what traits can be labeled "temperaments" and also due to the lack of an adequate instrument to measure these temperaments, particularly in questionnaire form. These two problems--defining the categories and creating a questionnaire measure--go hand in hand. Of course evidence of adult temperament should not be limited to measurement just through paper and pencil tests; adult temperament should also be measurable via naturalistic observation and

in laboratory measurement. However, the efficiency of a self-administered paper and pencil test would provide an essential research tool for the study of adult temperament. Guidelines for constructing such an adult temperament scale can be drawn from a theoretical consideration of temperament and a review of past attempts at the empirical study of temperament.

Theoretical Guidelines for Adult Temperament Questionnaire

While several scales with the word "temperament" in the title now exist, these and related scales, with one exception, are conceptually flawed, reflecting an ambiguity in the common operational definitions of temperament.¹ To understand the problem in available tests, consider Allport's (1961) definition of temperament, often cited as the definition most accepted by psychologists (Buss, 1973).

Temperament refers to the characteristic phenomena of an individual's nature, including his susceptibility to emotional stimulation, his customary strength and speed of response, the quality of his prevailing mood, and all peculiarities of fluctuation and intensity of mood, these being phenomena regarded as dependent on constitutional makeup and therefore largely hereditary in origin (p. 34).

Buss and Plomin (1975) point out three basic components in this definition: (1) temperament involves style rather than content; (2) temperament includes only broad personality dispositions rather than extremely specific or narrow acts; and (3) temperament has an hereditary component.

¹Of course this flaw might only invalidate these tests for studies of temperament; they could be reasonably valid for studying other facets of adult personality.

Most scales of adult temperament reflect only the first two components. Such scales focus on temperament as the how of behavior, the style or manner in which a person acts, as contrasted with the what or why of behavior, ability, or motives, respectively. These scales also try to tap dispositions or innate patterns as broadly defined rather than in specific acts. However, adult temperament scales have not been constructed with an hereditary component in mind. There is one exception--Buss and Plomin's EASI-III scale (1975); its title is an acronym for the four categories of temperament it measures: Emotionality, Activity, Sociability, and Impulsivity.

Buss and Plomin (1975) suggest that heritability is a critical component in order to distinguish temperament from other features of personality which derive from experiences; they suggest that any personality disposition be labeled temperament only if it contains an hereditary component. While beginning with inheritance as the critical criterion for labeling a personality disposition as being a temperament, they then suggest that heritability implies four additional criteria, namely: stability during childhood, retention into maturity, adaptive value, and presence in our animal forebearers.

Based on this reasoning, any scale of adult temperament should include as categories only categories that meet the following guidelines:

1. The temperament categories should be compatible with child development research.

2. The categories should be based on empirical evidence of long-term stability in adults.
3. The categories should reflect style rather than content.
4. The categories should include broad personality dispositions rather than specific acts.

There is one additional theoretical guideline that is independent of the above line of reasoning: if there are different categories representing different aspects of temperament, each category should be relatively uncorrelated with other categories. Not all categories need be totally uncorrelated all of the time; after all, there may be certain personality types or certain age periods in which two categories of temperament become moderately correlated with one another. However, if temperament is to be a meaningful concept, categories neither too large nor too small are needed. The heritability guidelines, #1 and #2 above, insure that categories are not too large. But unless categories are independent of one another, they can be infinitely subdivided into increasingly more minute components. When Sheldon (1942) first grappled with temperament, he assumed that potential temperaments could be found in any English-language adjective that was generally applied to people; hundreds of potential temperaments had to be considered. If there were really hundreds of temperaments, then the concept of temperament becomes so unwieldy as to be useless. Accepting categories only when they are uncorrelated with one another insures a meaningful level of analysis. Hence a fifth guideline:

5. The categories of temperament should be independent of one another.

Category Determination in Traditional Measures of Adult Temperament

While traditional scales of adult temperament meet the last three guidelines, they rarely meet the first two. Instead, as will be seen below in a review of some of these scales, the choices of temperament categories are based on factor analysis, criterion analysis, or some combination of the two.

Sixteen Personality Factor (16 PF) of Cattell and Eber (1962)

The "16 PF" scale was constructed using factor analysis. Cattell (1951) has argued that the factors he derived are the "primary source traits" or building blocks of personalities; Cattell also suggested that the "natural history" of these source traits should be investigated, including their life course and stability. However, many of the categories are clearly incompatible with child development temperament studies, so that even if long-term stability were demonstrated, the factors might still be primarily or entirely learned. Factor H, shyness, is said to be "largely hereditarily determined." However, while the high end of Factor H is said to be "shy, withdrawing, cautious"--all conceivably hereditary--this person is also said to "usually have inferiority feelings," which are unlikely to be inherited. Other factors appear to have an entirely learned component. Consider the following three:

Factor G--Expedient (evades rules, feels few obligations) vs. Conscientious (persevering, staid, rule-bound); Factor N--

Forthright (natural, artless, sentimental) vs. Shrewd (calculating, worldly, penetrating); Factor Q₁--Conservative (respecting established ideas, tolerant of traditional difficulties) vs. Experimenting (critical, liberal, analytical, free-thinking) (Cattell & Eber, 1962).

These categories could only evolve by analysing the behavior of highly socialized individuals; they would be meaningless in the study of infants. The technical title for Factor G is weaker superego strength vs. stronger superego strength; by any commonly accepted definition, superego, even if relatively stable, is based on experience rather than heredity.

California Psychological Inventory (CPI) (Gough, 1964)

Although the CPI was designed to define and then measure both social and personal descriptive personality constructs of a wide relevance, unlike the 16 PF scale, no claim was made that the 18 scales are the basis for the measurement of the total personality. The scales were constructed by an empirical technique, which followed the following three steps: (1) a criterion dimension is defined; (2) items which appear to bear on the defined criterion are then written; (3) the items are then validated on a relevant criterion group independently chosen, usually by a peer nomination method.

However useful the CPI may be for measuring adult personality, it is simply not useful as an adult temperament scale. First, its scales are not independent; Socialization and Dominance correlate +.65; Socialization and Self-Control about +.50; the Responsibility scale usually correlates with the Capacity for Status scale about +.35. Second, it shows only modest stability over even 1 year.

Test-retest correlation over a year's time with high school students ranged from .44 to .77 for females and from .38 to .74 for males. Eleven of the scales for females and 14 of the scales for males had test-retest correlations below .70. The most serious defect is that the scales have no relevance to the study of child temperament research: the purpose of the Tolerance scale is "to identify persons with permissive, accepting, and non-judgmental social beliefs and attitudes," while the purpose of the Sense of Well Being scale is "to identify persons who minimize their worries and complaints, and who are relatively free from self-doubt and disillusionment." The scales consist in part or in their entirety of learned behavioral styles.

The Guilford-Zimmerman Temperament Survey (Guilford & Zimmerman, 1949)

Using a strictly empirical approach, namely a factor analysis of items in a personality inventory, Guilford and Zimmerman obtained 14 factors of temperament; each factor was then divided into smaller, homogeneous units by a "rational approach" in which the items for a given factor were further regrouped using a combination of content inspection and statistical correlation.

The same criticisms which apply to Cattell's research apply here. Since the factors are based neither on a theory nor on any external longitudinal study, there is neither evidence for enduring stability nor a theory to suppose they should be enduring. Since the factors are derived by an empirical study of adult responses to adult test items--test items designed to measure features of adult

personality but having no developmental theoretical framework--there is no reason for the categories to have any relevance to the study of child temperament. Child temperament categories should at least have the theoretical possibility of preceding learning even if in adults they later show some additional learned component. The issue is one of degree: with too large a learned component the characteristic could no longer be labeled a "temperament" even if it shows life-long stability. An individual's native language, such as German rather than English, has life-long stability for most people; they retain their accent even when they have left their native land and rarely speak their native language. However, this primary language, being entirely learned, could not be considered a temperament.

Many of the categories used by Guilford and Zimmermann are wholly or primarily based on learned experiences. The Objectivity factor includes (1) egocentrism, (2) ideas of reference, (3) unwarranted sympathy, (4) hypersensitivity. The Masculinity vs. Femininity factor includes (1) fearfulness, (2) inhibition of emotional experience, (3) masculine vocational interest, (4) masculine avocational interest, (5) disgustfulness, and (6) sympathy. All of these, with the possible exception of fearfulness, would be entirely learned; they are thus unlikely to be temperaments.

Even when a Guilford-Zimmermann factor might appear to possibly be inheritable, it includes a subfactor which would be socialized. For example, in the General Activity factor, liking for action is a subfactor. Of course having adults show such a learned subfactor (liking for action) is compatible with the primary

factor (General Activity) being a temperament, if the subfactor does not comprise too large a component of the primary factor. This learned subfactor would not be a problem if studies revealed that in adults a learned subfactor later became attached to the heritable factor: in the above case, adults who had inherited a high active temperament might usually learn to accept their temperament and later report they "like action."

The learned subfactor is only a problem when the primary factor is defined as including a learned subfactor component, as in the above case: then there is no way of ever sorting out under what circumstances the learned behavior is or is not associated with the temperament. Intervening and antecedent variables cannot be explored. Consider the following analogy: imagine an I.Q. test where liking for reading was a subfactor of the test. Now it is possible that most people with high native intelligence will also show some degree of liking to read, while individuals with low native intelligence will not show such a preference. The proposition that intelligence has some partial genetic component could remain true even if liking to read were both entirely learned and frequently associated with high I.Q. After all, school and families tend to reward success in reading and punish failure, so individuals with high I.Q. scores might learn to "like to read." However, high I.Q. could be an antecedent variable for liking to read and liking to read need not always be associated with high I.Q.: there are many students in the public schools who do not perform up to their potential ability exactly because they do not like to read despite a high

I.Q. If liking to read were defined as a component of I.Q. tests, then the relation between the variables would be lost, since low scores on liking to read would lower overall I.Q. scores. Similarly, when liking for action is defined as part of General Activity, the possibility of someone showing high General Activity but not liking it is precluded from study.

Twin studies are the traditional means of measuring heritability and when the scales listed above have been used in twin studies investigating the heritability of personality features, the results have been uneven (Vandenberg, 1967; Buss & Plomin, 1975). These inconsistent findings are probably caused by the failure to explicitly consider heritability when the scales of the above tests were defined, so that even if a category could theoretically be inherited, the heritable component in the scale is diluted by items in the scale which clearly tap learned attitudes or behaviors.

Thorndike Dimensions of Temperament (Thorndike, 1963)

While the Thorndike Dimensions of Temperament test has many innovative features in format and structure, discussed later, the hypothetical traits that it attempts to measure are based on the factor analytic work of Guilford and Zimmerman, and the objections to their work noted above also apply to Thorndike.

Thurstone Temperament Schedule (Thurstone, 1953)

The seven categories of the Thurstone Temperament Schedule were based on a reanalysis of the schedule of Guilford. At the

same time, interest and personality questionnaires were surveyed and, after items relating to abnormal behavior and psychiatric categories were eliminated, a file of several thousand items was accumulated. By eliminating duplications and items that did not match the categories derived from the factor analysis, 320 items remained in a questionnaire which was filled out by 198 adults. The 20 most discriminating items for each scale were retained, for a total of 140 items.

The same problems noted for the Cattell 16 PF scale and the Guilford-Zimmerman Temperament Survey also apply here.

Johnson Temperament Analysis (Johnson, 1944)

The Johnson Temperament Analysis measures nine different traits, which are defined as "a constellation of behavior patterns and behavior tendencies sufficiently coherent to be measured and effectively used." The nine were chosen because they seemed useful to the areas where the test might be used, including vocational counseling, marriage counseling, diagnosis, and criminology.

The traits are not independent: Depressive correlates with Nervous at $+0.74$, Critical with Subjective at $+0.72$. A more serious problem is that the definition of traits does not include long-term stability as a criterion for choice; Nervousness, for example, "may be a temporary condition brought on by the onset of much worry, fatigue," etc. Finally, the traits are not relevant to child development research.

Empirical Guidelines for an Adult
Temperament Questionnaire

There have been two large-scale observational studies specifically aimed at the study of temperament. The first, a study of adult temperament by Sheldon (1942), had serious methodological flaws and has been generally discredited. The second, a study of child temperament by Thomas et al. (1963, 1968), provided the methodology and the results which spurred many of the recent investigations of infant temperament. Contrasting the two studies to see how the second avoided the problems of the first can provide additional guidelines for considering an adult temperament scale.

On the assumption that any adjective generally applied to people was a potential temperament, Sheldon began by sifting through long lists of such adjectives. This labor led to 650 alleged temperament traits, which, by assiduous sifting and classification, were then reduced to 50. Each of these 50 alleged traits was converted to a seven-point scale. Sheldon then began a systematic study of some 33 male graduate students and academicians. Each subject was extensively studied for 1 year, including 20 analytic interviews plus observation of the subjects in their daily routines and their social interactions. All ratings in this massive initial study were done by Sheldon. Finally intercorrelations were run for the entire series of traits--a total of 1,225 correlations.² A

²This was done before computers were available. Sheldon's comment on this immense labor is poignant: "The tedious element of the job did not lie in hunting and finding traits, however, nor in hunting subjects to be rated. . . . The pain lay in the statistical analysis of large masses of data that accumulated" (p. 16).

search for patterns in these intercorrelations led to the discovery of three basic factors; all traits within a factor had a positive correlation of at least $+0.60$ with other traits within the cluster and negative correlation of at least -0.30 with all other traits in the other two clusters. Of the original 50 traits, 20 met this criterion: six traits in Group I, seven in Group II, and nine in Group III; these groups were later renamed Viscerontonia, Somatotonia, and Cerebrotonia, respectively. These 22 traits became the core of the Scale of Temperament; however, 4 additional years were spent revising and expanding this scale 7 or 8 times eventually to arrive at 20 traits in each cluster which still met the initial correlation criterion.

Sheldon's large-scale study of temperament repeated the same methodology used to develop the scale: subjects had to be studied for over a year in as many situations as possible and in addition interviewed for no less than 20 separate 1-hour sessions, preceded by 1- or 2-hour-long sessions devoted to gaining rapport with the subject. In this study, Sheldon claimed to discover a relationship between body type and temperament.

The Thomas et al. New York Longitudinal Study attempted to trace the emergence of temperament patterns from birth in order to uncover the relation, if any, between infant temperament and child behavior disorders. Over a 12-year period, 141 children from 85 families were studied; the families were fairly homogeneous on most socioeconomic status variables. Most temperament data came from parent interviews beginning shortly after the infant's birth; later

additional data were collected from parent interviews, from direct observations of the infant in school at least once a year, and from observations made when the child was brought in for standard psychological testing at ages three and six. By inductive content analysis of parent interviews from the first 20 children, a total of nine temperament categories were obtained. To rate infants on these categories, a three-point scale was designed for each category.

While the Sheldon study was riddled with methodological flaws, the New York Longitudinal study included four procedures designed to insure objectivity, validity, and reliability of results. Reviewing these four procedures, and contrasting them with Sheldon's errors, provide additional guidelines for constructing an adult temperament scale.

First, to avoid "halo" effects, Thomas et al. made sure that different observers handled the phases of data collection for any particular infant. It is the charge of distortions due to halo effects that is at the heart of various critiques and studies which have largely discredited Sheldon. The relation between body type and temperament that Sheldon claimed to discover was actually an artifact of having the same rater, usually Sheldon, rate subjects on both body type and temperament. Even if the raters did not deliberately try to match temperaments to body types, they may have been influenced by their knowledge of the body type categories; indeed Sheldon specifically says that a well-trained temperament observer "should have the advantage of the constitutional analysis" (p. 56). When other studies attempted to duplicate Sheldon's work

while having different people rate a subject on the two scales, the correlations largely disappeared (Cortes & Gotti, 1965; Walker, 1962).

The second methodological safeguard of Thomas et al. was periodic checks during the course of the study on inter- and intra-observer and interviewer reliabilities. The only reliability study which Sheldon reports being conducted during the course of the study was a re-rating of 83 cases 1 year after the initial rating; though the test-retest correlation was $+ .96$, all the ratings were done by Sheldon, suggesting confounding effects due to both the "halo" effect and memory effects. Another reliability study of the instrument was conducted by having a class of graduate students rate one another on the three components of temperament; reliabilities ranged from $.17$ to $.94$. Sheldon notes that the four best raters had an average mean correlation of $+ .90$ with Sheldon and $+ .86$ with one another. Of course these were not the raters used in his study; had the four worst raters in his graduate class been the raters in his study, the average mean reliability would have been about $.27$.

The third safeguard used by Thomas et al. was a careful attempt to record both a child's first response to a stimulus and all subsequent exposures until a stable pattern was clear. In other words, developmental patterns were carefully scrutinized. Sheldon could not exactly repeat this safeguard, for adults, after all, have usually already had some exposure to most ordinary daily stimuli and their patterns of response are already set. However, Sheldon did not even consider either heritability or developmental possibilities from infancy when he picked his initial 50 traits for study. Of

course his final theory, relating body type to temperament, is highly developmental. But he claimed to build the theory from his observations, and not to make the observations to test his theory. The difference is critical. His three temperaments, later labeled Viscerotonia, Somatotonia, and Cerebrotonia, were each initially comprised of a total grab-bag of unrelated traits. Some traits in a cluster were physiologically based; others socially based. The cluster later called Viscerotonia began with six traits: relaxed posture, love of physical comfort, greed for affection, deep sleep, and need of people when troubled. If we ignore the later morphological theory and consider only these traits, there is no reason to assume the social behaviors stem from the physiological rather than the reverse. Perhaps a learned response causes all of these; for example, these traits might be a variant of what psychoanalytic theory calls the "oral" personality. Or consider the cluster later called Cerebrotonia, which initially stemmed from correlation among the following traits: restrained movements, fast reactions, sociophobia, inhibited social address, vocal restraint, poor sleep habits, youthful intent and manner, and need for solitude when in trouble. Perhaps individuals with youthful intent and manner are more likely to be physically immature and hence self-conscious: all the other traits may reflect being nervous and self-conscious. The point is that when a grab-bag of unconnected traits is correlated, there is no way of knowing a chronological sequence of connection; temperament--implying hereditary or constitutional origins--is only one possible explanation among many possibilities. Finally, the

fact that the 3 temperament clusters each eventually grew to include 20 traits instead of the original 5 or 8 provides no additional evidence for an underlying temperament; these additional traits are an artifact of the methodology, being largely but variants on the original traits in the cluster. Thus, to the pleasure of digestion is added a second trait called love of eating, and a third variant called socialization of eating; to greed for affection there is a second variant called orientation to people and a third called need for people when troubled.

The first three safeguards in the Thomas et al. methodology have some relevance to adult self-report scales. Self-report scales of adult temperament should also be based on developmental categories, as was argued earlier. Scales naturally should show respectable reliability too. And while self-report scales do not have the problem of "halo" effects, they should be designed to minimize response sets, such as social desirability (more on this below).

However, Thomas et al. developed a fourth safeguard and this safeguard provides a particularly critical guideline in considering adult temperament. Objectivity was sought by having interviews focus on details of ordinary daily living, such as eating, sleep, play, etc., with behaviors described in purely factual rather than judgmental terms whenever possible. The need for objective behavior descriptions rather than judgments or attitudes when attempting to judge temperament was already apparent to the earliest temperament researchers. Kretchmer (1925), the first scientist to attempt a temperament study, noted,

If we ask a peasant woman "Was your brother nervous and peace-loving, energetic, etc.?" we shall often get a vague and uncertain answer. If, on the other hand, we ask: "What did he do when he was a child, if he had to go alone in the dark hayloft?" or "How did he behave himself when there was a row up at the pub on a Sunday evening?" then perhaps this same woman will give us concise and unequivocal information, which . . . bears the stamp of trustworthiness. . . . I have laid particular stress on this point, that as much as possible [questions] should be asked in this concrete manner and that direct questions on characterology should only be used to fill out the picture, to fill in the time, and to serve as control questions scattered about among the concrete accounts (pp. 111-112).

Despite his weaknesses in so many other areas of methodology, Sheldon also occasionally strikes this same note. In explaining the trait called "Love of Privacy" he notes: ". . . Ignore any superficial statement of verbal attitude. Most people say they like to be alone. Study the individual's habits and his history" (p. 73).

This point is essential, as Thomas et al. (1968) also noted:

The parent and teacher interviews focused on the details of daily living during feeding, play, sleep, etc. Behavior was described in factual descriptive terms with a concern not only for what the child did but how he did it. Statements as to the presumed meaning of the child's behavior were considered unsatisfactory for primary data. When such interpretative statements were made by a parent or teacher, the interviewer pressed for an actual description. Thus, to a parental report that "the baby hated his cereal," or that "he loved his bath," the question was always posed, "What did he do specifically that made you think he loved or hated it?" Similarly, if a teacher commented that "this child always gets angry if he doesn't get his way," she was asked to give several examples with detailed descriptions of the manner in which the anger was expressed. If a staff observer reported that a child "was afraid to ask the teacher for help," she was instructed to spell out in detail the incidents she had observed and describe the behavior she had interpreted as "fear" (p. 15).

There are two key points to this methodological safeguard. First the unit of analysis, namely the rating on an abstract concept,

which in this case is a temperament category, is different from the unit of the primary data, which is a discrete objective behavior. Second, the behaviors chosen are neutral descriptions of common, everyday behaviors. Both points are also useful for an adult temperament questionnaire.

The first advantage of separating the unit of analysis from the unit of original data is that Thomas, Chess, and Birch were able to avoid alternative meanings offered by similar evaluative words describing behavior. While one mother might say "hated his bath" to mean the child sulked and showed no pleasure, a second might say "hated his bath" to mean the child actively struggled to get out of the water. Though both mothers initially described their child in similar ways, it would be meaningless to say the two children had the same styles of response to new stimuli. The same problem exists in querying adults: three adults who answer yes to "I often feel I am bursting with energy," an item from the EASI scale, might mean three very different things. For one it might mean very physical activities such as jogging a few miles, while for the second it might mean intellectual activity, such as working through a difficult mathematical or crossword puzzle, and for a third a great deal of small-muscle physical activity, such as carving a miniature in soap, where little physical energy is expended.

The second advantage in separating the unit of analysis from the unit of the data is that the analysis should be relatively free of attitude bias. Attitude bias could be either a very idiosyncratic attitude or a general "response set" bias. An example of an

idiosyncratic bias would be a mother who herself hated giving someone else a bath. Without deliberately distorting her response, she might focus on trivial variations in the child's behavior as proof that the child also "hated his bath." However, when the behaviors of the child are themselves described, there may be no evidence that this child "hated his bath" more than other children whose mothers described them as "liking" their bath.

The same distortion process might stem from a more general "response set," such as social desirability. If it is commonly thought that a child who is responsive when held does so because he has been well cared for, and if mothers like to think that they do a good job caring for their child--both attitudes which might reasonably exist--then mothers who say "my child is responsive when being held" may be influenced by a social desirability response set. Even though they are not deliberately lying these mothers might focus on small behaviors which give credence to their views, behaviors which are ignored as too trivial by mothers lacking this social desirability response set for child responsiveness. The distortions due to response sets or idiosyncratic value systems can be weeded out by having the interviewer make sure the descriptions are broken down into behaviors and then using the behaviors alone as the primary data from which to extract judgments about responsiveness, reactions to stimuli or whatever abstract temperament category is being considered.

In a temperament questionnaire, there is no interviewer available to extract specific behaviors from either evaluative

statements or generalities. Instead the items themselves have to be in the form of behaviors. The more general or abstract the behavior description, the less likely a person is to know how to answer, and hence the more likely he is to be influenced by attitudes or response sets. It may be difficult to respond to a statement describing a whole category of behaviors, such as this item from the EASI scale: "I like to be busy all of the time." However, when the behavior is very specific, people can probably decide whether in fact they do the behavior, as in this item: "If I were going from the first floor to the third floor in an office building, I would rather ride an elevator than take the stairs."

The difficulty in responding to a general description of a category of behaviors is further compounded when people are asked to compare themselves to others. One person's judgment about where he stands in relation to others might be quite different from another person's judgments although both display the same behaviors. Again, it should become easier to respond if the behaviors are specific enough. While it might be difficult for someone to make a judgment about this EASI item "I have fewer fears than most people my age," most people would know whether or not "I tend to be frightened during loud thunderstorms."

Of course since we do not actually see the respondents performing the behaviors which comprise the items of a questionnaire, there remains the possibility that we are only tapping the respondent's attitude toward the behaviors rather than an objective inventory of his/her performance. Once again response sets might distort

the results. However, this can at least be partially compensated for by choosing as behaviors neutral descriptions of common, everyday behaviors. A yes-no response to everyday behavior should be less likely to draw on a response set than a yes-no response to a general description of a category of behaviors. If the items themselves are then carefully chosen so as not to be laden with emotional or desirability meanings, then we at least reduce the probability that we are tapping an attitude toward behaviors more than the behaviors themselves.

All of the above arguments about controlling for the effects of attitude in response choices deal with cases where the respondent is not deliberately altering his response. In other words, the respondent is not deliberately lying. When the respondent is deliberately lying, then interviews provide no advantage over questionnaires. Both are at a disadvantage and the only way to discover if a temperament is present is by actually observing an individual in a host of different social and private situations--the arduous methodology developed by Sheldon.

However, if the lying stems not from a general hostility to the test-taking situation but rather from a dislike of the temperamental characteristics being measured, then once again there may be an advantage to a temperament test containing only behavior items. The advantage, which needs empirical proof, would be that subjects taking a test consisting only of specific behavioral items may be less likely to know exactly what the test is attempting to measure than would subjects taking a test such as the EASI, which asks general

behavioral questions. This argument is a variation on the social desirability response set argument but now applied to people who deliberately change their answers according to what they would like to be rather than to people whose responses to items where they lack definite answers are unconsciously affected by their attitudes.

Consider Sociability as a temperament: on the EASI Temperament

Survey the five items which measure this temperament are:

1. I make friends very quickly.
2. I am very sociable.
3. I tend to be shy.
4. I usually prefer to do things alone.
5. I have many friends.

When people read these items, they probably know that this is an attempt to measure sociability (though once again this assertion needs empirical verification). Imagine someone who is in fact not at all sociable but hates himself for being a recluse; such a person might deliberately distort his answers to the EASI scale. Now consider the following items which are behaviorally more specific:

1. People approach me to get acquainted before I approach them.
2. I would rather see a ball game alone than stay home alone and watch the game on TV.
3. When I am happy I sometimes smile or say hello to people I hardly know.
4. I enjoy telling my friends about an interesting experience.

Someone might answer these questions without immediately knowing that they are designed to tap sociability. If their purpose is less manifest, they are probably more likely to be answered truthfully by our recluse who hates himself for being asocial. Empirical support of this argument would involve two findings: (1) subjects have an

easier time guessing the purpose of a test when the items describe classes of behavior rather than specific behaviors; and (2) knowing the purpose of a test alters one's responses.

To make this argument clearer, consider the following analogy. There are many people with high mathematical aptitude who hate mathematics; this can be attested to by any psychology professor teaching an introductory graduate statistics course. If we asked such people, "Do you like mathematics?" they would probably answer "No," and this would be quite true, since it reflects their attitude toward mathematics. If we asked them, "Are you mathematical?" or "Are you mathematically inclined?" they might again answer "No"; in this second case, their answer would be a false answer. This could be simply demonstrated: if we gave them a mathematical aptitude test, a test consisting of mathematical problems, they might perform well above the mean, which is to say they would perform mathematical behaviors quite well. The analogy is to people who claim sociability when asked abstractly but would attest to no sociable behaviors when given a list of specifics.

There remains one additional and central argument for creating an adult temperament questionnaire consisting of specific everyday behaviors. If adult temperament is a meaningful concept, it should actually manifest itself in objective behaviors or ordinary daily life; if it is not so manifested, then either the concept lacks the broad personality disposition which we noted earlier was part of Allport's definition, or else temperament accounts for but a trivial percentage of the variance in behaviors. It would be like people

with high mathematical aptitude being no better at arithmetic or graph reading than people with low aptitudes; in this case, "high mathematical aptitude" becomes a meaningless concept. Similarly, everyone could probably rate themselves somewhere along a five-point Likert-type scale ranging from "very fearful" to "very calm."

However, if in an everyday fear-producing situation, such as an extremely violent storm, 10 people who rate themselves "very fearful" show no more signs of being afraid than 10 people who rate themselves "being calm," then being fearful is either a useless concept or not a temperament. Since the actual display of behaviors such as fearfulness is the key to the validity of the concept, a questionnaire to measure temperament is best off starting with the objective, specific behaviors rather than the abstract categories.

Such a position does not imply that all people who display a particular temperament, such as fearfulness, will all consistently show the same fear responses when in the same situations. If there are 10 behaviors which we expect fearful people to show, we might expect any one person to show only four such behaviors, for example, while a second person might show an entirely different set of four fearful behaviors. With 20 fearful people, each person might show a completely different set of four fearful behaviors. This would lower the correlation between the 10 items. However, despite a low correlation among any two items, the items remain useful items if 20 "calm" people show few or none of the 10 "fearful" behaviors.

Based on the methodological issues raised by the empirical studies of temperament reviewed above, the following guidelines

for an adult temperament questionnaire can be added to those cited previously:

6. The unit of temperament analysis, namely a measure of the presence or absence of stable behavior tendencies, should differ from the units of the questionnaire data.
7. The best units for questionnaire data are neutral and concrete descriptions of common, everyday behaviors.

EASI-III (1975): A Temperament Scale That Meets
the Theoretical Guidelines

The EASI-III (1975), developed by Buss and Plomin, uses categories which meet the theoretical guidelines outlined above but which fail to meet the empirical guidelines. The EASI-III attempts to measure four characteristics of temperament: activity, sociability, emotionality, and impulsivity. These four categories were chosen because on theoretical grounds they meet the heritability criteria and because some empirical evidence for their heritability exists (Buss, 1973). On purely theoretical grounds these four basic categories were each further subdivided as follows: activity into tempo and vigor; emotionality into general emotionality, fear, and anger; impulsivity into inhibitory control, decision time, sensation seeking, and persistence; and sociability into general sociability and affection.

The EASI-III has only been used once, this being a study in which husbands and wives rated themselves and their spouses. A test-retest reliability over 2 to 3 months was obtained for the self-report of 32 women; the reliability ranged from .57 to .95 with an

average reliability of .79. A factor analysis of the four basic categories demonstrated their complete orthogonality. Such complete orthogonality is unusual in a personality questionnaire which was not originally constructed around orthogonal factors, as was the 16 PF, for example. The orthogonality of the four basic scales stems from a high average correlation among the items of each scale; in part this high average correlation of items in a given scale is due to the use of a five-point Likert-type scale for rating each item, a method which effectively raises the average correlation among items (Nunnally, p. 534). This high average correlation may also reflect the wording of the items; rather than being specific behaviors, the items are general and abstract descriptions of behavior tendencies, often being almost the same statement slightly reworded. Consider the following sets of items:

I. General Sociability

1. My spouse makes friends very quickly.
2. My spouse is very sociable.
3. My spouse has many friends.
4. My spouse tends to be a loner.

II. General Emotionality

1. My spouse gets excited easily.
2. My spouse is somewhat emotional.
3. My spouse frequently gets upset.

With statements such as these, it is easy for subjects to uncover the common core of meanings in the items, and hence be consistent.

The factor analysis of these subdivisions did not consistently verify their theoretical independence. The tempo and vigor components of the activity scale consistently loaded on to one factor. Emotionality split into fear and anger components, with the third

theoretical category, general emotionality, splitting between these two. Three of the four impulsivity components were not well-defined factors, persistence being the exception. Although the two sociability factors were independent, Buss and Plomin (1975) concluded that the affection subdivision really measured need for affection and hence did not belong in a temperament survey.

A validity measure of this instrument can be obtained by comparing a subject's self-report score with the rating of this same subject by the spouse. For example, if a husband rates his wife, we can consider the husband to be nominating a subject at different levels of the temperaments being measured; if the self-reports of the wife match the husband's perceptions, then the self-report instrument is a valid measure of temperament style as perceived by others. Here the results were not as good: the rater agreement between a spouse's self-report and the rating by the other spouse ranged from .36 to .75, with an average of .51.

Plomin (1974) ran a further analysis comparing a subject's self-report with that same subject's rating of his or her spouse; for example, the wife's self-report was correlated with the wife's rating of the husband. These comparisons ranged from .45 to -.24, with an average correlation of almost zero, suggesting that a subject's rating of a spouse was not merely a projection of the subject's own personality.

However, it is still unclear why there is not a higher level of agreement between spouse and self report. Plomin suggests that this lack of agreement is due to two factors, the first being that

each spouse lacked an absolute standard for their ratings, and the second being method variance when a subject's self-report is correlated with another person's rating of that same subject.

There are two other possible contributing factors. First, the EASI-III may only be a crude instrument when measuring temperament differences in the middle ranges of temperament; if this is true, then a new validity study has to be run to demonstrate that the EASI-III can be least differentiate between extreme temperament groups. The second explanation is more central: the items of EASI-III are very general and abstract descriptions of behavior rather than descriptions of specific behaviors, and this can cause all of the difficulties reviewed above in the discussion of Empirical Guidelines for a temperament scale. If a social desirability response set is more likely to be activated by general and abstract descriptions of behavior than by neutral descriptions of specific behaviors, and if such a response set is also more likely to be activated during a self-description than during a description of someone else, then the distortions in the EASI-III remain a function of the instrument design even when a subject's rating of a second person is not merely a projection of the first subject's own personality. Furthermore, the error factor which Plomin attributes to the raters' lack of an absolute standard to measure temperament may be exaggerated by the use of general rather than specific behavior items.

In conclusion, although the Buss and Plomin EASI-III appears to hold some promise as an instrument for measuring adult temperament, it still needs to demonstrate its validity.

CHAPTER II

DEVELOPMENT OF THE TEMPERAMENT SCALE-ERMAN (TS-E)

The Temperament Scale-Erman (TS-E) was developed to meet the seven guidelines outlined above. The next section includes a discussion of how this scale was constructed.

Choosing the Categories

The first task in developing the TS-E was choosing categories which met the theoretical guidelines. To meet the heritability guidelines, longitudinal studies of adults were reviewed. Bronson (1969, 1971), using the Berkley longitudinal data, found two relatively stable dimensions of behavior. The first, Expressive-Outgoing versus Reserved Withdrawn, and referred to as "emotional expressiveness," is defined as (a) a continuum from ebullience to depression and (b) differences in the extent to which interactions with other people serve as a focus of involvement. The second dimension, Placid-Controlled vs. Reactive-Explosive and referred to as "reactivity control," is defined as (a) differences in the readiness to act or in prevailing tension and (b) tendency to contentiousness vs. phlegmatic behavior.

The first dimension had a mean persistence over ages 5 to 16 of .73 for boys and .65 for girls. The second dimension had a mean persistence over these ages of .55 for boys and .48 for girls.

These two dimensions accounted for over half of the variance in all rated behaviors over this time span.

Since these two factors are orthogonal, they meet guideline number five. These two factors are described as "Central orientations: characteristic sets of attitudes or response tendencies which affect to a large degree the individual's interactions with his environment." As such, these two dimensions meet guidelines number three and four. The stability noted above meets guideline number one. Since the research begins at age five, it is only tenuously related to child development research. However, Bronson notes that her categories correspond to categories used in the investigations of Escalona and Heider (1959) and Thomas, Chess, Birch, Hergzig, and Korn (1963).

A large component of Bronson's emotional expressiveness dimension is sociability. While Bronson notes that Scarr (1969) showed sociability has a large hereditary component, Bronson herself nevertheless believes that learning is more parsimonious than genetic mediation as an explanation for the stability of these two dimensions. Bronson specifically proposes that either: (1) these orientations receive constant reinforcement from the environment or (2) once adapted by an individual these orientations just tend to persist unless forced to alter by a major disruption of their adaptive value.

However, nothing in Bronson's work precludes a genetic component, and since the temperament scale currently designed is only meant to aid further research into whether such a genetic component exists, these categories still remain useful.

The second major source for temperament categories was Buss et al. (1973, 1975). Following an extended review of published temperament studies, they derived four categories which meet the theoretical guidelines for temperament. These four categories are Emotionality--the level of arousal, which corresponds roughly to intensity of reaction; Activity--the sheer amount of response output; Sociability--the tendency to approach others; and Impulsivity--quickness of response (Buss, 1973). Their twin study indicates these four dimensions are independent and have an heritable component. Buss's Emotionality and Sociability dimensions appear to correspond to Bronson's reactivity control and emotional expressiveness dimensions, respectively.

The work of Thomas et al. (1963, 1968) is the largest longitudinal study specifically aimed at exploring temperament. They used nine categories of temperament, which appear to meet the first four theoretical guidelines. However, their published data only cover the first 5 years of life, so conclusive evidence of stability remains unavailable. Their nine categories also fail to meet guideline five, namely that the categories of temperament should be independent of one another. The nine categories of temperament have intercorrelations which range from $-.49$ to $+.48$; the Intensity category has significant and positive correlations on 31 of the 40 correlations computed. A factor analysis of the categories revealed three major factors; no information is available on the variance accounted for by these three factors. Factor A was primarily comprised of the mood, intensity, approach/withdrawal, and adaptability

categories. Factor A is particularly important because children high in Factor A appear to be at high risk for later behavior disorder. Factor B was primarily comprised of threshold, rhythmicity, intensity, and adaptability. In Factor C, the largest components were activity and intensity.

Although Factor C may correspond to Buss's Activity dimension, Thomas et al.'s first two factors do not appear to correspond to the categories found by Bronson and by Buss et al. This may be a coding artifact due to Thomas et al.'s failure to include a specifically Sociable component. The approach-withdrawal category is defined as reaction to new stimulus, be it animate or inanimate. Examples include, on the one hand, loving new toys and disliking the first taste of orange juice, while on the other, crying when strangers approach and enjoying a visit to the doctor. Mood is defined as the amount of joyful and friendly behavior as opposed to crying and unfriendly behavior. Again animate and inanimate examples are mixed: smiling at strangers or hitting girls on the playground, and fussing before going to sleep or crying when food the baby does not like is presented. If a sociability component were added, Factors A and B might correspond more closely to Bronson's reactivity control and emotional expressiveness dimensions.

Scholom (1975) provides additional evidence that the nine Thomas et al. categories can be reduced in number. He studied temperament in parents and their children by using scales based on Thomas et al.'s nine categories along with the Thorndike Dimensions of Temperament as a second scale for adults. A factor analysis

uncovered three factors present in both adults and children; these he labeled Mood, Energy, and Consistency. Energy and Consistency appear consistent with Buss et al.'s Activity and Impulsivity Dimensions. Mood appears to be an amalgam of what Buss et al. call Sociability and Emotionality.

For the sake of research consistency, the final categories used for the Temperament Scale-Erman (TS-E) were four in number and given the same names as the Buss et al. categories. They are defined as Buss et al. define them, but also expanded in definition to include concepts from Bronson and Thomas et al. Table 1 gives the definitions of the four scales in the TS-E.

Buss et al. also postulated theoretical reasons for subheadings in each of the four categories. For reasons noted earlier (see pp. 5-11), only the fear and anger components in the Emotionality scale were retained. Actually the Bronson reactivity control dimension is closest to just the Buss et al. anger component; fear may eventually be best considered a separate temperament. However, in the current TS-E, fear was subsumed in the Emotionality scale.

Writing the Items

Once the four categories of temperament had been determined, items were written for each of the scales. Some items were drawn from existent scales. With the assistance of four undergraduate psychology majors, every item in the Thorndike Dimensions of Temperament scale, the Thurstone Temperament Survey, the Johnson Temperament Analysis, and the EASI-III was reviewed to see if it appeared to be related to one of our four categories of temperament.

Table 1: Definitions of the Four Categories of Temperament in the Temperament Scale-Erman (TS-E)

Definition of the Impulsivity Scale

Impulsivity: Planful and persistent vs. impulsive and distractible. Impulsivity measures the quickness of our response. On the planful end, it involves planning ahead what you do and then completing it on time. On the far impulsive end, it involves doing things on the spur of the moment, easily dropping one thing to move on to other activities, never completing tasks.

Definition of the Emotionality Scale

Emotionality: Placidity vs. explosiveness. Emotionality refers to differences in our readiness to act, in our prevailing level of tension. We range from being placid or phlegmatic on the one hand to being explosive or contentious on the other hand. These differences may also be differences in our level of arousal, which correspond roughly to the intensity of our reaction. On the explosive end of the scale are the people who, in a given situation, are most likely to express anger or fear.

Definition of the Sociability Scale

Sociability: Ebullient vs. reserved or depressed. This is a measure of our tendency to approach others. It measures differences in the extent to which interaction with other people serves as the focus of our involvement. This continuum ranges from ebullience on the one end and goes toward being reserved on the other, ending in depression or social withdrawal. These descriptive words are used within the social context: someone who is friendly when meeting others but depressed when alone would not be depressed on this scale.

Activity: Active vs. lethargic

The activity scale measures the sheer amount of response output. It refers to the level, tempo, and frequency of our motor and muscular activities. Some people are full of energy, on the go, quick to get things done, ready at a moment's notice. Others are slow, easily tired, less productive than others, and like to move at a leisurely pace.

At the same time, original items were written at more than a dozen small group meetings exclusively devoted to this endeavor. Membership in these small groups changed constantly from meeting to meeting, and included graduate and undergraduate students, both psychology and non-psychology majors, as well as professors¹ and spouses of graduate students and professors. Participants in a given session were read one of the definitions of temperament and then asked to spend a few minutes thinking of people they knew who fit at one of the far ends of the trait being defined. Participants were then asked to think of specific behaviors which made them place these people at the ends of the continuum. These sessions proved the most productive source of items.

A final source of items was Buss and Plomin's A Temperament Theory of Personality (1975). After defining a given temperament, Buss and Plomin provide a description of a hypothetical man who falls at an extreme end of the temperament continuum. These descriptions often consisted of concrete behaviors which could be transformed into TS-E items.

These three sources led to a pool of some several hundred items. Any item which was not behaviorally specific was discarded; most items drawn from existent scales failed this criterion and were therefore discarded. In addition, any items which were specific to the activities of only one sex as well as items that appeared to contain a social desirability element were discarded. All forced

¹Professor Charles Hanley independently supplied some very useful items.

items were carefully reviewed to make sure each of the two choices was equally desirable or undesirable. Decisions on sex role behavior, social desirability, behavioral specificity, and relevance to the appropriate scale were made by my four undergraduate assistants and myself; any item not receiving unanimous consent was discarded.

From the pool of items still remaining, the 20 items that intuitively best matched a given scale were chosen. These 20 items were then rewritten so that 10 items had to be answered true and 10 items answered false to achieve the maximum score in a given direction. Due to a coding error, the Emotionality scale was not quite equally divided; a maximum high score needed 11 true and 9 false answers. Finally, tables of random numbers were consulted to arrange the order in which the items appeared in the questionnaire.

Format of the Questionnaire

The basic decision in choosing the format for a personality questionnaire is whether response choice will be limited to just two possibilities, usually true-false or yes-no. The most common additional choice is a neutral response. Adding a neutral response supposedly reduces examinee resentment that arises when responding in only one direction or the other to a given item seems to present an impossible choice. Among the personality tests reviewed earlier, the CPI is the only test that has forced-choice items; the Guilford-Zimmerman, the 16 PF, the Thurstone Temperament Schedule, and the Johnson Temperament Analysis all provide a middle response choice, while the EASI-III attaches a Likert-type scale to each item that permits five response choices.

Thorndike also assumed that extended response freedom was needed to maintain examinee acceptance of the test, but he developed an innovative alternative in the Dimensions of Temperament scale. His scale consists of 15 sets of items, with each set containing 10 items. The items of a given set contain one item which relates to each of the 10 dimensions of temperament which the test measures; hence the 10 items per set. In a given set of items, examinees pick the four items which are most like them, and the four items which are most unlike them; two items are omitted. Though an ingenious format, the possibility remains that one or two scales may get significantly fewer responses from a given subject. Subject scores on these two scales would then be an inaccurate measure of temperament.

Although the problem of examinee resentment to forced-choice items is a sound reason for including at least a third, neutral response possibility, responses were limited to true or false. The decision was based on the following considerations: first, neutral responses are unscored, and since the original questionnaire is shorter than most traditional personality scales, if large numbers of items were unscored, too few items might be answered to assign a meaningful temperament score to a subject; second, scales were still in the developmental stages. If a number of items were consistently passed over even though the instructions included no such option, those items could always be dropped. Furthermore, a neutral response category could be added at a later time. The fourth reason was that the items were carefully constructed to be neutral descriptions of

behavior and to be free of social desirability. Care was given to balance the choices in a given item. Finally, a Likert-type scale to each item was rejected because such an approach would have made the scale unnecessarily cumbersome, in effect adding a second scale on top of the first. Since the TS-E was closer to a personality inventory than to an attitude scale, high inter-item correlations on a given scale were not an issue of concern.

There nevertheless remained the problem of the impossible choice--items describing behaviors in which the examinee never indulged. To take a rather blatant example, consider TS-E item #12:

If I am following a recipe, I sometimes have to interrupt my cooking because I discover I am out of an ingredient.

There are surely many men and women who have never followed a recipe. To circumvent this problem, two steps were taken. First, the instructions were worded so that items were answered "true" if they were true or mostly true and likewise "false" if false or mostly false. In this the instructions were modeled after the instructions for the MMPI. These instructions provide examinees with greater latitude. The second step was breaking up many items into two clauses, a dependent clause beginning "If..." followed by the independent clause. The dependent clause is underlined and subjects were instructed to assume that this part of the item was actually true whether or not it actually was true. Having assumed the underlined clause true, they were then free to answer the independent clause true or false. If they never actually performed the underlined behavior, they were asked to imagine that they were performing that

behavior and then to pick among the choices of the independent clause, perhaps by thinking of a behavior analogous to the underlined clause. In the example given above, subjects who had followed recipes would answer based on actual experiences. Subjects who had no such experiences would answer the statement by drawing on an analogous experience--perhaps following a manual to repair a car or instructions to construct a toy. The actual instructions for the TS-E were worded as follows:

Please answer the following questions true or false. Answer every question; answer true if the statement is true or mostly true for you. Answer false if false or mostly false for you.

Some of the statements will ask you to pretend that you are in a situation. These statements begin:

"If"

An example is the following:

"If I were to buy a car, I would buy a big car rather than a small car."

For these statements, try to pretend you are doing what is described in the underlined part of the sentence, then answer the rest of sentence true or false, based on how you typically act or would expect to act. In the above example, answer the statement even if you have never bought a car and do not plan to buy a car; pretend you are buying a car and think how you would act in such a situation.

These instructions remove the obstacle of impossible forced choices while at the same time preserving a zero-one response for any item, excluding those that a respondent chooses not to answer at all.

Such instructions have a potential drawback. Since respondents were asked to imagine how they would act for items in which they have never performed the action described in the dependent "If..." clause, some capacity for imaginative self-projection is needed for accurate answers. The relation of responses to some

capacity for imaginative thought would be a problem if this imaginative dimension accounted for a substantial portion of the response variance. The scales would then be unable to pass any validation study based on the nomination method.

CHAPTER III

A MULTI-TRAIT MULTI-METHOD STUDY OF TEMPERAMENT

A multi-trait multi-method study of temperament was conducted using the Temperament Scale-Erman; however, the entire multi-trait multi-method matrix was not generated because it would have placed excessive time demands on subjects. Multi-trait refers to the four types of temperament that the TS-E is supposed to measure. Two methods comprise the multi-method criterion. The first method is the paper and pencil TS-E questionnaire. If the four scales showed little or no correlation with one another, this would indicate that the four scales are measuring different features of personality; this would be a form of discriminant validity. This study also provides a measure of reliability of the TS-E; hence it is called the Reliability Study. The second method was a peer-nomination method using high and low nominated subjects for each scale; if the TS-E and the EASI-III are able to match the nominators in distinguishing high nominees from the lows, this would provide convergent validity for the existence of the four temperaments. This study simultaneously tested the criterion validity of the EASI-III and the TS-E; hence this second study is called the Validity Study.

The Reliability Study

Method

The reliability of the 80-item TS-E was examined by having 71 men and 111 women complete the test. All subjects were students in an undergraduate Abnormal Psychology class at Michigan State University. The test was completed during class time. Students were free to leave the class early and not take the test, but were urged to stay in the interest of research. All students were assured anonymity and no identifying information was included on the completed tests; however, as questionnaires were turned in, they were sorted according to sex of subject. At the time they took the test, students were not told what the test was attempting to measure but 3 weeks later, when results were compiled, the test was explained to them. Students were told they would not be given their individual test scores since the TS-E was still in the experimental stage and individual scores would be meaningless.

Results

The four scales were analyzed for inter-scale correlations. Men and women first were analyzed separately and then the data were pooled. The results appear in Table 2. All correlations were very low, indicating that each scale was measuring an independent personality characteristic. In only one case did a correlation exceed .15: for male subjects, Activity correlated with Emotionality at $-.34$.

Table 2: Reliability Study: Inter-Scale Correlations for Male, Female, and Pooled Subjects on TS-E

Scale	Activity	Impulsivity	Sociability	Emotionality
I. Men: N = 71				
Activity	-	-.06	-.13	-.34
Impulsivity		-	-.02	.05
Sociability			-	-.04
Emotionality				-
II. Women: N = 111				
Activity	-	.07	.06	.00
Impulsivity		-	.06	.09
Sociability			-	.14
Emotionality				-
III. Men & Women: N = 182				
Activity	-	.02	-.02	-.13
Impulsivity		-	.02	.07
Sociability			-	.08
Emotionality				-

Means, standard deviations, and internal reliability as measured by coefficient alpha were also obtained for each scale; Table 3 shows the results for men, women, and pooled subjects. In all cases the internal reliability was between .42 and .60. These modest internal reliabilities reflect the low correlations between items of a given scale; such low correlations would be expected, given the true-false (0,1) nature of the scale.

Table 3: Reliability Study: Means, Standard Deviations, and Internal Reliability for TS-E

Scale	Mean	S.D.	Alpha	Standardized Item Alpha
I. Males: N = 71				
Activity	10.90	2.70	.43	.44
Impulsivity	9.74	3.10	.59	.60
Sociability	10.81	3.02	.63	.59
Emotionality	7.09	2.87	.57	.51
II. Females: N = 111				
Activity	10.82	2.75	.47	.46
Impulsivity	9.44	2.98	.53	.53
Sociability	11.68	2.73	.56	.55
Emotionality	7.68	2.75	.44	.45
III. Males & Females: N = 182				
Activity	10.85	2.72	.44	.43
Impulsivity	9.56	3.02	.55	.55
Sociability	11.35	2.87	.60	.57
Emotionality	7.46	2.80	.48	.48

Maximum score = 20.

The Validity Study

Method

The four scales of the TS-E were each expanded by five items, creating a TS-E of 100 items. The expanded four scales were then validated by a peer nomination method. The Buss EASI-III scales were validated at the same time. For clarity, students who did the nominating are referred to as the "seekers"; the people who were nominated and filled out the scales are referred to as the

"subjects"; the complete EASI-III and TS-E are each "tests"; and each test is composed of four "scales."

Seekers were given one of the four scale definitions that appear in Table 1 and asked to think of two people whom they know, one of whom might fit on the very high end of the scale, and the other on the very low end of the scale. A given seeker thus only aided in the validation of one of the four scales. For example, a seeker would be given the definition of Sociability that appears in Table 1 and would try to think of someone "high" on Sociability and someone "low" on this dimension. If scores on the scale discriminated between subjects in the direction perceived by the seekers, then the scale would be validated.

To insure that subjects were being nominated to meet the scale definitions and not just to match the relevant behaviors described by items on the Temperament Scale-Erman, seekers were not initially shown the TS-E. The seekers were reapproached a few days after they received the scale definition. If the seeker could think of two appropriate subjects, the seeker was then given two copies of the entire TS-E and two copies of that Buss EASI-III scale that matched the scale definition used for the nomination. The seeker gave these instruments to the subject to complete. Thus though a subject nominated for high sociability would only aid in the criterion validation of the Sociability scale of the TS-E, he/she would fill out the entire TS-E; however, he/she would only fill out the Sociability scale of the EASI-III. The entire EASI-III was not used so that time demands did not over-burden subjects.

Some seekers could only think of one appropriate subject, either "high" or "low" on a scale. In such cases, seekers were asked to give the questionnaire to the one nominated subject, as well as to any other person they randomly chose, categorized as "other."

Seekers were asked to identify the initials and the sex of the subject they had nominated; the subject's name was not recorded, so confidentiality of results could be assured. All questionnaires were coded so the scale being validated by a subject and the subject's nominated position on that scale, "high," "low," or "other," could be identified. Seekers were told always to give questionnaires to "low" subjects. To make sure that the subject nominated had filled out the appropriate questionnaire, as seekers returned questionnaires they were always asked the initials of the subject who had completed the questionnaire.

The seekers who aided in the validation of the Sociability, Emotionality, and Impulsivity scales were all advanced undergraduate psychology majors. They received no course credit for nominating subjects. Seekers were explicitly asked not to participate rather than give the questionnaire to subjects whom the seekers did not think appropriate to the ends of the defined scale categories. The numbers of seekers originally used for these three scales were: 12 for Sociability, 13 for Impulsivity, and 9 for Emotionality. One seeker for the Emotionality scale and one for the Impulsivity scale only returned data from a single subject and therefore had to

be dropped from the analysis. In addition, one subject from the Sociability and Emotionality pools failed to complete the appropriate EASI-III scale, and had to be dropped from the analysis comparing the two scales. Thus in the analysis, the pairs of subjects used were as follows: 11 for Sociability, 12 for Impulsivity, and 7 for Emotionality. Finally, six seekers found "other" subjects instead of a "high" or a "low" subject; these "other" subjects replaced two "low" and one "high" Sociable subject and two "low" and one "high" Impulsive subject.

Since the definition of activity is easier to grasp, students in an introductory psychology course were used as seekers for the Activity scale. Students received extra credit for participating in the study as seekers. These students were told they would be allowed to participate in the study for full credit even if they could not think of people who could be placed on either the high or the low ends of the activity scales; in such a case, they could give the questionnaire to any two people, defined as "others." However, the need to clearly identify subjects as either "high" active, "low" active, or "other" was carefully stressed. Twenty-eight seekers were originally used to aid in the validation of the Activity scale; three seekers for the Activity scale returned data from only a single subject and therefore had to be dropped from the analysis.

The EASI-III was scored by assigning a score of zero if the Likert-type scale for an item was marked at the extreme "low" direction of the scale and four if at the extreme "high" end.

Intermediate positions, moving from low to high, were scored one, two, and three, respectively. The TS-E was scored by assigning a score of one to any item answered in the "high" direction and zero to any in the "low" direction. To make scores from the scales on each test comparable, all scores were converted to percentage scores based on the maximum possible raw score for a given scale. Two subjects did not complete every item on the TS-E; for these subjects, the maximum score consisted of the total completed items for a given scale.

The analysis was planned to answer a number of questions. The first question was:

1. Can a scale from a single test differentiate the high nominated subjects from the low nominated subjects?

To answer this question, separate matched t-tests were conducted for each scale of each test. The matched t-test for a given scale used only the pair of subjects nominated by the seeker for that scale. For the TS-E scales, the only subject scores considered were those scores on the scale for which the subject had been nominated. Thus for a pair of subjects nominated as high and low on the activity dimension, their TS-E scale scores for Emotionality, Impulsivity, and Sociability were ignored.

The next set of questions was the following:

2. Do the two tests together distinguish high nominated subjects from low subjects?
3. Is one test better than the other in distinguishing high nominated subjects from low subjects?
4. Do subjects tend to answer more items in a positive direction on one test than on the other?

Question 3 is a critical question for deciding the relative merits of an EASI-III scale versus a TS-E scale. To answer questions 2-4, a within subject two-way analysis of variance was conducted; it was based on two levels of subjects (high, low) and two tests (EASI-III, TS-E) for each seeker. The main effect of subject or Subject Differentiation provided the answer to question 2. The main effect of tests or Test Strength provided the answer to question 4. Finally, the interaction of subject and tests or Test Differences provided the answer to question 3. Again, only matched pairs were used.

Results

The means and standard deviations of each scale in the validity study are presented in Table 4. For the purposes of comparison, Table 4 also includes the results of the reliability study converted to percentage maximum scores. In every scale for each test, the mean of high subjects is above the mean for low subjects. For the TS-E, the reliability data provide some tentative norms of a scale in a random population: Table 4 reveals that the high mean was always above the reliability study mean, while the low mean was always below the reliability study mean. However, for individual subject pairs in the validity study, some high-low reversals occurred for every scale except the TS-E Sociability Scale.² To statistically

²When only high-low pairs, rather than high-other or low-other pairs are considered, then the TS-E Impulsivity scale also shows no crossovers. Furthermore, when high-low pairs alone are considered, the TS-E Impulsivity and Sociability scales had a domain of high subject scores that never intersected the domain of low subject scores. That is, the high subject who scored lowest still scored above all low subjects. This was not true for the other two TS-E scales or for any of the EASI-III scales.

Table 4: Means and Standard Deviations of Reliability and Validity Studies^a

Scale	Validity Study			Reliability Study
	Subjects N	TS-E Scale: Mean (Standard Deviation)	EASI-III: Mean (Standard Deviation)	TS-E (N=182): Mean
Impulsivity	High ^b Imp. 12	60.58 (20.60)	56.25 (18.64)	47.78
	Low Imp. 13	36.00 (15.75)	41.77 (14.82)	
	Remainder ^c 93	41.85	50.90	
	<u>118</u>	(16.94)	(17.87)	
Sociability	High Soc. 12	62.00 (8.94)	60.00 (10.66)	56.75
	Low Soc. 12	31.33 (17.88)	38.33 (16.14)	
	Remainder 94	52.83	50.77	
	<u>118</u>	(17.61)	(18.04)	
Activity	High Act. 27	62.81 (12.64)	64.11 (15.41)	54.25
	Low Act. 26	46.77 (13.85)	39.73 (11.67)	
	Remainder 65	58.89	49.05	
	<u>118</u>	(15.01)	(17.45)	
Emotionality	High Emot. 9	49.78 (15.76)	57.44 (17.62)	37.28
	Low Emot. 76	29.14 (11.25)	39.00 (12.65)	
	Remainder 102	39.08	50.61	
	<u>118</u>	(15.31)	(17.92)	

^aTo make the reliability and validity studies comparable, all means and standard deviations are based on percentage maximum scores. Thus in the validity study, raw scores were first divided by the 25 items per scale; for the reliability study, the results in Table 3: III have been here divided by 20. Scores are printed above as percentage scores.

^b"High" and "low" include the six subjects defined as "others," who replaced two low and one high Sociability subjects and two low and one high Impulsivity subjects.

^c"Remainder" consists of all subjects who participated in the validity study excluding those nominated for the scale listed in column 1. Therefore "remainder" is never a random population.

determine how well each scale worked, the statistical analysis outlined earlier is reported in Table 4.

Matched t-test for individual scales. Table 5 presents the results of the separate matched t-tests for each scale of each test in the validity study. This answers Question 1. In the Temperament Scale-Erman, all four scales were able to differentiate significantly ($p < .05$) the nominated high subjects from the nominated low subjects. Three scales of the EASI-III test were able to differentiate high subjects from low subjects. These were the Impulsivity, Activity, and Sociability scales.

Table 5: Results of Matched T-Tests for Individual Scales, Validity Study

Scale	Test	N Pairs	(Difference) Mean	Standard Error	T-Value	1-Tail Prob.
Emotionality	TS-E	7	-16.57	8.21	-2.02	.045
	EASI-III	7	-16.71	9.41	-1.78	.063
Impulsivity	TS-E	12	-26.08	7.69	-3.39	.003
	EASI-III	12	-15.58	7.18	-2.17	.027
Activity	TS-E	25	-13.76	3.14	-4.38	.000
	EASI-III	25	-21.00	3.81	-5.51	.000
Sociability	TS-E	11	-32.45	6.54	-4.96	.0005
	EASI-III	11	-21.36	7.78	-2.75	.0105

ANOVA results. Table 6 presents the results of the two-way analysis of variance for each scale in the validity study. For every scale, the two tests together were able to distinguish significantly ($p < .05$) high nominated subjects from low subjects (see

Table 6, column 1, Subject Differentiation); so Question 4 can be answered affirmatively for each scale. Questions 2 and 3 for each scale are discussed together above (see Table 6, columns 2 and 3, respectively).

Results for the Impulsivity scale showed no differences between the TS-E and the EASI-III on the level of scores obtained by nominated impulsive/planful subjects (Question 2). However, the Impulsivity scale of the TS-E is significantly better ($p < .05$) than its counterpart on the EASI-III in differentiating high impulsive subjects from low subjects (Question 3).

On the Sociability scale, a strong trend ($p < .06$) indicates that nominated sociable subjects receive a higher score on the TS-E than on the EASI-III. Comparing the two tests' abilities to differentiate high sociable subjects from low sociable subjects reveals that neither test is significantly better. There is a very weak trend ($p < .18$) indicating the superiority of the TS-E over the EASI-III.

On the Activity scales, nominated active subjects tend to score higher on the Temperament Scale-Erman than on the EASI-III; but this trend in Test Strength differences was very weak ($p < .12$). More importantly, a strong trend ($p < .07$) indicates the EASI-III is superior to the TS-E in differentiating high active subjects from low.

Finally, on the Emotionality scale, nominated Emotional subjects scored an average of 14 points higher on the EASI-III than on the TS-E ($p < .05$). However, the two tests were identical in their

Table 6: Results of Two-Way Analysis of Variance, Validity Study

Test Strengths ^a		Subject Differentiation ^b		Test Differences ^c	
Grand Mean (Standard Error)	F= (p <)	Grand Mean (Standard Error)	F= (p <)	Grand Mean (Standard Error)	F= (p <)
<u>I. Impulsivity Scale (N=12 pairs)</u>					
-2.47 (3.62)	F=.47 (p<.51)	-29.46 (10.07)	F=8.56 (p < .01)	-7.42 (3.03)	F=6.00 (p < .03)
<u>II. Sociability Scale (N=11 pairs)</u>					
-5.53 (2.63)	F=4.42 (p < .06)	-38.05 (8.61)	F=19.52 (p < .001)	-7.84 (5.40)	F=2.11 (p < .18)
<u>III. Activity Scale (N=25 pairs)</u>					
4.89 (3.02)	F=2.63 (p < .12)	-24.58 (4.15)	F=35.14 (p < .0001)	5.12 (2.68)	F=3.65 (p < .07)
<u>IV. Emotionality Scale (N=7 pairs)</u>					
-14.04 (4.54)	F=9.56 (p < .08)	-23.58 (9.49)	F=6.15 (p < .05)	.10 (8.12)	F= .0002 (p < .97)
<u>V. Anger Subscale (N=7 pairs)</u>					
4.75 (4.80)	F= .98 (p < .36)	-28.99 (9.65)	F=9.03 (p < .02)	5.35 (9.27)	F= .34 (p < .58)

^aMain effect of tests, answering question 4: Does one test show higher scores? + Grand Mean = higher TS-E score; - Grand Mean = higher EASI-III score.

^bMain effect of subjects, answering question 2: Can the two tests together distinguish high nominated subjects from low subjects?

^cInteraction, answering question 3: Is one test better than the other? + Grand Mean = EASI-III scale is superior; - Grand Mean = TS-E scale is superior.

ability to differentiate high Emotional subjects from low Emotional subjects. In a separate analysis, the Fear subscale was dropped from the TS-E Emotionality scale in order to compare a pure Anger TS-E with the EASI-III Emotionality scale. There were no significant differences in regard to question 2 or question 3.

Summary of Results

All of the results might be summarized as follows.

I. Reliability Study

1. The low correlation between individual TS-E scales, ranging from $-.13$ to $.08$, indicates that the four scales of the TS-E measure independent characteristics.
2. Internal reliability, as measured by coefficient alpha, ranged from $.44$ to $.60$ for the TS-E scales. These modest reliabilities reflect the relatively short length of the TS-E scales and the true-false (0,1) nature of scale items.

II. Validity study: the following findings were significant ($p < .05$).

1. All four TS-E scales demonstrated criterion validity.
2. The following scales of the EASI-III test demonstrated criterion validity: Impulsivity, Activity, and Sociability.
3. All four temperament scales showed criterion validity when an EASI-III scale was used in conjunction with a TS-E scale.
4. The TS-E Impulsivity scale was better than the EASI-III Impulsivity scale in differentiating high nominated subjects from low subjects.
5. TS-E Emotionality scale scores were higher than EASI-III Emotionality scale scores.

CHAPTER IV

DISCUSSION

This research has demonstrated the existence of the four personality traits which comprise the TS-E. It has also demonstrated criterion validity for two paper and pencil tests which meet the theoretical guidelines for temperament and which measure the four personality traits. Having two such measures should provide essential tools for future research aimed at determining how large, if any, is the genetic component of temperament, how strongly temperament affects general behavior, and how stable temperament remains over a lifetime.

The Temperament Scale-Erman was specifically designed to meet empirical as well as theoretical guidelines. These guidelines were derived by contrasting failed studies of temperament with successful studies (see pp. 12-26 of this thesis). The guidelines were the following:

1. The unit of temperament analysis, namely a measure of the presence or absence of stable behavior tendencies, should differ from the units of the questionnaire data.
2. The best units for questionnaire data are neutral and concrete descriptions of common, everyday behaviors.

Since the EASI-III also met criterion validity without meeting these empirical guidelines, an essential question remains: is a test which meets these empirical guidelines a better test of temperament than one which does not? In particular, is the TS-E better than the EASI-III?

Discussion of this question will be followed by two sections that discuss ways to improve the TS-E and important future tests of the TS-E.

TS-E or EASI-III?

Is the TS-E better than the EASI-III? For at least one of the four categories of temperament the answer is clear: the TS-E Impulsivity scale is significantly better than the EASI-III Impulsivity scale. In addition, while the TS-E Emotionality scale was able to differentiate significantly high subjects from low subjects, the EASI-III Emotionality scale was not.

Three other features of this study suggest that meeting the empirical guidelines, as the TS-E did, was quite useful. The first was the comments of students who took the reliability study. After the reliability study was completed, the author spoke with about two dozen students who had completed the questionnaire. While most students reported that the questionnaire was interesting to complete, none of them was able to identify accurately what the questionnaire was measuring. The most frequent guesses were that it was assessing obsessiveness, a topic recently covered in the class lectures, or sex differences. This combination of interest but uncertainty matched

one of the aims in constructing behavioral items: controlling social desirability. Most people probably would rather appear sociable than unsociable and active rather than inactive, for example. Attitude type measures, such as the EASI-III, may be affected by such factors. By using behavioral items, any such social desirability factors can be avoided.

The value of the behavioral items was also suggested by the peculiar drop-out pattern from the Emotionality validity scale. Before potential seekers looked at either of the two tests, they were first asked to think of their two subjects, a high and a low for a given scale. Every scale originally had a minimum of 15 seekers. The Activity and Sociability scales lost a few seekers who never returned the questionnaires, usually because the seeker forgot about the study or else because the seeker's subjects were unavailable. The Impulsivity scale suffered the largest drop-out rate by Seekers; partly this was because it was the last scale to be validated in the study and thus began to run into end of the term time conflicts for the students. In addition, the Impulsivity scale suffered from a self-selection factor: seekers reported that their high impulsive subjects never completed the questionnaire, probably because they were high impulsives.

However, the Emotionality scale was a peculiar case. Here many Seekers who had already thought of their two subjects dropped out of the study just after they first looked over the two questionnaires. These seekers were concerned that when their high Emotional subjects saw the EASI-III scale, their subjects would know they had

been chosen because they were explosive, and these seekers were afraid their subjects would explode at them. Though these incidents are only anecdotal, they suggest that subjects who fill out the EASI-III can immediately know what the test purports to measure. Such knowledge might lead to any of the distortions discussed on pp. 12-26, distortions which should not occur with the behavioral item TS-E.

However, the value of the TS-E format is best suggested by its having done so well the first time it was used. The EASI-III has already been pretested; as its name implies, it is a revision of an earlier form. In addition, it currently leaves little room for further improvement. Poor items already had been discarded and the current scales already have large internal reliabilities, large part-whole correlations, and most importantly, already form four independent factors. The EASI-III is close to the ceiling of its improvability. By comparison, while the TS-E is only at the floor of its improvability, the TS-E is already clearly superior to one EASI-III scale and at least as good as two others. If poor items are discarded and if the weaker scales are lengthened, then the TS-E should markedly improve.

Finally, some indirect evidence, referred to in footnote 2, p. 51, suggests that the presence of matched pairs using "other" subjects may underestimate the value of the TS-E Sociability scale. When these "other" pairs are eliminated, then the TS-E Sociability scale and the TS-E Impulsivity scales are the only two scales among all the TS-E and EASI-III scales where there are no high-low reversals

in a matched pair and where the domain of all low subject scores never intersects the domain of high subject scores.

The contention that it is item types--behavioral versus attitudinal--which differentiate the TS-E from the EASI-III would still need empirical proof. In particular, a simple Q-sort experiment needs to be run. Subjects would be given a deck of cards; on each card would appear either an EASI-III item for a given scale or a TS-E item for that scale. Subjects would then be asked to sort the cards according to whether the item described a specific behavior or a general description of behavior. If most TS-E item cards and few EASI-III item cards appeared in the specific behavior pile, this would clearly demonstrate that subjects in fact perceive differences between the two types of items.

Further Development of the TS-E

A test-retest study of the TS-E needs to be run. Such a study would probably reveal levels of reliability for the TS-E scales considerably above the internal reliabilities reported earlier, demonstrating that the first reliability study found the minimal or floor levels of reliability for each scale. At the conclusion of such a study, subjects could also be asked to guess what they thought the test was attempting to measure. If earlier predictions about the benefits of behavior-specific items are correct, then subjects should have a more difficult time guessing the exact purpose of the TS-E than of a non-behavior-specific test.

TS-E scales could also be improved. If the TS-E had two scales as good as the EASI-III scales and two scales significantly better than comparable EASI-III scales, then the TS-E might be the single instrument of choice for measuring its four categories of temperament. By concentrating on the Activity and Sociability scales, this level of achievement could likely be reached. Without too many changes the TS-E Activity scale could be strengthened until it is as good as the EASI-III Activity scale, while the TS-E Sociability scale could be developed until it is significantly better than its EASI-III counterpart.

The first step in improving the TS-E involves "weeding out" poor items, particularly from the Activity and Sociability scales. Items should only be discarded if they fail to differentiate high subjects from lows over a number of repeated studies. Instead of immediately running new studies, however, the responses to the first reliability and validity studies could be reanalysed using split-half methods.

Before any final decision on dropping items, a new reliability and validity study using a new set of subjects would then have to be run. In this new study, the "discarded " Activity and Sociability items would not actually be discarded. However, a new set of items equal in number to the "discarded" items would be added to the two scales. In addition, 5 to 10 new items would be added to all four scales. If any "discarded" item again failed, then these failed items could actually be discarded for good. New items which appeared unable to differentiate high subjects from low subjects could not

be discarded until this entire procedure for eliminating items was repeated.

Behavior Validation of the TS-E

When the TS-E has been improved, other kinds of studies need to be undertaken. These would determine whether scores on the TS-E can be used to predict behavior. Several areas need to be explored: would high or low scores on a scale predict specific behaviors in an experimental situation? Would groups known to be high or low on a given dimension consistently score in an expected direction?

For the first type of study, the TS-E could be given to a large sample. Then high and low scoring subjects for a given temperament could be called back to perform in a predetermined experimental situation. For example, in an acquaintance process study if subjects were asked to sit in a waiting room along with an experimenter accomplice, then those subjects who had scored high on the Sociability scale might be expected to initiate conversations with the accomplices while subjects who had scored low on Sociability would be expected to sit quietly. Similarly, in a situation inducing frustration, subjects scoring high on Emotionality would be expected to get more upset than subjects who had scored low on Emotionality. Various appropriate experiments could be conducted with Ss scoring at the extremes of each scale.

In the second type of study, the TS-E would be given to known populations. For example, Sociability would be tested by giving the TS-E to a group of public relations employees and a group of forest

fire watchers. The first group needs to be relatively more sociable to succeed at their job, while the second group needs to be relatively less social, so the first group ought to score higher than the second on the TS-E Sociability scale. In some ways this study is similar to the reliability study reported in this paper, but there is one crucial exception: in the present study, high and low categorization was based on the relative judgment of an individual seeker, while in the future study, comparing public relations employees with forest fire watchers, there are probably absolute levels of sociability above or below which one is no longer suitable for the chosen professions. Thus if the TS-E were accurate, no forest fire watcher should receive a higher Sociability score than any public relations employee. In the present study, when a "low" nominated subject scored above some other "high" nominated subject, there is no way of knowing whether the test is inaccurate or whether the two different seekers who did the nominating had very different ideas about "high" or "low."

Consider the following extreme example for the Activity seekers: if one seeker were a seminary student and the other were a football player, while each might in fact choose the highest and lowest active subjects from their circles of friends, it is possible that the seminary student's "high" active is in fact less active than the football player's "low" active. By using known populations, such as the public relations employees for sociability, it is possible to control for relative differences in seeker nominations.

When the TS-E can demonstrate behavior prediction, then clearly the vast range of temperament issues is open to simple and efficient study.

APPENDICES

APPENDIX A

TEMPERAMENT SCALE-ERMAN (TS-E)

Please answer the following questions true or false. Answer every question; answer true if the statement is true or mostly true for you. Answer false if false or mostly false for you.

Some of the statements will ask you to pretend that you are in a situation. These statements begin:

"If"

An example is the following:

"If I were to buy a car, I would buy a big car rather than a small car."

For these statements, try to pretend you are doing what is described in the underlined part of the sentence, then answer the rest of sentence true or false, based on how you typically act or would expect to act. In the above example, answer the statement even if you have never bought a car and do not plan to buy a car; pretend you are buying a car and think how you would act in such a situation.

1. I never buy clothes on the spur of the moment.
2. I will sometimes take out two hours to talk to someone.
3. If I hear a tornado warning, I rarely bother to take cover.
4. If someone cuts into a line I'm waiting on, I would be more likely to say nothing than to complain to them.
5. When I was in high school, I preferred to go out with friends and stir up excitement rather than to go to a movie with them.
6. If I hear strange noises downstairs, I am more likely to call a friend than to ignore the noises and go to sleep.
7. Sometimes I hit or kick vending machines that take my money and give me no product.
8. There are times of the day I need time to just sit and do nothing.
9. From the time I finished high school, I've know what career I wanted.
10. If I am with a group of friends and an old friend we have not seen for years joins us, I would be less likely than my friends to give him/her a big hug.
11. I am able to work long hours without feeling tired.
12. If I am following a recipe, I sometimes have to interrupt my cooking because I discover I am out of an ingredient.
13. I sometimes feel I have to get away from people for a while.
14. If I have dinner with friends, I find that I eat my meal more slowly than my friends eat.
15. If I am wakened from an afternoon nap by people repeatedly honking their car horn outside my window, I would be more likely to first yell at the people in the car than to first close my window.
16. If I started a garden, I would plant the seeds in precisely the time of year that is best for each type of seed.
17. If I walk with people my own height, I find I usually walk quicker than they do.
18. I tend to be free from stage fright in speaking or performing in public.

19. When I arise in the morning, I usually do my regular morning tasks in the same order. (Tasks would include washing, brushing teeth, eating breakfast, dressing, etc.)
20. If I had a little extra money, I would be more likely to buy a big meal than to save it.
21. If I played a violin, I would rather play alone than play in a quartet.
22. If I have a good idea, I like to mull it over before sharing it with others.
23. If I had volunteered to help a church or political campaign and had a choice between two equally boring jobs, I would prefer to sit alone and stuff envelopes rather than sit alone and do telephone canvassing.
24. I find I often hurry to get places even when there is plenty of time.
25. I tend to be annoyed by out of the ordinary noises.
26. If a TV program becomes extremely scary, I turn the channel.
27. If a group of friends gather in my room or apartment and begin to sing a song that irritates me, I am more likely to let them finish the song than to insist they stop.
28. I am able to rest when there are unexpected noises and movements about me.
29. If I play with a group of young children I like, I prefer to play a quiet card game rather than a running game such as tag.
30. If I am driving after a hard day is over, I will try to pass more cars on the highway than I usually do.
31. I rarely mind if people drop in on me without calling ahead first.
32. If I must lose weight, I would prefer some kind of exercise rather than diet.
33. I would rather do a crossword puzzle than play scrabble.
34. I have an easy time starting a conversation with strangers at a party.
35. I like to take a nap during the day if it is possible.
36. If I am selecting what to wear on a cold winter day, I decide based on what enhances my appearance rather than what protects my health.

37. If I hurry to catch a bus and just miss it, I'd be more likely to shrug it off than to do a "slow burn."
38. I cry very easily during sad movies.
39. If I have a letter to mail, I will sometimes carry it for a few days before I mail it.
40. I like to be off and running as soon as I wake up in the morning.
41. I am frightened during loud thunderstorms.
42. Six hours of sleep is enough for me most nights.
43. If I have to speak to a group, I tend to pace a great deal.
44. Considering my income and needs, I tend to go into debt more than I should.
45. If I had to pick a job, I would pick a job where I sit and count the number of cars passing through an intersection over a job where I clean every door handle in an office building.
46. If I have to leave the house for the day, I listen to weather reports so I will know what to wear.
47. If I go on a trip, I prefer having my itinerary planned ahead of time.
48. I often run up stairs taking two at a time.
49. If I have to file federal income tax, I file as early as January or February.
50. If I go to a movie and someone near by bothers me with their talking, I am more likely to ignore him than to ask that person to stop.
51. I enjoy spending an evening alone by myself.
52. If I were a good swimmer and were by a beach or pool, I would still prefer lying in the sun over going for a long swim.
53. If I play golf, I would rather play as a twosome than play a solitary game trying to beat par.
54. If, after completing a large grocery shopping, a check-out clerk yells at me because I've left my check book at home, I would be more likely to quietly apologize than to yell back.
55. When I do not have enough sleep, I become irritable.

56. If I were camping and saw a snake, I would be more likely to stay and see what kind it is than to run away.
57. I prefer interesting work where I sit at a desk over work in which there is vigorous activity.
58. I don't run out of toothpaste at home because I keep a spare tube.
59. If I drive alone in a car with a radio in it and I cannot find my favorite radio station, I am more likely to change stations repeatedly than to settle on one station.
60. I would rather go to a quiet party than watch an excellent television show alone.
61. If I were going from the first floor to the third floor in an office building, I would rather ride an elevator than take the stairs.
62. I prefer to ride rather than walk when the distance is moderate.
63. During my vacations I would prefer just resting to sports or sightseeing.
64. If I see a big dog barking, I will often cross the street to avoid it.
65. Having people around sometimes gets on my nerves.
66. If I have an alarm clock, I often forget to set the alarm clock on days when I should set it.
67. When I am happy I sometimes smile or say hello to people I hardly know.
68. I enjoy telling my friends about an interesting experience.
69. I often make up lists of tasks and errands to do.
70. If I have a library book checked out, I almost always return the book to the library before the day on which a fine is charged.
71. I tend to fidget at a long lecture or sermon even when I am interested in the subject.
72. When I have a problem I would rather talk to a friend than mull it over by myself.
73. If I won a thousand dollars (\$1,000) in a state lottery, I would spend most of it on a spree rather than save most of it.
74. If I change a light bulb, I always unplug the lamp even if it is already turned off.

75. People approach me to get acquainted before I approach them.
76. If I went to a football game, I would be more likely to sit quietly than to scream and yell aloud.
77. I usually punch elevator buttons several times in a row rather than just once.
78. If I go alone to a restaurant, I would prefer waiting fifteen minutes for an empty table over being seated immediately at a table with a stranger.
79. I would rather go to see a ball game alone than stay home alone and watch the game on TV.
80. If there is a good movie on TV, I would enjoy watching it with friends more than watching it alone.
81. I sometimes forget to bring along enough money or my checks when I go shopping.
82. I would rather listen to records alone than go to a concert alone.
83. If a traveling encyclopedia salesman came to my house, I would be more likely to listen to his complete sales pitch than to quickly tell him to stop and go away.
84. If I saw a pair of policemen come out of their car with guns drawn, I would be more likely to immediately duck for cover than I would be to first wait and see about what was going on.
85. During my leisure time, I quickly get bored just sunbathing or lying in a shaded, grassy area.
86. I usually say the first thing that pops into my head without first considering the consequences.
87. I often do a lot of physical exercise.
88. If I have been sick and confined to bed, I would be more likely to start my usual activities as quickly as possible rather than take an extra day to relax.
89. If I am on an express check-out lane for people purchasing fewer than ten items and someone with a full grocery cart is in front of me, I would be more likely to tell that person off than to say nothing.
90. If I were riding on a train, I would rather have a stranger sit next to me than sit alone for the whole trip.

91. If I had plants to care for, I would remember to water the plants every time they needed to be watered.
92. If I joined a group of five strangers gathered to talk about personal problems, I would be more likely to be the very last to talk than the very first.
93. I would rather relax and go fishing than go on a vigorous hike.
94. If I am about to leave on a trip, I will carefully plan exactly what clothing I will take along.
95. In a heated discussion, I am often the first person to raise my voice.
96. If I am about to buy something expensive, I will first check out the price at many different stores rather than buy at the most convenient store.
97. If a neighbor's party is too loud and is keeping me awake, I am more likely to complain to them (or to the police) than to just ignore the noise.
98. If I were lost in a strange city, I would rather ask directions from strangers than consult a simple map.
99. If I stopped to join a crowd which had gathered to hear a street band perform wonderful music, I would be more likely to share my enthusiasm with strangers than I would be to keep my enthusiasm to myself.
100. I prefer parties where people sit and talk over parties filled with activities.

APPENDIX B

TEMPERAMENT SCALE-ERMAN (TS-E) BY TEMPERAMENT CATEGORY

ACTIVITY

(True = High Activity)

1. (11) I am able to work long hours without feeling tired.
2. (17) If I walk with people my own height, I find I usually walk quicker than they do.
3. (24) I find I often hurry to get places even when there is plenty of time.
4. (32) If I must lose weight, I would prefer some kind of exercise rather than diet.
5. (40) I like to be off and running as soon as I wake up in the morning.
6. (42) Six hours of sleep is enough for me most nights.
7. (43) If I have to speak to a group, I tend to pace a great deal.
8. (48) I often run up stairs taking two at a time.
9. (71) I tend to fidget at a long lecture or sermon even when I am interested in the subject.
10. (77) I usually punch elevator buttons several times in a row rather than just once.
11. (85) During my leisure time, I quickly get bored just sunbathing or lying in a shaded, grassy area.
12. (87) I often do a lot of physical exercise.
13. (88) If I have been sick and confined to bed, I would be more likely to start my usual activities as quickly as possible rather than take an extra day to relax.

(False = High Activity)

14. (8) There are times of the day I need time to just sit and do nothing.
15. (14) If I have dinner with friends, I find that I eat my meal more slowly than my friends eat.
16. (29) If I play with a group of young children I like, I prefer to play a quiet card game rather than a running game such as tag.
17. (35) I like to take a nap during the day if it is possible.
18. (45) If I had to pick a job, I would pick a job where I sit and count the number of cars passing through an intersection over a job where I clean every door handle in an office building.
19. (52) If I were a good swimmer and were by a beach or pool, I would still prefer lying in the sun over going for a long swim.
20. (57) I prefer interesting work where I sit at a desk over work in which there is vigorous activity.
21. (61) If I were going from the first floor to the third floor in an office building, I would rather ride an elevator than take the stairs.
22. (62) I prefer to ride rather than walk when the distance is moderate.
23. (63) During my vacations I would prefer just resting to sports or sightseeing.
24. (93) I would rather relax and go fishing than go on a vigorous hike.
25. (100) I prefer parties where people sit and talk over parties filled with activities.

SOCIABILITY

(False = High Sociability)

1. (10) If I am with a group of friends and an old friend we have not seen for years joins us, I would be less likely than my friends to give him/her a big hug.
2. (13) I sometimes feel I have to get away from people for a while.

3. (21) If I played a violin, I would rather play alone than play in a quartet.
4. (22) If I have a good idea, I like to mull it over before sharing it with others.
5. (23) If I had volunteered to help a church or political campaign and had a choice between two equally boring jobs, I would prefer to sit alone and stuff envelopes rather than sit alone and do telephone canvassing.
6. (33) I would rather do a crossword puzzle than play scrabble.
7. (51) I enjoy spending an evening alone by myself.
8. (65) Having people around sometimes gets on my nerves.
9. (75) People approach me to get acquainted before I approach them.
10. (78) If I go alone to a restaurant, I would prefer waiting fifteen minutes for an empty table over being seated immediately at a table with a stranger.
11. (82) I would rather listen to records alone than go to a concert alone.
12. (92) If I joined a group of five strangers gathered to talk about personal problems, I would be more likely to be the very last to talk than the very first.

(True = High Sociability)

13. (2) I will sometimes take out two hours to talk to someone.
14. (18) I tend to be free from stage fright in speaking or performing in public.
15. (34) I have an easy time starting a conversation with strangers at a party.
16. (53) If I play golf, I would rather play as a twosome than play a solitary game trying to beat par.
17. (60) I would rather go to a quiet party than watch an excellent television show alone.
18. (67) When I am happy I sometimes smile or say hello to people I hardly know.

19. (68) I enjoy telling my friends about an interesting experience.
20. (72) When I have a problem I would rather talk to a friend than mull it over by myself.
21. (79) I would rather go to see a ball game alone than stay home alone and watch the game on TV.
22. (80) If there is a good movie on TV, I would enjoy watching it with friends more than watching it alone.
23. (90) If I were riding on a train, I would rather have a stranger sit next to me than sit alone for the whole trip.
24. (98) If I were lost in a strange city, I would rather ask directions from strangers than consult a simple map.
25. (99) If I stopped to join a crowd which had gathered to hear a street band perform wonderful music, I would be more likely to share my enthusiasm with strangers than I would be to keep my enthusiasm to myself.

IMPULSIVITY

(False = High Impulsivity)

1. (1) I never buy clothes on the spur of the moment.
2. (9) From the time I finished high school, I've known what career I wanted.
3. (16) If I started a garden, I would plant the seeds in precisely the time of year that is best for each type of seed.
4. (19) When I arise in the morning, I usually do my regular morning tasks in the same order. (Tasks would include washing, brushing teeth, eating breakfast, dressing, etc.)
5. (46) If I have to leave the house for the day, I listen to weather reports so I will know what to wear.
6. (47) If I go on a trip, I prefer having my itinerary planned ahead of time.
7. (49) If I have to file federal income tax, I file as early as January or February.
8. (59) I don't run out of toothpaste at home because I keep a spare tube.

9. (69) I often make up lists of tasks and errands to do.
10. (70) If I have a library book checked out, I almost always return the book to the library before the day on which a fine is charged.
11. (91) If I had plants to care for, I would remember to water the plants every time they needed to be watered.
12. (94) If I am about to leave on a trip, I will carefully plan exactly what clothing I will take along.
13. (96) If I am about to buy something expensive, I will first check out the price at many different stores rather than buy at the most convenient store.

(True = High Impulsivity)

14. (5) When I was in high school, I preferred to go out with friends and stir up excitement rather than to go to a movie with them.
15. (12) If I am following a recipe, I sometimes have to interrupt my cooking because I discover I am out of an ingredient.
16. (20) If I had a little extra money, I would be more likely to buy a big meal than to save it.
17. (31) I rarely mind if people drop in on me without calling ahead first.
18. (36) If I am selecting what to wear on a cold winter day, I decide based on what enhances my appearance rather than what protects my health.
19. (39) If I have a letter to mail, I will sometimes carry it for a few days before I mail it.
20. (44) Considering my income and needs, I tend to go into debt more than I should.
21. (59) If I drive alone in a car with a radio in it and I cannot find my favorite radio station, I am more likely to change stations repeatedly than to settle on one station.
22. (66) If I have an alarm clock, I often forget to set the alarm clock on days when I should set it.
23. (73) If I won a thousand dollars (\$1,000) in a state lottery, I would spend most of it on a spree rather than save most of it.

- 24. (81) I sometimes forget to bring along enough money or my checks when I go shopping.
- 25. (86) I usually say the first thing that pops into my head without first considering the consequences.

EMOTIONALITY

(True = High Emotionality)

- 1. (6) If I hear strange noises downstairs, I am more likely to call a friend than to ignore the noises and go to sleep.
- 2. (7) Sometimes I hit or kick vending machines that take my money and give me no product.
- 3. (15) If I am wakened from an afternoon nap by people repeatedly honking their car horn outside my window, I would be more likely to first yell at the people in the car than to first close my window.
- 4. (25) I tend to be annoyed by out of the ordinary noises.
- 5. (26) If a TV program becomes extremely scary, I turn the channel.
- 6. (30) If I am driving after a hard day is over, I will try to pass more cars on the highway than I usually do.
- 7. (38) I cry very easily during sad movies.
- 8. (41) I am frightened during loud thunderstorms.
- 9. (55) When I do not have enough sleep, I become irritable.
- 10. (64) If I see a big dog barking, I will often cross the street to avoid it.
- 11. (74) If I change a light bulb, I always unplug the lamp even if it is already turned off.
- 12. (84) If I saw a pair of policemen come out of their car with guns drawn, I would be more likely to immediately duck for cover than I would be to first wait and see about what was going on.
- 13. (89) If I am on an express check-out lane for people purchasing fewer than ten items and someone with a full grocery cart is in front of me, I would be more likely to tell that person off than to say nothing.

14. (95) In a heated discussion, I am often the first person to raise my voice.
15. (97) If a neighbor's party is too loud and is keeping me awake, I am more likely to complain to them (or to the police) than to just ignore the noise.

(False = High Emotionality)

16. (3) If I hear a tornado warning, I rarely bother to take cover.
17. (4) If someone cuts into a line I'm waiting on, I would be more likely to say nothing than to complain to them.
18. (27) If a group of friends gather in my room or apartment and begin to sing a song that irritates me, I am more likely to let them finish the song than to insist they stop.
19. (28) I am able to rest when there are unexpected noises and movements about me.
20. (37) If I hurry to catch a bus and just miss it, I'd be more likely to shrug it off than to do a "slow burn."
21. (50) If I go to a movie and someone near by bothers me with their talking, I am more likely to ignore him than to ask that person to stop.
22. (54) If, after completing a large grocery shopping, a check-out clerk yells at me because I've left my check book at home, I would be more likely to quietly apologize than to yell back.
23. (56) If I were camping and saw a snake, I would be more likely to stay and see what kind it is than to run away.
24. (76) If I went to a football game, I would be more likely to sit quietly than to scream and yell aloud.
25. (83) If a traveling encyclopedia salesman came to my house, I would be more likely to listen to his complete sales pitch than to quickly tell him to stop and go away.

LIST OF REFERENCES

- Allport, G. W. Pattern and growth in personality. New York: Holt, Rinehart and Winston, 1961.
- Birns, B., Barten, S., & Bridger, W. H. Individual differences in temperamental characteristics of infants. Transactions of the New York Academy of Science, 1969, 31(8), 1071-1082.
- Bronson, W. C. Stable patterns of behavior: The significance of enduring orientations for personality development. In Hill, J. P. (ed.), Minnesota Symposia on Child Psychology, Vol. 2. Minneapolis: The University of Minnesota Press, 1969.
- Bronson, W. C. Adult derivations of emotional expressiveness and reactivity-control. Developmental continuities from childhood to adulthood. In Jones, M. C., Mayle, N., MacFarlane, J. W., & Honzik, M. D. (Eds.). The course of human development. Waltham: Xerox College Publications, 1971.
- Buss, A. H., & Plomin, R. A temperament theory of personality development. New York: John Wiley & Sons, 1975.
- Buss, A. H., Plomin, R., & Willerman, L. The inheritance of temperament. Journal of Personality, 1973, 41(4), 513-524.
- Carey, W. B. A simplified method for measuring infant temperament. Journal of Pediatrics, 1970, 77, 188-194.
- Cattell, R. B. Personality: A systematic and factual study. New York: McGraw-Hill, 1951.
- Cattell, R. B., & Eber, H. W. Sixteen personality factor questionnaire "the 16 PF" manual for forms A and B. Champaign, Illinois: Institute for Personality and Ability Testing, 1962.
- Cortes, J. B., & Gott, F. M. Physique and self-description of temperament. Journal of Consulting Psychology, 1965, 29, 417-431.
- Escalona, S., & Heider, G. Prediction and outcome. New York: Basic Books, 1959.

- Freedman, D., & Keller, B. Inheritance of behavior in infants. Science, 1963, 140, 196-198.
- Gough, H. G. Manual for the California Psychological Inventory (CPI). Palo Alto, California: Consulting Psychologists Press, Inc., 1964.
- Graham, P., Rutter, M., & George, S. Temperamental characteristics as predictors of behavior disorders in children. American Journal of Orthopsychiatry, 1973, 43(3), 328-335.
- Guilford, J. P. Personality. New York: McGraw-Hill, 1959.
- Guilford, J. P., & Zimmerman, W. S. The Guilford-Zimmerman Temperament Survey manual of instructions and interpretations. Beverly Hills, California: Sheridan Supply Company, 1949.
- Johnson, R. H. Manual of the Johnson Temperament Analysis. Los Angeles: Test Bureau, 1944.
- Kagan, J., & Moss, H. A. Birth to maturity: A study in psychological development. New York: John Wiley, 1962.
- Kretchmer, E. Physique and character: An investigation of the nature of constitution and of the theory of temperament, translated by W. J. H. Sprott. New York: Harcourt, Brace and Company, Inc., 1925.
- Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.
- Plomin, R. A temperament theory of personality development: Parent-child interactions. Unpublished Ph.D. dissertation, The University of Texas at Austin, 1974.
- Scarr, S. Social introversion-extroversion as a heritable response. Child Development, 1969, 40, 823-832.
- Scholom, A. H. The relationship of infant and parent temperament to the prediction of child adjustment. Unpublished Ph.D. dissertation, Michigan State University, 1975.
- Sheldon, W. H. The varieties of temperament: A psychology of constitutional differences. New York: Harper & Brothers, 1945.
- Thomas, A., Chess, S., Birch, B. G., Hertzog, M. E., & Korn, S. Behavior individuality in early childhood. New York: New York University Press, 1963.
- Thomas, A., Chess, S., & Birch, H. G. Temperament and behavior disorders in children. New York: New York University Press, 1968.

- Thorndike, N. L. Thorndike Dimensions of Temperament manual. New York: Psychological Corporation, 1966.
- Thurstone, L. L. Examiner manual for the Thurstone Temperament Schedule. Chicago: Science Research Associates, Inc., 1953.
- Vandenberg, S. G. Hereditary factors in normal personality traits. In J. Wortis (ed.), Recent advances in biological psychiatry, vol. IX. New York: Plenum Press, 1967.
- Vandenberg, S. G. The hereditary abilities study: Hereditary components in a psychological test battery. American Journal of Human Genetics, 1962, 14, 220-237.
- Walker, R. N. Body build and behavior in young children; I. Body build and nursery school teachers' ratings. Monographs of the Society for Research in Child Development, 1962, 27(3).
- Wilson, C. D., & Lewis, M. Temperament: A developmental study in stability and change during the first four years of life. Research bulletin 74-3. Princeton, New Jersey: Educational Testing Service, 1974.

MICHIGAN STATE UNIV. LIBRARIES



31293103135913