1971

An Investigation of Interrater Reliability of the Behavior Classification Project: a Comparison of the Response of Parents, Teachers, and Children to the Behavioral Classification Project (Bcp).

William Allen Gilkey Jr
Louisiana State University and Agricultural & Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_disstheses

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_disstheses/2210

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Historical Dissertations and Theses by an authorized administrator of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
INFORMATION TO USERS

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

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

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

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

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

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

University Microfilms
300 North Zeab Road
Ann Arbor, Michigan 48106
A Xerox Education Company
GILKEY, Jr., William Allen, 1943-  
AN INVESTIGATION OF INTERRATER RELIABILITY OF THE  
BEHAVIOR CLASSIFICATION PROJECT: A COMPARISON  
OF THE RESPONSE OF PARENTS, TEACHERS, AND  
CHILDREN TO THE BCP.  

The Louisiana State University and Agricultural  
and Mechanical College, Ph.D., 1972  
Psychology, clinical  

University Microfilms, A XEROX Company, Ann Arbor, Michigan  

© 1972  

WILLIAM ALLEN GILKEY, JR.  

ALL RIGHTS RESERVED
AN INVESTIGATION OF INTRERRATER RELIABILITY OF THE
BEHAVIOR CLASSIFICATION PROJECT: A COMPARISON
OF THE RESPONSE OF PARENTS, TEACHERS, AND
CHILDREN TO THE BCP

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Psychology

by

William Allen Gilkey, Jr.
B.S., Xavier University, 1965
M.A., Louisiana State University, 1967
May, 1972
PLEASE NOTE:

Some pages may have
indistinct print.
Filmed as received.

University Microfilms, A Xerox Education Company
ACKNOWLEDGMENTS

The writer expresses his deepest appreciation to Dr. Felicia A. Pryor for her invaluable guidance and continued support and encouragement both in the completion of this study and throughout his graduate studies. He is grateful to Drs. Joseph G. Dawson, Ralph M. Dreger, A. Clinton Pereboom, and Miles E. Richardson for their time and most helpful suggestions.

The writer also wishes to thank the staff of the following schools and agencies without whose help this project could never have been completed: East Baton Rouge Parish School Board, especially Mr. Lorin Smiley; Glasgow Junior High School, Baton Rouge, Louisiana; University Laboratory School of Louisiana State University, Baton Rouge, Louisiana; Westdale Elementary School, Baton Rouge, Louisiana; Baton Rouge Mental Health Clinic, Baton Rouge, Louisiana; Consulting Center, Columbus, Indiana; Hammond Mental Health Clinic, Hammond, Louisiana; Marion County Child Guidance Clinic, Indianapolis, Indiana; Monroe County Mental Health Clinic, Bloomington, Indiana; Riley Child Guidance Clinic, Indianapolis, Indiana; Vigo County Adult and Child Guidance Center, Terre Haute, Indiana.

The writer is especially indebted to James Mackey, counselor, University Laboratory School, to Dr. Richard Snider, staff psychologist, Riley Child Guidance Clinic, and to Dr. Edward E. Cureton, Medical Director, Monroe County Mental Health Clinic for their unselfish and
continued effort in the collection of data for this project.

The writer also wishes to thank Dr. Steven Dingfelder, Mr. Paul Fredrickson, Dr. Thomas Orr, Dr. Thomas Stigall, and Mr. John Young for their assistance.

Special thanks are due to Mr. Roger Blashfield for his invaluable help.

The writer wishes to express his sincere gratitude to Mrs. Mevers for her personal interest and patience as well as for her typing of the completed dissertation.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>1</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>METHOD</td>
<td>20</td>
</tr>
<tr>
<td>Subjects</td>
<td>20</td>
</tr>
<tr>
<td>Procedure</td>
<td>22</td>
</tr>
<tr>
<td>RESULTS</td>
<td>24</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>36</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>62</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>65</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>71</td>
</tr>
<tr>
<td>VITA</td>
<td>76</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table                                      Page

I. Descriptive Data .......................... 21

II. Summary of Factor Score Correlations, Clinic Sample.. 26

III. Summary of Factor Score Correlations, Control Sample . 29

IV. Comparison of Factor Scores between Clinic and Control Groups by Respondent Category Using an Analysis of Variance Design ...................... 32

V. Mean Factor Scores for All Respondent Categories ........ 34
ABSTRACT

The purpose of this study was to investigate the question of interrater reliability as it pertains to the Behavioral Classification Project (BCP). The major question investigated was whether the people who knew a child well including the child himself could agree on what constituted the behavior he exhibited. The implications of this study are that if various groups of raters who have good opportunity to observe the child (mothers, fathers, teachers and the children themselves) agree on what behaviors are present then any of these observers may be safely substituted for another. If, however, these groups of observers do not agree; then consideration should be given to which rater provides the kind of picture of the child the users of the BCP can most realistically use.

The Ss were 60 white males 11 through 13 years of age. Ss were drawn from public school children and outpatient clinic children. The control group consisted of 29 Ss and the clinic group was composed of 31 Ss. The BCP was administered independently to the mother, to the father and to the teacher of each S as well as to the S himself.

All BCPs were scored for each of the two groups of Ss. Factor scores were computed for each S on each of the 25 BCP factors. For each of the two groups the factor scores obtained from each category of respondents were correlated using the Pearson Product Moment Correlation with those obtained from all other categories of respondents,
for example, mothers' factor scores were correlated with fathers',
mothers' scores with teachers', etc. The results indicate a moderate
level of agreement between parents for both groups of Ss. In general
the correlations between the other pairs of respondents were too low to
allow for predictive statements.

The implications of this study are that the average reliabilities between comparison groups are too low for one rater to be safely
substituted for another. Further studies with larger samples and more
representative populations would help clarify the question of inter-
rater reliability. On the basis of the present findings it appears
that mothers' ratings should not be mixed with the ratings of others.
In addition the validity of mothers' ratings should be carefully
established. The overall results further seem to point to the urgency
of developing a classification system for children's emotional dis-
orders based on behavioral data.
INTRODUCTION

Within the last decade there has been a growing dissatisfaction among mental health workers with the Standard Psychiatric Nomenclature (American Psychiatric Association, 1952) as an effective means of classifying emotional disorders. This classification system has been found to be especially deficient in the area of children's emotional disorders. Dreger, Lewis, Rich, Miller, Reid, Overlade, Taffel, and Flemming (1964) argue that, in a majority of cases, the standard nomenclature neither takes account of the many differences between adults and children nor differentiates adequately among children's emotional problems. These authors point out that 40% of the children from the 17 Florida mental health clinics which they surveyed, were diagnosed Adjustment Reaction of Childhood. Dreger, et al. correctly indicate that this categorization simply reiterates what is already known about the child, that he has a problem. Spivack and Levine (1965) also feel that use of the present classificatory system often leads to confusion and lack of precision in psychiatric diagnosis of children's emotional disorders. These authors feel that the major flaw in the current nomenclature is its emphasis upon inferred states, intervening variables, and theory to the relative exclusion of behavioral data. Other researchers who feel that a new approach to the classification of children's emotional disorders is needed include Borgatta and Fanshel (1965); Brewer (1962); Fanshel,
Hylton, and Borgatta (1963); Pritchard (1963); and Spivack and Spotts (1965).

One of the outgrowths of the dissatisfaction with standard psychiatric nomenclature was the Behavioral Classification Project (BCP) (Dreger, et al., 1964). The purpose of this project was to establish a systematic classification of children's emotional disorders based on strictly behavioral data. The authors developed a list of 229 items descriptive of children's behavior. These items were selected from many well known tests in the area of children's problems as well as from many other sources relating both to normal and abnormal children and adults. One of the most stringent criteria for the admission of an item to the BCP was that it be purely descriptive of behavior, not requiring the respondent to make high level inferences or judgments concerning the child's behavior (Dreger, et al., 1964). Most studies indicate that such a requirement for objectively defined behavioral data leads to increased reliability (Bonnardel, 1964; Katz and Cole, 1963; Lapousse and Monk, 1958; Marks, 1961; and Rutter, 1967).

Initially the list of 229 behavioral items along with 11 demographic items were administered to parents and parent surrogates of 372 clinical subjects and 90 control subjects between the ages of 6 and 13 years of age. The respondents were asked to answer yes or no to each item according to whether or not they had observed the activity in question during the last 6 months. The mean number of "yes" responses obtained from the clinical group significantly differed at
The BCP has not been the only attempt to look closely at the behavior of children, whether emotionally disturbed or normal. Ackerson (1931, 1942), in a pioneer effort, obtained ratings for over 3000 clinic children on 161 behavior traits using interview data given by mothers. He reports correlations between these traits and general personality characteristics. Jenkins and Hewitt (1944) posited the existence of three behavior syndromes through the study of child guidance clinic case records. Another early and extensive developmental investigation of childhood behavior is the study of Gesell and Amatruda (1947). Using primarily naturalistic observation techniques, they developed behavioral indices utilized in the detection of pediatric problems. The Gesell Developmental Schedule was later factor analyzed by Ball (1961) in an attempt to determine item meaning from age level to age level. Macfarlene, Allen, and Honzick (1954) in a major longitudinal study, interviewed the mothers of a large sample of normal children between the ages of 21 months and 14 years. They found that the occurrence of behavior problems in normal children in this age range is relatively high. Lapousse and Monk (1958) studied the behavioral characteristics of a carefully selected random sample of 193 Buffalo, New York children between the ages of 8 and 12 years. These authors also found that behaviors commonly regarded as pathological can be found in the repertoire of normal children. Perhaps the most comprehensive and long range research program in the area of children's personalities is that of Cattell and associates (Cattell,
1957; Cattell and Coan, 1957a; Cattell and Coan, 1957b; Cattell, Beloff, and Coan, 1958; Coan and Cattell, 1959; Peterson and Cattell, 1958; Peterson and Cattell, 1959; Peterson, 1961; and Porter and Cattell, 1960). The above work constitutes an attempt to study the development of children's personalities all the way from early school through the high school years. It has led to the development of several factor scales to measure personality characteristics in children.

Along with the BCP there have been a number of other attempts to devise a classificatory system for children's emotional disorders based on behavioral data. Some of these projects, like the BCP, depend on the observations of parent or parent surrogates for data collection. Other behavioral scales were designed more for use in inpatient settings and either professional or nonprofessional child care workers are relied upon to observe and record the behaviors to be measured. There are still other behavior scales or checklists specifically developed for use by teachers in the school situation. Although the latter usually serve as screening devices they often do attempt to classify the child according to a few general categories.

In the past decade the work of Brewer (1962) constitutes an early attempt to develop a classification system of children's emotional disorders based on the parent's observation of the child's behavior. Kasky, Krishnaick, and Azzari (1962) factor analyzed statements obtained from the parents of 199 child inpatients concerning the
presence or absence of certain behavioral symptoms. A symptom was considered present if either parent reported it present. Kasky, et al. felt that their results provided reasonably clear-cut criteria for diagnostic grouping. Mandell, Cooper, and Silberstein (1965) examined the literature on child psychopathology and constructed a 295 item child behavior rating scale, the SIBS. Ninety-six of these items were judged indicative of psychopathology by at least five of seven psychoanalysts. The 96-item SIBS was administered (Blackman and Goldstein, 1966) to the mothers of 270 children. A principal axis factor analysis resulted in 20 factors being retained for rotation. According to the authors, nine of these factors successfully discriminated among children seen for organic problems, pediatric problems, enuresis and inpatient and outpatient psychiatric diagnosis.

Another group of studies which attempt to classify on the basis of behavior use professional and nonprofessional child care workers as observers. In a study of a very specialized group, McKinney (1962) factor analyzed behavioral data in an attempt to identify dimensions of severely retarded boys. Using a time-sampling technique, McKinney and a research nurse observed 48 severely retarded boys for 24 five minute intervals over a 12-day period. A behavior checklist was used to record their observations. McKinney on the basis of the factors extracted, concluded that much of the behavior of the retarded boys was significantly related to the amount of physical and chemical restraint present in the daily environment. Pritchard
(1963) devised a rating scale of somewhat arbitrarily chosen behaviors which he felt should be observed by the ward personnel of a children's inpatient psychiatric unit. He reports adequate interrater and test-retest reliability. Patterson (1956, 1961) described a procedure for empirically constructing a classification system for children's disturbed behavior. The procedure involved the collection of referral and observational data in a clinic setting, factor analysis of this data, and the analysis of factor profiles to determine homogeneous classes. Psychologists rated 100 boys referred for diagnosis to four child guidance clinics on a large number of behavioral items. The application of factor analytic procedures resulted in the extraction of five oblique factors; hyperactivity, withdrawn, immature, aggressive, and anxious. Patterson reports good interrater reliability among four psychologist-raters. Spivack and Levine (1964) constructed a set of 68 rating scales to assess behavioral symptoms in latency age atypical children. The 68 rating scales were completed by a supervisor and houseparent for 140 children in residence at the Devereux schools. A factor analysis of the ratings resulted in 15 interpretable factors. Spivack and Levine reported interrater coefficients for factor scores ranging from .77 to .93 with a median coefficient of .83. In an attempt to refine the scale developed by Spivack and Levine (1964), Spivack and Spotts (1965) had houseparents assess 121 behavioral symptoms in 252 institutionalized latency age children. Factor analytic procedures were applied to the ratings and 20 interpretable
first order factors were extracted. The authors concluded that their results both reproduced and expanded on earlier factors found by Spivack and Levine (1964). Spivack and Spotts (1965) felt that the work of Dreger and Dreger (1962) came closest to paralleling their own. In an extension of their earlier work, Spivack and Spotts (1967) developed a 172 item behavior rating scale to be used with adolescents. The attempt was made to gather items which covered a broad spectrum of behaviors which might characterize adolescents in all diagnostic groups as well as normal individuals. Nurses, case aids, recreational staff, child care workers and houseparents rated 640 Ss sampled from residential treatment centers and child care agencies in Pennsylvania. Factor analytic procedures were applied to the ratings resulting in 18 interpretable factors. Fanshel, Hylton, and Borgatta (1963) studied behavioral characteristics in 316 children confined to residential treatment centers using a 76 item rating scale. The ratings were made by professional and nonprofessional staff who had close relationships with the children. When the ratings were subjected to a principal components factor analysis, 12 rotatable factors were extracted representing approximately 80% of the variance. The rating scale was also administered to large samples of children seen in psychiatric outpatient clinics and those assigned to foster homes (Borgatta and Fanshel, 1965; Borgatta and Cautley, 1966). Factor analysis of the ratings obtained from the latter two groups resulted in factor structures generally similar to that obtained from the residential treatment
center population. Bransford (1966) developed a 54 item rating instrument and five category classifications which were used as the basis for scoring 200 children referred to the Child Psychiatry Division of the University of Minnesota Hospital. Bransford, using a somewhat different methodology from other researchers, developed his categorization system using the apriori judgments of 21 clinicians. He found that children most often scored on several dimensions rather than falling into a single category. However, Bransford found, as have many others, that the individual training, experience, and clinical bias of his raters (second and third year psychiatry residents) resulted in unreliable ratings and categorizations.

Behavioral measures have also been developed for use by teachers. Hallworth and Morrison (1964) used teachers to rate 200 boys and girls in the first through third grades on 12 personality traits. A correlational matrix was obtained for each sex group. A principal components factor analysis resulted in six factors accounting for 69% of the variance in the boys analysis and 60% in the girls analysis. Four factors were generally comparable across both analyses. Ross, Lacy, and Parton (1965) constructed the Pittsburg Adjustment Survey Scales (PASS) for the purpose of evaluating the social behavior of elementary school age boys. Classroom teachers rated 214 boys aged 6 through 12 years on the 94 item PASS. Factor analytic procedures resulted in the extraction of four factors similar to those found by Quay and Peterson (1960) in their work with juvenile delinquents. A short 26 item behavior rating scale was developed by Rutter (1967) to be used by
teachers as a screening device for the selection of children with behavior problems. The scale discriminated between clinic and normal children between ages 9 through 13 years. No factor analysis was attempted. Polito (1967) constructed a 10 item behavior rating scale for use by teachers in evaluating a pupil's social adjustment. Teachers rated 292 sixth grade pupils from 10 randomly selected schools in a large system. All 10 items on the scale discriminated between high adjustment and low adjustment groups as defined by the California Personality Test. Spivack and Swift (1966) and Swift and Spivack (1968) studied disturbed classroom behaviors as they related to academic achievement. In the earlier work 579 kindergarten through elementary school aged children sampled from both emotionally disturbed and regular classes were rated by their teachers on an 111 item behavior rating scale. Separate factor analyses were applied to the data from each group. Fourteen factors were derived, 12 of which were common to both groups. Eleven of the factors were significantly correlated with academic achievement. The 1968 work essentially involved an extension and refinement of the scale developed by Spivack and Swift (1966). The results confirmed the scale's usefulness for the purpose developed. Perhaps the most ambitious attempt to classify problem behaviors in public school children using teachers as raters, was that of Peterson (1961). He derived a scale of 158 of the most frequent items found in the referrals to child guidance clinics. Factor analysis of teacher ratings of 831 pupils from kindergarten through the sixth grade yielded
two stable factors. The first factor related to conduct problems while the second was best described as an impulse inhibited personality factor. Related work on the study of juvenile delinquency by Peterson, Quay, and Cameron (1959); Peterson and Quay (1961); and Quay and Peterson (1960) found general neurotic, psychopathic, and subcultural factors.

Behavioral symptoms in hospitalized adult psychiatric patients were studied by Wittenborn (1952), Wittenborn, Holzberg, and Simon (1953), and Wittenborn (1962). Rating scales to assess adult behavior disorders were reviewed by Lorr (1954) and Lorr (1961). In *Syndromes of Psychosis*, Lorr, Klett, and McNair (1963) evaluated and summarized all available studies designed to isolate major psychotic syndromes using factor analytic methods. They concluded that ten major syndromes could be confirmed. Katz and Cole (1963) devised a phenomenological approach to the classification of schizophrenic disorders using close relatives of schizophrenics as informants. Recent attempts to assess adult psychiatric disorders using behavioral data include the work of Ellsworth, Kroeker, and Childers (1967); McReynolds (1968); Small, Small, and Gonzales (1965); and Vestre and Zimmerman (1969).

The above review of attempts to classify, categorize or assess emotional disorders in children on the basis of observable behavior indicates that a variety of different observers have been relied upon to rate behavior. While some scales rely on one type of observer, for example teachers, many scales typically make use of any of a number of
people who have had opportunity to observe the child. As mentioned earlier, the BCP relies upon parent or parent surrogates to indicate whether a child exhibits certain behaviors. However, Becker (1960a) felt that parents are biased by emotional involvement with their children and therefore likely to be defensive about their children's weaknesses. Cattell and Coan (1957a) concluded that to obtain objective ratings of the behavior of their sample of 6 to 8 year old children, it was necessary to rely on teacher ratings. They explained that, while parents have superior knowledge of their children, their inevitably strong biases are likely to render their ratings less useful than teacher ratings. Although both Becker and Cattell and Coan do not distinguish between parents of emotionally disturbed children and parents of normal children in their comments on parental bias, their stated preferences for teacher ratings seems most applicable to ratings of children not labeled as emotionally disturbed. Glass (1967) and Nijhawan and Nath (1967) have also testified to the efficacy of teacher ratings of the behavior of their pupils. Bransford (1966), on the other hand, argued that if one is trying to develop a psychiatric or diagnostic type of classification, it is reasonable to use persons with specific clinical training in the area. He concluded that in view of the difficulty of getting trained observers to agree on the presence of certain symptoms, it is unlikely that untrained observers will be able to do any better. It seems probable that the nature and purpose of the scale is important in determining which rater provides
the most reliable and valid information. Since a number of studies have made use of a variety of professional and nonprofessional raters of the behavior of others, it is possible to begin to get an idea of the nature of the relationship between observers.

Lapousse and Monk (1958) directed 193 children between ages 8 to 12 years and their mothers to choose from a list of behaviors those which they felt the child exhibited. Lapousse and Monk found that mothers and children agreed well on objective behavioral items such as bedwetting and thumbsucking, but agreed less well on less objective items such as restlessness and overactivity. The percentage of agreement between mother and child on specifically defined items ranged from 56 to 84 with a median of 68%. The range of agreement for implicit and nonspecific items was from 46% to 54% with a median of 53%. In a subsample of the mothers of 482 children Lapousse and Monk obtained a test-retest median reliability coefficient of .91 using the objective items and a coefficient of .78 for the less specific behavior. Becker (1960b) administered a 72 bipolar item rating scale to the mothers, fathers and morning and afternoon teachers of 64 kindergarten children. The scales were selected to sample the personality domain advanced by Cattell (1957). Factor analytic procedures applied to the ratings resulted in five interpretable factors. Becker found that the average correlation of factor scores for factors one through five was .76 between teachers, and .52 between parents. The average correlation between mothers and teachers was .31
and a .28 correlation was found when fathers and teachers were compared. Becker concluded that it is likely that different information is contained in mother and father evaluations of the same factors on the same children. He also points out that in his sample, parent and teacher ratings have very little in common. In another study Becker (1960a) confirmed his finding of poor agreement between the behavioral ratings of parents and teachers on the same children. Sarason, Davidson, Lighthall, and Waite (1960) asked the mothers and fathers of a group of elementary school children to rate their children on 16 personality traits. They reported generally low parental agreement in judgments of their children. Sarason, et al. also found that in some instances fathers are more willing to say negative things about their children than mothers. They concluded that developmental studies which rely on the ratings of mother alone may be overlooking a source of important information. However, in a recent study of a somewhat different nature, Marshall and Mowrer (1968) gave the revised men's form of the Strong Vocational Interest Blank (SVIB) to 30 male high school seniors in a college preparatory program. The SVIB was also administered to the parents of the subjects who were instructed to answer in the same way they felt their sons would answer. The median correlation coefficient between the responses of mothers and sons was .78 while the median correlation between father-son responses was .71. Mothers' responses were also correlated with fathers resulting in a median coefficient of .79. Marshall and Mowrer concluded that the
perception of both parents proved to be fairly accurate, but that mothers were slightly more accurate than fathers. It was felt that little additional information was gained by consulting both parents rather than just mother. Borgatta and Cautley (1966) using the 76 item behavior rating scale developed by Fanshel, Hylton, and Borgatta (1963) found that social workers continually rated foster children in the direction of more pathology than did foster mothers. In another study mothers were compared with teachers in regard to the way in which they perceived psychiatric disorders in a sample of 10 and 11 year old children (Rutter and Graham, 1966). They described the degree of overlap between parent and teacher perceptions as surprisingly small.

Several other researchers have compared the ratings of children's behavior by various professional and nonprofessional observers. The studies to follow have not included ratings by parents. Finney (1961) had clinicians rate three boys from interview using a multi-item rating scale. Ratings were also obtained from the boys' teachers. Finney reported only a fair degree of agreement between the clinicians and teachers. In a much larger study social workers, psychiatrists, psychologists, school teachers, nonprofessional child care workers, and administrators rated 316 children living in residential treatment centers on a 76 item behavior rating scale (Fanshel, Hylton, and Borgatta, 1963). For unexplained reasons the psychiatrists, psychologists and teachers were combined into a group called nonsocial work
professionals. The ratings by the social workers were used as the standard by which all other groups were compared. In general the non-social work professionals saw less pathological behavior than social workers particularly in the area of sexual preoccupation and self destructive activity. Although social workers noted more sexual activity than nonprofessional child care workers neither group saw more or less pathology than the other. Since the child care workers were with the child all day the fact that they did not see more behavioral problems than the social workers was somewhat unexpected. Administrators on the other hand generally indicated more pathological behavior than did social workers. The administrators frequently indicated the presence of behaviors which were directly or indirectly related to the efficient management of the institution. Small, et al. (1965) had nurses and spouses of 138 psychiatric inpatients complete the 127 item Katz Adjustment Scale (Katz and Lyerly, 1963). The scale was also completed by the patients themselves. After analyzing the data, Small, et al. concluded the relatives provided a more complete picture of the patients than did the patients themselves. In another study the behavioral adjustment of schizophrenics was rated by relatives and hospital staff in the hospital and in the community before, during, and after hospitalization (Ellsworth, Arthur, Kroecker, and Childers, 1968). The ratings made by close relatives of the patients were found to be as reliable and valid as those made by the staff. Vestre and Zimmerman (1969) also found that behavioral ratings made by relatives
close connection between teacher ratings and peer ratings. Self ratings, however, differed considerably from both teacher and peer ratings.

In one final study, Schaie (1966) had homeroom teachers and two guidance counselors rate 650 pupils enrolled in grades kindergarten through twelve on 42 bipolar traits designed to sample Cattell's 15 basic personality factors. The ratings were made on a 3 point scale. The percentage of agreement between teachers and guidance counselors on 10 randomly drawn eighth grade students was .72. This was considered acceptable. Schaie also found that girls were rated toward the more social end of the scale on all traits except dominant-submissive and adventurous-timid.

It seems clear from the above review of the literature relating to children's behavioral scales that many kinds of observers have been used to rate children using many different varieties of scales. Little attempt has been made to coordinate research in this area. Many authors, as mentioned earlier, have advanced arguments for the efficacy of certain types of observers to be relied on to provide behavioral data. The usefulness of mothers, fathers, teachers, peers, trained clinicians and nonprofessional child workers has been both upheld and disparaged by various researchers in the area. A review of the attempts to compare raters found measures of agreement ranging from poor to good depending on which raters were compared in what situations using which scales on what samples of children.
The BCP, as will be recalled, provides for the use of parent or parent surrogates to indicate the behaviors exhibited by their children. Theoretically, however, the BCP can be completed by anyone who has had sufficient opportunity to observe the child. Included in this group of knowledgeable observers would be all the groups mentioned in the previous paragraphs (mothers, fathers, teachers, etc.). Although any of these knowledgeable observers may complete the BCP, the question which arises is "Do these people provide essentially similar or different pictures of the child?"

The purpose of this study is to investigate this question of interrater reliability as it pertains to the BCP. The major question to be answered by this study is simply can the people who know a child well agree on what constitutes the behavior he exhibits. The relationship between what others say the child is like and what the child himself sees as his behavior will also be examined. The implications of this study are that if various groups of raters who have good opportunity to observe the child (mothers, fathers, teachers, and the children themselves) agree on what behaviors are present; then any of these observers may be safely substituted for another. If, however, these groups of observers do not agree; then consideration should be given to which raters or perhaps, which combination of raters, provide the kind of picture of the child the users of the BCP can best use.
METHOD

Subjects

Ss consisted of 60 white males 11 through 13 years of age. Ss were drawn from public school children and outpatient clinic children. Control Ss included 29 boys attending school at either Glasgow Elementary School or University Lab School in Baton Rouge, Louisiana. Clinic Ss consisted of 31 boys sampled from two community mental health clinics and two child guidance clinics located in central and southcentral Indiana. Since the children were required to read the items and fill out the questionnaires themselves, all Ss with IQs less than 75 were excluded. All Ss were informed of the research aspects of the study and volunteered to participate. It was necessary that all Ss have a mother and a father living in the home. It was not required that these individuals be the natural parents of the Ss.

In general the control sample can be characterized as primarily composed of bright children whose families represent the upper middle and upper socio-economic levels as measured in terms of father's occupation. The clinic Ss are more normally distributed in terms of intellectual and socio-economic levels. The classification of Ss' fathers by occupation and the mean IQ of the Ss are given in Table 1. Fathers were classified according to seven occupational levels as outlined by Hollingshead (1957) and found in Meyers and Bean (1968). The mean IQ for each group was computed from IQs obtained from both
### TABLE I
DESCRIPTIVE DATA

<table>
<thead>
<tr>
<th>Occupational Groupings</th>
<th>Clinic Sample</th>
<th>Control Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Executives, Proprietors of Large Concerns and Major Professionals</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>2. Managers, Proprietors of Medium Concerns and Minor Professions, i.e., Teachers, Engineers</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3. Administrative Personnel of Large Concerns and Semiprofessionals</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>4. Owners of Little Businesses, Clerical and Sales People</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>5. Skilled Workers</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>6. Semiskilled Workers</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>7. Unskilled Workers</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Clinic Sample</th>
<th>Control Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=30</td>
<td>N=25</td>
<td></td>
</tr>
<tr>
<td>Mean IQ Score</td>
<td>98.9</td>
<td>115.9</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>13.79</td>
<td>13.32</td>
</tr>
<tr>
<td>Mean Age Fathers</td>
<td>39.6</td>
<td>40.6</td>
</tr>
<tr>
<td>Mean Age Mothers</td>
<td>37.9</td>
<td>39.0</td>
</tr>
</tbody>
</table>
group and individually administered intelligence tests.

For the clinic group 12 Ss fall in the first three occupational groupings and 13 in the last three groupings. In comparison, 21 control Ss fall in the first three groupings and only 2 in the last three groupings. No control Ss fell in the semiskilled or unskilled categories. The mean IQ for the control group was 17 points higher than the mean IQ of the clinic group. Although the initial plan was to draw comparable samples, the two groups of Ss finally included in the project differed in several important aspects. In addition to the fact that some children were designated as children with problems and others as control children, the two groups of Ss came from different socio-economic backgrounds and lived in different geographical areas. The two groups also differed in terms of measured intelligence. The initial plan to include a black sample in the study was abandoned when only one S out of the first group of control black children contacted agreed to participate.

Procedure

The BCP was administered independently to the mother, to the father and to the teacher of each S as well as to the S himself. The standard instructions found on the BCP face sheet were used for the adults. In the case of children minor modifications of the instructions were made in order to insure that the child understood that he was to respond to the items as they applied to him. All feasible precautions were taken to insure that the various respondents had no
opportunity to collaborate on their responses. The necessity of completing the form without help from another family member was strongly stressed.

Letters were written to the parents of the control children explaining the nature of the project and asking their cooperation. A transcript of the letter used is included in the appendix. Teachers of the children who comprised the control sample were contacted at the participating school. Similarly a letter was also written to clinic parents requesting their participation. A transcript of the letter used with clinic parents is included in the appendix. In the case of the 13 boys collected at the clinic where E is employed, the letter was not used and the explanation and instructions were given verbally by the E or another staff member familiar with the details of the procedure. Teachers of the clinic children were contacted through the school social worker where one was available. Otherwise a letter was mailed to the principal of the school requesting that the S's teacher fill out a BCP. The letter sent to the principal is also included in the appendix.
RESULTS

All BCPs were scored for each of the two groups of Ss. Factor scores were computed for each S on each of the 25 factors. The factor scores obtained from each category of respondents were correlated using the Pearson Product Moment correlation with those obtained from all other categories of respondents, for example, mothers' factor scores were correlated with fathers', mothers' scores with teachers', etc. The resultant correlation matrix includes 150 individual correlation coefficients for each group. Since such a large number of coefficients are involved the level of significance was set at the .001 level. Using the .001 level of significance the probability of obtaining an individual correlation larger than .531 is .148. Correlations exceeding .531 were considered significant.

For each comparison, the correlations were averaged across all factors. Means were also computed for each comparison group using only data from the first 10 factors. The first 10 factors were used because they generally account for the major portion of variance in most factor analytic procedures. This 10-factor computation was done to insure that the inclusion in the averaging of less well defined and possibly more unreliable factors did not unduly influence the results. The resultant average correlations serve as an overall measure of the relationship between a particular pair of respondents across factors.
For each factor, correlations obtained between each of the six comparison groups (mother-father, mother-teacher, etc.) were averaged across the comparison groups. Since many of the factors either included items or consisted almost entirely of items that teachers are unable to observe, means were also obtained across comparison groups for each factor after deleting the three teacher comparison groups. For example, for factor one, a mean was computed using only coefficients from the mother-father, mother-child, and father-child comparison groups. A second mean was computed using all six comparison groups. This two-step procedure was repeated for each of the 25 factors. The resultant average correlations serve as an overall measure of agreement on a particular factor across comparison groups.

The results obtained for the clinic group are summarized in Table II. For the clinic sample, 14 correlations exceeded the cutoff point set at .531. One half of these significant coefficients are found in the mother-father comparison column. Twenty-four other correlation coefficients reached a level of .40 or higher. The average correlation for the six comparison groups range from .389 for the mother-father comparison to .129 for the teacher-child comparison. Averages taking into account only the first 10 factors, increase the means for five of the six comparison groups. The largest increase came in the mother-father correlation which increased to .528. The results indicate a moderate positive correlation overall between parents of clinic children. The average correlation for the other
TABLE II
SUMMARY OF FACTOR SCORE CORRELATIONS
CLINIC SAMPLE

<table>
<thead>
<tr>
<th>Factors</th>
<th>MF</th>
<th>MT</th>
<th>MC</th>
<th>FT</th>
<th>FC</th>
<th>TC</th>
<th>Mean All Groups</th>
<th>Mean MF,MC,FC Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.413</td>
<td>-.177</td>
<td>.146</td>
<td>-.069</td>
<td>.297</td>
<td>.106</td>
<td>.1193</td>
<td>.2853</td>
</tr>
<tr>
<td>2</td>
<td>.485</td>
<td>.471</td>
<td>.560*</td>
<td>.619*</td>
<td>.590*</td>
<td>.416</td>
<td>.5235</td>
<td>.5450</td>
</tr>
<tr>
<td>3</td>
<td>.494</td>
<td>.210</td>
<td>.052</td>
<td>.357</td>
<td>.194</td>
<td>.038</td>
<td>.2242</td>
<td>.2367</td>
</tr>
<tr>
<td>4</td>
<td>.441</td>
<td>.239</td>
<td>-.068</td>
<td>.073</td>
<td>.239</td>
<td>.011</td>
<td>.1557</td>
<td>.2040</td>
</tr>
<tr>
<td>5</td>
<td>.508</td>
<td>.325</td>
<td>-.031</td>
<td>.281</td>
<td>.032</td>
<td>.276</td>
<td>.2318</td>
<td>.1697</td>
</tr>
<tr>
<td>6</td>
<td>.672*</td>
<td>.313</td>
<td>.442</td>
<td>.444</td>
<td>.462</td>
<td>.308</td>
<td>.4602</td>
<td>.5253</td>
</tr>
<tr>
<td>7</td>
<td>.771*</td>
<td>.317</td>
<td>.472</td>
<td>.334</td>
<td>.427</td>
<td>.376</td>
<td>.4482</td>
<td>.5567</td>
</tr>
<tr>
<td>8</td>
<td>.201</td>
<td>.004</td>
<td>.115</td>
<td>.103</td>
<td>.140</td>
<td>.253</td>
<td>.1360</td>
<td>.1520</td>
</tr>
<tr>
<td>9</td>
<td>.825*</td>
<td>-.089</td>
<td>-.006</td>
<td>-.106</td>
<td>.170</td>
<td>-.097</td>
<td>.1161</td>
<td>.3297</td>
</tr>
<tr>
<td>10</td>
<td>.476</td>
<td>.150</td>
<td>.440</td>
<td>.683*</td>
<td>.338</td>
<td>.287</td>
<td>.3823</td>
<td>.4180</td>
</tr>
<tr>
<td>11</td>
<td>.365</td>
<td>-.000</td>
<td>.539*</td>
<td>-.146</td>
<td>.367</td>
<td>-.053</td>
<td>.1787</td>
<td>.4237</td>
</tr>
<tr>
<td>12</td>
<td>.259</td>
<td>.219</td>
<td>.518</td>
<td>.156</td>
<td>.425</td>
<td>-.030</td>
<td>.2578</td>
<td>.4007</td>
</tr>
<tr>
<td>13</td>
<td>.523*</td>
<td>-.065</td>
<td>.511</td>
<td>.200</td>
<td>.547*</td>
<td>.254</td>
<td>.3283</td>
<td>.5270</td>
</tr>
<tr>
<td>14</td>
<td>.183</td>
<td>.452</td>
<td>.302</td>
<td>.281</td>
<td>.150</td>
<td>.228</td>
<td>.2660</td>
<td>.2117</td>
</tr>
<tr>
<td>15</td>
<td>.266</td>
<td>.096</td>
<td>.255</td>
<td>.133</td>
<td>.213</td>
<td>-.104</td>
<td>.1432</td>
<td>.2447</td>
</tr>
<tr>
<td>17</td>
<td>.619*</td>
<td>.350</td>
<td>.484</td>
<td>.388</td>
<td>.355</td>
<td>.068</td>
<td>.3773</td>
<td>.4860</td>
</tr>
<tr>
<td>18</td>
<td>.086</td>
<td>.005</td>
<td>-.168</td>
<td>.030</td>
<td>.291</td>
<td>.187</td>
<td>.0718</td>
<td>.0697</td>
</tr>
<tr>
<td>19</td>
<td>.084</td>
<td>.020</td>
<td>.065</td>
<td>.127</td>
<td>.120</td>
<td>-.031</td>
<td>.0218</td>
<td>.0897</td>
</tr>
<tr>
<td>20</td>
<td>.185</td>
<td>-.063</td>
<td>-.180</td>
<td>.025</td>
<td>.300</td>
<td>.273</td>
<td>.0817</td>
<td>.1017</td>
</tr>
<tr>
<td>Factors</td>
<td>MF</td>
<td>MT</td>
<td>MC</td>
<td>FT</td>
<td>FC</td>
<td>TC</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>---------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td>All Groups</td>
<td>MF, MC, FC Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>.040</td>
<td>.328</td>
<td>.061</td>
<td>.430</td>
<td>.046</td>
<td>-.084</td>
<td>.1368</td>
<td>.0490</td>
</tr>
<tr>
<td>22</td>
<td>.403</td>
<td>-.067</td>
<td>.212</td>
<td>.143</td>
<td>.050</td>
<td>.250</td>
<td>.1652</td>
<td>.2217</td>
</tr>
<tr>
<td>23</td>
<td>-.006</td>
<td>-.040</td>
<td>.199</td>
<td>-.030</td>
<td>-.062</td>
<td>-.166</td>
<td>-.0175</td>
<td>.0437</td>
</tr>
<tr>
<td>24</td>
<td>.149</td>
<td>.144</td>
<td>.209</td>
<td>.161</td>
<td>.446</td>
<td>.149</td>
<td>.2098</td>
<td>.2680</td>
</tr>
<tr>
<td>25</td>
<td>.544*</td>
<td>.104</td>
<td>.434</td>
<td>-.036</td>
<td>.258</td>
<td>.053</td>
<td>.2262</td>
<td>.4120</td>
</tr>
</tbody>
</table>

Mean 25 | .3887 | .1532 | .2378 | .1880 | .2662 | .1291 |

Standard Deviation | .232 | .198 | .226 | .226 | .161 | .164 |

Mean 10 | .528 | .176 | .212 | .272 | .289 | .197 |

Standard Deviation | .176 | .195 | .228 | .256 | .160 | .163 |

* sig. .001
five comparison groups were all positive but low. There was little
difference between them.

The average correlation for the 25 factors across the six com-
parison groups range from .524 for factor 2 to -.018 for factor 23.
Seven of the 25 factors had average correlations of .30 or higher.
When the three comparisons involving teachers were deleted eleven of
the 25 factors had average correlations of .30 or higher. Average
correlations of .50 or higher were obtained on factors 2, 6, 7, and 13,
and correlations of .40 or higher were obtained on factors 10, 11, 12,
16, 17, and 25. An average correlation of .32 was obtained on factor
9. Correlations of .20 or higher were obtained on factors 1, 3, 4, 14, 15,
22, and 24, and correlations of below .20 were obtained on factors
5, 8, 18, 19, 20, 21, and 23.

The results obtained for the control group are summarized in
Table III. For the control group 16 correlation coefficients reached
the significant level. The mother-father comparison column contained
11 of these 16 significant correlations. Fourteen other correlations
reached a level of .40 or higher. The average correlation for the six
control comparison groups range from .418 for the mother-father compari-
son to a .065 for the father-teacher comparison. Averages computed
across the first 10 factors raised the average coefficient for three of
the comparison groups and lowered it for the remaining three. When
data from just the first 10 factors are considered, one negative
coefficient results. An average coefficient of -.108 was obtained for
<table>
<thead>
<tr>
<th>Factors</th>
<th>MF</th>
<th>MT</th>
<th>MC</th>
<th>FT</th>
<th>FC</th>
<th>TC</th>
<th>Mean All Groups</th>
<th>Mean MF,MC,FC Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.001</td>
<td>.223</td>
<td>.580*</td>
<td>.079</td>
<td>.176</td>
<td>.246</td>
<td>.2175</td>
<td>.2523</td>
</tr>
<tr>
<td>2</td>
<td>.707*</td>
<td>.426</td>
<td>.255</td>
<td>.578*</td>
<td>.252</td>
<td>.326</td>
<td>.4240</td>
<td>.4047</td>
</tr>
<tr>
<td>3</td>
<td>.505</td>
<td>.474</td>
<td>.119</td>
<td>.006</td>
<td>.109</td>
<td>-.120</td>
<td>.1822</td>
<td>.2710</td>
</tr>
<tr>
<td>4</td>
<td>.623*</td>
<td>.000</td>
<td>-.073</td>
<td>.000</td>
<td>-.051</td>
<td>.000</td>
<td>.0832</td>
<td>.1663</td>
</tr>
<tr>
<td>5</td>
<td>-.030</td>
<td>-.137</td>
<td>-.073</td>
<td>.088</td>
<td>-.071</td>
<td>.323</td>
<td>.0170</td>
<td>-.0573</td>
</tr>
<tr>
<td>6</td>
<td>.515</td>
<td>.207</td>
<td>-.214</td>
<td>.294</td>
<td>-.161</td>
<td>-.079</td>
<td>.0937</td>
<td>.0467</td>
</tr>
<tr>
<td>7</td>
<td>.504</td>
<td>.034</td>
<td>-.029</td>
<td>.397</td>
<td>.143</td>
<td>.329</td>
<td>.2297</td>
<td>.2060</td>
</tr>
<tr>
<td>8</td>
<td>.756*</td>
<td>.108</td>
<td>.023</td>
<td>.143</td>
<td>-.020</td>
<td>.348</td>
<td>.2263</td>
<td>.2530</td>
</tr>
<tr>
<td>9</td>
<td>-.073</td>
<td>.643*</td>
<td>-.070</td>
<td>-.037</td>
<td>-.226</td>
<td>-.116</td>
<td>.0202</td>
<td>-.1230</td>
</tr>
<tr>
<td>10</td>
<td>.571*</td>
<td>.012</td>
<td>.055</td>
<td>-.164</td>
<td>-.259</td>
<td>.088</td>
<td>.0505</td>
<td>.1223</td>
</tr>
<tr>
<td>11</td>
<td>.659*</td>
<td>-.133</td>
<td>.400</td>
<td>-.158</td>
<td>.255</td>
<td>-.151</td>
<td>.1453</td>
<td>.4380</td>
</tr>
<tr>
<td>12</td>
<td>.148</td>
<td>.037</td>
<td>.079</td>
<td>-.001</td>
<td>.231</td>
<td>.283</td>
<td>.1295</td>
<td>.1527</td>
</tr>
<tr>
<td>13</td>
<td>.602*</td>
<td>-.102</td>
<td>.013</td>
<td>.331</td>
<td>.026</td>
<td>.216</td>
<td>.1810</td>
<td>.2137</td>
</tr>
<tr>
<td>14</td>
<td>-.062</td>
<td>.000</td>
<td>-.060</td>
<td>.000</td>
<td>-.104</td>
<td>.000</td>
<td>-.0343</td>
<td>-.0753</td>
</tr>
<tr>
<td>15</td>
<td>.660*</td>
<td>-.078</td>
<td>.264</td>
<td>-.080</td>
<td>.411</td>
<td>-.141</td>
<td>.1727</td>
<td>.4450</td>
</tr>
<tr>
<td>16</td>
<td>.808*</td>
<td>-.102</td>
<td>.439</td>
<td>-.116</td>
<td>.391</td>
<td>.212</td>
<td>.2553</td>
<td>.5460</td>
</tr>
<tr>
<td>17</td>
<td>.651*</td>
<td>-.028</td>
<td>.438</td>
<td>.043</td>
<td>.214</td>
<td>.080</td>
<td>.2360</td>
<td>.4343</td>
</tr>
<tr>
<td>18</td>
<td>.454</td>
<td>-.165</td>
<td>.454</td>
<td>.148</td>
<td>.534*</td>
<td>.019</td>
<td>.2407</td>
<td>.4807</td>
</tr>
<tr>
<td>19</td>
<td>.468</td>
<td>.532*</td>
<td>-.021</td>
<td>.050</td>
<td>.305</td>
<td>.026</td>
<td>.2267</td>
<td>.2507</td>
</tr>
<tr>
<td>20</td>
<td>.276</td>
<td>.206</td>
<td>.239</td>
<td>-.145</td>
<td>.128</td>
<td>.062</td>
<td>.1277</td>
<td>.2143</td>
</tr>
<tr>
<td>Factors</td>
<td>MF</td>
<td>MT</td>
<td>MC</td>
<td>FT</td>
<td>FC</td>
<td>TC</td>
<td>Mean All Groups</td>
<td>Mean MF,MC,FC Groups</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-----------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>21</td>
<td>.834*</td>
<td>.000</td>
<td>.062</td>
<td>.000</td>
<td>.121</td>
<td>.000</td>
<td>.1695</td>
<td>.3390</td>
</tr>
<tr>
<td>22</td>
<td>-.004</td>
<td>.030</td>
<td>-.077</td>
<td>-.069</td>
<td>-.153</td>
<td>-.194</td>
<td>-.0778</td>
<td>-.0780</td>
</tr>
<tr>
<td>23</td>
<td>.593*</td>
<td>-.101</td>
<td>.377</td>
<td>.152</td>
<td>.286</td>
<td>-.060</td>
<td>.2078</td>
<td>.4187</td>
</tr>
<tr>
<td>24</td>
<td>-.123</td>
<td>-.063</td>
<td>.435</td>
<td>-.069</td>
<td>-.154</td>
<td>-.079</td>
<td>.0088</td>
<td>.0527</td>
</tr>
<tr>
<td>25</td>
<td>.401</td>
<td>.107</td>
<td>-.048</td>
<td>.163</td>
<td>-.279</td>
<td>.078</td>
<td>.0703</td>
<td>.0247</td>
</tr>
</tbody>
</table>

Mean 25  | .4178| .0852| .1416| .0653| .0842| .0699|
Standard Deviation | .301 | .218 | .220 | .178 | .221 | .167 |

Mean 10  | .4079| .1990| .0575| .138 | -.1080| .1345|
Standard Deviation | .300 | .240 | .212 | .212 | .128 | .190 |

* sig. .001
the father-child comparison. The results indicate a moderate positive correlation overall between control parents. The correlations for the other five comparison groups were extremely low.

The average correlation for the 25 factors across the six comparison groups range from .424 for factor 2 to -.078 for factor 22. The average correlation for factor 2 was the only coefficient to exceed .30. When the three comparisons involving teachers were deleted 8 of the 25 factors had average correlations of .30 or higher. A correlation of .55 was obtained on factor 16 and correlations of .40 or higher were obtained on factors 2,11,15,17,18, and 23. An average correlation of .33 was obtained on factor 21. Average correlations of .20 or higher were obtained on factors 1,3,7,8,13,19, and 20, and correlations of below .20 were obtained on factors 4,5,6,9,10,12,14,22,24, and 25.

Since the data includes mean factor scores on all factors for all groups of Ss, it is possible to compare the mean factor scores between clinic and control Ss using an analysis of variance design (ANOVA). A 3-factor design, mixed models was used with Ss nested in clinic versus controls. Clinic mothers were compared with control mothers; clinic fathers were compared with control fathers; clinic teachers were compared with control teachers and clinic children were compared with control children. The results are summarized in Table IV. All F-tests were significant at the .01 level for all four comparison groups. Thus the BCP factor scores differentiated between the clinic and control groups regardless of which pair of respondents was compared. The mean factor scores for each respondent group are presented in Table V.
## TABLE IV

**COMPARISON OF FACTOR SCORES BETWEEN CLINIC AND CONTROL GROUPS BY RESPONDENT CATEGORY USING AN ANALYSIS OF VARIANCE DESIGN**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANOVA Mothers' Factor Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinic vs. Controls</td>
<td>29218.6766</td>
<td>1</td>
<td>29218.6766</td>
<td>64.7*</td>
</tr>
<tr>
<td>Factors</td>
<td>2133019.8028</td>
<td>24</td>
<td>88875.8251</td>
<td>660.3*</td>
</tr>
<tr>
<td>Subjects</td>
<td>24388.7283</td>
<td>56</td>
<td>435.5130</td>
<td></td>
</tr>
<tr>
<td>Interaction (Clinic vs.nonclinic x factors)</td>
<td>54284.4097</td>
<td>24</td>
<td>2261.8504</td>
<td>16.8*</td>
</tr>
<tr>
<td>Subjects x factors</td>
<td>180973.5476</td>
<td>1344</td>
<td>134.6529</td>
<td></td>
</tr>
<tr>
<td><strong>ANOVA Fathers' Factor Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinic vs. Controls</td>
<td>10344.9166</td>
<td>1</td>
<td>10344.9166</td>
<td>24.5*</td>
</tr>
<tr>
<td>Factors</td>
<td>2182644.0993</td>
<td>24</td>
<td>90943.5041</td>
<td>651.9*</td>
</tr>
<tr>
<td>Subjects (Error term 1)</td>
<td>23576.8469</td>
<td>56</td>
<td>421.0151</td>
<td></td>
</tr>
<tr>
<td>Interaction (Clinic vs.nonclinic x factors)</td>
<td>24457.1007</td>
<td>24</td>
<td>1019.0459</td>
<td>7.3*</td>
</tr>
<tr>
<td>Subjects x factors</td>
<td>187447.3600</td>
<td>1344</td>
<td>139.4698</td>
<td></td>
</tr>
<tr>
<td>Error term 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ANOVA Teachers' Factor Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinic vs. Controls</td>
<td>6765.1200</td>
<td>1</td>
<td>6765.1200</td>
<td>12.7*</td>
</tr>
<tr>
<td>Factors</td>
<td>3723339.6372</td>
<td>24</td>
<td>155139.1516</td>
<td>1209.2*</td>
</tr>
<tr>
<td>Subjects (Error term 1)</td>
<td>29794.3172</td>
<td>56</td>
<td>532.0414</td>
<td></td>
</tr>
<tr>
<td>Interaction (Clinic vs.nonclinic x factors)</td>
<td>20582.6386</td>
<td>24</td>
<td>857.6099</td>
<td>6.7*</td>
</tr>
<tr>
<td>Subjects x factors</td>
<td>172494.9241</td>
<td>1344</td>
<td>128.3444</td>
<td></td>
</tr>
<tr>
<td>Error term 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE IV (Continued)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANOVA Children's Factor Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinic vs. Controls</td>
<td>6623.3110</td>
<td>1</td>
<td>6623.3110</td>
<td>12.2*</td>
</tr>
<tr>
<td>Factors</td>
<td>2351668.9959</td>
<td>24</td>
<td>97986.2082</td>
<td>700.7*</td>
</tr>
<tr>
<td>Subjects (Error term 1)</td>
<td>25723.9228</td>
<td>56</td>
<td>459.3558</td>
<td></td>
</tr>
<tr>
<td>Interaction (Clinic vs. nonclinic x factors)</td>
<td>8860.1545</td>
<td>24</td>
<td>369.1731</td>
<td>2.86*</td>
</tr>
<tr>
<td>Subjects x factors</td>
<td>187715.2496</td>
<td>1344</td>
<td>139.6691</td>
<td></td>
</tr>
<tr>
<td>Error term 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* significant .01
<table>
<thead>
<tr>
<th>Factors</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>*CIM</td>
<td>85.81</td>
<td>54.38</td>
<td>22.65</td>
<td>1.16</td>
<td>2.48</td>
<td>67.29</td>
<td>9.84</td>
<td>15.90</td>
<td>3.90</td>
<td>4.16</td>
<td>5.45</td>
<td>3.23</td>
<td>21.81</td>
<td>1.00</td>
<td>2.71</td>
</tr>
<tr>
<td>CLF</td>
<td>84.13</td>
<td>43.00</td>
<td>18.71</td>
<td>.87</td>
<td>1.74</td>
<td>60.32</td>
<td>9.68</td>
<td>14.52</td>
<td>4.09</td>
<td>4.45</td>
<td>8.45</td>
<td>3.61</td>
<td>14.58</td>
<td>1.65</td>
<td>1.48</td>
</tr>
<tr>
<td>CLT</td>
<td>45.58</td>
<td>46.90</td>
<td>6.10</td>
<td>.45</td>
<td>1.42</td>
<td>35.81</td>
<td>5.03</td>
<td>8.65</td>
<td>1.16</td>
<td>2.84</td>
<td>1.03</td>
<td>2.32</td>
<td>11.77</td>
<td>.48</td>
<td>1.90</td>
</tr>
<tr>
<td>CLC</td>
<td>89.42</td>
<td>36.87</td>
<td>21.16</td>
<td>1.35</td>
<td>1.68</td>
<td>39.38</td>
<td>6.65</td>
<td>10.16</td>
<td>5.16</td>
<td>2.94</td>
<td>6.94</td>
<td>4.90</td>
<td>16.00</td>
<td>2.48</td>
<td>2.77</td>
</tr>
<tr>
<td>**CM</td>
<td>128.55</td>
<td>25.06</td>
<td>11.00</td>
<td>.38</td>
<td>.65</td>
<td>30.66</td>
<td>2.03</td>
<td>6.79</td>
<td>2.21</td>
<td>1.59</td>
<td>5.38</td>
<td>.93</td>
<td>4.66</td>
<td>.10</td>
<td>.58</td>
</tr>
<tr>
<td>CF</td>
<td>120.58</td>
<td>25.31</td>
<td>16.14</td>
<td>.17</td>
<td>1.03</td>
<td>39.03</td>
<td>3.62</td>
<td>9.79</td>
<td>2.24</td>
<td>1.72</td>
<td>4.93</td>
<td>1.34</td>
<td>7.90</td>
<td>.62</td>
<td>.55</td>
</tr>
<tr>
<td>CT</td>
<td>33.75</td>
<td>21.17</td>
<td>3.28</td>
<td>.00</td>
<td>.27</td>
<td>17.97</td>
<td>.76</td>
<td>4.21</td>
<td>.17</td>
<td>.24</td>
<td>.59</td>
<td>.41</td>
<td>2.55</td>
<td>.00</td>
<td>.44</td>
</tr>
<tr>
<td>CC</td>
<td>94.06</td>
<td>20.17</td>
<td>11.97</td>
<td>.31</td>
<td>1.41</td>
<td>26.83</td>
<td>6.03</td>
<td>6.14</td>
<td>3.97</td>
<td>.72</td>
<td>6.83</td>
<td>2.52</td>
<td>7.34</td>
<td>.97</td>
<td>1.59</td>
</tr>
</tbody>
</table>
TABLE V (Continued)

<table>
<thead>
<tr>
<th>Factors</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>*CLM</td>
<td>9.61</td>
<td>12.52</td>
<td>7.39</td>
<td>29.90</td>
<td>1.32</td>
<td>1.13</td>
<td>4.23</td>
<td>2.45</td>
<td>2.03</td>
<td>3.48</td>
</tr>
<tr>
<td>CLF</td>
<td>8.13</td>
<td>11.58</td>
<td>4.74</td>
<td>24.74</td>
<td>1.45</td>
<td>0.58</td>
<td>3.94</td>
<td>1.71</td>
<td>2.10</td>
<td>3.10</td>
</tr>
<tr>
<td>CLT</td>
<td>4.65</td>
<td>12.16</td>
<td>4.13</td>
<td>30.93</td>
<td>2.10</td>
<td>1.06</td>
<td>3.42</td>
<td>2.09</td>
<td>1.61</td>
<td>1.87</td>
</tr>
<tr>
<td>CLC</td>
<td>8.61</td>
<td>10.58</td>
<td>6.39</td>
<td>25.19</td>
<td>2.87</td>
<td>2.65</td>
<td>5.81</td>
<td>3.13</td>
<td>6.42</td>
<td>4.45</td>
</tr>
<tr>
<td>**CM</td>
<td>3.21</td>
<td>2.52</td>
<td>3.51</td>
<td>7.31</td>
<td>0.72</td>
<td>0.34</td>
<td>0.69</td>
<td>0.48</td>
<td>0.37</td>
<td>2.45</td>
</tr>
<tr>
<td>CF</td>
<td>3.76</td>
<td>3.07</td>
<td>3.62</td>
<td>11.93</td>
<td>0.55</td>
<td>0.90</td>
<td>1.66</td>
<td>1.41</td>
<td>0.65</td>
<td>3.07</td>
</tr>
<tr>
<td>CT</td>
<td>2.38</td>
<td>1.76</td>
<td>1.07</td>
<td>9.82</td>
<td>0.41</td>
<td>0.00</td>
<td>0.66</td>
<td>0.62</td>
<td>0.14</td>
<td>1.24</td>
</tr>
<tr>
<td>CC</td>
<td>6.17</td>
<td>4.83</td>
<td>4.38</td>
<td>7.31</td>
<td>1.21</td>
<td>0.55</td>
<td>1.31</td>
<td>1.38</td>
<td>1.51</td>
<td>4.10</td>
</tr>
</tbody>
</table>

* CL = Clinic
** C = Control
DISCUSSION

In beginning the discussion of the results of this study it might be recalled that the project was carried out with males in a narrowly restricted age range. Therefore, generalizing to groups of children of different ages and sex should be done with caution. The results indicate a moderate level of agreement between parents for both groups of Ss. In general the correlations between parent-child, parent-teacher and child-teacher were too low to allow for predictive statements. Although there did appear to be a slight tendency toward greater consistency and agreement among some clinic Ss neither sample showed a higher overall level of agreement than the other.

In general the average reliabilities between comparison groups are too low for one rater to be safely substituted for another, i.e., child for mother. The data suggest that to a large degree the different raters are providing different information about the child. The results indicate that the child's behavior apparently is seen from three different points of view; that of the parents, the teachers, and the child himself. However, these points of view appear to converge in significant but unpredictable ways. For example on factor 9, an incontinence versus continence factor, the correlation between clinic parents is .825 indicating a high level of agreement. The highest correlation for the other five comparison groups on this factor is .170 and four of the five correlations carry negative signs. It seems,
therefore, that clinic parents can agree between themselves whether or not their child is continent, but do not agree with anyone else. The correlation between mother and child for example is negative. However, when the data on factor 9 is examined for normal Ss a correlation of .623 is found for the mother-teacher comparison. The correlations for the other five comparison groups are all negative. The mother-father correlation is -.073. In another example, on factor 10, a temper tantrum factor, the correlation between clinic fathers and teachers is .683. On the other hand the mother-teacher correlation for this same factor is .150. Thus, while there is little overall or general agreement between raters there are points of concurrence.

While the overall interrater reliabilities appear somewhat low they are in keeping with the results of earlier research on this problem. Becker (1960b) correlated the factor scores derived from mother and father ratings of children and obtained an average correlation of .52 across five factors. Mother and father correlations in the present study approach this level. Becker also found low correlations between parent and teacher scores. The explanation for low interrater reliabilities likely lies in the fact, noted by many authors, that mother, father and teacher generally base their ratings of children's behavior on different samples of behavior. Fathers generally have less opportunity to observe their children than mothers; teachers see children in a qualitatively different setting; and children may have difficulty being objective about their own behavior. A clinician then cannot.
obtain a BCP from a child's mother and assume he has the same or nearly the same information he would have if the instrument was filled out by the child's father or by the child himself.

In respect to the BCP the data confirmed the instrument's ability to distinguish between a clinic and nonclinic at school sample. Test, retest reliabilities as reported by Dreger (1964) are quite adequate. The low interrater reliabilities likely result from raters observing different samples of behavior as mentioned earlier, and from the fact that some BCP factors, for example factors 4, 8, 14, 15, 18, 19, 20, 21, 22 and 23 in the clinic sample have average correlations so low that they pull down the overall average. In some cases these factors seem to be defined by too few items, a fact which tends to lower reliability. In other cases the lack of interrater agreement is less easily explained. A subsequent discussion of each factor individually will help identify the less reliable factors. Further studies with larger samples and more representative populations would shed additional light on the question of interrater reliability. However, on the basis of the present findings it would seem that those who attempt to develop a classification system using the BCP need to be aware of the fact that mothers' ratings do not agree well with the ratings of other important figures in the child's life, including the child himself. The implication is that mothers' ratings should not be mixed with the ratings of others. In addition the validity of mothers' ratings should be carefully established. The overall results further
seem to point to the urgency of developing a classification system for children's emotional disorders based on behavioral data. If close observers have the difficulty indicated by this study in agreeing on the presence or absence of objectively defined behaviors, it is easy to understand the lack of reliability when classification is made based on inferred states and higher order judgments.

Before beginning an examination of interrater reliability by factor, the data found in the table of mean factor scores will be briefly discussed. The data might be interpreted to suggest a defensiveness on the part of control mothers which is not seen in clinic mothers. At least it appears that control fathers seem more willing to say negative things about their children than do control mothers. For example, in 20 of 24 factors control fathers have higher mean factor scores than control mothers indicating a greater willingness for fathers to admit to pathology in their children. Factor one was not included in the comparison since it is mainly concerned with "good" rather than problem behaviors. In addition control children have higher mean factor scores than control mothers on 18 of 24 factors also indicating a greater willingness on the part of the children to admit to engaging in problem behaviors.

However in the clinic group, fathers have higher mean factor scores than mothers on only 7 of 24 factors indicating that mothers' BCPs tend to give a picture of greater behavioral disturbance than do fathers' BCPs. One further interesting observation in the clinic sample
is that on factors 20 through 25 children's mean factor scores exceed those of both parents. These factors are made up of items which in general reflect more severe and unusual types of pathology than indicated by factors 1 through 19. Thus, it seems that clinic children are more willing to admit to behaviors which are grossly pathological than are their parents. The above observation does not hold true for the control group.

In summary an observation of mean factor scores indicates that control fathers and children seem more willing to admit to problem behaviors on the part of the children than are control mothers. However in the clinic sample mothers' mean factor scores generally tend to be higher. The most notable exception to the tendency for mothers' scores to be higher is a seemingly greater willingness on the part of clinic children to admit to more severely pathological behaviors.

As described earlier, the data may be examined in terms of agreement by factors across comparison groups. A discussion of the agreement between the comparison groups by factor should help to clarify the points at which the various raters both come together and diverge in their BCP ratings. This discussion should in addition shed light on the adequateness of an individual factor's reliability. A brief description of each factor will be given followed by a discussion of the findings.

**Factor One** is called appreciative, concerned, obedient social orientation versus unappreciative, aggressive disobedience. A high
score indicates an obedient and thoughtful child while a low score suggests a child who does not obey, follow directions, or express appreciation for kindness. For the clinic sample none of the correlations reached the significance level. A correlation of .413 was obtained between clinic parents and was the highest correlation obtained from the six comparison groups. Low negative correlations were obtained between mother-teacher and father-teacher, and low positive correlations were obtained between the two parent-child comparisons. These results indicate a tendency toward a moderate level of agreement between clinic parents. The low level of the other coefficients indicates no tendency toward agreement in the other five comparison groups.

On the other hand, the highest correlation on factor one obtained from the control group was .580 between mother-child. Low positive nonsignificant correlations were obtained between the other five comparison groups. The correlation between the mother-father comparison was .001 as opposed to the .413 obtained between the clinic parents. This correlation indicates that knowing one control parent's score on factor one does not aid in predicting the score of the other parent. The fact that control parents apparently do not agree in the ratings of their child along the obedience-disobedience dimension is quite an unexpected and unexplained finding. The moderately high correlation between the mother-child comparison on factor one may indicate that agreement between mother and child as to what constitutes socially acceptable and desirable behavior is one of the factors which differentiates nonclinic from clinic children.
An overall look at factor one indicates rather disappointingly low levels of agreement between raters. Only one correlation coefficient (mother-child comparison in the control group) reached significance. The zero correlation between control parents is difficult to explain. For factor one the average correlations across comparison groups are low and do not differentiate the clinic from the control group. Ratings obtained from mother on factor one appears to lack adequate interrater reliability.

**FACTOR TWO** is called intellectual and scholastic retardation versus alert socialized scholastic achievement. As the name suggests the items measure school performance and achievement. For the clinic sample correlations for the six comparison groups all exceed .40 and three are significant. The significant correlations were obtained from the mother-child, father-teacher and father-child comparisons and are respectively .560, .619 and .590. These findings indicate that clinic parents and their children tend to agree on the child's school performance. The tendency to agree on school performance was relatively consistent across the clinic comparison groups. The control sample showed greater variability between comparison groups. Two significant correlations were obtained. The correlation between the mother-father comparison is .707 and the father-teacher correlation is .578. The other correlations are mother-teacher, .426; mother-child, .225; father-child, .252 and teacher-child, .326. Perhaps the most notable finding is the lack of agreement between parents and children on scholastic
performance in the control group as compared to the moderately high level of agreement in the clinic sample.

An overall look at factor two indicates greater consistency between the raters in the clinic sample than between the control group raters. The average correlation across comparison groups for the clinic sample is .523 and is higher than the .424 obtained from the control sample. Factor two appears to be one of the more reliable BCP factors and scores on this factor can be used with some confidence.

**FACTOR THREE** is called disturbed sleep and dreams versus undisturbed sleep. No significant correlations were obtained from the clinic comparison groups. The mother-father correlation is .494 and does reflect a tendency on the part of clinic parents to agree. The other correlations were extremely low and nonsignificant. Much the same pattern exists in the control sample. All correlations are nonsignificant although the mother-father correlation of .505 approaches significance. Again, there is some indication of agreement between parents, but there appears to be a lack of relationship between the other comparison groups.

In general, for factor three there is a tendency toward agreement between the parents of both samples, but a lack of agreement between parents and children. Since the items making up factor three do constitute quite objectively defined behaviors, higher reliabilities might have been expected. It may simply be that some parents sleep so soundly themselves that they are just unaware of and therefore unable to rate the sleeping pattern of their children.
FACTOR FOUR is called obsessive sado-masochism. High scores reflect such behaviors as attempting suicide, fearing the loss of one's mind and threatening to kill someone. For the clinic sample no significant correlations were obtained. The mother-father coefficient is .444 indicating some tendency toward parental agreement. The correlations between the other five comparison groups are extremely low and the mother-child correlation is in the negative direction. However, in the control sample the mother-father coefficient of .623 is significant. Thus, in the control sample knowing one parent's score on factor four helps predict the score of the other parent. The correlations for the other five comparison groups are essentially zero.

Overall there is a tendency toward parental agreement on factor four with the mother-father correlation for the control groups reaching the significance level. There is no tendency toward agreement in any of the other comparison groups for either sample. The factor is identified by only four items permitting only a narrow range of possible scores. The small number of items may, thus, be partially responsible for low reliabilities in the clinic group.

FACTOR FIVE is described as self-derogating school phobia. It is made up of items related to missing school and making self-depreciating comments. No significant correlations were obtained on factor five for the clinic sample. The mother-father correlation is .508 and is the highest correlation. The mother-teacher, father-teacher and teacher-child correlations are respectively .325, .281, and .276. The
parent-child correlations approach zero. Thus in addition to the trend toward parental agreement on factor five there is a very slight tendency toward agreement between teachers and all other respondent categories. One might expect that teachers, parents, and children could agree on items relating to school attendance. The data taken from the control group provides a different picture. Five of the correlations approach zero and three of these are in the negative direction. The mother-father correlation is -.030 indicating no relationship between the responses of control parents on factor five. The only correlation approaching moderate size was a .323 obtained between teachers and children.

In general there appears to be a greater tendency toward consistency in the responses of the clinic raters to the items loading on factor five. The average correlations for the six clinic comparison groups is .232 as opposed to .083 for the control group. In summary no significant correlations were obtained on factor five indicating that the information obtained from one respondent on factor five is likely unrelated to the information obtained from another category of respondents.

**FACTOR SIX** is named disobedient, sullen, hyperactive aggressiveness. The items measure behaviors such as lying, fighting, arguing, and not getting along with peers. A remarkable degree of consistency was noted in the ratings of the clinic respondents. A significant correlation of .672 was obtained between parents' scores. For the
other five comparison groups, the mother-child, father-child and father-
teacher correlations are between .40 and .50; the mother-teacher and
teacher-child correlations are approximately .31. Thus for the clinic
group there is moderately high parental agreement coupled with a trend
toward agreement in the other five comparison groups. Less adequate
interrater reliability is found in the control sample. No correlations
reached the significance level although the mother-father correlation
is .515. However, correlations for three of the other five comparison
groups are in the negative direction. For example the correlation
between mother-child is -.214 possibly suggesting a slight tendency
toward disagreement. Thus, there are indications that parents can
agree on whether their children exhibit the behaviors measured by
factor six, but there is no agreement between the other five comparison
groups.

In summary the factor scores of clinic parents showed a moder­
ately high level of agreements. In addition, all clinic respondents
tended to slightly agree on the presence or absence of the behaviors
measured by factor six. The average correlation for the clinic com­
parison groups is .523 as opposed to a correlation of .047 for the
control group. This difference indicates greater overall consistency
in the clinic scores. Factor six appears to be one of the better de­
fined and more reliable factors of the BCP.

**FACTOR SEVEN** is named anti-social aggressiveness. The
behaviors which characterize this factor include stealing, setting
fires, damaging property and keeping bad company. A significant correlation of .771 between clinic parents indicates that clinic parents were able to agree on whether their child exhibited the above types of behaviors. A tendency toward agreement was noted in all clinic comparison groups. Mother-child and father-child correlations are .472 and .427 respectively. The three teacher related correlations are between .30 and .40. To an extent therefore all clinic respondents tend toward agreement on the presence or absence of the anti-social behaviors tapped by factor seven. Again there is less consistency found in the control group. No correlations are significant although the mother-father coefficient is .504. The father-teacher and teacher-child correlations are .397 and .329 respectively. The other three correlations approach zero. There is therefore, an indication of parental agreement and to a lesser extent a tendency for fathers and teachers to agree.

In general the results indicate that parents, especially clinic parents, can agree on the highly objective behaviors sampled by factor seven. The average correlation across the clinic comparison groups is .557 as opposed to the average correlation of .227 for the control group. Factor seven appears to provide reliable information about important aspects of children's behavior.

**FACTOR EIGHT** is named negativism versus peer-aggressive obedience to authority. This factor samples behaviors such as losing possessions, not answering when spoken to and slowness in self-grooming
activities. None of the correlations obtained from the clinic sample exceed .253. Thus, one cannot maintain that the information obtained from one category of respondents is significantly related to information obtained from the other respondent categories. However, a significant correlation of .756 was obtained between the factor scores of control parents. This finding indicates a high level of agreement between control parents on the behaviors sampled by factor eight. With the exception of the .325 correlation between teacher-child scores, the other correlations between control comparison groups approached zero.

In general on factor eight little consistency between raters is found in the clinic sample. The only significant agreement seen is between control parents. Factor eight is another factor defined by a small number of items and scores on this factor lack adequate inter-rater reliability.

**FACTOR NINE** is called sadistic incontinence versus continence. All items with significant loadings on this factor relate to continence versus incontinence with the exception of one item identifying cruelty to animals. In the clinic sample the mother-father correlation is .825 which is significant. The correlations between the other five comparison groups are minimal and nonsignificant. Four of these correlations carry negative signs. The parents of clinic children, therefore agree at a high level on whether or not their children are continent. On the other hand neither parent agrees with the child and neither children or
parents agree with teachers. The correlation between control parents is -.073. A significant correlation of .643 is found in the mother-teacher column. All other correlations are low and four correlations are negative. Mothers and teachers, then, agree on the presence or absence of continence related behaviors, but no agreement is seen between any other pair of respondents.

In summary the scores of clinic parents on factor nine are highly related. No explanation is given for the difference noted between the clinic and control samples. The high parental reliability helps make factor nine a useful factor for the clinician.

**FACTOR 10** is called temper tantrums. It is identified by the typical tantrum behaviors of screaming, shouting, kicking, cursing, and other indicators of anger. All correlations obtained from the clinic sample are positive and one is significant. The father-teacher correlation is .683. In addition the correlations between mother-father and mother-child are between .40 and .50 and, thus, indicate some tendency toward agreement. The highest level of agreement on tantrum behaviors, then, is between fathers and teachers. On the other hand the lowest agreement is between mothers and teachers (r = .150).

In the control sample a significant correlation of .571 was obtained between parents but the father-teacher correlation is -.164. Furthermore, the father-child correlation is -.259. Thus, in the control group the trend is for parental agreement coupled with a slight tendency toward disagreement between fathers and children.
In summary there is a trend toward parental agreement on tantrum behaviors in both samples. The major difference between samples is found in the father-teacher category as noted above. The average correlation across comparison groups in the clinic sample is .418 versus an average correlation of .123 for the control group which indicates greater overall consistency in the clinic sample. Factor 10 appears to have at least average interrater reliability.

**FACTOR 11** is named phobic, negativistic, finicky-eating versus positive eating. The items refer to phobic behaviors, eating habits and school refusal. For the clinic sample the mother-child correlation is a significant .539; the mother-father and father-child correlations are between .30 and .40 and the correlations between the other three comparison groups approach zero. The relevant finding seems to be that mother-child agreement exceeds parental agreement. It may be that fathers are generally not observant of their children's eating habits. In the control sample the correlation between parents is .659 indicating a moderately high level of agreement. A tendency toward agreement is again found in the mother-child category indicated by a correlation of .400. The other coefficients are low.

In summary the average correlation across comparison groups with teacher comparison groups deleted for both samples is between .40 and .45. The significant feature appears to be the tendency in both groups toward mother-child agreement. Overall interrater reliability may be considered adequate.
FACTOR 12 is called ruminative obsequiosness. The items appear to reflect behaviors such as being overly apologetic, self-conscious and unassertive. In the clinic group the highest, although nonsignificant correlations, were obtained in the mother-child and father-child comparison groups. The coefficients are .518 and .425 respectively. The other correlations including the mother-father coefficient are low. Thus, in the clinic groups there appears to be a tendency toward parent-child agreement. All correlations obtained from the control group are quite low and no trends toward agreement are noted.

Overall, if teacher related correlations are deleted, the average correlation across comparison groups in the clinic sample exceeds the average coefficient in the control sample ($r = .401$ versus $r = .153$). In the clinic sample there is a trend toward parent-child agreement but low agreement between parents. The factor is identified by few items which may partially account for the lack of significant reliability coefficients.

FACTOR 13 is named immature, neurasthenic paranoid reactions. One group of items identify behaviors such as feeling misunderstood, being without friends, and thinking others are against one. Other behaviors sampled are statements of being tired, sick, worried, and sad. Two of the comparison groups in the clinic sample have significant correlations and another correlation approaches significance. The mother-father, father-child and mother-child correlations are .523,
.547, and .511 respectively. These correlations indicate a moderate level of agreement between parents as well as agreement between parents and children. The three teacher related correlations are low. Moderately high agreement was also found between control parents (r = .602). The other correlations are too low to interpret.

In summary on factor thirteen significant correlations were obtained from mother-father groups in both samples and a trend toward parent-child agreement was noted in the clinic sample. The average correlation across comparison groups for the clinic sample with teacher correlations deleted, is .527 and exceeds the average correlation on factor thirteen obtained from the control group (average r = .214). Thus, in the clinic group good interrater reliability exists among parent and child raters.

**FACTOR 14** is called homoerotic affectation versus hysteric apprehensiveness. The items refer to behavior such as asking about sex, behaving like the opposite sex, and vomiting when worried. None of the correlations obtained from the clinic sample are significant. The highest correlation, a .452, is found between mothers and teachers and suggests a tendency toward agreement. Little relation is seen between mother-father factor scores (r = .183). In the control group all correlations approach zero indicating no tendency toward agreement between any of the pairs of raters. Factor 14 from the standpoint of interrater reliability appears to be one of the weaker factors of the BCP.
FACTOR 15 is named negativistic, aggressive sexuality. The factor is identified by such items as writes dirty words, has sexual intercourse, runs away from home, bites other children, and is bothered by skin problems. None of the correlations between clinic raters reach significance. The highest correlation, a .266 between parents, is too low to interpret. However, in the control group the mother-father correlation is .660 indicating a moderately high level of agreement. In addition, a correlation of .411 was obtained between fathers and children suggesting a tendency toward agreement.

In summary factor 15 is one of the few factors in which more consistency is noted in the control group. The average correlation across comparison groups in the control sample, after deleting teacher related correlations, is .445 versus the average correlations of .245 obtained from the clinic sample. Interrater reliability among clinic Ss is quite inadequate.

FACTOR 16 is called dirty-mindedness versus clean speech. The items sample the behaviors of cursing, uttering dirty words and smoking. In the clinic sample two significant correlations were obtained. The correlation between clinic parents is .732 indicating a high level of agreement. In addition a moderate level of agreement between mothers and teachers is indicated by a correlation of .581. A tendency toward agreement is also noted between the father-teacher group (r = .424). The other three correlations all exceed .230 but are too low to interpret. High parental agreement is also found in the
control group (r = .808). The correlation between mother-child is .439 indicating a tendency toward agreement. The other correlations are too low to interpret and two are negative coefficients.

In summary although there is somewhat greater overall consistency in the clinic sample (r = .437 versus r = .236) there is high parental agreement in both samples. It appears that reliable data concerning the "cleanness" of a child's speech is provided by factor sixteen.

**FACTOR 17** is named fearful, desurgent seclusiveness versus sociableness. In general the items appear to differentiate the dependent child who tends to withdraw from the more independent, interpersonally oriented youth. The correlation between clinic parents is .619 indicating a moderately high level of agreement. A tendency toward agreement between mothers and children is also found (r = .484). In addition the mother-teacher, father-teacher and father-child coefficients are all between .35 and .40. Good agreement is seen between control parents (r = .651). There is also evidence of a tendency toward agreement between mothers and children (r = .438). The other coefficients are too low to interpret.

In summary factor 17 like factor 16 possesses good overall parental agreement. Both samples also show a trend toward mother-child agreement. Factor 17, thus, seems to provide reliable data concerning a child's tendency to withdraw.

**FACTOR 18** is called masochistic psychoid reactions. Items reflect self-injurious behavior, self concern, excessive bathing, and
failure to stand up for one's self. None of the correlations obtained from the clinic group are significant and none are high enough to interpret. The mother-father coefficient is .086. However, in the control group the father-child correlation reaches the significance level ($r = .534$). The mother-father and mother-child correlations are both .454 indicating a tendency toward agreement. Factor 18 is therefore another of the factors in which greater consistency is seen in the control groups. Interrater reliability in the clinic sample is inadequate.

**FACTOR 19** is named verbal psychoid reactions. The items appear to reflect some difficulty and confusion in speaking and in thinking. In the clinic sample all correlations are below .150 indicating no tendency toward agreement. However, in the control group the mother-teacher correlation of .532 is significant and indicates a moderate level of agreement. A tendency toward agreement is also found between control parents ($r = .468$). The other control group correlations are too low to interpret.

In summary on factor 19 a greater consistency is again found in the control group with most agreement seen between mothers and teachers. The correlations in the clinic group are remarkably low with the correlation between clinic parents being essentially zero. Reliability is therefore inadequate.

**FACTOR 20** is called anxious organicism. Items reflect behaviors often felt to indicate neurological dysfunction. Correlations
between comparison groups in the clinic sample are all low and non-significant. The above statement also holds true for the control group. The interrater reliability on factor 20 is, therefore, extremely poor. No tendency toward agreement is found between any of the pairs of raters in either sample.

**FACTOR 21** is named sexualized psychoid organicism. This factor also includes behaviors frequently thought to indicate neurological disturbance but is also made up of several items pertaining to sexual behavior. None of the correlations between clinic comparison groups are significant. There is, however, a trend toward agreement between fathers and teachers \((r = .430)\). The mother-father correlation is .040. On the other hand the correlation between control group parents is .834 indicating a high level of agreement. The other correlations in the control group on factor 21 all approach zero.

In summary, therefore, the only significant correlation is found between control group parents. Interrater reliabilities between clinic comparison groups are inadequate.

**FACTOR 22** is called aggressive, psychoid organicism. This factor also includes behaviors frequently thought to indicate neurological disturbance. None of the correlations between clinic comparison groups are significant. There is however a tendency toward agreement between clinic parents \((r = .403)\). All correlations in the control group approximate zero and five of the six coefficients are negative. Although there seems to be some tendency for clinic parents to agree, the interrater reliability on factor 22 is inadequate.
FACTOR 23 is named clumsiness and visual problems. The items reflect behaviors such as stumbling, falling, having accidents, and difficulty in seeing. All correlations between clinic comparison groups are extremely low and nonsignificant. A moderate level of agreement between control group parents is indicated by a correlation of .593. The correlation between the other control comparison groups are too low to interpret.

In summary, the average correlation across control comparison groups after teacher related correlations are deleted is considerably higher than the average correlation in the clinic group (average $r = .419$ versus $r = .044$). Again, there is no tendency toward agreement in the clinic group.

FACTOR 24 is called organic psychosis. Items include such behaviors as hearing voices, believing one is controlled by machines, feeling crawling insects and drooling and slobbering. No clinic correlations reached the significance level although there was a tendency toward agreement in the father-child group ($r = .446$). The mother-father correlation is .149. Similar results are found in the control group. There are no significant correlations but some tendency toward agreement is seen in the mother-child group ($r = .435$). The mother-father correlation is -.123.

In summary no parental agreement is found on factor 24. There is some tendency toward parent-child agreement. Interrater reliabilities are inadequately low.
FACTOR 25 is called functional psychosis. Items indicate behaviors such as laughing at serious events, excessive perspiring, rapid eating, and excessive self interest. A significant correlation of .544 is found between clinic parents indicating a moderate level of agreement. In addition, a tendency toward agreement is seen in the mother-child group (r = .434). No significant correlations are found in the control group although the correlation between control parents is .401. The remaining coefficients are too low to interpret.

Overall the average correlation across clinic comparison groups with teacher related correlations deleted, is .412 versus an average correlation of .025 in the control. This finding indicates a greater tendency toward consistency in the clinic sample. The significant correlations between clinic parents suggests adequate reliability.

In the final analysis if the data from the clinic sample is examined after teacher related correlations are deleted some factors appear to have adequate interrater reliability. Factors 2, 6, 7, 10, 11, 12, 13, 16, 17, and 25 all have reliability coefficients between .40 and .55. The indications, therefore, are that if a clinician obtains a BCP from a mother, father or child he will have some fairly reliable information on how well the child is doing in school, whether he is aggressive, throws temper tantrums and acts out, or is fearful and withdrawn (Factors 2, 6, 7, 10, 12, 16, 17). In addition, factors 11, 13, and 25 will provide information on whether the child is obsequious, suspicious or exhibiting psychotic like behaviors.
In summary then there is evidence, as might be expected, that some BCP factors have much higher interrater reliabilities than others. This finding should be kept in mind by clinicians making use of the instrument. Furthermore, it does not appear that, in general, one rater can safely be substituted for another. The results indicate only a moderate level of agreement between parents for both groups of Ss, while correlations between the other comparison groups are too low to interpret. Although interrater reliability studies with larger samples and more representative populations are needed, it is expected that correlations between parents will not greatly exceed the .50 level. If this expectation does hold true, it means that for purposes of developing a classification system of children's emotional disorders the responses of one category of respondents, i.e., mothers should not be mixed with those of others. At present it seems that mothers' responses may provide the most useful information. Teachers observe children in structured and very particular kinds of settings. They have the opportunity to observe a somewhat narrowly restricted range of the child's behavior. It might be noted at this point that teachers simply do not have the opportunity to observe many of the behaviors described in the BCP. Better estimates of interrater reliabilities involving teachers could be obtained by deleting for purposes of comparison those items which teachers are unable to observe. Children of course are aware of the behaviors in which they engage but greater variability in children's responses might be expected. Such things as
understanding instructions and correctly reading an item are practical problems which effect results. Fathers as mentioned throughout the paper generally have a more limited opportunity than mothers to observe their children. However, it was suggested that control fathers were more willing to admit that their children engaged in pathological type behaviors than were control mothers. This result is in accord with the finding of Sarason, et al. (1960) that fathers of elementary school children are more willing to say negative things about their children than mothers. This hypothesized reluctance on the part of mothers to say negative things about their children appears to be limited to non-clinic populations. It however might be worthwhile to investigate the possibility of greater defensiveness on the part of mothers as opposed to fathers in responding to BCP items.

One method of estimating validity of information provided by a particular category of respondents is to compare that information with the responses given by a criterion group. Clinically trained professional observers, i.e., psychologists, psychiatrists, and social workers can provide the criterion measure. Items which professionals have no opportunity to observe could be deleted for purposes of comparison. Correlating the factor scores obtained by a particular category of respondents with those obtained by a clinical professional group would provide a validity measure. This type of comparison should be a part of follow-up studies.

It is assumed that the BCP itself will undergo revision as more
data is accumulated. Some factors as noted earlier appear to need better definition. Increasing the reliability of individual factors will of course lead to greater reliability, validity and usefulness of the scale as a diagnostic instrument.

Perhaps the most basic implication of the results is that emphasizing behavioral data as a basis for diagnostic classification does not automatically insure reliability, at least interrater reliability. Such reliability is a function of many factors. However, both logic and the results of numerous previous studies indicate that the requirement for objectively defined behavioral data as a basis for classification does lead to increased reliability. Furthermore, if classification can be made on the basis of behavioral data there is hope that the diagnostic process may begin to take on more meaningful implications for treatment. In addition evaluating the successfulness of the treatment process could then consist of a re-evaluation of the child's behavioral repertoire. In its present form the BCP may provide a valuable pre- and post-therapy measure.

Finally, the overall results of the present study emphasizes the urgency of developing a classification system of children's emotional disorders based on behavioral data. Although much research is needed, the BCP appears to have the potential to provide the kind of systematization which is lacking in the area of children's emotional disorders.
SUMMARY

The purpose of this study was to investigate the question of interrater reliability as it pertains to the Behavioral Classification Project (BCP). The major question investigated was whether the people who knew a child well including the child himself could agree on what constituted the behavior he exhibited. The implications of this study are that if various groups of raters who have good opportunity to observe the child (mothers, fathers, teachers and the children themselves) agree on what behaviors are present then any of these observers may be safely substituted for another. If, however, these groups of observers do not agree; then consideration should be given to which rater provides the kind of picture of the child the users of the BCP can most realistically use.

The Ss were 60 white males 11 through 13 years of age. Ss were drawn from public school children and outpatient clinic children. The control group consisted of 29 Ss and the clinic group was composed of 31 Ss. The BCP was administered independently to the mother, to the father and to the teacher of each S as well as to the S himself.

All BCPs were scored for each of the two groups of Ss. Factor scores were computed for each S on each of the 25 BCP factors. For each of the two groups the factor scores obtained from each category of respondents were correlated using the Pearson Product Moment Correlation with those obtained from all other categories of respondents,
for example, mothers' factor scores were correlated with fathers', mothers' scores with teachers', etc. The results indicate a moderate level of agreement between parents for both groups of Ss. In general the correlations between the other pairs of respondents were too low to allow for predictive statements. In addition some BCP factors have much higher interrater reliabilities than others, a finding which must be kept in mind by clinicians making use of the instrument.

The implications of this study are that in general, one rater cannot be safely substituted for another. Further studies with larger samples and more representative populations would help clarify the question of interrater reliability. However, it is expected that correlations between parents will not greatly exceed the .50 level. If this expectation holds true, it means that for purposes of developing a classification system for children's emotional disorders, the responses of one category of respondents, i.e., mothers should not be mixed with those of others. At present it seems that mothers' responses provide the most usable information. The validity of mothers' responses should be carefully established. The first step might involve comparing the responses of the various respondent categories with those of trained clinical observers. Those items which professionals have no opportunity to observe should be deleted for purposes of comparison. A highly desirable procedure is to have professionals spend time observing children in their home setting. Such a situation is obviously enormously difficult to arrange.
Perhaps the most basic implication of the results is that emphasizing behavioral data as a basis for diagnostic classification does not automatically insure reliability, at least interrater reliability. Such reliability is a function of many factors. However, both logic and the results of numerous previous studies indicate that the requirement for objectively defined behavioral data as a basis for classification does lead to increased reliability. Furthermore, if classification can be made on the basis of behavioral data there is hope that the diagnostic process may begin to take on more meaningful implications for treatment. In addition evaluating the successfulness of the treatment process could then consist of a re-evaluation of the child's behavioral repertoire. In its present form the BCP may provide a valuable pre- and post-therapy measure.

Finally, the overall results of the present study emphasizes the urgency of developing a classification system of children's emotional disorders based on behavioral data. Although much research is needed, the BCP appears to have the potential to provide the kind of systematization which is lacking in the area of children's emotional disorders.
REFERENCES


Becker, W. C. The matching of behavior rating and questionnaire personality factors. *Psychological Bulletin,* 1960, 57, 201-212. (a)

Becker, W. C. The relationship of factors in parental ratings of self and each other to the behavior of kindergarten children as rated by mothers, fathers and teacher. *Journal of Consulting Psychology,* 1960, 24, 507-527. (b)


Cattell, R. B., & Coan, R. W.  Personality factors in middle childhood as revealed in parents' ratings.  *Child Development,* 1957, 28, 439-458.  (b)


APPENDIX
I would like to take this opportunity to thank you in advance for the time and effort you have given to this project.

Sincerely yours,
Dear Mr.

We are writing to you for the purpose of enlisting your cooperation in a study to develop a questionnaire for use with families who bring their children to child guidance clinics. It is hoped that the information gained from this questionnaire will help us better understand and deal with children's problems.

If you are a parent, please read the directions on the first page of the questionnaire, and fill out the questionnaire as directed. In the blank marked "date of birth" please put your own birthdate. It is very important for the success of this study that each parent fill out their questionnaire without discussing their answers with the other parent.

If you are a child, read each item carefully and ask yourself, "Have I done this within the last 6 months?" If you have, check the column marked "true," if you have not, check the column marked "false." On the first page write your birthdate in the blank marked "date of birth."

When both parents and child have completed their questionnaires, place them in the manila envelope and return them to the secretary at your next appointment.

We would like to thank you in advance for the time you have given this study.

Sincerely yours,
Dear __________:  

We are conducting a study using some of our current clinic cases. Enclosed is a questionnaire designed for use with families who bring their children to child guidance clinics. We are requesting that this questionnaire be filled out by one of the teachers of ____________________________ who is presently being seen at this clinic.

You will note that for many of the items in the questionnaire, the teacher will not have had the opportunity to observe the behaviors in question. These items should simply be marked false. The completed questionnaire may be returned in the enclosed self-addressed, stamped envelope.

Your co-operation in this matter will be greatly appreciated.

Sincerely yours,
VITA

William Allen Gilkey, Jr. was born in Maysville, Kentucky on April 2, 1943. He graduated from Saint Patrick's High School, Maysville, Kentucky in 1961. In June, 1965 he received the Bachelor of Science degree from Xavier University in Cincinnati, Ohio. In September, 1965 he entered the Graduate School of Louisiana State University to pursue graduate studies in Psychology. Mr. Gilkey received his Master of Arts degree in Clinical Psychology in May, 1967. He served his internship in Clinical Psychology at the Indiana University Medical Center, Indianapolis, Indiana from September, 1968 until September, 1969. He completed course work for his Doctor of Philosophy degree in Clinical Psychology in May, 1970. Mr. Gilkey is presently working as Director of Psychological Services at the Monroe County Mental Health Clinic in Bloomington, Indiana and as Adjunct Assistant Professor of Psychology at Indiana University, positions which he assumed in September, 1970. Mr. Gilkey is married to the former Jacqueline Marie Rayner.
Candidate: William Allen Gilkey, Jr.

Major Field: Psychology

Title of Thesis: An Investigation of Interrater Reliability of the Behavior Classification Project: A Comparison of The Response of Parents, Teachers, and Children to The BCP

Approved:

[Signatures]

Major Professor and Chairman

Dean of the Graduate School

EXAMINING COMMITTEE:

[Signatures]

Date of Examination:

February 4, 1972