A NEW LOOK AT HUMAN FIGURE DRAWINGS: RESULTS OF A META-ANALYSIS AND DRAWING SCALE DEVELOPMENT

by

Bryan Acton

B.A., University of Saskatchewan

M.Sc., Memorial University of Newfoundland

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the department
of
Psychology

SIMON FRASER UNIVERSITY

June, 1995

APPROVAL

NAME: Bryan Vincent Acton
DEGREE: Doctor of Philosophy (Clinical Psychology)

TITLE OF THESIS: A New Look at Human Figure Drawings: Results of a Meta-Analysis and Drawing Scale Development

Examining Committee:

Chair: Dr. Elinor Ames

Marlene Moretti, Ph.D.
Senior Supervisor
Associate Professor
Department of Psychology

Patricia Kerig, Ph.D.
Assistant Professor
Department of Psychology

Ray Koopman, Ph.D.
Associate Professor
Department of Psychology

Jack Naglieri, Ph.D.
External Examiner
Professor and Coordinator
School of Psychology Program
Ohio State University

Marilyn Bowman, Ph.D.
Internal/External Examiner
Associate Professor
Department of Psychology
Simon Fraser University

Date Approved June 30, 1995

PARTIAL COPYRIGHT LICENSE

Title of Thesis/Project/Extended Essay

A New Look at Human Figure Drawings: Results of a Meta-Analysis and Drawing Figure Development

Author:

(signature)

Bryan Acton

(name)

July 4/1994

(date)

July 4, 1995

To whom it may concern,

I am including this letter with my dissertation to state that the drawing manual included in the body of the text was also written by myself and that I give permission for it to be reproduced under the same guidelines as the dissertation itself.

Sincerely Yours,

Bryan Acton

Abstract

Three studies were carried out to examine the empirical validity of individual features of human figure drawings as measures of specific forms of psychopathology. The first study reviewed 40 years of empirical research on these features. This study, unlike several comprehensive and influential past reviews, grouped study findings by construct and employed meta-analytic techniques to determine effect sizes and test their significance. The results suggest that the validity of such features may be greater than had been previously thought. The second study used the results of the previous meta-analysis to develop drawing scales for four specific constructs of psychopathology: anger/hostility, anxiety, social maladjustment, and thought disorder. Total scores from these scales were correlated with independent measures from the MMPI, the Jesness Inventory, and the WAIS-R or the WISC-R in a sample of young offenders. The drawing scales demonstrated valid patterns of convergent and discriminant validity. Observed correlations, however, were small to moderate in size. The third study was a full replication of the second using a new sample of young offenders. Only one of the scales continued to demonstrate the expected pattern of convergence. Results suggest some potential for aggregates of individual drawing features to provide valid measures of specific forms of psychopathology.

Acknowledgments

Foremost among those whose contribution toward this dissertation I would like to acknowledge is Dr. Marlene Moretti, my senior supervisor. Marlene has played many roles for me during my time as a doctoral candidate. She has supervised my research and clinical work, she has taught me both the knowledge of psychology and the knowledge of how to be a good psychologist, and she has provided guidance and support as I developed to be a psychologist. For each of these things I thank her.

My committee is also to be thanked. Dr. Patricia Kerig, like Marlene, has played numerous roles during my time at Simon Fraser University. I have appreciated both her ability to think critically and her warmth and collegial attitude. Dr. Ray Koopman's role has been restricted to my dissertation. However, I do not think the work would have gained as high a standard if it were not for invaluable discussions with him regarding correlation, regression, and significance. I thank him for his time and the insight he provided.

I would also like to thank the staff of Youth Court Services. Without them I would have no data to analyze. More than that, however, the staff provided me the opportunity to see and to work along side highly qualified and professional staff striving to meet the needs of a demanding population. I appreciate all of the opportunities that these people have afforded me. In particular, I would like to thank Roy O'Shaughnessy, Clinical Director of Youth Court Services, for supporting my research there and Dr. Nicole Aube who sparked an interest in human figure drawings and paved the way for my work and research at Youth Court Services.

I would be remiss if I did not also thank the members of my research group: Alice Bush, Julie Carswell, Nicole Fairbrother, Jocelyne Lessard, Carolyn Nesbitt, Amy Rein, and Inna Vlassev. I thank you for your patience when listening to often very dry

discussions of psychometrics, your critical insights, your assistance in practicing with my scoring manual, and, last but not least, your support and camaraderie. Most of all I would like to thank Jocelyne for twice being the second rater in the meta-analysis.

Finally, I would like to thank my family, Ursula, Christopher, Sean, and Peter. Graduate school demands a lot of a person's time. It is also stressful and can easily become one's only focus. I thank my family for abiding by me when I had little time for them, being there for me when things were stressful, and giving me perspective so that I could see the world beyond graduate school and my dissertation. Whether they know it or not, these things have likely been the greatest assistance I have had during almost a decade of graduate work. My wife Ursula deserves special thanks for her assistance as my second rater in Study 2, her comments on drafts of this dissertation, and her willingness to take on the role of a single parent when perspective was lost.

Table of Contents

List of Tables

List of Tables Cont'd

List of Tables Cont'd

List of Figures

The study of psychopathology as represented in artistic productions, such as the drawing of a human figure, has been with us for some time. Anastasi and Foley (1941), in their review of research on artistic performance in abnormal groups, found articles examining the relationship between human figure drawings and abnormality dating as far back as 1871. More recently, surveys of test use have documented the continued interest of psychologists in such drawings. Sunberg (1961) found Machover's (1949) Draw-A-Person test to rank number two among 63 objective and projective instruments surveyed, while Buck's (1948) House-Tree-Person test ranked twelfth. Piotrowski, Sherry, and Keller (1985), Wade and Baker (1977), and Wade, Baker, Morton, and Baker (1978) found instruments using human figure drawings to consistently rank among the top ten employed by psychologists. Further, use does not appear to depend on the orientation of the psychologist. Surveys indicate that employment of drawings is as high among counseling (Watkins & Campbell, 1989) and behaviorally oriented psychologists (Piotrowski & Keller, 1984), as among psychologists in general. Both of the latter groups have not been traditionally identified as projective test employers. Use of human figure drawings is common (i.e., among the top ten tests used) even in assessment areas where the possibility of litigation is high, for example, in child custody assessments (Keilin & Bloom, 1986).

Despite both a long history and high popularity, it is not clear that assessments based on human figure drawings validly predict psychopathology. Four highly influential and comprehensive reviews of the empirical literature regarding the measurement of psychopathology via human figure drawings have been carried out (i.e., Kahill, 1984; Roback, 1968; and Swensen, 1957, 1968). The consensus by these authors is that, while such instruments might have promise, studies to date had not established their validity. A similar guarded prognosis was given by Sims, Dana, and Bolton (1983), though they felt

some potential was demonstrated for the measurement of anxiety in stress induction studies. Only one review, Handler and Reyher (1965), provides a generally positive evaluation. Their review, however, was quite circumspect, assessing only those studies which measured the association between individual features of human figure drawings and independent measures of anxiety.

One approach to improving the validity of assessments based on human figure drawings has been the development of scales composed of aggregations of individual drawing features. Kahill (1984), Roback (1968), and Swensen (1957, 1968) all conclude that this research has been the most successful of any in the area. However, with a number of notable exceptions (i.e., Hiler & Nesvig, 1965; Holzberg & Wexler, 1950; Koppitz, 1966c; Naglieri & Pfeiffer, 1992; Sturner, Rothbaum, Visintainer, & Wolfer, 1980), few studies which have developed such drawing scales have also achieved high levels of validity. Further, authors such as Naglieri and Pfeiffer (1992) and Tharinger and Stark (1990) have voiced the opinion that even if valid drawing scales can be developed for human figure drawings, they are likely only to provide a general screening for psychopathology.

Even from this brief overview it is clear that much controversy exists regarding the employment of human figure drawings in the assessment of psychopathology. The present research is intended to provide some clarification in the ongoing debates surrounding the validity of assessments based on human figure drawings. The goals of the research are to respond to two broad questions. The first question is whether individual drawing features are valid as measures of psychopathology and the second is whether aggregate drawing scales can be developed which provide highly valid measurement of specific psychopathological states or conditions. The first question addresses the basic issues of the validity of human figure drawings and the utility of individual drawing features. If

individual drawing features are found to correlate with other measures of psychopathology, that support would validate the key elements proposed to link human figure drawings to various psychological ills. These findings can also inform choices regarding how best to employ such features in assessment systems designed for the interpretation of human figure drawings. The second question regards the utility of developing drawing scales composed of aggregations of individual features and asks what methods are sufficient to produce measurement systems valid enough to be used to diagnose psychopathology.

The present research was comprised of three studies. The first was a meta-analysis of over forty years of empirical research which examined the correlations between individual drawing features and independent measures of three constructs: anger/hostility, anxiety, and thought disorder. This analysis provided an assessment of the validity of these features, as well as supplying basic information regarding their psychometric potential. The second study investigated the development of four scales derived through the aggregation of individual drawing features (IDFs). The results of the previous meta-analysis were used in this study to assist in the selection of drawing features, which was expected to increase rigor and enhance the validity of the drawing scales. The third study assessed whether the results of the second study could be replicated on an independent sample. The closing discussion of the paper looks at the limitations of the present studies, examines what the studies tell us about the two broad questions this research hoped to address, proposes directions for future research, and warns of the hazards of employing any current measurement system in making inferences regarding specific forms of psychopathology from human figure drawings.

Study 1

A central issue in the study of human figure drawings is the validity of drawing features as measures of psychopathology. Numerous authors (Buck, 1948, DiLeo, 1973; Hammer, 1958; Jolles, 1971; Koppitz, 1968, 1984; Machover, 1949; Wenck, 1977) have proposed or perpetuated hypotheses relating specific features, such as the presence of a smile or the bizarreness of the figure, to pathological states and conditions. The meaning of these individual drawing features holds such an important role in this field of study that the seminal interpretive text in the area (Machover, 1949) dedicates all but 28 of 155 pages to the description and the interpretation of these features.

The validity of individual drawing features (IDFs) as measures of psychopathology, however, is quite controversial. A number of comprehensive reviews have been published (Kahill, 1984; Roback, 1968; Swensen, 1957, 1968) which have concluded that there is little empirical support for these features. Swensen (1968) concludes that IDFs have proved to be unreliable and are "not likely to provide any improvement in the clinicians' judgmental accuracy" (p. 40). Roback (1968) sums up his review with the statement, "the studies reviewed in this manuscript generally failed to support ... hypotheses" (p. 16). Kahill's (1984) findings are somewhat more favorable in that she finds IDFs to be reliable, but she too concludes that the evidence for IDFs is largely not supportive. Yet renowned authors whose publications show good understanding of the psychometric issues underlying the validation of IDFs continue to publish texts proposing largely unsubstantiated relationships between specific features and psychopathological states (e.g., Koppitz, 1984; Wenck, 1977).

Previous empirical research was reviewed once again in the present study in order to re-examine the validity of individual drawing features. The rationale for this re-examination rests on the assumption that review methods commonly employed by highly

influential past reviews may not have been capable of detecting the potential of IDFs. The introduction to the study begins with an overview of past review findings and a critique of the methods they employed. An alternative approach is then proposed for the aggregation and the analysis of past research findings.

Selection of past literature reviews. Four reviews (Kahill, 1984; Roback, 1968; and Swensen, 1957, 1968) were selected for detailed analysis. These reviews illustrate the results of previous surveys and the conclusions generally drawn regarding IDFs. Other reviews of the empirical literature on IDFs have been written, but there are good reasons for restricting the current analysis to these four papers, at least initially. First, because of their depth and breadth, these four papers have been highly influential in determining generally held beliefs regarding the validity of human figure drawings, particularly the earliest three. Each review selected a period of between 10 and 15 years and attempted to review all studies published within that time. Kahill was slightly less ambitious than the other authors and restricted her review to studies having adult samples. This limitation not withstanding, these four works cover the vast majority of studies published on human figure drawings between 1947 and 1982.

Second, all four studies employ a featural hypothesis testing approach in the aggregation and evaluation of study findings. As a result, the four reviews take a common perspective on the human figure drawing literature. This is a benefit in many ways. Combination of the findings across reviews is easier and statements about the reviews as an aggregate can be made with greater confidence. The reviews are also largely contiguous, with the exception of Swensen (1968) and Roback (1968) who cover largely the same time period. This means that there are no significant gaps in their coverage. The limitation of sharing a methodology, however, is that if that method overlooks or obscures certain findings, then important information may be lost. Given the influence of these

reviews, an alternative approach that reveals what is hidden to these methods will increase our understanding of this literature.

The featural hypothesis testing method. The key aspect of this method is its reliance on featural hypotheses. Featural hypotheses are proposed relationships between IDFs and specific psychopathological states which are based on the experience of clinicians familiar with the interpretation of human figure drawings. Kahill (1984), Roback (1968), and Swensen (1957, 1968) made use of the hypotheses of two influential theorists in the area, Machover (1949) and Hammer (1954, 1958). The decision to rely on the work of these authors was based largely on their wide acceptance by researchers and clinicians.

Collection and aggregation of research findings using the hypothesis testing method focuses on the performance of IDFs within the interpretive framework of the clinical hypotheses. The reviewer extracts the findings associated with each IDF from every study. He or she then groups these results by featural hypotheses. For example, a study may have measured three IDFs--head size, presence of eyes, and amount of shading--and correlated the scores for these features with the results of an independent scale of anxiety. The author would look at each IDF and determine whether the correlation with the measure of anxiety provided a test of one of the hypotheses proposed by either Machover (1949) or Hammer (1954, 1958). If the study result was believed to reflect a test of a hypothesis, then it would be grouped with other research findings testing this same hypothesis.

Once the relevant study findings are grouped, box-score methods are used to summarize the results. For each featural hypothesis the authors count the number of statistically significance findings. A hypothesis is deemed to be supported if all or a majority of the results have reached significance. Where results are mixed, the feature is

described as partially supported. Where all or most of the results are found to be nonsignificant, the hypothesis is said to be unsupported. Unfortunately, cut-off values used in making such decisions are seldom reported (see Swensen, 1968 for an exception).

Review findings. The level of support found for each featural hypothesis studied in the four reviews is reported in Table 1. The table is divided into four sections, one for each review. The headings across the top of the table refer to the results of the count of significant findings and correspond to the categories of support described in the previous paragraph. The "not tested" column was included to reflect those IDFs noted by the authors to have been untested at the time of their review. Featural hypotheses are listed by the name of the associated IDF.

Very few of the featural hypotheses were judged to be supported in any of the reviews. Only 1 of 41 hypotheses were supported in Swensen (1957) and only 3 of 51 in Swensen (1968). Roback (1968) found supportive evidence for only 3 of 33 hypotheses he tested. Kahill (1984) found support for only 2 of 30 hypotheses. In three of the four reviews the "not supported" column held the largest number of features.

The consistency of strong support across reviews provides yet another estimate of the validity of featural hypotheses. The performance across different reviews of hypotheses which achieved supportive evidence is reported in Table 2. The values in the table refer to the number of these hypotheses which achieved each level of support in at least one of the other three reviews. For example, the value 1 beside the row label "mixed support in at least 1 other review" and under the column heading "Swensen (1957)" indicates that one supported hypothesis from Swensen (1957) (the only such one) received mixed support in at least one of the other three reviews. Looking at the table as a whole, none of the hypotheses were found to be supported in any other review, four of the nine

hypotheses gained mixed support in at least one other review, while three were judged to be not supported in at least one other review.

Very few, if any, of the featural hypotheses proposed by Machover (1949) and Hammer (1954, 1958) were supported in the reviews of Kahill, Roback, and Swensen. Based on these findings the authors concluded that there was little evidence of the validity of such features. The current examination of the consistency of findings across the four reviews provides further evidence of the lack of support for featural hypotheses.

Limitations of the featural hypothesis testing method. One of the difficulties the three authors faced, because of their choice of review methodology, was determining the featural hypotheses tested by the specific findings within each study. This problem arose because the tests of IDFs carried out in the research often diverged from the originally proposed links between features and psychopathological states. Machover (1949) and Hammer (1954, 1958) proposed complex, idiosyncratic, and dynamic featural hypotheses which did not easily lend themselves to operationalization. Researchers, when translating these hypotheses into group comparisons or other experimental manipulations, often sacrificed the intricacies of these hypotheses for simpler formulations because of these difficulties. In so doing, the researchers were substituting their own more easily measured constructs for those proposed by the theorists. This shift away from the original hypotheses increased with time as new studies began to adopt different strategies for selecting features; for example, selection based on the performance of features in previous studies. The end result of these design-based deviations from theory was further and further movement away from the originally proposed relationships.

The greatest impact on the review process of such deviations has been on the selection and aggregation of study findings. The reviewers have been faced with the problem of having to select and group studies as tests of the original hypotheses which

few, if any, have directly assessed. Had the reviewers chosen to use only those studies which clearly tested featural hypotheses, they would have had to set aside much of the research they had collected. Instead they adopted liberal criteria, relying on their own best judgment as to whether a study constituted a test of a hypothesis. This subjective approach introduces the potential for erroneously selecting and aggregating study results which reflect a mixed bag of findings, reflecting associations between individual IDFs and numerous different constructs.

A good example of how rational selection can lead to a heterogeneous mix of study findings is Kahill's (1984) examination of the hypothesis that nudity in a human figure reflects schizoid tendencies, narcissistic self-absorption in one's own body, and sexual maladjustment. Kahill gathered together nine study findings to test the above hypothesis. Among the findings were a comparison of psychiatric subgroups on the amount of clothing drawn, correlations between nudity and a second drawing feature in samples of male felons and in a sample of job applicants, and comparisons of the rate with which nude figures were drawn in college students versus recent undergraduates, abused versus nonabused adolescents, and adolescents scoring high versus low on the Mf subscale of the MMPI. The only seeming constant is the study of nudity. The samples are quite divergent, representing everything from the mainstream (job applicants and college graduates) to psychiatric inpatients. Further, only some of these samples suggest that they may be measuring the traits of interest (i.e., schizoid tendencies, narcissism, and sexual maladjustment) and none report manipulation checks or independent evidence that these are the constructs under study. Finally, some of the comparisons are tautological. In two of the studies the authors correlated nudity with the sex of the first drawn figure, providing no independent evidence that either measures the traits of interest.

Mixing findings in the above manner risks confounding study selection with the performance of the drawing feature. The criterion for determining support for a featural hypothesis is consistently significant correlations between the IDF and independent measures across studies. Failure to achieve this criterion may be the result of poor performance of the drawing feature. This is the conclusion drawn by Kahill (1984), Roback (1968), and Swensen (1957, 1968). However, the degree to which a group of studies selected for aggregation differ from each other also contributes to inconsistent results, especially where these differences extend to the construct under investigation.

This confound is best illustrated in the case in which the aggregated findings lead to the judgment that there is conflicting evidence. Kahill's (1984) review of the IDF of nudity once again provides an excellent example. Kahill found significant results in several studies, one of which compared the rate of drawing a nude figure in abused and nonabused adolescents, another which compared psychiatric inpatients with different diagnoses, and a third which correlated nudity with the likelihood that male felons would draw a female figure first. She found nonsignificant results in an almost equal number of studies, one of which correlated nudity and the drawing of a female figure first in job applicants, another which compared the frequency of drawing a nude figure in sexual and nonsexual offenders, a third which compared that same variable in adolescents with high vs. low Mf scores on the MMPI, and a fourth which examined these rates in college graduates and compared their findings to those found in previous empirical studies. The mix of study samples aggregated by Kahill make it impossible to know whether the inconsistent results reflected the lack of validity of the IDF or differing potential of the IDF across numerous constructs.

The studies combined by the review authors also appear to be heterogeneous in terms of the age of samples and the quality of their research designs. Both of these factors

are important because they may moderate the relationship between IDFs and independent variables. Numerous authors (Goodenough, 1929; Koppitz, 1968; Thomas & Silk, 1990) have observed that there are developmental trends in the use and meaning of IDFs. An IDF which is significant and meaningful at one age may be normative at another. For example, Bieliauskas (1960) observed that only 5% of fourteen year olds produced figures which were poorly differentiated as to sex, a feature which has been hypothesized to measure maladjustment, while 35% of five year olds did so. Unless 35% of Bieliauskas' five year olds were maladjusted, this feature has very different meanings at these two ages. Study quality can have similar effects. The quality of research may determine whether a significant effect is observed or not. A poorly designed study may fail to find significance not because the association was not present, but rather because the method used did not adequately test the relationship among the measures.

A closer examination of the practice of grouping study findings by featural hypothesis would suggest that this method has the potential to bring together unrelated findings. Study findings grouped informally using featural hypotheses may in fact be a mixed bag of results reflecting the effects of different constructs and ages, and differing quality of research design. As a result, findings arising out of this method of study review are likely questionable, particularly where mixed support is observed.

An alternative method of review. An alternative approach to assessing the performance of IDFs in past research is to group studies according to the construct investigated. The potential advantage of aggregation in this manner is that it will lead to a more homogeneous collection of studies. As already mentioned, researchers began to drift away from the featural hypotheses which Kahill (1984), Roback (1968), and Swensen (1957, 1968) proposed to test. In so doing they sought more common and easily testable constructs which had agreed upon meanings and appeared robust. Rather than studying

such features as the conflict over one's hands produced by feelings of dextral inadequacy, researchers studied phenomena such as depression, anxiety, thought disorder, and behavioral problems. As research progressed, valid measures of these common phenomena were developed which enhanced the degree to which different studies measured the same construct. The end result has been increasingly larger numbers of studies assessing largely the same constructs.

Two reviews employing aggregation by construct suggest that this method does lead to a greater number of drawing features being considered valid. Handler and Reyher (1965) reviewed 51 studies assessing the relationship between IDFs and anxiety. Twenty-one IDFs were investigated in all. The authors concluded that nine of the IDFs studied were consistently supported, while a further six received mixed support. Sims, Dana, and Bolton (1983) also reviewed the literature pertaining to anxiety, choosing those studies published after Handler and Reyher. Their review was divided into those studies using stress-induction techniques and those using correlational methods. While declining to make any firm statements regarding the validity of IDFs, the authors concluded that a number of features investigated using stress-induction methods received consistent support.

A further improvement over the featural hypothesis testing method can be gained by employing meta-analytic techniques in the review process. Past reviews may have been confounded not only by the grouping of studies assessing different constructs, but also by the aggregation of studies using subjects of different ages and studies of differing quality. Meta-analytic techniques allow for the assessment of such moderating variables as subject age and study quality (Rosenthal, 1991). While such techniques do not remove or control for the effects of such variables, they allow for the assessment of their effects and the potential removal of studies which may confound the final analysis.

Meta-analytic techniques also provide improvements over box-score methods employed in previous reviews. Box-score methods rely on the number of significant findings as a metric of the strength of the combined results. This approach can lead to an undervaluing of the potential of IDFs. Individual drawing features, like any single item, tend to be less reliable than scale scores or global ratings and, as a result, show small effect sizes at best. Box-score methods can underestimate the significance of findings when effect sizes are small, particularly when the sample sizes of contributing studies are low. Lack of significant results in studies employing small samples can lead to the erroneous conclusion that findings are mixed, even when the magnitude of effect across studies is similar. Meta-analytic techniques compensate for this potential oversight by measuring the magnitude of effects as well as significance levels (Rosenthal, 1991). Further, meta-analysis can be used to test the overall significance of a finding using the combined samples from all of the contributing studies, providing for a more sensitive test of significance.

In summary, the alternative method for study review proposed herein would include aggregation of study results by construct, rather than by featural hypothesis, and employment of meta-analytic techniques for the assessment of study findings. This alternative method has two potential benefits which may lead to a stronger demonstration of the potential of IDFs as measures of psychopathology. First, this method is likely to lead to the testing of more homogeneous groups of studies. This improvement is achieved through selection by construct. Homogeneity of the sample of studies is likely also to be increased by testing for moderating variables such as sample age and study quality. Second, the employment of meta-analytic techniques will lead to more sensitive and fine grained analyses of IDF performance.

The present study. The purpose of the present study was to provide an assessment of the validity of IDFs. The study gathered results of over 40 years of empirical research on IDFs and analyzed the correlations between scores on these features and those of independent measures of specific forms of psychopathology. The study also provided a test of the review method prescribed in the previous section. If aggregation by construct using meta-analytic techniques is superior to featural hypothesis testing methods, then more IDFs should be found significant in this review than in those of Kahill (1984), Roback (1968), or Swensen (1957, 1968).

Three constructs were studied in the present review. The decision to restrict the scope of the study was largely pragmatic. There is a large literature on human figure drawings and so many constructs have been studied that to summarize all of them would have been extremely effortful. Further, many constructs have been tested in only one study and others in only a few. These small bodies of work would have tested few features and garnered little additional knowledge. Potentially more reliable and meaningful results could be derived from research areas in which numerous studies had tested a large number of drawing features. Yet another consideration was that the goal of this study was not to provide an exhaustive summary of all past research findings, but to test the potential of IDFs. This goal could be achieved by assessing a limited number of constructs. Following the rationale just presented, the constructs which received the greatest amount of research, as determined in an initial survey of the literature, were chosen for further study. These constructs were anger/hostility, anxiety, and thought disorder.

Given the nature of this study it was not possible to provide specific hypotheses for each IDF. Generally speaking, however, it was expected that a number of drawing features would demonstrate significant correlations with independent measures of the

three constructs. Further, it was expected that more features would be validated in this study than in the reviews of Kahill (1984), Roback (1968), or Swensen (1957, 1968).

Method

Study retrieval. The search for empirical studies of human figure drawings involved a survey of published studies, dissertations, and books. The Psychological Abstracts were searched from 1950 to 1992 to find research articles and dissertations. The reference sections of these works were then examined to see whether they could provide further references to as yet undiscovered articles, dissertations, or books.

Recording effect size and significance. The first step in the process of recording effect size and significance was to identify results which assessed the relationship between IDFs and independent measures of the three constructs. It became apparent early in this process that reports containing a single, unambiguous test of the above relationship made up only a small portion of the studies reviewed. The remaining studies included either multiple comparisons and/or correlations, or reported comparisons which may or may not have provided an adequate test of the relationship between IDFs and the construct of interest. A set of decision rules was developed and placed in a manual (see Appendix A) for use by raters to facilitate reliable identification and coding of study findings.

Statistics reflecting effect size (e.g., *t*, *r*, chi-square, etc.) and significance levels for the identified statistics were recorded as originally reported in each study. Significance levels were recorded as exactly as possible. For some studies this meant that a *p* value could be recorded to the second or third decimal place. For other studies, where reporting was less exact, the level of significance was recorded either as a cut-off value, e.g., $p < .05$, or simply as a statement of significance, i.e., significant or nonsignificant.

The findings and their significance levels were then converted to a common metric. This step facilitated calculations and the reporting of results. The correlation coefficient

was selected as the statistic for use when working with measures of effect size. A correlation was chosen as the effect size indicator because it is readily interpretable as a measure of the degree of relationship between two variables. The correlations were then transformed to Fisher's z_rs for the purpose of computing average effects and completing computations in the meta-analysis. Significance levels were all converted to *p* values. Conversion to such values was complicated in a number of studies by the style in which the authors reported significance. In these more problematic studies results were usually reported simply as significant or nonsignificant, with no accompanying *p* value or value for the statistic used in the study (e.g., $t = 2.98$). To ensure that raters were coding results in the same fashion and to provide a consistent method for translating such results, a set of conversion rules was developed. The conversion rules stipulated that: 1) results reported only as significant would be assigned the value of $p = .05$; and 2) results reported only as nonsignificant would be assigned the value of $p = .50$. These rules are in keeping with recommendations by Mullen (1989) and Rosenthal (1991). Values for *p* were then converted to *z* scores for the purposes of computing the overall significance of observed results and completing computations in the meta-analysis.

To ensure as much as possible that each study finding contributed independently to the meta-analysis, only a single effect size and significance level was recorded for each IDF in each study. This required that, for studies reporting more than one relevant finding, a single value would be calculated to represent the findings of each feature in each study. The recommendation from Rosenthal (1991) is to average effect sizes and significance levels in the study, using the values of z_r and *z* in the computations. However, in the present analysis the calculation of averages seemed inappropriate for some studies. Several studies reported mixed results, but gave no more information than whether a finding was significant or not. Using the decision rules proposed herein the

mean of the findings would have led to a very conservative estimate of the average correlation in these studies. Further, because of these rules, no feature with even a single nonsignificant finding could achieve the $p < .05$ level of significance. To avoid this overly conservative conversion of findings, it was decided that in cases in which studies reported results in a general manner, such as significant/nonsignificant, only the significant results would be included in determining the average effect size and significance levels.

Recording other variables. Four other variables were coded from each of the studies selected for the meta-analysis. These variables included the construct investigated, the age group of the study sample, the reliability of the IDF scoring, and the validity of the independent measure(s). The last three variables were assumed to be moderator variables, as defined by Rosenthal (1991).

The first task of the raters was to determine whether the construct investigated was anger/hostility, anxiety, thought disorder, or other. A study was deemed to measure a particular construct if: 1) the author(s) stated that the study was of that construct; 2) the authors provided evidence that the measures used assessed the construct; and/or 3) the sample of the study suggested that the construct was being assessed. As an example of the latter condition, if a study compared the responding of student nurses and schizophrenics, it would be deemed a study of thought disorder.

The raters then determined whether the majority of the study sample fell in one of three age groups: children, adolescents, or adults. The age brackets for the three categories were: for children, four to twelve years of age; for adolescents, 13 to 18 years of age; and, for adults, 19 years of age or older.

The raters then assessed whether the scoring of features was reliable or not. Scoring was deemed to be reliable if an average interrater correlation of .70 or greater was observed, or if the average agreement between raters was 80% or greater. If interrater

agreement fell below this level, scoring was classified as unreliable. Two kinds of features were assumed to be very high in reliability. These were mechanical measures, such as height, width, placement on page, etc., and sex of figure drawn, if drawn by a subject 10 years of age or older. The latter rule was developed based on empirical evidence that sex of the figure is accurately and reliably coded after this age (Thomas & Silk, 1990).

Reliability and validity were initially recorded separately for independent measures because it was thought that some of the studies reviewed would provide evidence of reliability without evidence of validity. For example, a study which used psychiatric diagnosis as the independent measure may have included an assessment of reliability in the form of a measure of agreement between two separate diagnosticians and yet give no indication of the validity of the diagnosis. However, no study uncovered in the present review provided evidence of reliability without evidence of validity. As a result, coding of this variable was made based on evidence of validity alone. An independent measure was deemed valid if: 1) the measure was a well known measure of the construct being investigated; 2) the measure was a well known diagnostic interview used to assess the construct; or 3) the study referenced independent research which had confirmed the validity of the scale/diagnostic tool used.

The rules for scoring these four variables and additional details useful in making scoring judgments were described in the coding manual developed for this study (see Appendix A).

Reliability of study coding. A subset of sixteen studies was selected for the reliability analysis. Twelve of these studies were chosen from those used in the meta-analysis, one set of four for each of the three constructs being investigated. Each set of studies was randomly chosen from among the pool of studies selected by the first rater as testing each construct. A further four studies were chosen randomly from all the empirical

studies not included in the meta-analysis. A second rater then coded the sixteen studies following the study coding manual.

Interrater reliability was assessed using product-moment correlations and kappa coefficients for the four variables of construct studied, age of sample, IDF scoring reliability, and validity of the independent measure. Such analyses could not be used to determine the reliability of the coding of effect size and level of significance because there was no meaningful scale on which to compare certain differences in coding between the two raters. This is best exemplified by looking at the kinds of differences which could occur between raters. One difference could occur if both raters located the same information but read or recorded the information incorrectly. As an example, both reviewers could have found a t-test, but one recorded the finding as $t = 3.58$ and the other as $t = 8.58$. Standard methods of correlation could have been applied to data including such an error, but the meaning of the resulting coefficient would not have been clear. This is because the magnitude of the difference between the two results is in some ways unrelated to the degree of error in recording. The same error, misrecording one digit by a value of five units, could have likely led to the writing of $t = 3.53$. In both cases only a single error is made, but the results of a correlational analysis would have been markedly changed. An even more problematic difference is the recording of results from different analyses or comparisons within a study. Where one rater records one result and the second yet another, it is not clear what scale should be used to compare the two. For example, the first rater could record the result of a comparison between a normal and a pathological group while the other records findings of a comparison between two pathological groups. In both cases described above the problem is that the recording of the results is not along a common scale. The only commonality held by the two raters is whether they agree or not. Any other scale could grossly inflate or artificially erase

differences between raters for reasons other than lack of reliability in coding. Reliability for the coding of these two variables, therefore, has been reported as the percentage of times the first and second rater selected the same results from the studies.

Analysis of IDFs. Individual drawing features were analyzed by blocks of study findings defined by the IDF used and the construct investigated. For example, if shading was the IDF and anxiety and thought disorder were the constructs, then there would be two blocks of studies: those using shading and investigating anxiety, and those using shading and investigating thought disorder. When the study findings were grouped in this way it became apparent that the blocks consisted of only one or two studies for many of the IDFs. Blocks consisting of one finding were treated as single study estimates of the validity of the individual drawing feature.

Analysis progressed through a series of stages for those features having two or more relevant study findings, as recommended by Rosenthal (1991). The first step in each analysis was to determine whether the results gathered for the feature were sufficiently similar in magnitude of effect and level of significance to be combined in subsequent analyses. The homogeneity of the results was assessed by diffuse tests. Diffuse tests for effect size were calculated using unweighted Fisher's z_r, while those for level of significance employed unweighted z scores.

When the effect sizes and significance levels of a block of study findings were determined to be homogeneous (i.e., the diffuse tests were nonsignificant), they were combined to provide an estimate of the overall effect size. The overall effect size has been suggested to be somewhat akin to the true score for the population of study findings (Bangert-Drowns, 1986). Combination of significance levels provided an indicator of the overall significance of the observed relationship between a feature and independent measures of a construct (Rosenthal, 1991). Combination of effect sizes and significance

levels was carried out using, respectively, unweighted z_rs and unweighted z scores. Results were then retransformed to product-moment correlations and p values for ease of interpretation.

Blocks of studies which were found to be heterogeneous (i.e., one of the diffuse tests was significant), were subsequently analyzed to test for the effects of specific moderators using focused test procedures. The order of testing of the three moderator variables followed the plan outlined below. Figure 1 provides a flow chart depiction of the progress of testing for moderator effects. The age of subjects was tested first. If this moderator variable proved nonsignificant, the analysis continued with a test of the effects of the next variable--reliability of feature scoring. If the effects of this moderator variable were found to be significant, the block of studies was subdivided along sample lines. For example, the block of studies could be broken into child and adult groupings or sub-blocks. Diffuse tests were then carried out on sub-blocks to determine whether the new groups of studies were homogeneous. If the studies were homogeneous, and the sub-block was of interest, the findings would be combined at that point. If not, the effects of the next moderator variable would be tested. Analyses continued until all moderator variables were exhausted or a homogeneous sub-block of study findings was found.

Selection of sub-blocks for interpretation. Not all sub-blocks of study findings were expected to produce results which would be useful for inferring the validity of IDFs. For instance, a sub-block containing poor quality studies assessing the relationship between anxiety and shading was considered less desirable, for interpretive purposes, than a sub-block of good quality studies. Thus, the only sub-blocks interpreted were those which evidenced no differential effects of study quality, or were comprised of high quality studies. Similarly, because the sample to be employed in the second study was made up of adolescents, when age of sample was a factor, interpretation of sub-blocks of studies

assessing adolescents was preferred. However, when effects for the more desirable sub-blocks could not be found (i.e., because there was no block of studies meeting the criteria), results of the remaining sub-blocks were used.

Results

This results section is divided into three parts. The first examines the reliability with which the raw material for the meta-analysis was extracted from the component studies. Reliability was determined for recordings of significance levels and effect sizes, and for judgments of the construct investigated, the age of sample, the reliability of drawing feature scoring, and the validity of the independent measures. The second section describes the studies located for the meta-analysis. The third section provides the results of the meta-analysis for the three constructs of anger/hostility, anxiety, and thought disorder.

Reliability of study coding. Product-moment correlations, kappa coefficients, and percent agreement are reported in Table 3 for the four variables of construct studied: age, reliability of feature scoring, and validity of the independent measure. Percent agreement between the first and the second rater for recordings of effect size and significance level are also reported in Table 3. Product-moment correlations measured the degree to which the ratings of the two coders varied together. Correlations between ratings of the first and the second rater for the three moderator variables and construct investigated ranged from .75 to 1.00, with three of four variables falling above .85. Kappa coefficients measured the agreement between two raters (Cohen, 1960, 1968). Once again the magnitude of the resulting coefficients was high, with no coefficient falling below .75 and with three out of four values falling at or above .85. These results suggest that scoring of these four variables was reliable.

The agreement observed between the two raters for effect size and level of significance was adequate, with 88% agreement for both variables. This level of percentage agreement was roughly in keeping with that observed for the other four variables.

Study sample. Listed in Table 4 are the 36 studies located for the meta-analysis. Important descriptors of these studies are also reported in Table 4, including: the type of source (i.e., dissertation versus journal article); the construct studied, the sample characteristics, the design of the study, the evidence for the validity of the independent measure or measures, the number of features studied, and the reliabity with which the drawing features were scored.

Eight studies were deemed to assess anger/hostility, 20 studies to assess anxiety, and 12 to assess thought disorder. These studies were chosen from a much larger pool of empirical studies of human figure drawings measuring a variety of constructs ranging from alcoholism to sexuality.

Examination of the results presented in Table 4 indicates that the distribution of the different levels of the moderator variables varied widely across the three groups of studies. The percent of studies assessing children, adolescents, and adults, respectively, were 13%, 50%, and 27% for anger/hostility studies, and 20%, 15%, and 65% for anxiety studies. Adult subjects were employed in 100% of the studies of thought disorder. Reliability of feature scoring was also somewhat variable, with reliable scoring reported in 100% of the anger/hostility studies, 65% of the anxiety studies, and 75% of the thought disorder studies. Validity of independent measures was most variable, with valid measures employed in 63% of anger/hostility studies, 75% of anxiety studies, and 17% of thought disorder studies.

Results of the meta-analysis. Results of the complete analyses for each IDF can be found in Appendix B grouped according to the construct investigated. Summary findings from the meta-analysis are reported in Table 5. The top-most entries in the table indicate the number of features for which only a single study result could be found and total approximately two thirds of the findings. The proportion of features whose results are based on a single study varied widely across constructs, from approximately 33% of the findings for anger/hostility, to nearly 50% for anxiety, and almost 90% for thought disorder.

The number of features found to be homogeneous is indicated in the second row. Homogeneous study findings were those found not to have significant differences among their effect sizes or significance levels across studies in the diffuse tests. (Note that the IDF results based on single studies are not included in these figures.) Findings were combined at this point in the analysis to determine the average magnitude of effect and the overall significance of the relationship with the independent measures. The majority of the study findings were determined to be homogeneous for all three constructs. Those findings determined to be nonhomogeneous (i.e., with at least one significant diffuse test) contained study findings which differed significantly. These findings were subsequently tested for the moderating effects of age of sample, reliability of feature scoring, and validity of the independent measure.

The number of features which demonstrated significant correlations with independent measures at some point during the analysis are reported in the final row. Results are reported separately for findings based on a single finding and those based on more than one. One hundred and eight features correlated significantly with independent measures. This corresponds to approximately 28% of all features tested. The percentage of significant correlations varied across constructs from a low of 21% for IDFs assessing

anger/hostility to a high of 37% for IDFs assessing anxiety. Of the significant findings, two anger/hostility features, six anxiety features, and one thought disorder feature achieved significance only after being tested for effects of moderator variables. Fourteen features, enumerated under the heading "untestable", were significantly heterogeneous and not explainable using the moderating variables. None of these were accepted for further analysis.

The average effect sizes for the three constructs were very small for the study as a whole. Average correlations of .10, .16, and .07 [1] were found for anger/hostility, anxiety, and thought disorder features, respectively. When only significant features were used in calculating the mean, the average correlations grew to .24, .35, and .25, respectively. The latter correlations, while larger in magnitude than their whole study counterparts, are still considered small.

Discussion

The present study found numerous features that correlated significantly with independent measures of anger/hostility, anxiety, and thought disorder. The numbers of significant features are much higher than those reported in Kahill (1984), Roback (1968), or Swensen (1957, 1968), both absolutely and as a proportion of the total number of features studied. The largest number of significant features reported in any one of these previous reviews was three, and the lowest number was one. These values fall below the range of 15 to 56 significant features observed when each construct is considered separately and lie far below the study total of 108 significant effects. Similarly, the largest proportion of results attributable to significant findings in any of the four reviews was 3 of 20 findings, or 15%. The range across the three constructs investigated in the present study was from 22% to 37%, with an average of 28%.

In comparison with previous reviews which focused on constructs rather than featural hypotheses, the proportion of validated features is about the same. Handler and Reyher (1965) determined that 9 of 21 anxiety features (approximately 43 percent of those studied) were consistently supported. Sims, Dana, and Bolten (1983) did not give exact numbers in their study; however, they suggested that many of the features they reviewed as measures of anxiety achieved consistent support. The results of the present research appear to fall between these two studies.

The present results indicate that IDFs have greater validity than is suggested by the reviews of Kahill (1984), Roback (1968), and Swensen (1957, 1968). These findings do not contradict the conclusion of these past reviewers, however, but augment them. In choosing an alternative review method, the present study looked at the human figure drawing literature from a different perspective. Kahill, Roback, and Swensen examined the validity of hypotheses regarding IDFs. The present study examined the potential of IDFs to measure three relatively clearly defined constructs. Each form of review provides slightly different information and together they provide a more complete picture of the validity of IDFs. Previous conclusions were that there was little support for the validity of IDFs. With the inclusion of the present findings, the conclusion becomes there is little support for featural hypotheses, but there is some support for the validity of IDFs when assessing robust constructs, such as anger/hostility, anxiety, and thought disorder.

The present study also adds to the work of previous reviewers by reinforcing the utility of a construct-focused approach. Results of the present analysis suggest that a more profitable means of reviewing past research on IDFs is to group and test study findings according to the construct investigated, rather than by drawing feature hypotheses.

Several important qualifications need to be made regarding the present analysis. First, some of the features may have been incorrectly identified as validly correlating with independent measures of the three constructs. In all, 382 correlations or average correlations were examined in the meta-analysis. The large number of analyses carried out make it very likely that at least some of the observed significant effects were a product of chance. This possibility is mitigated when a number of study findings have been collected regarding a drawing feature, as the repeated significance increase the likelihood that the observed effect is valid. Where single studies are the basis for conclusions, however, there is no evidence of replicability, and confidence in the veracity of findings should be low.

A correction for the number of correlations could have been used to control for chance effects. This could have been accomplished by setting a stricter value of p, requiring larger effects before significance would be said to have been achieved. There were several reasons for not doing so in the present analysis. First, a number of studies included in this analysis reported findings only in terms of significance (i.e., significant or nonsignificant). In transforming these results significance was set at $p = .05$. A family-wise error correction would have unreasonably excluded these results or any group of findings aggregating like results, not based on the true performance of the feature, but based on the decision rule developed for the study. Second, no such correction was used in previous reviews of this literature. Results of the present study were more comparable to past findings because similar criteria were employed in interpreting the results of the review. Finally, the study itself was exploratory in nature and sought to discover how many features would be found significant if results were organized according to construct rather than according to figure drawing hypotheses.

It is also possible that some features were not identified which, in truth, correlate with independent measures of the constructs studied. Once again the small number of

study findings for many of the features may have been a factor. Chance factors underlying spurious results can lead to erroneously nonsignificant results, as well as significant ones. Other aspects of the study may also have played a role in the under-identification of valid IDFs. One of these aspects may have been conservative decision-making. One decision was to substitute $p = .50$ for all study findings reported simply as nonsignificant. This is an extremely conservative approach to study coding and has likely led to the over-exclusion of drawing features by artificially lowering the overall significance of findings for some features. This rule could also have reduced the number of selected IDFs because it contributed to the formation of a greater number of unresolvable significant diffuse tests. IDF findings which were found to be nonhomogeneous (i.e., which produced a significant diffuse test) and whose nonhomogeneity could not be explained by the moderating variables used in the study were excluded from further analysis. The decision to assess the effects of moderator variables if the diffuse test for either effect size or statistical significance was significant may also have led to the rejection of an increased number of IDFs, because the moderator variable analyses did not explain the differences between study findings.

Finally, while numerous IDFs in the present study have shown potential to measure the three constructs investigated, the magnitude of observed correlations would suggest that few could provide highly valid measures of these constructs. Correlations determined to be significant in the present study ranged in magnitude from .07 to .63, with an average of .28. While correlations of this size are sufficient to achieve significance, they are not large enough to reliably do so in other studies. Further, larger correlations with independent criteria would be desired before a measure would be deemed to be highly valid. As an example, validity coefficients obtained by correlating self-report inventories of anxiety are routinely in the range of .60 to .80 (Clark & Watson, 1991; Watson &

Clark, 1984). For an IDF to be considered a good test of anxiety, therefore, it too would have to demonstrate equally high correlations with independent assessments of anxiety. The low correlations obtained in the present study suggest that most single drawing features have weak validity and, therefore, lack the necessary psychometric qualities to function as independent measures.

The effects of small samples and weak correlations may be mitigated by the use to which the present analysis is put. Most of the qualifications expressed above are relevant when the results of the meta-analysis are employed to support the use of IDFs as independent tests. For the reasons provided above and several others, to be discussed in study 2, drawing features do not make good tests. An alternative use for IDFs is as items on drawing scales. Scale items have much weaker correlations with criterion measures. Anastasi (1982) states that the average correlation between an item on a self-report scale and its criterion measure is .25, which is very similar in magnitude to the average correlation observed here. The possibility also presents itself that the results of the meta-analysis could be used as the initial stage of item generation and selection in the process of developing drawing scales.

In conclusion, numerous drawing features were found to demonstrate significant correlations with independent measures of anger/hostility, anxiety, and thought disorder. These results support the use of a construct-focused approach in the review of human figure drawing studies and suggest that many IDFs are potentially valid measures of these constructs. The small number of study findings underlying the analyses for numerous IDFs and the conservative practices used in the meta-analysis, however, render many of the conclusions regarding the potential validity of individual features tenuous. Further, the magnitude of the observed correlations between IDFs and independent measures are too small to consider employing them as tests. An alternative and more appropriate use for

drawing features is as items on drawing scales. The present meta-analysis can then be used as a means of generating and selecting items for such scales. Subsequent analyses to determine which features contribute to such drawing scales would provide further confirmation of the validity of the features identified in the present analysis. Study 2 describes the development of four drawing scales using the results of the current meta-analysis to generate and select IDFs to be used as items on the scales.

Study 2

Another issue in the study of human figure drawings is the validity of the technique of aggregating IDFs into drawing scales which measure specific forms of psychopathology. This issue contains two separate but related questions. The first is whether IDFs can be combined into scales which are more meaningful and useful than IDFs used on their own, and the second is whether these drawing scales can be designed to measure specific psychopathological states.

The first question, whether the aggregation of IDFs will produce scales which are more meaningful and useful than IDFs, is not much debated. For example, Kahill (1984), Roback (1968), and Swensen (1968) all provide positive reviews of research employing drawing scales. Each author stated that the evidence indicates that such scales show far more promise than the IDFs used on their own.

The second question, whether drawing scales can measure specific psychopathological states, has become contentious. Some authors, such as Tharinger and Stark (1990), have begun to argue that scoring systems for human figure drawings should not seek to measure specific states at all, but should focus on the global assessment of broad emotional conditions, such as psychological well-being and adjustment. This argument has been fueled by the relative success of instruments such as the Draw-A-Person: Screening Procedure for Emotional Disturbance (DAP:SPED; Naglieri, McNeish,

& Bardos, 1991). Numerous authors, however, have found significant correlations between drawing scales and several different constructs, including anxiety (Engle & Suppes, 1970; Sturner, Rothbaum, Visintainer, & Wolfer, 1980), body-image (Carlson, Quinlan, Tucker, & Harrow, 1973; Shaffer, Duszynski, & Thomas, 1984), conventionality (Shaffer et al., 1984), self-esteem (Calhoun, Ross, & Bolton, 1988; Calhoun, Whitley, & Ansolabehere, 1978; Spiga, Mindingall, Long-Hall, & Blackwell, 1986), sexual elaboration (Carlson et al., 1973), and thought disorder (Lapkin, Hillaby, & Silverman, 1968; Wexler & Holzberg, 1952). These latter findings suggest that drawing scales devised by aggregating IDFs are potentially valid as measures of specific types of psychopathology.

This second study developed and assessed the validity of four scales for human figure drawings. The specific constructs measured by these scales were anger/hostility, anxiety, social maladjustment, and thought disorder. The introduction to the study begins with an examination of the psychometric potential of IDFs, which also serves to argue for the adoption of drawing scales. The next section reviews the literature pertaining to drawing scales comprised of aggregated IDFs and highlights the importance of employing rigorous methods when developing these scales. The third section discusses the use of the meta-analytic results from Study 1 to generate and select IDFs for use in the drawing scales to be developed. The final section re-examines the research on anger/hostility used in the previous meta-analysis. Based on that re-examination, it is argued that the studies may also be interpreted as assessing the construct of social maladjustment.

The limitations of IDFs. While not much discussed, the question, "Why aggregate IDFs into drawing scales?" needs to be answered in order to justify any greater effort spent on developing scales. It is also important to set forth the arguments against

independently employing IDFs in order to clarify the hazards of drawing inferences regarding psychopathology from them.

In the discussion section of the previous study it was argued that the magnitude of observed correlations between IDFs and independent measures made these features more suited to the role of scale item than to the role of test. The small size of these correlations is also evidence of the psychometric weakness of such features. Individual drawing features simply do not correlate highly with other tests of constructs such as anger/hostility, anxiety, and thought disorder. As such, they are not likely to comprise strong predictors of these states nor do they explain much of the variance in these constructs.

It is not surprising that IDFs do not correlate highly with independent measures of psychological constructs. Individual drawing features are not tests, but rather samples of behavior. Like any sample of behavior, be it a test item or a simple action, drawing features can be expected to be unreliable as measures of larger constructs (Nunnally, 1978). The unreliability of these features has profound effects on any assessment of validity, because the reliability of a measure sets an upper limit on the magnitude of the obtained validity coefficient (Cronbach & Meehl, 1955). Measures with poor reliability will produce, at best, low validity coefficients.

The lack of strong correlations between IDFs and independent measures can also be attributed to the large percentage of features which possess highly skewed distributions. Product-moment correlations, the most commonly used correlations in validity studies of IDFs, are affected by any difference in the distributions of the two variables being correlated (Nunnally, 1978, p. 141). The magnitude of the difference between distributions has a direct functional relationship to the maximum product-moment

correlation. The greater the difference between distributions the lower the upper limit on the resulting coefficient.

Numerous authors have documented that a large number of IDFs possess skewed distributions. Marzolf and Kirchner (1970) examined the incidence of 32 dichotomously scored IDFs in a sample of 850 undergraduate students and found that 22 of them occurred in either fewer than 10% or more than 90% of drawings. An even larger percentage of skewed features was observed by Soccolich and Wysocki (1967), who found that 11 of the 14 dichotomously scored features they studied occurred in fewer than 10% of drawings made by university students. Many features with skewed distributions are also observed in the drawings of children. Snyder and Gaston (1970) examined the incidence rate of Koppitz's (1968) dichotomously scored emotional indicators in a sample of 324 first grade students. They found that 15 of 31 features investigated were present in either fewer than 10% or more than 90% of drawings. Groves and Freid (1991) found 62% of Koppitz's drawing features to possess similarly extreme distributions in samples of five, six, and seven year olds. These frequencies are consistent with Koppitz's (1968) results from her sample of over 1,800 children aged six to twelve.

Independent measures, in contrast to the IDFs with which they are correlated, are likely to be near normal in distribution. This is because studies attempting to validate IDFs will seek to employ established measures which are more likely to have normal distributions than the features with which they will be compared. Once again it is not surprising that small correlations are observed.

The psychometric arguments presented here suggest that IDFs would be unreliable, capable of only weak demonstrations of validity, and not likely to be of much use in the prediction or the explanation of the construct under study. Results of the

previous meta-analysis confirm these expectations. Both theory and empirical evidence suggest that IDFs lack the psychometric qualities necessary to be used on their own.

Drawing scales. A common approach to increasing the reliability and validity of a measure is to collect the less reliable individual items into a scale. This aggregation produces a measurement instrument with greater reliability than that possessed by the individual items (Nunnally, 1978). Increasing the reliability of the measure also increases the ceiling of the validity coefficient, allowing the new scale to produce higher correlations with independent measures. Direct empirical evidence that aggregation of IDFs results in higher validity coefficients than the individual features alone comes from a meta-analysis reported in Acton and Moretti (1991). We employed meta-analytic techniques to directly compare the correlations between independent criteria and drawing feature measures of anxiety with correlations between independent criteria and drawing scale measures of anxiety. Each of the eight studies analyzed employed validated self-report measures of anxiety as their independent criteria. We found that the average correlation was .09 for IDFs and .41 for drawing scales. The difference between the two average correlations was statistically significant.

The utility of pursuing drawing scale measures is also supported by the number of authors who have found significant correlations between aggregate drawing scales and independent measures of constructs of interest (Calhoun, Ross, & Bolton, 1988; Calhoun, Whitley, & Ansolabehere, 1978; Carlson, Quinlan, Tucker, & Harrow, 1973; Engle & Suppes, 1970; Lapkin, Hillaby, & Silverman, 1968; Naglieri & Pfeiffer, 1992; Shaffer, Duszinski, & Thomas, 1984; Sturner, Rothbaum, Visintainer, & Wolfer, 1980; Wexler & Holzberg, 1952). The consistency with which these authors have achieved significant findings tends to confirm that aggregation of drawing features leads to more reliable and valid measurement.

Drawing scales, while relatively more successful than individual features, are still somewhat variable in their performance. Carlson, Quinlan, Tucker, and Harrow (1973), while achieving significant correlations with some measures, failed to find significant correlations between their body-image drawing scale and four of their independent measures of body image. Similarly, Spiga, Mindingall, Long-Hall, and Blackwell (1986) and Tharinger and Stark (1990) failed to find significant differences between normal and pathological groups when employing Koppitz's (1968) emotional indicators, despite the past success of other authors with the measure (Koppitz, 1966a, 1966b, 1966c; Sturner, Rothbaum, Visintainer, & Wolfer, 1980). Further, among studies which have found significant correlations or differences between groups (Calhoun, Ross, & Bolton, 1988; Calhoun, Whitley, & Ansolabehere, 1978; Carlson et al., 1973; Hiler & Nesvig, 1965; Koppitz, 1966c; Naglieri & Pfeiffer, 1992; Shaffer, Duszinski, & Thomas, 1984) effect sizes have ranged from small (e.g., $\underline{r} = .29$) to large (e.g., $\underline{r} = .82$.)

One possible reason that the results of past studies employing drawing scales have varied to the degree that they have is that the authors have used practices which have differed in their efficacy. Studies have employed widely diverging methods for generating and selecting scale items. Procedures used to generate items for scales have included reviewing projective hypotheses such as those proposed by Machover (1949), selection based on the previous use of items in empirical studies, and combinations of these two practices. Researchers have also employed different means of evaluating which items to retain in their scales. Some studies keep all of the features which have been generated (e.g., Handler & Reyher, 1966). Other studies assess whether items appear to contribute to a scale (e.g., Carlson et al., 1973). Still others select items based on criteria such as base rates in a normative sample and ability to discriminate between criterion groups (e.g., Naglieri, McNeish, & Bardos, 1991).

Of these various approaches to item selection, those which make use of empirical methods to select drawing scale members appear to find the strongest results. Naglieri and Pfeiffer (1992) obtained a correlation of .82 between the Draw A Person: Screening Procedure for Emotional Disturbance (DAP:SPED; Naglieri, McNeish, & Bardos, 1991) and group membership, in a study comparing normally achieving high school students and adolescents in an outpatient psychiatric program. The items for the DAP:SPED were selected based on rational criteria combined with base rates in a large normative sample. Koppitz (1966b) was able to discriminate between normal children and children referred to a child guidance clinic, achieving an effect size of $\underline{r} = .67$. She used a scale of emotional indicators (Koppitz, 1968) which had been selected such that: a) the items differentiated between normal and emotionally disturbed adolescents; b) the items had a low rate of occurrence in her normative sample of 1800 children; and c) the items were independent of age effects. Hiler and Nesvig (1965) found that they could distinguish between adolescents hospitalized in a psychiatric facility and normal school students. The effect size associated with their results was $\underline{r} = .72$. Their scale was comprised of drawing features identified by practicing clinicians which were also found to differentiate between hospitalized and nonhospitalized subjects in a pre-test sample.

None of the studies which have used rigorous item selection procedures have also chosen to assess narrowly specified psychological disorders. Koppitz (1966a, 1966b, 1966c, 1968) and Naglieri and his associates (1991, 1992) have focused on developing screening measures which function only to indicate the presence or absence of emotional problems. Similarly, Hiler and Nesvig (1965) were interested only in establishing that reliably measured combinations of drawing features could differentiate between psychiatric patients and normal controls. As a result, while these authors have provided good evidence of the potential validity of aggregate drawing scales, their work does not indicate

whether scales having strong psychometric qualities can be developed for the assessment of specific forms of psychopathology.

Item generation and selection using meta-analysis. One approach to item selection which may prove nearly as rigorous as the methods used above and may also provide for the measurement of specific forms of psychopathology is meta-analysis. Drawing features which demonstrate an overall significant relationship to independent measures of specific constructs across a number of studies are likely to contribute to drawing scales of those same constructs. Study 1 reported a meta-analysis which assessed the correlation between a large number of drawing features and independent measures of several constructs. This meta-analysis provides the kind of basic information about the function of individual items which one would attempt to derive when selecting IDFs for measurement scales.

A re-examination of the research on anger/hostility. Many of the independent measures employed by the authors of the research assumed to assess anger/hostility appear also to assess antisocial values and delinquency. For example, the Psychopathic Deviate (Pd) subscale of the MMPI used in Wainwright (1970) possesses a number of items reflecting anger, but also has items indicative of past delinquent behavior, family conflict, and criminal values. The scale was, in fact, initially validated on a sample of delinquent adolescents (Greene, 1980). Histories of verbally or physically aggressive acts which constituted behavioral indicators of anger/hostility in many studies also may permit more than one interpretation. Such acts are commonly found in individuals who do not conform to societal expectations. It is not surprising, therefore, that the samples employed in some of the studies consisted of individuals with conduct problems (Griffith & Lemley, 1967; Koppitz, 1966c) and delinquents (Daum, 1983). In all of four of these studies it could be argued that it is not anger/hostility that is measured, but social maladjustment in the form of aggressive behavior and antisocial attitudes.

Not all of the studies have assessed conduct disordered adolescents or used measures of delinquency. Greenberg and Fisher (1971), Kurdek and Darnell-Goetschell (1987), and Shafranske (1981) employed university students as subjects and used measures validated solely as tests of hostility. These studies do not appear to measure social maladjustment. Manning (1987) used a sample of child victims of physical abuse. The homes of the victims were described as violent and all children were currently living in a victim's shelter. There is the possibility in this sample that because of exposure to violent models these subjects may have internalized antisocial values or begun to engage in delinquent behavior. This possibility was not explored in the study, however, and so the degree to which these findings reflect social maladjustment rather than anger or hostility is unknown.

This review of the studies of anger/hostility suggests that approximately half may have been assessing social maladjustment. This makes it highly likely that at least some of the significant results attributed to anger/hostility in the previous meta-analysis actually reflect effects of social maladjustment and that both should be assessed.

The present study. The present investigation developed drawing scales for the constructs anger/hostility, anxiety, social maladjustment, and thought disorder. The results from the meta-analysis carried out in Study 1 were used to select potential items for each drawing scale. A multi-trait approach was used to test the validity of the newly developed scales, with independent measures selected to provide evidence of both convergent and discriminant validity. It was expected that drawing scales would correlate most highly with independent self-report measures of the same construct, and poorly or not at all with discriminant measures.

Method

Participants. Subjects were selected from among 1,000 consecutive admissions to a young offender program, Youth Court Services, in Burnaby, B. C. between 1986 and 1990. All subjects were between 12 and 18 years of age, had been charged with at least one criminal offense, and were referred for psychological or psychiatric services. To be selected for the study a subject had to have completed the core tests of the psychological assessment battery used at the center and have drawn a scorable human figure. The core tests included the House-Tree-Person (H-T-P; Buck, 1948), from which the drawing of the person was taken, the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1967), the Jesness Inventory (JI; Jesness, 1983), and either the Wechsler Intelligence Scale for Children - Revised (WISC-R; Wechsler, 1974) or the Wechsler Adult Intelligence Scale - Revised (WAIS-R; Wechsler, 1981). To be valid a human figure drawing needed to meet six criteria: 1) it must be drawn on an eight by eleven sheet of paper; 2) it must not be a stick figure; 3) it must not be a cartoon figure; 4) it must be a full figure (drawings of a head or only the upper body were excluded); 5) the figure in the drawing must not be obscured by other objects drawn; and 6) there must be only one figure, unless multiple figures are drawn with one clearly central.

Five hundred and fifteen of the potential subjects were excluded from the present study. Of these, 421 were rejected because their files did not contain the complete core battery. There were several reasons why a potential subject might not have completed the battery. First, many of the cases referred to the Youth Court Services did not receive the assessment battery used in this study. These adolescents were directly referred for individual or group treatment and received testing specific to their treatment. Second, a large number of the adolescents, as is often seen in a setting such as this, refused to complete one or more of the standardized tests. Another 69 potential subjects completed

the battery, but did not complete a scorable drawing of a human figure. The results of a further 25 participants were removed from the sample to provide a training sample on which to practice scoring. The final sample consisted of 410 males and 75 females.

Feature selection and the scoring manual. Selection of the IDFs to be scored and the design of a scoring manual were completed prior to the rating of the human figure drawings. Individual drawing features were chosen from among those analyzed in Study 1, using $p \leq .10$ as the selection criterion for the majority of cases. The cut-off point was set at this level in order to provide a liberal selection of features. A further set of features was chosen from among the IDFs which were found to be nonhomogeneous (i.e., had a significant diffuse test) and for which differences between findings could not be explained using the predefined moderator variables. Findings for these features either did not differ along the dimensions assessed by the moderator variables or the analyses proved nonsignificant at all levels. For these features a further set of selection rules was devised. To be selected, the majority of correlations observed for a feature had to have been as large or larger than those of features which had already been selected for further study at other points in the analysis. When there were only two findings, one of which was of large magnitude and the other of zero magnitude or near zero, the feature was not selected.

Criteria provided in the scoring manual were adapted from a number of previous works which attempted to provide reliable coding schemes for drawing features. The use of many sources led to the selection of numerous inconsistent and, sometimes, contradictory criteria for scoring the same feature. A set of rules was devised to decide which schemes to employ. A scoring scheme was retained if it was more reliable than comparison schemes, provided multi-point rather than dichotomous scoring, and/or was less subjective in nature than a rival approach. Two or more scoring schemes were kept for a single feature when they employed different enough criteria but were similar enough

in decision rules to suggest that either could be a viable form of rating. In addition, a number of feature scoring schemes were eliminated because they were too difficult to reconstruct. Scoring criteria for these were either presented in too little detail in the original work or made use of materials or practices not employed at Youth Court Services. When these decision rules were used, 71 scoring schemes were dropped from the study. Fortunately, because of the high degree of redundancy across scoring criteria, these eliminations resulted in only 14 features being removed from the pool of possible items. The resulting manual provided detailed scoring instructions for 102 drawing features and a further seven feature scores were computed from these original scorings. The scoring manual containing the names of the features and their criteria can be found in Appendix C.

Scoring and assessment of reliability. All drawings were scored by the first author following the manual developed for the study. A second rater then scored a subsample of the human figure drawings to provide an estimate of interrater reliability. Both raters were trained on an independent sample of 25 human figure drawings.

Selection of drawings on which to test interrater reliability proved problematic because of differences in the distribution of scores across drawing features. For features such as physical measurements or scoring schemes which possessed a large number of increments, a reliability sample could be selected randomly from the total sample scored by the first author. The remaining features possessed between two and five levels of scoring and many of them, when initially rated, produced highly skewed distributions. For example, only 3 of 485 drawings were considered to be without legs in the dichotomously scored item "omission of legs". A simple random selection of drawings may have produced a sample which either varied little or not at all for some of these features. As a result, estimates of reliability would have been greatly biased or impossible to derive. A

process was developed which guaranteed selection of a sample of drawings from all levels of scoring in order to avoid the possibility of choosing an invariant or highly skewed sample of drawings. An equal number of drawings was randomly selected at each level of scoring for features which possessed five or fewer levels. For example, if a feature was scored dichotomously, 20 drawings would be selected, 10 of which had been scored as 0 by the first rater and 10 of which had been scored as 1 by the first rater. The scores of the first rater were always used to select the sets of drawings for the reliability analysis. For 7 of these features, however, the number of scorable drawings was too low even for this form of equalization. In these cases an imbalanced selection of features was used that incorporated all of the drawings for the level with the low count and enough other drawings to form a sample of 20 drawings. Twenty drawings were selected because all of these features were dichotomous and that number of drawings was used for all other dichotomously scored features.

Product-moment correlations between the scores of the two raters served as the initial estimates of interrater reliability in order to facilitate comparison of the present results with those of past studies. Obtained correlations between raters ranged from .30 to 1.00, with an arithmetic average of .83 for untransformed rs. Almost 25 percent of the features scored achieved a perfect correlation. Further, the ratings of the two scorers correlated at .80 or better for 74% of the features and at .70 or better for 85% of the features. These results are consistent with other correlational findings for interrater reliability. Kahill (1984), in her review of interrater reliabilities, found that among the studies she examined the majority of the correlations were over .80, with over three quarters of the coefficients over .70. The range of correlations in Kahill's review was from -.13 to 1.00 for what she termed content-based features and from .29 to 1.00 for what she

called structural features. Few of the features reported in Kahill's study achieved a correlation of 1.00.

Cohen (1960, 1968) has recommended that statistics used to calculate interrater reliability should reflect agreement rather than simply correlation. He recommends the use of kappa for categorical variables, while other authors (Fleiss & Cohen, 1973; Shrout & Fleiss, 1979; Landis & Koch, 1977) recommend the intra-class correlation for continuous data. The scores of the two raters were therefore further analyzed using these statistics.

Estimates of inter-rater reliability based on the above statistics differed from those based on correlational analyses for only some of the variables. Results for 64 of the 102 features were unchanged when the new statistics were applied. Perfect agreement in the scoring of many features and the presence of equal or nearly equal distributions of scores by the two raters (referred to as marginals) are the factors responsible for these high rates of agreement across assessments. When marginals are equal in comparisons of two variables, the kappa coefficient is identical in magnitude to the product-moment correlation (Fleiss & Cohen, 1973). For the remaining results, kappa coefficients and intra-class correlations were in the same range as the product-moment correlations, although for 8 variables there were noticeable declines in the size of the resulting coefficient (i.e., a fall of more than .20 points.) Despite the relatively lower kappa coefficient for some variables, none fell below the lower bound of .30 found in the correlational analysis. Further, the overall distribution of interrater reliability scores was not much different from those found in the correlational analyses. It appeared that the majority of items were reliably scored.

Product-moment correlations, intra-class correlations, kappa coefficients, and the sample size upon which each value was based are reported in Appendix D.

Independent measures. The seven scales employed as independent measures in the present study came from four psychological tests: the MMPI (Hathaway & McKinley, 1967), the JI (Jesness, 1983), the WAIS-R (Wechsler, 1981), and the WISC-R (Wechsler, 1974). Four of the scales served as criterion measures and provided evidence of convergent validity. Three of the scales served as discriminant measures and provided evidence of discriminant validity.

All four of the criterion measures came from the two self-report inventories, the MMPI and the JI. The three scales used from the MMPI were the Hostility (Ho) scale (Cook & Medley, 1954), the Manifest Anxiety (MAS) scale (Taylor, 1953), and the Schizophrenia (Sc) scale (Hathaway, 1956). The Hostility scale contains 50 items that were chosen to assess the presence or absence of attitudes and beliefs which reflect dislike for and distrust of others. Item selection was made based on expert judgments and the potential of items to distinguish teachers who thought pupils were dishonest and untrustworthy and those who regarded their students positively. The internal consistency of the scale was found to be .86 (Cook and Medley, 1954). More recent studies have found scores on the Hostility scale to correlate with independent measures of distrust, frequent anger, resentment, and suspiciousness (Greenglass & Julkunen, 1989; Pope & Smith, 1990; Smith & Saunders, 1990), but not with overtly aggressive behavior (Smith & Frohm, 1985). This scale is believed to assess the construct of anger/hostility.

The Taylor Manifest Anxiety scale contains 50 true/false items that assess the presence of fears and somatic complaints associated with anxiety, and feelings of over-excitement, nervousness, upset, and worry. Scale items were selected via content evaluation by expert judges and magnitude of item-total correlations in a small college sample. Taylor (1953) reported test-retest correlations of .89, .82, and .81 at intervals of three weeks, five months, and seven to nineteen months, respectively. Split-half reliability

of .92 has also been reported (Hilgard, Jones, and Kaplan, 1951). Significant and high correlations have been observed with the psychasthenia scale of the MMPI (Brackbill & Little, 1954) and the IPAT Anxiety scale (Bull & Strongman, 1971). This scale served as the independent measure of anxiety.

The Sc scale contains 78 items which were found to empirically differentiate between schizophrenic patients and normal controls (Hathaway, 1956). Test-retest correlations ranged from .74 to .95, at two weeks, and from .37 to .64, at intervals of up to one year (Dahlstrom, Welsh, & Dahlstrom, 1975). Subsequent studies of validity have confirmed the scale's potential to measure schizophrenia in adults (Molden & Gottesman, 1991; Patrick, 1988; Walters, 1984) and adolescents (Newmark & Gentry, 1983). This scale served as the independent measure of the construct thought disorder.

The Social Maladjustment (SM) scale was the only scale used from the Jesness Inventory (Jesness, 1983) as a criterion measure. This scale contains 63 true/false items which reflect attitudes expressed by antisocial adolescents. Items for this scale were selected based on their ability to differentiate between groups in a sample of 2,045 delinquent and nondeliquent males. Jesness (1983) reported split-half reliability of .84, when corrected for reductions in test length, and test-retest reliability which ranged from .79 at an eight month interval, to .65 at a one year interval. Numerous subsequent studies have confirmed that the SM scale differentiates youths in trouble with the law from non-offenders (Graham, 1981; Martin, 1981; Rothermal, 1985; Saunders & Davies, 1976; Singh, 1983; Vallance & Forrest, 1971). This scale served as the independent measure of social maladjustment.

The discriminant measures were the Repression (Rep) scale of the JI (Jesness, 1983), and the Full Scale IQ (FSIQ) score from either the WAIS-R (Wechsler, 1981) or the WISC-R (Wechsler, 1974). The Repression scale contains 15 true/false items which

reflect lack of awareness or failure to admit to feelings of anger, dislike, or rebellion, and a generally uncritical view towards self and others. This scale was developed from a cluster analysis of responding by 2,045 delinquent and non-delinquent males. Jesness (1983) reported a split-half reliability of .64 when corrected for test-length, and test-retest reliability of .55 at an eight month interval, and .50 at a one year interval. Several studies have shown that scores on the Rep scale are largely independent of those on the SM scale (Martin, 1981; Rothermal, 1985; Saunders & Davies, 1976). Scores on the Rep scale also do not correlate with measures of psychological well-being (Munson & Blincoe, 1984; Roberts, Shmitz, Pinto, & Cain, 1990). These findings support the discriminant potential of the scale.

The WAIS-R is a well established test of intelligence, normed on a representative sample of 1,880 adults and adolescents. The FSIQ, which is the age-corrected, standard score equivalent of the total of all of the tests on the WAIS-R, is the most reliable estimate of IQ from this measure. Wechsler (1981) reported that the average split-half reliability for the FSIQ was .97 and that the test-retest reliability ranged from .95 to .96. Correll (1985) provides evidence of the independence of the FSIQ score and measures of psychopathology. He correlated the scores from the MMPI clinical subscales, including the Sc subscale, and the MAS with the subtest scores, the Verbal IQ, the Performance IQ, and the FSIQ of the WAIS-R. He found that the number of significant correlations among the measures did not exceed chance levels and that the magnitude of the correlations were uniformly small.

The FSIQ from the WISC-R possesses similar levels of reliability and validity as it's WAIS-R counterpart. The WISC-R was normed on a representative sample of 2,200 children who ranged from six-and-a-half to sixteen-and-a-half years of age. Wechsler (1974) reported an average split-half reliability of .96 and an average test-retest reliability

of .95 across a three to five week interval. Concurrent validity of the FSIQ has been well established (Sattler, 1982). These results support the use of the FSIQ as a discriminant measure in this study.

Design. The analysis of Study 2 was carried out in three steps. The first two steps explored the relationship between the independent measures of anger/hostility, anxiety, social maladjustment, and thought disorder and sets of drawing features using multiple regression techniques. The decision to use multiple regression analyses hinged on the argument that this statistical approach takes into account the combined effects of groups of variables. While the performance of IDFs was of interest, the ultimate goal of the study was to find the best aggregation of these features.

The third step of the analyses assessed the intercorrelations among the scores of the drawing scales and the independent measures. The intercorrelation of the independent measures was tested first to identify an expected pattern of intercorrelation. The pattern of intercorrelations among the drawing scales was then assessed to see how closely these scales mirrored relationships observed among the independent measures. The third and last set of correlations calculated were between the scores of the drawing scales and those of the six independent measures, which provided a direct test of convergent and discriminant validity.

Results

The goal of the first step of the analysis was to ascertain whether sets of IDFs chosen based on their performance in the previous meta-analysis would correlate with their respective criterion measures. The sets of IDFs were chosen such that: 1) the significance of their correlation with independent measures in the meta-analysis was $p \leq .10$; and 2) all IDFs in a set were found to correlate with measures of the same construct. For example, all features correlating at $p \leq .10$ with independent measures of anxiety

would be in the same set. A p value of $\leq .10$ was used instead of the more common $p \leq .05$ in order to select a larger sample of features, in keeping with the exploratory nature of this study.

Four step-wise multiple regression analyses were carried out, one for each combination of drawing features and criterion measure. No transformations of the feature scores were made at this point in the analysis. Each set of drawing features was initially entered as a block. Results of this initial step in the analysis were used to determine whether the features selected by the meta-analysis could predict the criterion measures. Drawing features from the initial block were then removed in a stepwise fashion using $p > .10$ as the criterion for removal and $p \leq .05$ as the criterion for re-entry. The step-wise procedure identified subsets of the total blocks of features which best predicted the criterion measures.

The total number of features selected from the meta-analysis were 16 for anger/hostility, 31 for anxiety, 16 for social maladjustment, and 61 for thought disorder. The overall results of forcing these sets of IDFs into the regression analyses with the four criterion measures are reported in Table 6. The magnitude of the observed multiple correlations for all constructs were small to moderate. Even the thought disorder scale, which contained 61 variables, produced a multiple correlation of only .38. None of the multiple correlations exceeded the $p \leq .05$ level of significance and the adjusted $\underline{R}^2$s were zero or near zero for all scales, suggesting that findings were at chance levels when corrected for the number of variables.

The results of the stepwise regression procedure are reported in Table 7. The final equations for which statistics are reported were comprised of only those variables which significantly contributed to the regression equations at the $p \leq .10$ level of significance. The multiple correlations for three of the four regression equations exceeded a p value of

$\leq .05$. The multiple correlations, however, were small. Further, the fact that the adjusted R^2s for three of the four equations were near zero suggests that those results were likely a product of chance.

It is also important to note that, when only significantly contributing features were used, the number of features included dropped appreciably. One, five, two, and nine features were used in the final regression equations for anger/hostility, anxiety, social maladjustment, and thought disorder, respectively. These results suggest that the expected features combined to produce only weak drawing scales and that very few of the features expected to contribute to the scales actually showed a relationship to the constructs they were hypothesized to measure.

Exploratory regression analyses: Regression using all features. To further examine the potential of the drawing features, a series of exploratory multiple regression analyses were carried out to determine the best combination of features for predicting the criterion measures from among all of the IDFs selected for Study 2. Before these analyses were completed a large number of the drawing features were transformed from their original scoring format. These transformations were carried out to minimize differences in the number of levels of scoring across features and to produce more normal distributions in the scored features. Efforts to reduce the number of increments included recoding features with large numbers of possible values, i.e., reducing those with more than five to four levels of measurement. The choice to use four levels was made because the majority of categorically scored, nondichotomous drawing features used four levels. Features scored with two, three, or five levels of scoring were, for the most part, not transformable to four levels. Scores assigned to the dichotomous features were changed to 0 and 3 in order to make the weight of their positive contribution as high as that of the highest point of the four level scoring schemes.

Step-wise backwards entry multiple regression was used as before. This time, however, all 109 IDFs scored in the study were used in every analysis. All features were initially entered in a block and then removed and re-entered in a step-wise fashion. Once again, the criterion for removal was $p > .10$ and the criterion for re-entry $p \leq .05$.

The statistical results for the four regression equations are reported in Table 8. The multiple correlations for all of the regression equations exceeded a p value of $\leq .05$ and the adjusted R^2s were larger than zero. The magnitude of the observed multiple correlations were also larger than in the previous analyses.

Exploratory regression analysis: Age effects. At this stage in the analysis the sample was divided into two age groupings, those 15 years of age and younger and those 16 years of age and older. The respective sample sizes were n=253 and n=232. Analysis by age groupings was carried out because age might have had an effect on drawing scale content as it has had on scale content of self-report measures used with other adolescent samples, such as the Child Behavior Checklist (Achenbach & Edelbrock, 1986).

Once again stepwise multiple regression was used to assess which combinations of drawing features best predicted the independent measures. As with the previous analyses, an initial block of features was entered first. Four separate blocks of features were used in the age analyses, one for each of the constructs studied. Each block was comprised of the collection of features which had been selected in the previous analyses as best predicting their respective criterion measures. The four blocks were entered into eight multiple regression analyses, one for each construct by each of the two age levels. Once the initial blocks had been input, any features remaining from the total sample of 109 were entered in a step-wise fashion. A feature was added only if it produced a significant change in the F ratio for the regression equation at the $p \leq .05$ level. It should be noted that the drawing features were not entered with the regression weights from the overall analyses. Instead

the features were simply entered as a block, allowing the regression program to set new weights.

Very few features were added as a result of these analyses. The number of additional features ranged from one to five, with the modal number equal to two. The results of the regression analyses employing these additional features are reported in Table 9. All of the multiple correlations had an associated p value of $\leq .05$. The division of the sample into age groups and the addition of more features to the scales produced increases in the magnitude of the multiple correlation for all of the regression equations over that observed in the previous analyses. However, a corresponding increase in the magnitude of the adjusted R^2s was not as reliably observed. Only seven of eight regression analyses demonstrated increases in significance.

Convergent and discriminant validity. The next stage of the analyses was to provide evidence of the validity of the drawing scales by examining the pattern of correlations among the drawing scales and independent measures. Separate correlational analyses were completed to compare the independent measures, the drawing scales, and the drawing scales with the independent measures. The first analysis established the expected pattern of correlations among the independent measures. This analysis was carried out because evidence suggests that self-report measures of negative feeling states can be highly intercorrelated (Gotlib, 1984; Clark & Watson, 1991; Watson & Clark, 1984). In the absence of such comparison data, if one were to observe generally high correlations among drawing scales, and among drawing scales and independent measures of different constructs, it might be construed that drawing scales lacked specificity. However, if high intercorrelations are expected, then a similar pattern of effects observed with drawing scales would tend to support their validity. The second analysis assessed whether a pattern of correlations similar to that observed with the independent measures

would be found among the drawing scales. The last analysis provided a direct test of the degree to which drawing scales converged with independent measures of the same construct and were discriminable from independent measures of other constructs.

Intercorrelations of the independent measures. It was expected that the criterion measures would show strong intercorrelations because of the presence of the shared factor, negative affectivity (Watson & Clark, 1984). Smaller nonsignificant correlations were expected between these measures and the discriminant measures of intelligence and repression.

The results, reported in Tables 10 and 11, for subjects 15 years of age and younger and those 16 years of age and older, respectively, were largely what was expected from past research. The pattern of convergence for both age groups showed the four criterion measures achieving moderate to high correlations with each other. Coefficients ranged in size from .47 to .83, with the largest correlations observed between the anxiety and the thought disorder scales. The Rep subscale and the FSIQ measures showed consistently lower correlations with criterion measures than such measures exhibited among themselves. Further, one third of the observed correlations between convergent and discriminant measures were nonsignificant.

The results of this analysis indicated that there was an expected pattern of convergence and discrimination among the independent measures used in the study. Criterion measures were highly correlated with each other. It was expected that drawing scales would also be highly intercorrelated. The criterion measures, however, produced only small and occasionally nonsignificant correlations with discriminant measures. To keep with this pattern of results, the drawing scales would also have to show low or nonsignificant correlations with the discriminant measures.

Intercorrelations of drawing scales. Total drawing scale scores used in the correlational analyses reported in this and the subsequent section were based on the unweighted aggregation of the transformed feature scores. The correlations between drawing scale scores are reported in Table 12 for both age groups. All but a few of the correlations between drawing scales were significant. The magnitude of the correlations varied somewhat, with the majority falling in the small to moderate range. Significant correlations ranged in size from .13 to .49, in the younger age group, and from .17 to .52, in the older one. These correlations were clearly smaller than those observed between the criterion measures. Lower correlations suggest greater independence between scores on the drawing scales than on the criterion measures.

While the differences in magnitude of correlations among measures for both the criterion and the drawing scales was not large, some notable similarities may be observed between the two correlation tables. The highest correlations were observed between the anxiety and thought disorder measures in both analyses, while the lowest correlations were observed between the measures of anger/hostility and social maladjustment. These similarities provide further evidence of concordance between the two sets of measures. The relatively lower magnitude of correlation, however, suggests that the strength of association among the drawing scales is not as great as it is among the self-report measures.

Intercorrelations of drawing scales and independent measures. Correlations between drawing scales and independent measures are reported for younger subjects in Table 13 and for older ones in Table 14. Drawing scales exhibited the expected pattern of uniform correlations across the different criterion measures and correlated more highly with these measures than with discriminant ones. Additionally, correlations were highest between drawing scales and criterion measures of the same construct. The only exception

to this tendency was the higher correlation observed between drawing scales measures of anxiety and self-report measures of thought disorder than between drawing scale and self-report measures of anxiety. Such high correlations between measures of anxiety and thought disorder were not surprising given the level of intercorrelation noted between such measures in both of the previous correlational analyses.

Evidence for discriminant validity was good. Correlations of drawing scales with scores from the intelligence scales and the Rep subscale were expected to provide the clearest evidence of discrimination. None of the correlations between the drawing scales and the Rep subscale were significant for either age group. Correlations with the intelligence test scores were low or very low in all cases, with only two achieving statistical significance.

The low magnitude of the observed correlations with the criterion measures is of concern, however. The largest convergent correlation between a drawing scale and a criterion measure of the same construct was .38. The average of such correlations across both age groups was .33. Correlations of this size indicate only weak relationships between the two sets of measures.

Discussion

None of the combinations of features selected based on the results of the previous meta-analysis produced significant multiple correlations when regressed onto the criterion scales in this study. Attempts to identify more predictive subsets of these features led to increases in the magnitude of the multiple correlations and the adjusted R^2s. However, the observed increases were small and the resulting sets of features contained few items.

Subsequent exploratory analyses employed all 109 of the drawing features scored in the study. These analyses focused on identifying subsets of drawing features that would better predict the criterion measures. Several subsets of features were identified which

produced small multiple correlations when regressed onto the criterion measures. While small, the magnitude of the resulting correlations was on average greater than that observed in the previous analyses. These feature sets also included sufficient numbers of items to be considered useful as scales.

The total sample was then split into two age groups. Multiple regression analyses were repeated on the separate samples to determine whether new scale items would be selected. Additional items did contribute to the regression equations in both groups, suggesting that age affects the content of the drawing scales. The small number of additional features selected would indicate that, at least for the present study, any such effects were minimal.

The newly developed drawing scales were then assessed in a series of correlational analyses looking at patterns of convergence and discrimination between measures. The performance of the drawing scales needed to match the observed pattern of correlations among the independent measures employed in the study (i.e., high intercorrelations among criterion measures and low or nonsignificant correlations between criterion and discriminant measures) to demonstrate their potential validity. The drawing scales produced a very similar pattern of correlations with one another and with the independent measures.

As with previous analyses, the magnitude of the correlations between drawing scales and independent measures were small, ranging from .14 to .38. To interpret these correlations, it is important to first compare these results with findings from previous studies which have developed drawing scales measuring specific forms of psychopathology. If the size of correlation found in the present study is comparable to that in other studies, then the present correlations might be interpreted as the expected size of effect for such scales. Acton and Moretti (1991) found the average correlation

between drawing scales and independent measures of anxiety in three studies to be .46, with correlations ranging from .28 to .76. Other drawing scale studies have generally performed in this range. Correlations, or average correlations if multiple independent measures were used, range from a low of .13 for a drawing scale measure of body disturbance (Carlson, Quinlan, Tucker, & Harrow, 1973) to a high of .64 for a drawing scale measure of schizophrenia (Holzberg & Wexler, 1950). The results reported in the present study fall within this range, suggesting that they are consistent with results obtained by other authors seeking to develop aggregate drawing scales of specific psychopathological conditions.

This comparison with previous findings suggests that the present research has been no more successful than past studies at developing measures of specific forms of psychopathology. Use of meta-analytic techniques in the selection of IDFs has not provided the improvement in rigor necessary to produce drawing scales possessing superior psychometric properties. Perhaps this is because the low number of findings per feature precluded reliable testing for moderating effects for age of sample and study quality. Alternatively, meta-analytic techniques may be no more able to appropriately assess and screen drawing features than most past research. Thus, while the meta-analysis of past studies may prove useful in generating features for study, it does not appear to provide the necessary rigor for selecting drawing scale members.

As a final point, it is important to acknowledge the context in which the present results should be interpreted. Study 2 is a study of scale development. Cureton (1950) makes a strong argument that results of an initial investigation involving scale development cannot be interpreted. Using the results from an experiment in which chance was allowed to determine the selection of scale items, he was able to show that the methods commonly used in selecting items for a scale can lead to apparently strong results

in the absence of meaningful relationships between items and criteria. He concluded that, "[w]hen a validity coefficient is computed from the same data used in making an item analysis, this coefficient cannot be interpreted uncritically... [it] cannot be interpreted 'with caution'... [in fact] [t]here is only one clear interpretation for all such validity coefficients. This interpretation is -- 'Baloney'" (Cureton, 1950, p. 96).

Cureton's (1950) arguments produce a logical dilemma. If validity coefficients in initial scale development studies cannot be interpreted, how does one justify further research? One cannot. The escape from this dilemma is the recognition that statements of validity are probabilistic, reflecting the likelihood that a scale is valid, not whether it is valid. Conclusions drawn about validity are made with more or less confidence depending on the degree to which chance is expected to determine results. Cureton has not proven that interpretations of validity coefficients in initial scale development research are "baloney". What he has done is to provide evidence that chance can be a large determining factor in the results of such studies.

The message from Cureton's (1950) data is clear: there is always the possibility that initial findings are a product of chance. In the case of Study 2 many of the items chosen for the scales had already established themselves as valid measures of the constructs in previous research. Their continued valid performance as scale members is a replication of past performance and, therefore, increases the confidence that results were not simply due to chance. However, while confidence is increased in the present set of studies because of such replication, the degree to which chance has contributed to the findings is still unknown. The only way to determine whether the scales developed herein are robust is to repeat the analyses on an independent sample of subjects. Study 3 reports a confirmatory analysis with a second sample of subjects from Youth Court Services.

Study 3

The previous study developed eight drawing scales. Multiple regression analyses used to select the items produced small to moderate multiple correlations with criterion measures which achieved p values of ≤ .05. Examination of the adjusted R^2s associated with these regression analyses suggests that the observed multiple correlations are not solely a product of the number of items employed. The correlational analyses of the drawing scales and the independent measures showed reasonable convergent and discriminant validity. These results, however, are still open to criticism. Statistics such as the adjusted R^2 are only estimates of the potential of a set of variables to meaningfully correlate with an independent measure. The "true" ability of these scales to correlate with the independent measures is suggested, but not confirmed, by that statistic. Further, both the item analyses and the validation of the drawing scales were carried out on the same sample, increasing the likelihood that the research has capitalized on chance.

Stronger evidence of the validity of the drawing scales would come from a confirmatory analysis which replicates the correlational results of the previous study on an independent sample. According to Anastasi (1982), retesting a new scale on a new group of subjects provides good evidence of the robustness of the initial findings. This is because the chance factors which may have inflated results in the initial study are unlikely to perform in the same manner in the new one. If the results of the first study are replicated in the second, then one can conclude that it is unlikely that chance was responsible for the initially observed results.

In the present study a second sample of young offenders gathered around the same time period as the original was used to confirm the observed relationships between drawing scales and independent measures. It was expected that the correlations among

the independent measures, the drawing scales, and the drawing scales with the independent measures would replicate the results of the second study.

Method

Participants. A power analysis was carried out to determine the number of subjects needed to achieve significance in Study 3. For this analysis a two-tailed test was employed, with alpha set at .05 and power at .80. The anticipated true correlation was assumed to range around .25. Based on this analysis a sample size of 150 was chosen for each of the two age groups.

Subjects were selected from among 686 admissions to the same young offender program as was used in Study 2, Youth Court Services. All subjects were between 13 and 19 years of age, had been charged with at least one criminal offense, and were referred for psychological or psychiatric services. Selection criteria for subjects were largely the same as those employed in Study 2, i.e., subjects had to have completed the core tests of the psychological assessment battery used at the center and have drawn a valid human figure. The additional constraint of creating equal sized groups, however, led to a somewhat greater percentage of files being rejected in this study than in Study 2, because the quota of 150 files was filled more quickly for the younger age sample than the older one. This meant that a number of complete files for subjects 15 years of age and younger had to be excluded along with the usual number of incomplete files. As in the previous study, however, the majority of subjects were rejected because their files contained unscorable drawings or were incomplete.

Procedure. Drawings used in the present study were scored using the manual developed in Study 2. The same independent measures were employed in this confirmatory analysis as were used in the previous study. These independent measures included the four criterion scales: the Ho scale (Cook & Medley, 1954); the MAS (Taylor,

1953); and the Sc subscale (Hathaway, 1956) from the MMPI (Hathaway & McKinley, 1967); and the SM subscale from the JI (Jesness, 1983). Discriminant scales included the Rep subscale from the JI (Jesness, 1983) and FSIQ scores from either the WISC-R (Wechsler, 1974) or the WAIS-R (Wechsler, 1981). The analysis in the present study replicated the correlational comparisons carried out in Study 2.

Results

As in Study 2, separate correlational analyses were carried out for the independent measures, the drawing scales, and the drawing scales with the independent measures. The correlational analysis was repeated for the independent measures to document that the same pattern of relationships would be observed in this study as in the previous one. The expected pattern was replicated (see Tables 15 and 16).

The next series of analyses examined the pattern of correlations among the drawing scales. It was expected that the drawing scales would mirror the results observed among the criterion measures and that, based on the results of Study 2, the magnitude of correlations among drawing scales would be small. The observed pattern of correlations among drawing scales for both age groups is shown in Table 17. In the previous analysis all but two of the drawing scales were significantly correlated with criterion measures and the magnitude of statistically significant correlations fell in the range of .15 to .52. As can be seen from the results presented in the table, eight of the correlations expected to be positive and significant were so, only two less than that observed in Study 2. Further, the magnitude of significant correlations ranged from .16 to .45, indicating similar effect sizes in the two studies.

The final analysis examined the correlations among the drawing scale scores and scores from the criterion and the discriminant measures for the two age groups. These results are reported in Tables 18 and 19. The correlations in the two tables are, for the

most part, very small in size and of lower magnitude than those found in Study 2. The largest significant correlation was between the drawing scale measure of thought disorder and the Sc subscale of the MMPI, at $r = .22$. Only 3 of 32 correlations among drawing scales and criterion measures achieved significance. These findings are quite different from the previous analysis, in which all of the correlations among these measures were significant, and only partially replicate the pattern of intercorrelations observed among the criterion measures.

A further step in analyzing the present results was to see how many of the correlations between measures of like constructs reached significance. Of the eight correlations--four each at the two age levels--one correlation was significant, that between the drawing scale and self-report measures of thought disorder. The thought disorder drawing scale had the highest correlation with its criterion measure in the previous analysis, though at that time it was for the 15 and younger age group, while the present finding was in the 16 and older age group.

Discussion

The results of Study 3 provide a partial replication of the findings of Study 2. The pattern of intercorrelations among the self-report measures replicated that observed between these measures in the previous study. The criterion measures of psychopathology correlated highly with each other, while demonstrating low or nonsignificant correlations with discriminant measures of IQ and repression. Intercorrelations among drawing scales were significant in eight out of a total of 12 comparisons, which is a reduction of two from the results of Study 2. Further, the magnitude of observed correlations was roughly equal in magnitude across the two studies. These findings suggest that Study 3 was able to replicated the pattern of intercorrelation among drawing scales. It was most important to replicate results regarding the correlations between the drawing scales and the independent

measures. Results of the present study indicate that three of the drawing scales correlated significantly with criterion measures, while two of the scales correlated significantly with discriminant measures. This is a marked drop in the number of statistically significant correlations.

Ideally, the few significant correlations in this study would all have been between drawing scales and criterion measures of the same construct. Only one of the significant correlations, however, was between two measures of the same construct. Even this lone finding, however, should be viewed positively. The items in this scale have been found to validly measure thought disorder in past research and the scale itself has achieved significance in two studies.

The significant correlations among drawing scales and criterion measures of constructs other than the one targeted by the drawing scale should also be considered; for example, the correlation between the drawing scale measure of social maladjustment and the Hostility Scale. The review of the literature on drawing measures of anger/hostility suggested that it might be antisocial values and not anger that was being measured in this group of studies. The items on the drawing scales of anger/hostility and social maladjustment are also very similar. Given the high potential for overlap between the two constructs, and certainly between the two drawing scales, it is not surprising that one of the observed significant correlations is between the drawing scale measure of anger/hostility and the self-report measure of social maladjustment. Similar arguments could be made for the observed correlation between the drawing scale measure of thought disorder and the Manifest Anxiety Scale in the 16 year old and older group. Measures of thought disorder and anxiety demonstrated high intercorrelations throughout Study 2.

The magnitude of the correlations in Study 3 are very small. Correlations of this magnitude suggest that, even if the results represent robust phenomena, the observed

effects are so small as to make the scales of little utility. There are many other measures of the four constructs studied in this dissertation which have much higher validity coefficients and are much more likely to assist in making diagnostic statements. The results reported here suggest that even though the scales may have some validity, they would not prove to be useful psychometrically.

To understand why this confirmatory analysis failed to produce a larger number of significant correlations it may be informative to look at the work of Holzberg and Wexler (1950) and Wexler and Holzberg (1952). These authors provide the only previous confirmatory analysis of aggregate drawing scales developed to test a specific form of psychopathology, and their replication was successful. Initial correlations between drawing scales and group membership in Holzberg and Wexler (1950) were .64 for a schizophrenia scale, .60 for a paranoid schizophrenia scale, and .47 for a hebephrenic schizophrenia scale. When replicated in Wexler and Holzberg (1952), the observed correlations were .58 and .29 for the schizophrenia and paranoid schizophrenia scales, respectively.

The superior results found in the work of Holzberg and Wexler (1950; 1952) were likely a product of the empirical approach employed by the authors when developing their scales. The initial investigation by these authors involved the generation of 174 possible drawing features. These features were then tested individually using criterion-keying methods to determine which IDFs discriminated between criterion groups. Those features which differentiated between groups at the $p \leq .05$ level of significance were chosen for scale membership. This method is very similar in rigor to the techniques employed by authors such as Hiler and Nesvig (1965), Koppitz (1968), and Naglieri, McNeish, and Bardor (1991), all of whom have developed aggregate drawing scales which have reliably demonstrated their ability to screen for psychopathology. It is likely that the present study

did not achieve similar levels of performance because its grounding in the meta-analysis simply did not provide sufficient rigor, despite expectations to the contrary.

It is interesting to note, given the methods employed by Holzberg and Wexler (1950), that the thought disorder drawing scale which was replicated in the present study derived many of its features from the scales originally developed by these authors. Research methods, such as meta-analysis, are dependent on the quality of past studies (Rosenthal, 1991). It is possible that, in the case of the thought disorder scale, the analysis carried out in Study 1 may have capitalized on the previous work of Holzberg and Wexler, thereby creating a more valid scale.

The present findings suggest that the development of scales for human figure drawings is best carried out using rigorous empirical methods. Even the use of meta-analytic techniques to improve the process of selecting results from previous studies is not sufficient for the task of selecting potential drawing scale items. Scales developed without the benefit of the increased rigor of such methods are not likely to prove highly valid. Certainly, the drawing scales developed in the current work have not done so.

General Discussion

Two broad questions were addressed in the three studies reported in this dissertation. The first was whether IDFs are valid measures of psychopathology. The second was whether aggregate drawing scales could be developed which provide highly valid measurement of specific psychopathological states or conditions.

The answer to the first question is a qualified yes. The meta-analysis of previous research carried out in Study 1 showed that a large number of features correlated with independent measures of several constructs (i.e., anger/hostility, anxiety, and thought disorder). These findings suggest that IDFs have greater validity than had previously been thought (Kahill, 1984; Roback, 1968; Swensen, 1957, 1968). The most likely reason that

the current results are more favorable is that previous reviewers tested the potential of IDFs via featural hypotheses, while Study 1 aggregated and tested drawing features according to the construct investigated. Differences between the results of the two review methods should not, however, be considered evidence of contradiction. The two review methods provide different perspectives. Combining these perspectives, I would conclude that while there is little evidence for the validity of traditional human figure drawing hypotheses, IDFs do correlate with measures of relatively robust and well established constructs.

The small size of the average correlations in the meta-analysis leads to a further qualification regarding the validity of IDFs. To be considered a highly valid test of psychopathology, correlations with independent measures should be large, or at least in the upper end of the moderate range. Correlations with criterion measures in the meta-analysis ranged as high as .67, which falls within this range, but were on average .28. This magnitude of correlation suggests that IDFs will not make highly valid tests. Taking into account the psychometric properties of IDFs, however, it would seem that they would make valid items on aggregate drawing scales.

The answer to the second question, whether aggregate drawing scales can be developed which provide highly valid measurement of specific psychopathological states or conditions, is also a qualified yes. One drawing scale, the thought disorder scale for subjects age 16 and older, correlated significantly with its corresponding criterion measure in Studies 2 and 3. Two other drawing scales correlated with criterion measures in both studies, though not with measures of the constructs they were designed to assess. This is not a strong show of convergent validity, and the small number of significant results suggests that few valid scales may have been developed. It is even possible that these results were simply chance findings and that none of the scales are valid. While this latter

possibility must be entertained, the replication of findings across two studies and the selection of IDFs based on significant findings in the meta-analysis suggests that at least one or two scales are valid. The observed validity coefficients are so small, however, that the scales developed are not likely to prove useful in assessing psychopathology. Thus, while the results of the second and third studies suggest that valid scales may be developed, the performance of these scales suggests they are of little utility.

Magnitude of Effect Size in Studies of Drawing Scales

Few studies that have developed drawing scales to measure specific forms of psychopathology have found large effect sizes. The magnitude of effect size for most studies, including Study 2, falls in the small to moderate range. One interpretation of these findings is that drawing scales, while more reliable than IDFs, are still inherently low in validity when measuring specific constructs. This interpretation would be in keeping with the perspective presented by Tharinger and Stark (1990) who stated that drawing scales enable us to make only gross distinctions, such as the presence or absence of psychopathology.

Small effect sizes in studies of drawing scales may also reflect common methodological problems, particularly regarding feature selection. Practices employed by researchers seeking to develop drawing scales for the assessment of specific psychopathological states are usually informal. The present study sought to use meta-analysis to improve on the process of item selection, thereby producing psychometrically superior drawing scales. The failure of the present research to improve on past findings suggests that the use of the meta-analysis does not lead to sufficient increases in reliability and validity and that, in fact, reliance on past research may be useful only for the purposes of item generation.

It is possible that use of even more rigorous methods in the selection of IDFs would lead to the desired increases in validity. The researchers who initially proposed that better quality item selection would lead to stronger performance, Hiler and Nesvig (1965), Koppitz (1968), and Naglieri, McNeish, and Bardos (1991), employed quite rigorous methods. Hiler and Nesvig (1965) selected only those items which differentiated between adolescents receiving psychiatric treatment and nonpatient high school students. Koppitz (1968) chose features which were statistically unusual in a sample of 1,800 children, were unrelated to normal cognitive development, and predicted whether or not a child had been seen in a mental health clinic. Naglieri et al. (1991) selected their features based on past performance in research studies, frequency of occurrence in a sample of over 2,300 children and adolescents, and psychometric appropriateness. It would appear that to reliably achieve similar levels of success when developing drawing scales for specific psychopathological constructs, research methods will have to approach more closely those employed in these relatively more successful measures of general psychopathology.

The work of Holzberg and Wexler (1950) and Wexler and Holzberg (1952) re-affirms the conclusion that more rigor is needed in scale development. These authors used criterion-keying in selecting IDFs for their schizophrenia scales. Their studies produced initial correlations with independent criteria ranging from .47 to .64 and subsequent correlations in a partial replication ranged from .29 to .58. These results, while reflecting the work of only one set of authors, suggest that development of drawing scales for specific states or conditions could be enhanced by the adoption of more rigorous psychometric practices.

It is also possible that a method factor has not been taken into account when studying drawing scales. All of the five studies discussed above which successfully developed drawing scales employed comparison groups to test the validity of their

measures. Further, none of the comparison groups used to test validity were derived based on self-report measures. In contrast, the vast majority of studies which have sought to develop tests of specific forms of psychopathology have employed such measures extensively.

The differences in the magnitude of effect size observed between the two groups of studies may be a result of such methodological differences. Human figure drawings are nonverbal tools which seek to measure psychopathology through the graphic productions of self-projection (Koppitz, 1968; Machover, 1949). Drawings differ from self-report measures in several important ways. For one, expressing one's psychological difficulties nonverbally through drawings may differ in important ways from reading and responding to self-descriptive statements. Drawings are also projective devices and are expected to be influenced by unconscious factors more than conscious ones. Again this is very different from the reading and decision-making required by a self-report measure which relies on effortful, conscious processing. A multi-trait, multi-method study employing other projective measures, or measures of different kinds such as behavioral surveys or interviews, would provide one means of assessing the impact of these method factors.

There is a third possible explanation for the observed magnitude of correlations. The method used in the present study, of analyzing the drawing of a single human figure, potentially produces a measure of weak reliability which provides a poor demonstration of the validity of such tests. Two of the four studies listed above which demonstrated strong correlations between their scales and independent measures employed more than one drawing in their assessment (i.e., Hiler & Nesvig, 1965; Naglieri et al., 1991). Interestingly, given this observation, Koppitz (1966c) and Holzberg and Wexler (1950) reported the lowest effect sizes of the four groups of authors. In contrast, the majority of authors reporting research demonstrating only small to moderate correlations with

independent measures employ single drawings in their research protocols. The results of the present study and those of other authors studying specific drawing scales may not be more robust because the protocols used produced insufficient material for valid scoring.

Future Research

The previous section suggests several future directions for drawing scale research. One direction is towards more rigorous, empirically-based methods of scale development. Researchers who make use of methods such as criterion-keying in selecting IDFs for inclusion on drawing scales will hopefully form drawing scales with sounder psychometric properties. Methods used in construct validation research could also be adopted. Such methods involve the clear definition of what it is that is being measured, multiple independent and dependent measures from several methodological and constructional domains, and clear statements about how the different forms of assessment should inter-relate. A research design such as this could provide rigorous item selection, test for method factors such as the differential effect of employing drawing scale versus self-report measures, and improve our understanding of how projective measures relate to psychopathological constructs.

Future research could also investigate changes in method which might enhance the reliability and validity of drawing scales. Increasing the number of drawings used in the basic protocol is one means to increase the reliability of drawing scales. Another option is to use human figure drawings with different demand characteristics. Some authors have included a third drawing in their protocol. Handler & Rehyer (1964, 1966) used a drawing of an automobile as their third drawing. Naglieri, McNeish, & Bardos (1991) used the drawing of oneself, a particularly interesting variation as it may increase the projective value of the test and thereby improve validity.

Another possible change in protocol would be the inclusion of a post-drawing interview. Buck (1948) included a structured interview following the completion of the drawings as part of the administration of the House-Tree-Person test. Similarly, it is commonly recommended that an interview follow the Draw-A-Person test (Groth-Marnat, 1984). To test whether such changes in the drawing protocol would increase the validity of the drawing assessment, information from the interview could be scored and added to the test results of the drawings. The utility of adding an interview to the test could then be assessed directly by measuring the increase in reliability or validity afforded by the inclusion of the interview data.

Conclusions

The present findings suggest that there may be ways to develop valid measures of specific psychopathological conditions through human figure drawings. Results of the meta-analysis carried out in Study 1 indicated that IDFs are more valid than they have been thought in the past, though their utility is likely limited to that of items on drawing scales. Findings in Studies 2 and 3 indicate that drawing scales assessing specific psychopathological states can be formed. Weak replication of the key results in Study 3, however, would suggest that the present methods of scale development are not likely to produce highly valid measures.

The use of meta-analytic techniques to select drawing features does not improve the validity of drawing scales above that of other studies employing less rigorous methods. Speculation regarding the reasons for the failure to find stronger effect sizes in both the present and past studies suggests that more research is needed. Too little is known about the factors which determine the reliability and validity of human figure scoring systems. Performance of drawing scales developed to screen for psychopathology, such as Naglieri, McNeish, and Bardos' (1991) DAP:SPED, and Holzberg and Wexler's (1950)

schizophrenia drawing scales would suggest, however, that the adoption of empirical methods of feature selection may lead to the production of drawing scales with greater validity. Good quality research, possibly employing methodology borrowed from construct validation studies, is needed to provide strong criteria with which to select IDFs for scale membership and to determine whether method factors may lead to underestimations of the potential of drawing scales.

References

Note: References marked with an asterisk indicate studies included in the meta-analysis.

Achenbach, T. M., & Edelbrock, C. S. (1986). Child Behavior Checklist and Youth Self-Report. Burlington, VT: Author.

Acton, B. V., & Moretti, M. M. (1991, June). A meta-analytic study of diagnostic indicators in figure drawings. Presented at the Annual Conference of the Canadian Psychological Association, Calgary, Alberta.

Anastasi, A. (1982). Psychological Testing (5th ed.). New York: Macmillan.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. Psychological Bulletin, 99, 388-399.

Bieliauskas, V. J. (1960). Sexual identification in childrens' drawings of human figures. Journal of Clinical Psychology, 39, 1033-1034.

Brackbill, G. & Little, K. B. (1954). MMPI correlates of the Taylor scale of manifest anxiety. Journal of Consulting Psychology, 18, 433-436.

Buck, J. N. (1948). The H-T-P technique: A qualitative and quantitative scoring manual. Journal of Clinical Psychology, 4, 317-396.

Bull, R. J. C., & Strongman, K. T. (1971). Anxiety, neuroticism, and extroversion. Psychological Reports, 29, 1101-1102.

Calhoun, F., Jr., Ross, J. L., & Bolton, J. A. (1988). Relationship between human figure drawings and self-esteem. Perceptual and Motor Skills, 66, 253-254.

Calhoun, F., Jr., Whitley, J. D., & Ansolabehere, E. M. (1978). An investigation of the Goodenough-Harris Drawing Test and the (Coopersmith) Self-Esteem Inventory. Educational and Psychological Measurement, 38, 1229-1232.

Carlson, K., Quinlan, D., Tucker, G., & Harrow, M. (1973). Body disturbance and sexual elaboration factors in figure drawings of schizophrenic patients. Journal of Personality Assessment, 37, 56-63.

* Cauthen, N. R., Sandman, C. A., Kilpatrick, D. G., & Deabler, H. L. (1969). DAP correlates of Sc scores on the MMPI. Journal of Projective Techniques and Personality Assessment, 33, 262-264.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. Journal of Abnormal Psychology, 100, 316-336.

* Clodfelter, D. L., & Craddick, R. A. (1970). Variance in size of drawing in a psychotic population. Perceptual and Motor Skills, 30, 110.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.

Cook, W. W., and Medley, D. M. (1954). Proposed hostility and pharisaic-virtue scales for the MMPI. Journal of Applied Psychology, 38, 414-418.

Correll, R. E. (1985). Relationship of anxiety and depression scores to WAIS performance of psychiatric patients. Psychological Reports, 57, 295-301.

* Craddick, R. A., Leipold, W. D., & Cacavas, P. D. (1962). The relationship of shading on the Draw-A-Person test to manifest anxiety scores. Journal of Consulting Psychology, 26, 193.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cureton, E. E. (1950). Validity, reliability, and baloney. Educational and Psychological Measurement, 10, 94-96.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). An MMPI handbook: Vol II. Research applications (Rev. ed.). Minneapolis MN: University of Minnesota Press.

* Daum, J. M. (1983). Emotional indicators in drawings of aggressive or withdrawn male delinquents. Journal of Personality Assessment, 47, 243-249.

DiLeo, J. H. (1973). Children's drawings as diagnostic aids. New York: Brunner/Mazel.

* Doubros, S. G., & Mascarenhas, J. (1967). Effect of test produced anxiety on human figure drawings. Perceptual and Motor Skills, 25, 773-775.

Engle, P. L., & Suppes, J. S. (1970). The relation between human figure drawing and test anxiety in children. Journal of Projective Techniques and Personality Assessment, 34, 223-231.

* Exner, J. E., Jr. (1962). A comparison of the human figure drawings of psychoneurotics, character disturbances, normals, and subjects experiencing experimentally-induced fear. Journal of Projective Techniques, 26, 392-397.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33, 613-619.

* Fox, C., Davidson, K., Lighthall, F., Waite, R., & Sarason, S. B. (1958). Human figure drawings of high and low anxious children. Child Development, 29, 297-301.

* Goldstein, H. S. (1972). Gender identity, stress and psychological differentiation in figure-drawing choice. Perceptual and Motor Skills, 35, 127-132.

* Goldstein, H. S., & Faterson, H. F. (1969). Shading as an index of anxiety in figure drawings. Journal of Projective Techniques and Personality Assessment, 33, 454-456.

Goodenough, F. (1929). The measurement of intelligence by drawings. Yonkers, NY: World Book.

Gotlib, I. H. (1984). Depression and general psychopathology in university students. Journal of Abnormal Psychology, 93, 19-30.

Graham, S. A. (1981). Predictive and concurrent validity of the Jesness Inventory Asocial Index: When does a delinquent become a delinquent? Journal of Consulting and Clinical Psychology, 49, 740-742.

* Greenberg, R. P., & Fisher, S. (1991). Some differential effects of music on projective and structured psychological tests. Psychological Reports, 28, 817-818.

Greene, R. L. (1989). The MMPI: An interpretive manual. Orlando, FL: Grune and Stratton.

Greenglass, E. R., & Julkunen, J. (1989). Construct validity and sex differences in Cook-Medley Hostility. Personality and Individual Differences, 10, 209-218.

* Griffith, A. V., & Lemley, D. W. (1967). Teeth and threatening look in the Draw-A-Person test as indicating aggression. Journal of Clinical Psychology, 23, 489-492.

* Griffith, A. V., & Peyman, D. A. R. (1959). Eye-ear emphasis in the DAP as indicating ideas of reference. Journal of Consulting Psychology, 23, 560.

Groth-Marnat, G. (1984). Handbook of Psychological Assessment. New York: van Nostrand Reinhold.

Groves, J. R., & Freid, P. A. (1991). Developmental items on children's human figure drawings: A replication and extension of Koppitz to younger children. Journal of Clinical Psychology, 47, 140-148.

Hammer, E. F. (1954). Guide for qualitative research with the H-T-P. Journal of General Psychology, 51, 41-60.

Hammer, E. F. (1958). The clinical application of projective drawings. Springfield, IL: Thomas.

* Handler, L., & Rehyer, J. (1964). The effect of stress on the Draw-A-Person Test. Journal of Consulting Psychology, 28, 259-264.

Handler, L., & Rehyer, J. (1965). Figure drawing anxiety indexes: A review of the literature. Journal of Projective Techniques and Personality Assessment, 29, 305-313.

* Handler, L., & Rehyer, J. (1966). Relationship between GSR and anxiety indices in projective drawings. Journal of Consulting Psychology, 30, 60-67.

Hathaway, S. R. (1956). Scale 5 (masculinity-femininity), 6 (paranoia), and 8 (schizophrenia). In G. S. Welsh, & W. G. Dahlstrom (eds.), Basic reading on the MMPI in psychology and medicine (pp.). Minneapolis: University of Minnesota Press.

Hathaway, S. R., & McKinley, J. C. (1967). MMPI manual (Rev. ed.). New York: Psychological Corporation.

* Haworth, M. R. (1962). Responses of children to a group projective film and to the Rorschach, CAT, Despert Fables and D-A-P. Journal of Projective Techniques, 26, 47-60.

Hiler, E. W., & Nesvig, D. (1965). An evaluation of criteria used by clinicians to infer pathology from figure drawings. Journal of Consulting Psychology, 29, 520-529.

Hilgard, E. R., Jones, L. V., & Kaplan, S. J. (1951). Conditioned discrimination as related to anxiety. Journal of Experimental Psychology, 42, 94-99.

Holzberg, J. D., & Wexler, M. (1950). The validity of human form drawings as a measure of personality deviation. Journal of Projective Techniques, 14, 343-361.

* Hoyt, T. E., & Baron, M. R. (1959). Anxiety indices in same-sex drawings of psychiatric patients with high and low MAS scores. Journal of Consulting Psychology, 23, 448-452.

Jesness, C. (1983). The Jesness Inventory (Rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

* John, K. B. (1974). Variations in bilateral symmetry of human figure drawings associated with two levels of adjustment. Journal of Clinical Psychology, 30, 401-404.

* Johnson, J. H. (1971). Note on the validity of Machover's indicators of anxiety. Perceptual and Motor Skills, 33, 126.

* Johnson, J. H. (1971). Upper left hand placement of human figure drawings as an indicator of anxiety. Journal of Personality Assessment, 35, 336-337.

Jolles, I. (1971). A catalogue for the qualitative interpretation of the H-T-P (Revised). Los Angeles: Western Psychological Services.

Kahill, S. (1984). Human figure drawing in adults: An update of the empirical evidence, 1967-1982. Canadian Psychology, 25, 269-292.

* Kay, S. R. (1978). Qualitative differences in human figure drawings according to schizophrenic subtype. Perceptual and Motor Skills, 47, 923-932.

Keilin, W. G., & Bloom, L. J. (1986). Child custody evaluation practices: A survey of experienced professionals. Professional Psychology: Research and Practice, 17, 338-346.

* Kokonis, N. D. (1972). Body image disturbance in schizophrenia: A study of arms and feet. Journal of Personality Assessment, 36, 573-575.

Koppitz, E. M. (1966a). Emotional Indicators on human figure drawings and school achievement of first and second graders. Journal of Clinical Psychology, 22, 481-483.

Koppitz, E. M. (1966b). Emotional Indicators on human figure drawings of children: A validation study. Journal of Clinical Psychology, 22, 313-315.

* Koppitz, E. M. (1966c). Emotional Indicators on human figure drawings of shy and aggressive children. Journal of Clinical Psychology, 22, 466-469.

Koppitz, E. M. (1968). Psychological evaluation of children's human figure drawings. New York: Grune and Stranton.

Koppitz, E. M. (1984). Psychological evaluation of human figure drawings of middle school pupils. New York: Grune and Stratton.

* Kurdek, L. A., & Darnell-Goetschell, G. (1987). Young adolescents' human figure drawings as indicators of psychopathology. Journal of Adolescent Research, 2, 69-74.

Lapkin, B., Hillaby, T., & Silverman, L. (1968). Manifestations of the schizophrenic process in figure drawings of adolescents. Archives of General Psychiatry, 19, 465-468.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Machover, K. (1949). Personality projection in the drawings of the human figure. Springfield, IL: Thomas.

* Maloney, M. P., & Glasser, A. (1982). An evaluation of the clinical utility of the Draw-A-Person Test. Journal of Clinical Psychology, 38, 183-190.

Manning, T. M. (1987). Aggression depicted in abused children's drawings. The Arts in Psychotherapy, 14, 15-24.

Martin, R. D. (1981). Cross-validation of the Jesness Inventory with delinquents and non-delinquents. Journal of Consulting and Clinical Psychology, 49, 10-14.

Marzolf, S. S., & Kirchner, J. H. (1970). Characteristics of House-Tree-Person drawings by college men and women. Journal of Projective Techniques and Personality Assessment, 34, 138-145.

• Mogar, R. E. (1962). Anxiety indices in human figure drawings: A replication and extension. Journal of Consulting Psychology, 26, 108.

Molden, S. O. & Gottesman, I. I. (1991). Replicated psychometric correlates of schizophrenia. American Jounral of Psychiatry, 148, 762-767.

Mullen, B. (1989). Advanced BASIC meta-analysis. Hillsdale, NJ: Erlbaum.

Munson, R. F., & Blincoe, M. M. (1984). Evaluation of a residential treatment center for emotionally disturbed adolescents. Adolescence, 19, 253-261.

Naglieri, J. A., McNeish, T. J., & Bardos, A. N. (1991). Draw A Person: Screening Procedure for Emotional Disturbance. Austin, TX: ProEd.

Naglieri, J. A., & Pfeiffer, S. I. (1992). Performance of disruptive behavior disordered and normal samples on the Draw A Person: Screening Procedure for Emotional Disturbance. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 4, 156-159.

Newmark, C. S. & Gentry, L. (1983). Utility of MMPI indices of schizophrenia with adolescents. Journal of Clinical Psychology, 39, 170-172.

Nunnally, J. C. (1978). Psychometric Theory. New York: McGraw Hill.

Patrick, J. (1988). Concordance of the MCMI and the MMPI in the diagnosis of three DSM-III Axis I disorders. Journal of Clinical Psychology, 44, 186-190.

Piotrowski, C., & Keller, J. W. (1984). Attitudes toward clinical assessment by members of the AABT. Psychological Reports, 55, 831-838.

Piotrowski, C., Sherry, D., & Keller, J. W. (1985). Psychodiagnostic test usage: A survey of the Society for Personality Assessment. Journal of Personality Assessment, 49, 115-119.

Pope, M. K., & Smith, T. W. (1990). Cognitive, behavioral, and affective correlates of the Cook and Medley Hostility Scale. Journal of Personality Assessment, 54, 501-514.

* Prytula, R. E., & Hiland, D. N. (1975). Analysis of general anxiety scale for children and Draw-A-Person measures of general anxiety level of elementary school children. Perceptual and Motor Skills, 41, 995-1007.

* Reznikoff, M., & Dies, R. R. (1969). The use of clothing in human figure drawings. Journal of Clinical Psychology, 25, 80-81.

* Reznikoff, M., & Tomblen, D. (1956). The use of human figure drawings in the diagnosis of organic pathology. Journal of Consulting Psychology, 20, 467-470.

* Ries, H. A., Johnson, M. H., Armstrong, H. E., Jr., & Holmes, D. S. (1966). The Draw-A-Person Test and process-reactive schizophrenia. Journal of Projective Techniques and Personality Assessment, 30, 184-186.

Roback, H. B. (1968). Human figure drawings: Their utility in the clinical psychologist's armamentarium for personality assessment. Psychological Bulletin, 70, 1-19.

Roberts, G., Schmitz, K., Pinto, J., & Cain, S. (1990). The MMPI and Jesness Inventory as measures of effectiveness on an inpatient conduct disorders treatment unit. Adolescence, 25, 989-996.

Rosenthal, R. (1991). Meta-analytic procedures for social research. Newbury Park, CA: SAGE Publications.

Rothermal, R. D., Jr. (1985). A comparison of the utility of the Personality Inventory for Children and the Jesness Inventory for assessing juvenile delinquents. (Doctoral Dissertation, Wayne State University, 1985). Dissertation Abstract International, 46 (5-B), 1740.

* Royal, R. E. (1949). Drawing characteristics of neurotic patients using a drawing-of-a-man-and-woman technique. Journal of Clinical Psychology, 5, 392-395.

Sattler, J. M. (1982). Assessment of children's intelligence and special abilities (2nd ed.). Boston: Allyn and Bacon.

Saunders, G. R., & Davies, M. B. (1976). The validity of the Jesness Inventory with British delinquents. British Journal of Social and Clinical Psychology, 15, 33-39.

Shaffer, J. B., Duszynski, K. R., & Thomas, C. B. (1984). A comparison of three methods for scoring figure drawings. Journal of Personality Assessment, 42, 363-369.

*Shafranske, E. P. (1981). A concurrent validity study of hostility indicators on the Draw-A-Person test of adolescents. Unpublished doctoral dissertation, United States International University, San Diego.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

* Silverstein, A. B. (1966). Anxiety and the quality of human figure drawings. American Journal of Mental Deficiency, 70, 607-608.

Sims, J., Dana, R. H., & Bolton, B. (1983). The validity of the Draw-A-Person Test as an anxiety measure. Journal of Personality Assessment, 47, 250-257.

Singh, A. (1983). Validity of Jesness Inventory with Indian delinquents. Indian Journal of Clinical Psychology, 10, 485-489.

Smith, T. W. & Frohm, K. D. (1985). What's so unhealthy about hostility? Construct validity and psychosocial correlates of the Cook and Medley Ho Scale. Health Psychology, 4, 503-520.

Smith, T. W., & Sanders, J. D. (1990). What does the Cook and Medley Hostility scale measure? Affect, behavior, and attributions in the marital context. Journal of Personality and Social Psychology, 58, 699-708.

Snyder, R. T., & Gaston, D. S. (1970). The figure drawing of the first grade child -- item analysis and comparison with Koppitz norms. Journal of Clinical Psychology, 26, 377-383.

Soccolich, C., & Wysocki, B. A. (1967). Draw-A-Person protocols of male and female college students. Perceptual and Motor Skills, 25, 873-879.

Spiga, R., Mindingall, M. P., Long-Hall, C., & Blackwell, M. W. (1986). Effects of differences between self- and other's ratings on children's human figure drawings. Perceptual and Motor Skills, 62, 956-958.

Sturner, R. A., Rothbaum, F., Visintainer, M., & Wolfer, J. (1980). The effects of stress on children's human figure drawings of man, woman, and self. Journal of Clinical Psychology, 36, 324-341.

Sunberg, N. D. (1961). The practice of psychological testing in clinical services in the United States. American Psychologist, 16, 79-83.

* Swartz, J. D., Laosa, L. M., & McGavern, M. L. (1976). Spatial placement of human figure drawings as an indicator of cognitive and personality characteristics among normal young adolescents. Journal of Consulting and Clinical Psychology, 44, 307-308.

Swensen, C. H. (1957). Empirical evaluations of human figure drawings. Psychological Bulletin, 54, 431-466.

Swensen, C. H. (1968). Empirical evaluations of human figure drawings: 1957-1966. Psychological Bulletin, 70, 20-44.

Taylor, J. A. (1953). A personality scale of manifest anxiety. Journal of Abnormal and Social Psychology, 48, 285-290.

Tharinger, D. J., & Stark, K. (1990). A qualitative versus quantitative approach to evaluating the Draw-A-Person and Kinetic Family Drawings: A study of mood- and anxiety-disorder children. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2, 365-375.

Thomas, G. V., & Silk, A. M. J. (1990). An introduction to the psychology of children's drawings. New York: Harvester Wheatsheaf.

Vallance, R. C., & Forrest, A. R. (1971). A study of the Jesness Personality Inventory with Scottish Children. British Journal of Educational Psychology, 41, 338-344.

Wade, T. C., & Baker, T. B. (1977). Opinions and use of psychological tests: A survey of clinical psychologists. American Psychologist, 32, 874-882.

Wade, T. C., Baker, T. B., Morton, T. L., & Baker, L. J. (1978). The status of psychological testing in clinical psychology: Relationships between test use and professional activities and orientations. Journal of Personality Assessment, 42, 3-10.

Walters, G. D. (1984). Identifying schizophrenia by means of Scale 8 Sc of the MMPI. Journal of Personality Assessment, 48, 390-391.

• Wainwright, B. B. (1970). Quantitative scales for scoring human figure drawings. Doctoral dissertation, University of California, Los Angeles.

Watkins, C. E., & Campbell, V. L. (1989). Personality assessment and counseling psychology. Journal of Personality Psychology, 53, 296-307.

Watson, D. C., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. Psychological Bulletin, 96, 465-490.

Wechsler, D. (1974). Manual for the Wechsler Intelligence Scale for Children - Revised. New York: The Psychological Corporation.

Wechsler, D. (1981). Manual for the Wechsler Adult Intelligence Scale - Revised. New York: The Psychological Corporation.

Wenck, L. S. (1977). House-Tree-Person drawings: An illustrated diagnostic handbook. Los Angeles: Western Psychological Services.

Wexler, M., & Holzberg, J. D. (1952). A further study of the validity of human form drawings in personality evaluation. Journal of Projective Techniques, 16, 249-251.

Footnotes

[1] Average correlations reported for all features and for the subset of significant features are product-moment correlations transformed from an unweighted mean z_r calculated from the relevant results.

Table 1

Summary Conclusions Regarding Drawing Features by Kahill (1984), Roback (1968), and Swensen (1957, 1968)

Author	Supportive Evidence	Conflicting Evidence	Not Supported	Not Tested
Swensen (1957)	Neck	Action Buttons Eye Facial Expression Hair Hands and Arms Lips Mouth Perspective Size Stance Type of lines Waist	Anatomy Breasts Ears Erasure Fingers Head Joints Legs and Feet Mid-line Nose Placement Sexual treatment Shading Succession Toes	Chin Clothing Eyebrows Hips and buttocks Pockets Shoe and hat Shoulders Symmetry Theme Tie Trunk
Swensen* (1968)	Delineation line Distortion Line Pressure	Arms Belt Body Breasts Buttons Detail in Figure Elbow Erasures Eyelashes Eyes Fingers Hair Hands Head length Head size Head/body ratio Heels Height Hips and buttocks	Asymmetry Ears Face Feet Head Perspective Sex symbols or organs Shading	

Table 1 Cont'd

Author	Supportive Evidence	Conflicting Evidence	Not Supported	Not Tested
Swensen (1968) Cont'd		Knee Legs Limb size Line discontinuity Line emphasis Line heaviness Line sketchy Mouth Neck Nose Nude Omission Placement Reinforcement Sex drawn first Shoulders Size Stance Teeth Toes Transparency		
Roback (1968)	Hair Joints Type of line	Clothing Ears Eyes Head Sexual treatment Shading Size	Erasure Eyebrows Facial Expression Hands and arms Hips and buttocks Legs and Feet Lips Nose Placement Waistline	Actions Anatomy Breasts Chin Midline Mouth Perspective Shoulders Stance Succession Trunk

Table 1 Cont'd

Author	Supportive Evidence	Conflicting Evidence	Not Supported	Not Tested
Kahill (1984)	Color Trunk	Ambiguous figures Breasts Clothing/nudity Contact features Detailing Eyes Facial Expression Head size Line characteristics Mouth and teeth Shading Size Symmetry	Anatomy indicators Distortion Ears Erasure Eyebrows Face Hair Omission Perspective Placement Props and themes Sex of first drawn figure Shoulders Stance Transparency	

*Studies were deemed supportive if three of every four studies or more found the feature to be significant, deemed largely unsupportive if three of every four studies or more found the feature not to have reached significance, and deemed to be inconsistent if the ratio of studies fell somewhere between these two extremes.

Table 2

The Number of Supported Features by Level of Support in Other Reviews by Review

	Studies			
	Swensen (1957)	Swensen (1968)	Roback (1968)	Kahill (1984)
Supported in at Least 1 Other Review	0	0	0	0
Inconsistently Supported in at Least 1 Other Review	1	0	2	1
Not Supported in at Least 1 Other Review	0	1	2	0
Not Tested in at Least 1 Other Review	1	2	2	1

Note. The number of drawing features achieving supportive evidence in each study is: Swensen (1957) - 1, Swensen (1968) - 3, Roback (1968) - 3, and Kahill (1984) - 2.

Table 3

Product-Moment Correlations and Percent Agreement Between Raters One and Two for Construct Studied, Moderator Variables, Effect Size, and Significance

	Statistic		
Variable	Correlation	Kappa	Percent Agreement
Construct	.98	.92	94%
Moderator variables			
Age of sample	1.00	1.00	100%
Reliable IDF coding	.87	.86	94%
Valid Independent Measure	.75	.75	88%
Inferential Variables			
Effect Size			88%
Significance			88%

Table 4

Characteristics of Studies Selected for Inclusion in the Meta-Analysis

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
1.	Cauthen, Sandman, Kilpatrick & Deabler (1960)	Article	Thought Disorder	36 adult psychiatric patients	Compared drawing feature scores of patients scoring above 70 on Sc subscale (MMPI) and those showing no scale elevation.	Valid scale	1	inter-class corr. = .85
2.	Clodfelter & Craddick (1970)	Article	Thought Disorder	155 adult males: Psychotics, Normals	Compared drawing feature scores of psychotic patients and normals.	No evidence of validity was given for diagnoses	1	Assumed reliable
3.	Craddick, Leipold & Cacavas (1962)	Article	Anxiety	272 male and female college freshmen	Compared drawing feature scores of males and females, where females had significantly higher scores on the Manifest Anxiety Scale.	Valid scale	1	80.5% agreement between raters

Table 4 cont'd.

Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
4. Daum (1983)	Article	Anger/ Hostility	400 adolescent delinquents	Compared drawing feature scores of groups of delinquents (aggressive, withdrawn, and undifferentiated) and non-delinquents, defined based on case reviews.	No evidence of validity was provided for the review process	16	avg. $\underline{r}$ = .92
5. Doubros & Mascarenhas (1967)	Article	Anxiety	203 14 year old, high school students	Compared drawing feature scores before and after an exam.	No manipulation check done	10	None established
6. Exner (1962)	Article	Anxiety	80 self-referred adults suffering from anxiety or depression	Compared drawing feature scores of 1) anxious & depressed clients, 2) anxious clients, 3) subjects with induced anxiety, and 4) controls.	No evidence of validity was provided for the self-referred clients.	12	None established

Table 4 cont'd.

Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
7. Fox, Davidson, Lighthall, Waite, and Sarason (1958)	Article	Anxiety	747 children in grades 1 to 4	Compared drawing feature scores of subjects in the first and fourth quartile of scores on the Test Anxiety and General Anxiety scales.	Valid scale	6	inter-rater r ranged from .77 to .92; avg. r = .85
8. Goldstein (1972)	Article	Anxiety	28 adult nightshift workers	Compared drawing feature scores of subjects who viewed an anxiety provoking versus a neutral film.	Manipulation check using Nowlis Mood Checklist & Gottschalk ratings	1	Assumed reliable
9. Goldstein and Faterson (1969)	Article	Anxiety	28 adult nightshift workers.	Compared drawing feature scores of subjects which viewed either an anxiety provoking film or a neutral film.	Manipulation check using Nowlis Mood Checklist & Gottschalk ratings	2	Avg. inter-rater r = .90

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
10.	Greenberg & Fisher (1971)	Article	Anger/ Hostility	40 adult female respondents to an ad for volunteers	Compared drawing feature scores of subjects which listened to either exciting or calm music.	Manipulation check using semantic differential ratings of music	2	Assumed reliable
11.	Griffith & Lemley (1967)	Article	Anger/ Hostility	90 educably mentally retarded group home members, age 8 to 20 years; avg. age 14 years	Compared ratings of aggressivity of students who drew 1) no aggressive features, 2) drew teeth, 3) drew slash mouth, or 4) drew teeth & slash mouth.	No evidence of validity was provided for the staff ratings	2	Avg. percent agreement = 94%
12.	Griffith & Peyman (1959)	Article	Thought Disorder	76 adult psychiatric patients	Compared rating of ideas of reference by psychologist for patients who had 1) eye or ear emphasis or 2) no eye or ear emphasis.	No evidence of validity was provided for the ratings of the psychologist	1	Used only those drawings on which raters agreed

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
13.	Handler & Rehyer (1964)	Article	Anxiety	57 male college students	Compared drawing feature scores of subjects which received a stress induction and subjects which did not.	Observed signs of anxiety after manipulation	42	Percent agreement ranged from 67% to 100%; avg. = 87%
14.	Handler & Reyher (1966)	Article	Anxiety	96 male college students	Correlated drawing feature scores with GSR frequency and average conductance.	Assumed to reflect anxiety	46	Percent agreement ranged from 67% to 97%; avg. = 87%
15.	Haworth (1962)	Article	Anxiety	30 children in grades 1 and 2	Compared drawing feature scores of the 15 most anxious and the 15 least anxious students. Level of anxiety was determined using the Rock-A-Bye Baby projective film.	No evidence of the validity of the film was provided	11	None established

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
16.	Holzberg & Wexler (1950)	Article	Thought Disorder	78 female student nurses and 38 females with a diagnosis of schizophrenia	Compared drawing feature scores of student nurses with hebephrenic, paranoid, and undifferentiated schizophrenics.	No evidence of the validity of the diagnoses was provided	174	Used only those drawings on which raters agreed
17.	Hoyt & Baron (1959)	Article	Anxiety	112 female neurotic and psychotic patients between 20 and 65 years of age	Compared drawing feature scores of patients in the top and bottom quartiles of scores on the Manifest Anxiety Scale.	Valid scale	10	None established
18.	John (1974)	Article	Thought Disorder	60 nonpatients, avg. age = 23 years, and 60 schizophrenic inpatients, avg. age = 28 years	Compared drawing feature scores of nonpatients and patients.	No evidence of the validity of the diagnoses was provided	9	Inter-rater r ranged from .84 to .99

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
19.	Johnson (1971a)	Article	Anxiety	103 first year university students	Correlated drawing feature scores with IPAT Anxiety scale scores.	Valid scale	3	Used only those drawings on which raters agreed
20.	Johnson (1971b)	Article	Anxiety	103 first year university students	Correlated drawing feature scores with IPAT Anxiety scale scores.	Valid scale	1	Used only those drawings on which raters agreed
21.	Kay (1978)	Article	Thought Disorder	136 schizophrenic inpatients, 18 to 50 years of age	Compared drawing feature scores of the 3 groupings of schizophrenics: paranoid vs. not; acute vs. chronic; nuclear vs. schizophreniform. The acute vs. chronic results were employed in the meta-analysis.	Confirmed psychiatric diagnosis, but without measures of concordance	7	Inter rater $\underline{r}$ ranged from .56 to .96; avg. $\underline{r}$ = .75

Table 4 cont'd.

Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
22. Kokonis (1972)	Article	Thought Disorder	104 normal males and 128 schizophrenic males, 21 to 39 years of age	Compared drawing feature scores of normals and schizophrenic patients.	No evidence of validity of the diagnoses was provided	2	None established
23. Koppitz (1966c)	Article	Anger/ Hostility	62 children attending a child guidance clinic, 5 to 12 years of age	Compared drawing feature scores of patients with a history of aggression and patients with a history of shyness. Patients in the 2 groups were matched for age, sex, & IQ.	No evidence of validity of the history was provided	30	Found reliable in past study
24. Kurdek and Darnell-Goetschell (1987)	Article	Anger/ Hostility; Anxiety	29 grade 7 students, avg. age = 12.8 years, and 15 grade 9 students, avg, age =15.1 years	Correlated drawing feature scores with scores on the Symptom Checklist 90 - Revised.	Valid scale	21	Avg. percent agreement = 94%; all agreement above 90%

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
25.	Maloney & Glasser (1982)	Article	Thought Disorder	5 groups of 7 subjects: first year and grad. students, and nonpsychotic, psychotic, and retarded patients; avg. age 21 to 32	Compared drawing feature scores across groups. Comparisons between the student groups and the psychotic patient group were used in the meta-analysis.	No evidence of the validity of the diagnoses was provided	9	Inter-rater r ranged from .42 to .78; avg. r = .67
26.	Mogar (1962)	Article	Thought Disorder	123 adult male psychiatric inpatients	Correlated drawing feature scores with Manifest Anxiety scale & Rorschach Content scale scores.	Valid scale	10	Inter-rater r ranged from .84 to 1.00
27.	Prytula & Hiland (1975)	Article	Anxiety	300 grade 5 and 6 students	Compared drawing feature scores of 30 hi anxious and 30 low anxious students, based on scores on the General Anxiety Scale for Children.	Valid scale	12	Some assumed reliable, none established as reliable

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
28.	Reznikoff & Dies (1969)	Article	Thought Disorder	100 inpatients, 15 to 65 years of age, in 3 groups: 1) personality disordered, 2) neurotic, & 3) schizophrenic	Correlated drawing feature scores with scores on the clinical and the special scales of the MMPI. Compared drawing feature scores of the 3 psychiatric groups.	Valid scale (MMPI), but no evidence of the validity of the diagnoses was provided	1	Percent agreement = 80%
29.	Reznikoff & Tomblen (1956)	Article	Thought Disorder	75 inpatients, 19 to 55 years of age (avg. = 35.4), in 3 groups: 1) neurotic, 2) organic, and 3) schizophrenic	Compared drawing feature scores of the 3 patient groups. The neurotic vs. schizophrenic results were used in the meta-analysis.	No evidence of the validity of the diagnoses was provided	15	Avg. percent agreement = 91%

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
30.	Ries, Johnson, Armstrong, & Holmes (1966)	Article	Thought Disorder	32 normal males, 32 reactive schizophrenics, and 32 process schizophrenics, between 19 and 50 years of age	Compared drawing feature scores of normals, reactive schizophrenics, & process schizophrenics. Subjects were matched on age, education, vocabulary, & SES. The Phillips Prognostic scale was used to differentiate process and reactive schizophrenics.	Valid scale, but no evidence of the validity of the diagnoses was provided	80	Described as reliable, but no evidence of reliability given in study
31.	Royal (1949)	Article	Anxiety	100 non-anxious dental clinic patients and 80 veterans with anxiety disorders	Compared drawing feature scores of veterans with anxiety problems and controls.	No evidence of the validity of the diagnoses was provided	22	None established

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
32.	Shafranske (1981)	Article	Anger/ Hostility	75 students 16 to 17 years of age, divided into 2 groups which received an anger induction and 2 groups which served as controls	Compared drawing feature scores of subjects receiving an anger induction or not. Drawing feature scores correlated with Hostility scales (Taylor-Johnson Temperament Analysis and the Holtzman Inkblot test).	Valid scale	7	Percent agreement ranged from 61% to 91%; avg. = 79%
33.	Silverstein (1966)	Article	Anxiety	80 hospitalized mentally retarded children, with an avg. age of 13 years	Compared drawing scale scores of children scoring above or below the median on the Test Anxiety Scale for Children, in tough vs. tender examiner conditions.	Valid scale, but no manipulation check was carried out	1	Found reliable in past study

Table 4 cont'd.

	Study	Source	Construct Studied	Sample Characteristics	Study Design	Validity of Ind. Meas.	# of Features	Reliability of Scoring
34.	Swartz, Laosa, & McGavern (1976)	Article	Anxiety	129 13 year old students	Compared Sarason Test Anxiety Scale for Children scores of students at different levels of scoring of the drawing feature.	Valid scale	1	Assumed reliable
35.	Wainwright (1970) study 1	Disser-tation	Anger/ Hostility, Anxiety, Thought Disorder	95 first year university students	Correlated drawing feature scores with clinical subscale scores of the MMPI.	Valid scale	30	Avg. inter-rater r = .99
36.	Wainwright (1970) study 2	Disser-tation	Anger/ Hostility, Anxiety, Thought Disorder	166 first year university students	Correlated drawing feature scores with clinical subscale scores of the MMPI.	Valid scale	30	Avg. inter-rater r = .99

Table 5

Number of Individual Drawing Features by Construct Investigated for Which Results are Based on a Sample of One, for Which Homogeneity was Found, and for Which Significance was Found Reported by Sample Size

	Constructs		
	Anger/Hostility	Anxiety	Thought Disorder
Sample Size			
n of 1	20	45	193
n > 1	49	52	22
Homogeneity			
Homogeneous	40	36	20
Nonhomogeneous	9	16	2
Significance			
n of 1			
Significant	5	18	48
Nonsignificant	15	27	144
n > 1			
Significant	10	19	8
Nonsignificant	34	26	14
Untestable	6	7	1

Table 6

Results of the Regression Analyses for Drawing Features Selected by the Meta-Analysis

Independent Measure	N	# of Features	Multiple R	R-Square	Adjusted R-Square	F(Eqn)	p
Anger/ Hostility	485	16	.17	.03	-.00	.87	ns
Anxiety	485	31	.25	.06	-.00	.99	ns
Social Mal-adjustment	485	16	.16	.03	-.01	.74	ns
Thought Disorder	485	61	.38	.15	.02	1.18	ns

Table 7

Results of the Regression Analyses for Drawing Features Selected Based on Meta-Analytic Findings Which Significantly Contribute to the Regression Equation by Construct Studied

Independent Measure	N	# of Features	Multiple R	R-Square	Adjusted R-Square	F(Eqn)	p
Anger/ Hostility	485	1	.09	.01	.01	3.63	.06
Anxiety	485	5	.20	.04	.03	3.85	.00
Social Maladjust.	485	2	.12	.01	.01	3.39	.03
Thought Disorder	485	9	.26	.07	.05	3.83	.00

Table 8

Results of the Regression Analyses Using All Drawing Features by Construct Studied

Independent Measure	N	# of Features	Multiple R	R-Square	Adjusted R-Square	F(Eqn)	p
Anger/ Hostility	485	10	.30	.09	.07	4.58	.00
Anxiety	485	13	.32	.10	.08	4.19	.00
Social Maladjust.	485	16	.35	.12	.10	4.14	.00
Thought Disorder	485	17	.36	.13	.10	4.12	.00

Table 9

Results of the Regression Analyses Using All IDFs by Age Group by Construct Studied

Independent Measure	N	# of Features	Multiple R	R-Square	Adjusted R-Square	F(Eqn)	p
15 Years and Younger							
Anger/Hos.*	253	14	.36	.13	.08	2.50	.002
Anxiety	253	14	.35	.13	.07	2.43	.003
Soc. Malad.	253	23	.48	.23	.16	2.93	.000
Tht. Dis.	253	17	.41	.17	.11	2.75	.000
16 Years and Older							
Anger/Hos.	232	13	.47	.22	.17	4.66	.000
Anxiety	232	15	.39	.15	.10	2.61	.001
Soc. Malad.	232	18	.47	.22	.16	3.35	.000
Tht. Dis.	232	20	.43	.19	.11	2.45	.00

Note. Anger/Hos. = Anger/Hostility, Tht. Dis. = Thought Disorder, and Soc. Malad. = Social Maladjustment.

Table 10

Correlations Among Independent Measures for 15 Year Old and Younger Subjects

	Independent Measures				
Independent Measures*	MAS	Sc	SM	Rep	FSIQ
Anger	.62*	.59*	.47*	-.15*	.00
MAS		.82*	.58*	-.04	-.15*
Sc			.61*	-.04	-.13*
SM				.06	.04
Rep					-.17*

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .05$

Table 11

Correlations Among Independent Measures for 16 Year Old and Older Subjects

	Independent Measures				
Independent Measures*	MAS	Sc	SM	Rep	FSIQ
Anger	.66*	.73*	.60*	-.30*	-.13
MAS		.83*	.69*	-.25*	-.23*
Sc			.74*	-.22*	-.24*
SM				-.19*	.18*
Rep					-.17*

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .05$

Table 12

Correlations Among Human Figure Drawing Scales by Age Group

	Drawing Scales		
Drawing Scales*	Anxiety	Thought Disorder	Social Mal-adjustment
15 and Younger			
Anger/Hostility	.13*	.15*	-.03
Anxiety		.49*	.08
Thought Disorder			.16*
16 and Older			
Anger/Hostility	.17*	.35*	.17*
Anxiety		.52*	.18*
Thought Disorder			.36*

*$p < .05$

Table 13

Correlations Among Self-Report Measures and Human Figure Drawing Scales for 15 Year Old and Younger Subjects

	Independent Measures*					
Drawing Scales	Hos	MAS	Sc	SM	Rep	FSIQ
Anger/ Hostility	.36*	.21*	.20*	.16*	-.07	-.08
Anxiety	.14*	.31*	.28*	.19*	.03	-.10
Thought Disorder	.16*	.25*	.38*	.20*	-.01	-.04
Social Mal-Adjustment	.15*	.15*	.16*	.31*	.05	.07

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .05$

Table 14

Correlations Among Self-Report Measures and Human Figure Drawing Scales for 16 Year Old and Older Subjects

	Independent Measures*					
Drawing Scales	Hos	MAS	Sc	SM	Rep	FSIQ
Anger/ Hostility	.37*	.15*	.18*	.15*	-.03	-.04
Anxiety	.17*	.28*	.22*	.20*	.01	-.22*
Thought Disorder	.29*	.32*	.31*	.26*	-.12	-.21*
Social Mal-Adjustment	.16*	.14*	.21*	.34*	-.04	-.04

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .05$

Table 15

Correlations Among Independent Measures for 15 Year Old and Younger Subjects in the Confirmatory Sample

	Independent Measures				
Independent Measures*	MAS	Sc	SM	Rep	FSIQ
Anger	.63*	.64*	.45*	-.14	-.05
MAS		.81*	.64*	-.21*	-.14
Sc			.66*	-.16*	-.17*
SM				-.14	-.17*
Rep					-.16*

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .05$

Table 16

Correlations Among Independent Measures for 16 Year Old and Older Subjects in the Confirmatory Sample

	Independent Measures				
Independent Measures*	MAS	Sc	SM	Rep	FSIQ
Anger	.64*	.66*	.61*	-.14	-.05
MAS		.88*	.67*	-.10	-.09
Sc			.72*	-.15	-.13
SM				-.13	.01
Rep					-.28*

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .05$

Table 17

Correlations Among Human Figure Drawing Scales in the Confirmatory Analysis by Age Group

	Drawing Scales		
Drawing Scales*	Anxiety	Thought Disorder	Social Mal-adjustment
15 and Younger			
Anger/Hostility	.08	.16*	.04
Anxiety		.45*	.18*
Thought Disorder			.32*
16 and Older			
Anger/Hostility	.22*	.36*	-.09
Anxiety		.36*	.31*
Thought Disorder			.12

*$p < .05$

Table 18

Correlations Among Self-Report Measures and Human Figure Drawing Scales for 15 Year Old and Younger Subjects in the Confirmatory Analysis

	Independent Measures*					
Drawing Scales	Hos	MAS	Sc	SM	Rep	FSIQ
Anger/ Hostility	-.13	-.05	-.06	-.04	.15	-.07
Anxiety	.02	.01	.02	.01	-.09	-.33*
Thought Disorder	-.04	-.03	-.09	-.01	.05	-.15
Social Mal-adjustment	.05	.10	.08	.04	-.01	.03

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .001$; this value is still significant when bonferroni corrections are applied.

Table 19

Correlations Among Self-Report Measures and Human Figure Drawing Scales for 16 Year Old and Older Subjects in the Confirmatory Analysis

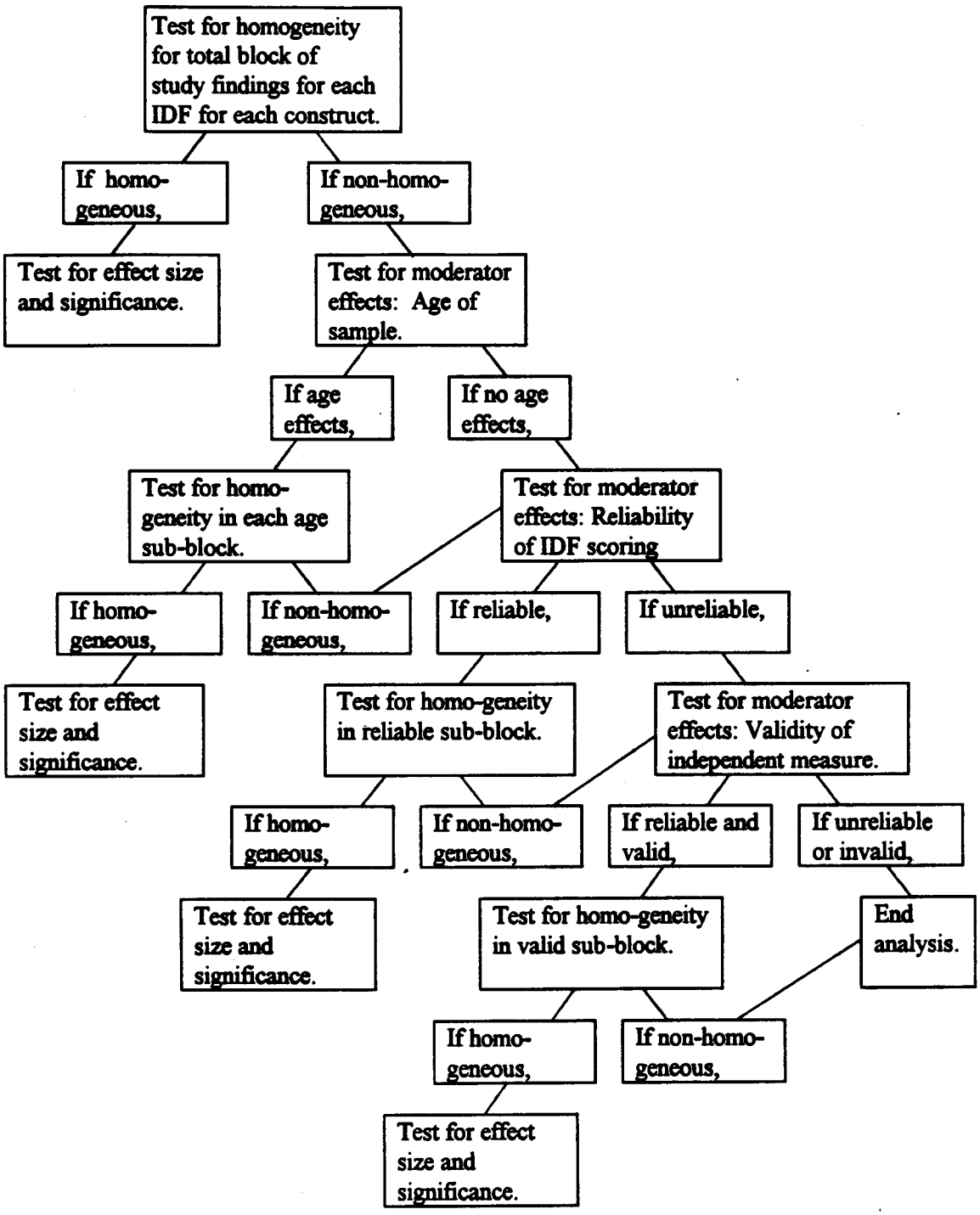
	Independent Measures*					
Drawing Scales	Hos	MAS	Sc	SM	Rep	FSIQ
Anger/ Hostility	.06	.15	.14	.19*	.03	.00
Anxiety	.04	.06	.08	.09	.04	-.03
Thought Disorder	-.01	.16*	.22*	.14	.04	-.23*
Social Mal-adjustment	.02	.10.	.11	.10	-.07	-.01

Note. Hos = Hostility Scale; MAS = Manifest Anxiety Scale; Sc = Sc Subscale (MMPI); SM = Social Maladjustment subscale; Rep = Repression Subscale; FSIQ = Full Scale IQ.

*$p < .05$; No correlations were significant after bonferroni corrections were applied.

Figure Caption

Figure 1. Sequence of analysis for moderator variables.


Test for homogeneity for total block of study findings for each IDF for each construct.
If homo-geneous,
If non-homo-geneous,
Test for effect size and significance.
Test for moderator effects: Age of sample.
If age effects,
If no age effects,
Test for homo-geneity in each age sub-block.
Test for moderator effects: Reliability of IDF scoring
If homo-geneous,
If non-homo-geneous,
If reliable,
If unreliable,
Test for effect size and significance.
Test for homo-geneity in reliable sub-block.
Test for moderator effects: Validity of independent measure.
If homo-geneous,
If non-homo-geneous,
If reliable and valid,
If unreliable or invalid,
Test for effect size and significance.
Test for homo-geneity in valid sub-block.
End analysis.
If homo-geneous,
If non-homo-geneous,
Test for effect size and significance.

Appendix A

Study Coding Manual: Human Figure Drawing Meta-Analysis

The following is the coding manual employed in Study 1 to code drawings for the meta-analysis. It reflects the decision process and criteria employed in the coding of the studies employed in the meta-analysis and that used in the reliability analysis carried out on a subsample of the studies. The manual is broken into sections describing coding for each of the variables coded in the meta-analysis. There are also subsections with information relevant to the reliability analysis specifically.

Construct

The term construct refers to the theoretical entity which the study purports to measure. For the purposes of the Study 1 only three constructs are of interest; anger/hostility, anxiety, and thought disorder. It is also necessary for the reliability analysis to code for a fourth class of variables -- all other constructs. This latter class is simply referred to as "other" constructs, or as "other".

Several, assumedly, interchangeable terms can be used to identify each of the three constructs of primary interest. The terms aggression, anger, and hostility are descriptors which indicate that a study is assessing the construct anger/hostility. Terms such as delinquents or conduct disorder are not seen as indicating the assessment of anger/hostility, unless subsequent statements regarding groups described with these terms also indicates that they are selected for acting out in an aggressive or angry fashion. The terms anxiety (and its derivatives) and neuroticism (and its derivatives) are considered to be descriptors which implied the study of the construct anxiety. The terms thought disorder, schizophrenia (and its derivatives), and psychoticism (and its derivatives) are considered to be descriptors of the construct thought disorder.

Studies will usually be coded as measuring one of anger/hostility, anxiety, thought disorder, or other. However, the occasional study uses a complex design which assesses

more than one construct. All relevant constructs should be coded (e.g., anxiety/ thought disorder/ other). To determine which constructs are being studied the coder should look first to the authors' stated focus. For example, the title or introduction indicates that the study focuses on anxiety in children. If the focus of the study is still unclear the coder should look next at the types of comparisons or correlations being made. For example, a study comparing schizophrenics with student nurses would be deemed a study of thought disorder. Independent measures can also be useful here. If the author is using the State-Trait Anger Inventory, the construct is likely anger/hostility.

One type of research design will present particular problems to the rater. This unusual case is that of research which purports to study psychopathology in general and uses several pathological comparison groups. Particularly problematic in this regard are studies which compare neurotics and psychotics, and those which study psychotics and organic patients. The difficulty is that it is not clear whether differences may be attributed to neuroticism, psychoticism, or organic problems, and, thus, it is impossible to say which of the constructs is the focus of the study. For these studies the rule of thumb is that if the comparison is psychotics/schizophrenics with neurotics or personality disordered groups, the study is of thought disorder, while any comparison between organics and any other sample is deemed a study of organicity. One of the assumptions here is that anxiety pervades neurotic and psychotic disorders, while thought disorder distinguishes the latter. Any differences, therefore, are most likely attributable to that difference. Organicity is chosen as the pre-eminent construct because of the pervasive effects of such a condition and the problems of distinguishing the results of organics and psychotics. The rule of thumb used in the present research is that groups with assumedly lesser pathology should serve as controls for more pathological groups. Following these rules, neurotics would

serve as controls for psychotics, who would themselves serve as controls for organic patients.

Age of Subjects

An important moderator variable is age of the sample. In the meta-analysis the age range of the subjects is broken into three groups: children (age 4 to 12), adolescents (age 13-18), and adults (age 19 and older). This information is usually found in the Method section under the subtitle subjects. Occasionally, the age of the subjects will range across the boundaries of the groupings listed here. When this is the case, the average of the group and the relative amount of subjects falling within a category should be taken into account when selecting one age group.

Reliability of Drawing Features

Reliability of drawing features is scored for the features used as a whole. That is, the rater must judge whether the features as a body have been scored reliably. Fortunately, where results are reported they are either given as a global score (e.g., the average percent agreement was never less than ...) or individual reliability or percentage agreements are provided and readily averaged. In most cases the coder's decision is not whether features are reliably rated or not, but whether reliability is known. Unknown reliability is treated as an absence of reliability.

Some features may not have reported reliabilities, but be assumed to reliable. These include physical measurements (e.g., height, width, area, placement on page, etc.) and sex of the first drawn figure. The latter is considered reliable only if the average age of subjects is 10 years of age or older.

Reliability and Validity of the Independent Variable

The assessment of the reliability and the validity of an independent measure can be set out clearly and easily. A potentially more difficult task is the identification of the independent measure. Many studies include numerous measures and multiple comparisons. The coder must determine based on their assessment of which construct is being measured, which of the authors' manipulations or other measures assesses the construct. For the purposes of the reliability analysis the identified manipulation or instrument is recorded along with its reliability and validity.

Once the independent measure is identified its reliability and validity should be assessed using the following decision rules.

Reliability

An independent measure is coded as reliable, if

1) it is a commonly used standardized test (e.g., the MMPI), or

2) it is a less commonly used test for which the authors provide references indicating the test's reliability, or

3) it is ratings/diagnoses based on a standardized instrument with known reliability, or

4) it is ratings/diagnoses for which the reported inter-rater reliability within the article is .70 or greater.

Validity

An independent measure is coded as valid, if

1) it is a commonly used standardized test known to assess the construct(s) of interest in the study, or

2) it is a less commonly used standardized test for which the authors provide references indicating the test's validity, or

3) it is ratings/diagnoses based on a standardized instrument with known validity as an assessment of the construct(s) of interest in the study, or

4) it is ratings/diagnoses for which the study provides independent confirmation of their validity, in the form of concurrent or predictive validity coefficients which are moderate to large in size.

Study Results

The most difficult task facing the coder is the selection of study results. Findings in a study are often numerous. Moreover, results reported may involve a number comparisons or correlation with independent measures which may vary in the validity with which they test hypotheses regarding the construct of interest. Finally, authors sometimes include several analyses of the same basic findings, or test several converging pieces of data which are all relevant to the construct of primary interest.

In selecting results the coder must first decide which comparisons or correlations are relevant tests of the relationship between drawing features and the construct of interest in the study. Relevant tests are comparisons of groups high and low on the construct of interest, comparisons of diagnostic groups high on the construct of interest with normal controls and correlations of drawing feature scores with independent measures of the construct.

There can be several complications within the design of a study which make identification of relevant tests difficult. One complication is the inclusion of several pathological comparison groups without a clear rationale as to how or why they might differ. This has already been discussed as a problem when trying to identify the construct studied. The best rule of thumb is to decide which construct is being studied using the decision rules in that section and then look for the comparison between groups which is

most likely to produce differences along the dimension tapped by the construct. For example, in a study assessing acutely psychotic, neurotic, and chronic schizophrenic groups, the best comparison would likely be that between acutely disturbed and neurotic patients because differences between those groups are most likely to be as a result of thought disorder. However, if the groups were acutely psychotic, chronic schizophrenic, and undifferentiated schizophrenics (falling somewhere in between the other two groups), the choice would be to use the results from comparisons of the acute and chronic groups. The rule, therefore, is to select differences which are most likely to tap the widest possible differences between groups in terms of the construct studied.

Complications are also introduced when more than one potentially meaningful test is carried out. The following rules can be used in deciding which results to use. The common theme across these rules is to use the largest estimate of the relationship, except where the sample used in separate tests is identical.

1) If there is more than one relevant value for the same sample, then the reported values are transformed to correlations or z-scores and averaged.

2) If there is more than one relevant value for different but potentially equally meaningful comparisons, then the largest reported value is used. For example, a study using two schizophrenic groups and a normal comparison group might find the t-test for height to be largest in comparisons with the first schizophrenic group, while the t-test for width is largest with the second schizophrenic group. The coder would use the value from the first comparison for height and for the second for width.

3) In either of the above cases, when one or more of the results, but not all, are reported as non-significant and a specific level of significance or effect-size is not provided with these results, they are not used. For example, if one result is

reported to be at the $p<.04$ and two others are reported as non-significant, then only the .04 value is transcribed.

For the purposes of the reliability analysis, the identified results are recorded along with the results themselves. This allows for a test of the results selection procedure which is more meaningful than the simple transcription of results from the study onto a coding sheet.

Recording Effect Size

Record the exact effect size of all drawing features reported in the study in their original form following the selection procedures described above. In the case where results are reported as simply "significant" or "non-significant", significance is transformed to the correlational equivalent of $p=.05$ one tailed and non-significance is transformed to a correlation of .00. Transform all effect-size indicators to r, z_r, and r^2.

Recording Significance

Record the exact level of significance of all drawing features reported in the study. Transform all significance levels to z-score equivalents. The reporting of significance with no attached value is assumed to be $p=.05$ and non-significance to be $p=.50$.

Appendix B

Tables of the Results of the Meta-Analysis for Each Individual Drawing Feature by Construct

Table B.1

Summary of Meta-Analytic Results for Individual Drawing Features Assessed as Indicators of Anger/Hostility

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z^*	z_r	z	p	z_r	r	
Angle of Extremities	ns	ns	.000	.500	.000	.000	2
Arm:Leg	ns	ns	.000	.500	.000	.000	2
Arms							
(Clinging)	ns	ns	.704	.242	.076	.076	2
(Location)	ns	ns	.000	.500	.000	.000	2
(Long)	ns	ns	1.593	.06	.207	.205	2
(Muscular)	ns	ns	.000	.500	.000	.000	1
(Short)	ns	ns	.000	.500	.000	.000	2
Asymmetry of Limbs	ns	ns	2.147	.016	.277	.270	1
Body Separations	ns	ns	.000	.500	.000	.000	2
Buttons (# of)	ns	ns	.000	.500	.000	.000	2
Chin (Emphasized)	ns	ns	.000	.500	.000	.000	1
Clothing (Amount of)	ns	ns	.000	.500	.000	.000	2
Continuity	ns	ns	.000	.500	.000	.000	2
Details (Total #)	ns	ns	.000	.500	.000	.000	2
Ears (Size of)	ns	ns	.000	.500	.000	.000	2
Eyes							
(Crossed)	ns	ns	.422	.337	.045	.045	2
(Size of)	ns	ns	.552	.291	.040	.040	2
(Type of)	sig	sig					2
Face Dim	ns	ns	2.008	.018	.141	.140	1
Feet							
(Emphasized)	ns	ns	.000	.500	.000	.000	1
(Pointed)	ns	ns	2.345	.009	.172	.170	1
Fingers (Talon-Like)	ns	ns	.857	.195	.070	.070	2
Genitals	ns	ns	1.549	.061	.203	.200	1

Table B.1 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z	z_r	z	p	z_r	r	
Hands							
(Big)	ns	ns	1.882	.003	.183	.182	2
(Size of Head)	ns	ns	.000	.500	.000	.000	1
(Fist)	ns	ns	.490	.312	.040	.040	2
Head							
(Large)	ns	ns	.000	.500	.000	.000	1
(In Profile)	sig	sig					3
(Small)	sig	sig					2
Head:Body	ns	ns	.000	.500	.000	.000	2
Height	ns	ns	2.639	.004	.203	.200	4
Integration (Poor)	ns	ns	.000	.500	.000	.000	1
Legs (Together)	ns	ns	.000	.500	.000	.000	1
Line Solidity	ns	ns	.000	.500	.000	.000	2
Monster	ns	ns	.422	.337	.045	.045	2
Mouth							
(Straight)	ns	ns	3.145	.001	.196	.195	3
(Type of)	ns	ns	.000	.500	.000	.000	2
Movement	ns	ns	.000	.500	.000	.000	2
Number of Figures (3+)	ns	ns	.000	.500	.000	.000	1
Objects							
(Clouds)	ns	ns	.469	.319	.050	.050	2
(Weapons)	ns	ns	2.252	.012	.266	.260	1
Omission							
(Arms)	ns	sig	1.341	.090			3
(Body)	ns	ns	.000	.500	.000	.000	1
(Eyes)	ns	ns	1.266	.102	.139	.138	2
(Facial Features)	ns	ns	2.184	.015	.151	.150	1
(Feet)	ns	ns	.891	.187	.096	.096	2
(Hands)	ns	ns	2.245	.012	.216	.215	2
(Legs)	ns	ns	.862	.195	.131	.130	1
(Mouth)	ns	ns	3.618	.000	.391	.375	2
(Neck)	ns	ns	.000	.500	.000	.000	1

Table B.1 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	$\underline{z}$	$\underline{z}_T$	$\underline{z}$	$\underline{p}$	$\underline{z}_T$	$\underline{r}$	
Omission (Nose)	ns	ns	2.526	.006	.257	.250	2
Placement	sig	ns			.349	.335	2
Posture	ns	ns	.000	.500	.000	.000	2
Sex of First Drawn Figure	ns	ns	.000	.500	.000	.000	2
Sex Rating	sig	ns			.205	.200	2
Shading (General)	sig	sig					2
(Body)	ns	ns	.798	.212	.086	.086	2
(Face)	ns	ns	.187	.425	.020	.020	2
(Hands)	ns	ns	1.407	.079	.155	.154	2
Shoulders (Squared)	ns	ns	2.066	.019	.111	.111	2
Size	sig	ns			.284	.275	2
Slanting Figure	ns	ns	.465	.319	.035	.035	4
Stance	sig	sig					4
Teeth	ns	ns	2.368	.009	.140	.139	5
Toes on Shoes	ns	ns	.000	.500	.000	.000	1
Transparency	ns	ns	.000	.500	.000	.000	1
Trunk:Leg	ns	ns	.000	.500	.000	.000	2
Width	ns	ns	.000	.500	.000	.000	1
Width of Foot	ns	ns	.000	.500	.000	.000	2
Wrinkles (# of)	sig	sig					2

* - Complete results are presented for non-significant diffuse tests only.

Table B.2

Summary of Meta-Analytic Results for Individual Drawing Features Assessed as Indicators of Anxiety

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	Z^*	Z_r	Z	p	Z_r	r	
Age of Figure Discrepant	sig	sig					2
Angle of Extremities	ns	ns	.000	.500	.000	.000	2
Arm:Leg	ns	ns	.000	.500	.000	.000	2
Arm (Length)	ns	ns	.652	.258	.057	.057	3
Arms							
(Down)	ns	ns	1.835	.033	.146	.145	3
(Location)	ns	ns	.000	.500	.000	.000	2
(Long)	ns	ns	.332	.371	.050	.050	1
(Short)	ns	ns	.796	.212	.121	.121	1
Ascendancy	ns	ns	.000	.500	.000	.000	1
Base-Line Drawing	ns	ns	.000	.500	.000	.000	1
Body							
(Area of)	ns	ns	.295	.382	.017	.017	3
(# of separations)	sig	sig					2
(Simple)	ns	sig	2.681	.004			2
Buttons (# of)	ns	ns	.000	.500	.000	.000	2
Clothing (Amount of)	ns	ns	.000	.500	.000	.000	3
Delineation Line Absent	ns	ns	2.414	.008	.216	.213	2
Detail							
(Loss)	ns	ns	3.485	.000	.497	.460	1
(Physical)	ns	ns	.000	.500	.000	.000	1
(Total #)	ns	ns	.000	.500	.000	.000	3
Distorted Figure	ns	ns	3.182	.001	.301	.292	2
Ears (Size of)	ns	ns	.000	.500	.000	.000	2
Erasures	ns	sig	2.584	.005			10
Eyes							
(Crossed)	ns	ns	.531	.298	.080	.080	1

Table B.2 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z	z_r	z	p	z_r	r	
Eyes							
(Size of)	sig	sig					2
(Type of)	sig	sig					2
Hands (Big)	ns	ns	.199	.421	.030	.030	1
Head							
(Large)	ns	ns	.492	.312	.050	.050	1
(Simple)	sig	sig					2
(Size)	ns	ns	.898	.184	.042	.042	5
(Tiny)	ns	ns	2.322	.010	.365	.350	1
Head:Body	ns	ns	1.691	.046	.164	.162	7
Height							
(General)	ns	ns	1.251	.106	.052	.052	6
(Large)	ns	ns	.833	.203	.090	.090	1
(Small)	ns	ns	2.649	.004	.194	.192	2
Humor	ns	ns	.000	.500	.000	.000	1
Incomplete	ns	ns	.000	.500	.000	.000	1
Line							
(Discontinuous)	ns	ns	1.390	.082	.075	.075	5
(Darkness)							
- Heavy Line	ns	ns	2.565	.005	.219	.216	2
- Light Line	ns	ns	1.418	.078	.088	.088	2
(Emphasis)	ns	ns	4.699	.000	.450	.430	2
(Pressure)							
- General	ns	ns	.000	.500	.000	.000	1
- Head Only	ns	ns	3.162	.001	.549	.500	1
- Heavy	ns	ns	1.470	.071	.151	.150	1
- Increasing	ns	ns	2.915	.002	.406	.385	1
- Light	ns	ns	1.274	.102	.131	.130	1
- Lower Body	ns	ns	3.668	.000	.662	.580	1
- Upper Body	ns	ns	2.909	.002	.497	.460	1
(Regularity)	ns	ns	.000	.500	.000	.000	1
(Reinforced)	ns	ns	5.487	.000	.266	.260	5

Table B.2 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	Z	Z_T	Z	p	Z_T	r	
Line							
(Sketchy)							
- General	ns	ns	.000	.500	.000	.000	1
- Head	ns	ns	2.087	.018	.343	.330	1
- Lower Body	ns	ns	4.111	.000	.775	.650	1
- Upper Body	ns	ns	3.733	.000	.678	.590	1
(Solidity)	ns	ns	.000	.500	.000	.000	2
Monster	ns	ns	2.123	.002	.332	.320	1
Mouth (Type of)	ns	ns	.000	.500	.000	.000	2
Movement	ns	ns	.158	.436	.013	.013	4
Mutilation	ns	ns	2.050	.020	.266	.260	1
Nudity	ns	ns	.000	.500	.000	.000	1
Objects	sig	sig					3
Omission							
(Arms)	ns	ns	.066	.472	.010	.010	1
(Eyes)	ns	ns	2.189	.014	.343	.330	1
(Facial Features)	ns	ns	.000	.500	.000	.000	1
(Feet)	ns	ns	.281	.390	.030	.030	2
(General)	ns	ns	2.367	.009	.101	.101	6
(Hands)	ns	ns	.115	.452	.010	.010	3
(Legs)	ns	ns	.199	.421	.030	.030	1
(Limbs)	ns	ns	.000	.500	.000	.000	1
(Mouth)	ns	ns	2.189	.014	.343	.330	1
(Nose)	ns	ns	2.322	.010	.365	.350	1
Only Head	ns	ns	.000	.500	.000	.000	1
Placement	sig	sig					9
Playful Figure	ns	ns	2.330	.010	.299	.290	1
Posture	ns	ns	.000	.500	.000	.000	2
Profile Drawn	ns	ns	4.364	.000	.848	.690	1
Proportion	ns	ns	.000	.500	.000	.000	1
Quality	ns	ns	1.455	.072	.117	.116	2
Rigid Figure	sig	ns			.150	.149	2

Table B.2 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z	z_r	z	p	z_r	r	
Sex of First Drawn Figure	ns	ns	3.084	.001	.135	.134	5
Sex Rating	ns	ns	.000	.500	.000	.000	2
Sexual Differentiation	ns	ns	1.960	.025	.255	.250	1
Shading							
(General)	sig	sig					10
(Hands)	ns	ns	4.179	.000	.741	.630	1
(Head)	ns	ns	4.225	.000	.280	.273	5
(Lower Body)	sig	sig					3
(Upper Body)	sig	sig					3
Shape of Head/Trunk	ns	ns	.000	.500	.000	.000	1
Smile	sig	ns			.150	.149	2
Stance	sig	sig					2
Stick Figure	ns	ns	1.530	.063	.110	.110	1
Teeth	ns	ns	.066	.472	.010	.010	1
Transparency	ns	ns	.687	.245	.040	.040	5
Trunk:Leg	sig	sig					2
Vertical Imbalance	ns	ns	2.084	.019	.114	.113	6
Width	ns	ns	.000	.500	.000	.000	1
Width of Foot	ns	ns	.000	.500	.000	.000	2
Wrinkles (# of)	ns	ns	.000	.500	.000	.000	2

* - Complete results are presented for non-significant diffuse tests only.

Table B.3

Summary of Meta-Analytic Results for Individual Drawing Features Assessed as Indicators of Thought Disorder

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z^*	z_T^*	z	p	z_T	r	
Adam's Apple	ns	ns	.000	.500	.000	.000	1
Arms							
(Asymmetry Length)	ns	ns	4.491	.000	.436	.410	1
(Asymmetry Width)	ns	ns	4.382	.000	.424	.400	1
(Behind Back)	ns	ns	2.330	.010	.224	.220	1
(Bent at Elbow)	ns	ns	.000	.500	.000	.000	1
(Down)	ns	ns	.000	.500	.000	.000	1
(Emphasized)	ns	ns	1.750	.040	.161	.160	1
(Length)	ns	ns	1.640	.050	.131	.130	1
(Long)	ns	ns	.000	.500	.000	.000	1
(Misplaced)	ns	ns	.000	.500	.000	.000	1
(Muscular)	ns	ns	.000	.500	.000	.000	1
(Out)	ns	ns	.000	.500	.000	.000	1
(Over Head)	ns	ns	.000	.500	.000	.000	1
(Perpendicular)	ns	ns	.000	.500	.000	.000	1
(Poor Proportion)	ns	ns	.000	.500	.000	.000	1
(Short)	ns	ns	.000	.500	.000	.000	2
(Sticks)	ns	ns	.000	.500	.000	.000	1
(Width)	ns	ns	4.382	.000	.424	.400	1
Bending	ns	ns	.000	.500	.000	.000	1
Bizarreness	ns	ns	1.560	.059	.175	.175	2
Body							
(Bulky)	ns	ns	.000	.500	.000	.000	1
(Simple)	ns	ns	1.160	.123	.236	.230	2
(Square)	ns	ns	1.160	.123	.107	.107	2
(Thin)	ns	ns	.000	.500	.000	.000	1
Breasts							
(Delineated)	ns	ns	2.330	.010	.224	.220	1

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	$\underline{z}$	$\underline{z}_r$	$\underline{z}$	$\underline{p}$	$\underline{z}_r$	$\underline{r}$	
Breasts							
(Nipples Delineated)	ns	ns	.000	.500	.000	.000	1
(Nude)	ns	ns	.000	.500	.000	.000	1
(Small)	ns	ns	1.640	.051	.213	.210	1
Buttons (# of)	ns	ns	1.640	.051	.213	.210	1
Chest							
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(Irrelevant Markings)	ns	ns	.000	.500	.000	.000	1
(Narrow)	ns	ns	.000	.500	.000	.000	1
(Internal Organs Seen)	ns	ns	.000	.500	.000	.000	1
(Simple)	ns	ns	.000	.500	.000	.000	1
Chin							
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(Long)	ns	ns	.000	.500	.000	.000	1
Clothing							
(Amount of)	ns	ns	3.253	.001	.179	.177	4
(Emphasized)	ns	ns	2.330	.010	.255	.250	1
(Inadequate Clothing)	ns	ns	.000	.500	.000	.000	1
(Minimal Clothing)	ns	ns	.000	.500	.000	.000	1
(Overcoat)	ns	ns	.000	.500	.000	.000	1
(Shoelaces)	ns	ns	.000	.500	.000	.000	1
(Unusual Clothing)	ns	ns	.000	.500	.000	.000	1
Cosmetic Effect	ns	ns	.000	.500	.000	.000	1
Distorted Figure	ns	ns	2.330	.010	.741	.630	1
Ear							
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(Large)	ns	ns	.000	.500	.000	.000	1
(Misplaced)	ns	ns	.000	.500	.000	.000	1
(None, None Expected)	ns	ns	.000	.500	.000	.000	1
(One Ear)	ns	ns	.000	.500	.000	.000	1

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z	z_r	z	p	z_r	r	
Ear							
(Two Ears in Profile)	ns	ns	.000	.500	.000	.000	1
Earings	ns	ns	.000	.500	.000	.000	1
Erasures							
(General)	ns	ns	1.345	.089	.075	.075	3
(Arm and Hand)	ns	ns	.000	.500	.000	.000	1
(Noticeable)	ns	ns	.000	.500	.000	.000	1
Eyes							
(Circles)	ns	ns	2.330	.010	.255	.250	1
(Crosses)	ns	ns	.000	.500	.000	.000	1
(Dashes or Curves)	ns	ns	.000	.500	.000	.000	1
(Dots)	ns	ns	2.050	.020	.224	.220	1
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(One Eye in Front)	ns	ns	.000	.500	.000	.000	1
(Slit)	ns	ns	1.025	.152	.112	.112	2
(Two Eyes in Profile)	ns	ns	.000	.500	.000	.000	1
Eye or Ear (Emphasis)	ns	sig	2.785	.003			3
Eyebrow (Emphasis)	ns	ns	1.750	.040	.161	.160	1
Eyelash (Emphasis)	ns	ns	.000	.500	.000	.000	1
False Starts	ns	ns	.000	.500	.000	.000	1
Feet							
(Bare)	ns	ns	.000	.500	.000	.000	1
(Emphasis)	ns	ns	2.330	.010	.224	.220	1
(Jewelry on Ankle)	ns	ns	.000	.500	.000	.000	1
(Large)	ns	ns	.000	.500	.000	.000	1
(Penis-Like)	ns	ns	.000	.500	.000	.000	1
(Poor Form)	ns	ns	.000	.500	.000	.000	1
(Single Dimensional)	ns	ns	.000	.500	.000	.000	1
(Small)	ns	ns	1.640	.051	.131	.130	1
(Small and Pointed)	ns	ns	.000	.500	.000	.000	1
(Toenails)	ns	ns	.000	.500	.000	.000	1

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	$\underline{z}$	$\underline{z}_r$	$\underline{z}$	$\underline{p}$	$\underline{z}_r$	$\underline{r}$	
Feminine Figure	ns	ns	.000	.500	.000	.000	1
Fingers							
(Absence of a Finger)	ns	ns	.000	.500	.000	.000	1
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(Finger Nails)	ns	ns	.000	.500	.000	.000	1
(Long Finger Nails)	ns	ns	.000	.500	.000	.000	1
(Only One Finger)	ns	ns	.000	.500	.000	.000	1
(Petal-like)	ns	ns	.000	.500	.000	.000	1
(Pointing)	ns	ns	.000	.500	.000	.000	1
(Ring)	ns	ns	.000	.500	.000	.000	1
(Single Dimensional)	ns	ns	.000	.500	.000	.000	1
(Unshapely)	ns	ns	2.330	.010	.255	.250	1
Floor Line	ns	ns	.000	.500	.000	.000	1
Genitals							
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(Present)	ns	ns	.000	.500	.000	.000	1
Gloves	ns	ns	.000	.500	.000	.000	1
Hair							
(Adequate)	ns	ns	2.330	.010	.224	.220	1
(Carefully Styled)	ns	ns	.000	.500	.000	.000	1
(Masculine)	ns	ns	.000	.500	.000	.000	1
(None)	ns	ns	.000	.500	.000	.000	1
Hands							
(Distorted)	ns	ns	.000	.500	.000	.000	1
(Emphasized)	ns	ns	2.330	.010	.245	.240	1
(Hidden)	ns	ns	2.330	.010	.255	.250	1
(Objects In)	ns	ns	.000	.500	.000	.000	1
Head							
(Large)	ns	ns	.947	.171	.047	.047	3
(Simple)	ns	ns	2.330	.010	.741	.630	1
(Small)	ns	ns	.000	.500	.000	.000	1
(Tilted)	ns	ns	.000	.500	.000	.000	1

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z	z_r	z	p	z_r	r	
Height							
(Large)	ns	ns	.000	.500	.000	.000	2
(Small)	ns	ns	2.330	.010	.224	.220	1
High Heels	ns	ns	2.050	.020	.224	.220	1
Knee Joint	ns	ns	2.330	.010	.255	.250	1
Kneeling	ns	ns	.000	.500	.000	.000	1
Knuckles	ns	ns	.000	.500	.000	.000	1
Legs							
(Asymmetry Length)	ns	ns	4.710	.000	.460	.420	1
(Asymmetry Width)	ns	ns	4.820	.000	.472	.430	1
(Closed)	ns	ns	.000	.500	.000	.000	1
(Emphasis)	ns	ns	1.880	.030	.192	.190	1
(Hidden)	ns	ns	.000	.500	.000	.000	1
(Length)	ns	ns	1.640	.050	.131	.130	1
(Long)	ns	ns	.000	.500	.000	.000	1
(Poor Proportion)	ns	ns	.000	.500	.000	.000	1
(Short)	ns	ns	.000	.500	.000	.000	1
(Stick-Like)	ns	ns	.000	.500	.000	.000	1
(Width)	ns	ns	4.820	.000	.472	.440	1
Line							
(Broken)	ns	ns	.000	.500	.000	.000	1
(Darkness)							
- Heavy Line	ns	ns	1.160	.123	.066	.066	2
- Light Line	ns	ns	.000	.500	.000	.000	1
- Mixed	ns	ns	.000	.500	.000	.000	1
(Sketchy)							
- Part of Drawing	ns	ns	.000	.500	.000	.000	1
- Whole Drawing	ns	ns	.407	.341	.025	.025	1
Lips (Emphasis)	ns	ns	.000	.500	.000	.000	1
Midline (Emphasis)	ns	ns	2.330	.010	.224	.220	1
Mouth							
(Emphasis)	ns	ns	2.330	.010	.224	.220	1

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z	z_r	z	p	z_r	r	
Mouth							
(Frown)	ns	ns	.000	.500	.000	.000	1
(Line)	ns	ns	.000	.500	.000	.000	1
(Open)	ns	ns	.000	.500	.000	.000	1
(Small)	ns	ns	1.640	.051	.141	.140	1
Neck							
(Emphasis)	ns	ns	1.880	.030	.182	.180	1
(Jewelry)	ns	ns	.000	.500	.000	.000	1
(Large)	ns	ns	1.640	.051	.131	.130	1
(Long)	ns	ns	.000	.500	.000	.000	1
(Shoulders Unlinked)	ns	ns	.000	.500	.000	.000	1
(Stick-Like)	ns	ns	.000	.500	.000	.000	1
Nose							
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(Large)	ns	ns	1.160	.123	.066	.066	2
(Simple Form)	ns	ns	.000	.500	.000	.000	1
(Two Dot Nostrils)	ns	ns	.000	.500	.000	.000	1
(U-shaped)	ns	ns	.000	.500	.000	.000	1
Nudity	ns	ns	2.050	.020	.213	.210	1
Objects							
(General)	ns	ns	2.702	.004	.153	.152	3
(In Mouth)	ns	ns	.000	.500	.000	.000	1
Omission							
(Arms)	ns	ns	2.330	.010	.224	.220	1
(Arms and Hands)	ns	ns	2.429	.008	.161	.160	1
(Body)	ns	ns	.000	.500	.000	.000	1
(Chin)	ns	ns	.000	.500	.000	.000	1
(Ears)	ns	ns	1.880	.030	.182	.180	1
(Eyebrows)	ns	ns	.000	.500	.000	.000	1
(Eyelashes)	ns	ns	.000	.500	.000	.000	1
(Feet)	ns	ns	2.330	.010	.224	.220	1
(Fingers)	ns	ns	1.160	.123	.066	.066	2

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	z	z_T	z	p	z_T	r	
Omission							
(Hands)	ns	ns	.000	.500	.000	.000	1
(Head)	ns	ns	.000	.500	.000	.000	1
(Facial Features)	ns	ns	.000	.500	.000	.000	1
(General)	ns	ns	2.074	.019	.291	.285	2
(Legs)	ns	ns	2.330	.010	.224	.220	1
(Legs and Feet)	ns	ns	4.306	.000	.200	.197	3
(Neck)	ns	ns	2.050	.020	.192	.190	1
(Shoulders)	ns	ns	2.330	.010	.224	.220	1
(Waist)	ns	ns	2.050	.020	.213	.210	1
Paper Bottom Drawing	ns	ns	.000	.500	.000	.000	1
Paper Topped Drawing	ns	ns	.000	.500	.000	.000	1
Parts							
(Loosely Joined)	ns	ns	.000	.500	.000	.000	1
(Misplaced)	ns	ns	.000	.500	.000	.000	1
Placement	ns	ns	2.570	.005	.099	.099	5
Posture	ns	ns	.000	.500	.000	.000	2
Profile Drawn	ns	ns	.000	.500	.000	.000	2
(Head in Profile)	ns	ns	.000	.500	.000	.000	1
(Body in Profile)	ns	ns	.000	.500	.000	.000	1
(Feet in Profile)	ns	ns	.000	.500	.000	.000	1
(Back Drawn)	ns	ns	.000	.500	.000	.000	1
Quality of Drawing	ns	ns	2.330	.010	.741	.630	1
Running	ns	ns	.000	.500	.000	.000	1
Sexual Differentiation	ns	ns	1.640	.051	.472	.440	1
Sexual Elaboration	ns	ns	.000	.500	.000	.000	1
Shading							
(Chest)	ns	ns	2.330	.010	.224	.220	1
(Chin)	ns	ns	.000	.500	.000	.000	1
(Ears)	ns	ns	.000	.500	.000	.000	1

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	$\underline{z}$	$\underline{z}_r$	$\underline{z}$	$\underline{p}$	$\underline{z}_r$	$\underline{r}$	
Shading							
(Eye, Whole)	ns	ns	.000	.500	.000	.000	1
(Feet)	ns	ns	.000	.500	.000	.000	1
(Finger Nails)	ns	ns	.000	.500	.000	.000	1
(General)	ns	ns	2.017	.022	.112	.112	2
(Hands)	ns	ns	.000	.500	.000	.000	1
(Legs)	ns	ns	.000	.500	.000	.000	1
(Lips)	ns	ns	.000	.500	.000	.000	1
(Mouth)	ns	ns	2.330	.010	.224	.220	1
(Neck)	ns	ns	.000	.500	.000	.000	1
(Nose)	ns	ns	.000	.500	.000	.000	1
(Pupils)	ns	ns	.000	.500	.000	.000	1
(Shoulders)	ns	ns	.000	.500	.000	.000	1
(Waist)	ns	ns	2.330	.010	.224	.220	1
Sitting	ns	ns	.000	.500	.000	.000	1
Shoulders							
(Asymmetry Width)	ns	ns	5.587	.000	.563	.510	1
(Emphasis)	ns	ns	2.330	.010	.255	.250	1
(Narrow)	ns	ns	.000	.500	.000	.000	1
(Wide)	ns	ns	2.330	.010	.245	.240	1
Skirt (Wide)	ns	ns	2.330	.010	.224	.220	1
Smile	ns	ns	2.330	.010	.224	.220	1
Stance	ns	ns	1.640	.051	.141	.140	1
Stick Frame	ns	ns	.000	.500	.000	.000	1
Teeth	ns	ns	.000	.500	.000	.000	1
Transparency	ns	ns	1.576	.057	.072	.072	3
Unshapely Parts	ns	ns	.000	.500	.000	.000	1
Vertical Imbalance	sig	ns			.128	.127	2
Waist							
(Straight Lines)	ns	ns	.000	.500	.000	.000	1

Table B.3 Cont'd

Individual Drawing Features	Diffuse Tests		Combined Tests				# of Studies
	Z	Z_r	Z	p	Z_r	r	
Waist							
(Emphasis)	ns	ns	.000	.500	.000	.000	1
(See Internal Organs)	ns	ns	.000	.500	.000	.000	1
Width	ns	ns	1.753	.040	.161	.160	1
Wrist (Jewelry)	ns	ns	.000	.500	.000	.000	1

* - Complete results are presented for non-significant diffuse tests only.

Table B.4

Significance of the Diffuse Tests for Moderator Variables and Results of Combined Tests for Individual Drawing Features Assessed as Indicators of Anger/Hostility Whose Initial Diffuse Test was Significant

Individual Drawing Features	Moderator Variables*			Combined Tests**	
	Age of Subjects	IDF Scoring	Independent Measure	p	r
Eyes (Type of)***	sig	sig	sig		
Head (Profile)	ns			.500	.000
Head (Small)	ns			.008	.360
Omission (Arms)	sig	ns		.010	.350
Placement	sig	sig	sig		
Sex Rating	sig	sig	sig		
Shading (General)	sig	sig	sig		
Size	sig	sig	sig		
Stance	ns			.291	.041
Wrinkles (# of)	sig	sig	sig		

* - No results are indicated after a non-significant diffuse test.

** - Reported results are for adolescents, if the diffuse test for age of sample is non-significant, for reliable scoring, if the IDF scoring diffuse test is non-significant, and for reliable and valid measures, if the diffuse test for independent measures is non-significant.

*** - Sig. means that the diffuse test was significant and ns. means the diffuse test was non-significant.

Table B.5

Significance of the Diffuse Tests for Moderator Variables and Results of Combined Tests for Individual Drawing Features Assessed as Indicators of Anxiety Whose Initial Diffuse Test was Significant

Individual Drawing Features	Moderator Variables*			Combined Tests**	
	Age of Subjects	IDF Scoring	Independent Measure	p	r
Discrepant Age ***	sig	sig	sig		
Body (# of Separations)	sig	sig	sig		
Body (Simple)	sig	ns		.031	.190
Erasures	sig	ns		.187	.043
Eyes (Size of)	sig	sig	sig		
Eyes (Type of)	sig	sig	sig		
Head (Simple)	sig	sig	sig		
Objects	ns			.015	.145
Placement	ns			.500	.000
Rigid Figure	sig	ns		.010	.290
Shading (General)	ns			.127	.080
Shading (Lower Body)	ns			.000	.610

Table B.5 Cont'd

Individual Drawing Features	Moderator Variables*			Combined Tests**	
	Age of Subjects	IDF Scoring	Independent Measure	p	r
Shading (Upper Body)	ns			.000	.610
Stance	sig	sig	sig		
Trunk:Leg	sig	sig	sig		

* - No results are indicated after a non-significant diffuse test as the analysis stopped at this point.

** - Reported results are for adolescents, if the diffuse test for age of sample is non-significant, for reliable scoring, if the IDF scoring diffuse test is non-significant, and for reliable and valid measures, if the diffuse test for independent measures is non-significant.

*** - Sig. means that the diffuse test was significant and ns. means the diffuse test was non-significant.

Table B.6

Significance of the Diffuse Tests for Moderator Variables and Results of Combined Tests for Individual Drawing Features Assessed as Indicators of Thought Disorder Whose Initial Diffuse Test was Significant

Individual Drawing Features	Moderator Variables*			Combined Tests**	
	Age of Subjects	IDF Scoring	Independent Measure	p	r
Eye or Ear (Emphasis) ***	sig	sig	sig		
Vertical Imbalance	sig	sig	ns	.003	.250

* - No results are indicated after a non-significant diffuse test as the analysis stopped at this point.

** - Reported results are for adolescents, if the diffuse test for age of sample is non-significant, for reliable scoring, if the IDF scoring diffuse test is non-significant, and for reliable and valid measures, if the diffuse test for independent measures is non-significant.

***- Sig. means that the diffuse test was significant and ns. means the diffuse test was non-significant.

Appendix C

Human Figure Drawing Scoring Manual

Human Figure Drawing Scoring Manual

Bryan Acton, M.Sc.

Simon Fraser University
Burnaby, B.C.

Human Figure Drawing Scoring Manual

The following is a manual of scoring criteria for human figure drawings. The reader should familiarize themselves with each set of criteria prior to scoring. Fortunately the majority of scoring criteria are straightforward and easily translated into a rating of the figure drawing. However, some of the criteria are more subjective. These latter criteria rely much more on the individual scorer's judgment. Where possible descriptive formulations have been provided to give a common reference for scorers and some reference charts have been designed.

The majority of the scoring criteria are self-contained. That is, all the information that is needed to derive a score is found within the description of coding for the variable. One exception is the division of body parts. For the purposes of this manual body parts are divided into 7 groups. These include 1) the head, 2) the neck, 3) one or both hands, 4) one or both feet, 5) one or both legs, 6) one or both arms, and 7) the trunk or body. The head is defined as the region of the figure including the hair, face, and apparel usually found on the head, e.g., a hat. The neck is defined as the region between the chin and the body of the figure. The lower reaches of the neck are usually demarcated by a neckline on the clothing of the figure. The hands are defined as the region from the wrist to the end of the fingers. The wrist is usually indicated by a line marking the cuff, but may also be determined by observing where the line of the arm widens to indicate the hand. Feet are defined as the region from the ankle to the toes. Usually feet are drawn as shoes or boots. When apparel is used as indicating the foot, the outside bounds of that region are marked by the lines indicating a separation between the footwear and the pants or leg. The arms are defined as the region between the wrist and the point where arm lines are clearly differentiated from the body. The legs are defined as the region between the top of the feet and the waist, including the hip area. The waist is usually indicated by either a line marking

the division between shirt and pants or skirt, or by a narrowing of the figure between the chest and hips. The body is the region between the neck, the legs, and the arms. The body's boundaries are usually indicated by the demarcation line at the waist, the neck line, and the point at which the arms separate from the body.

There are also certain practices or definitions which are applied across a series of scoring criteria, e.g., shading. Where this is the case these general practices or definitions are provided at the outset of the section describing the scoring of the feature series.

To score a human figure drawing using this manual several pieces of equipment are required. You will need a pencil, a twelve inch ruler (preferably in clear plastic), a protractor, a transparent sheet divided into four equal size quadrants, and a pocket calculator. The use of these pieces of equipment will be explained at the appropriate points in the manual. Scoring is also assisted by referring to the Quick Reference Sheet, which provides abbreviated statements of a number of scoring criteria which have been found to require frequent referencing.

As a final point, you will notice that references can be found within a number of the item definitions. These references refer to the articles which provided descriptions of feature scoring which either constitute or are the basis for the definition used within the manual.

Angle, Height, and Width Measurements

To prepare the figure for scoring of height, width, and angle features several measurements need to be taken and a midline drawn onto the figure. The first step is to draw the midline. To draw the midline place a ruler on the figure such that it passes through the crotch and the center point of the neck. Where possible try to align the ruler so that it passes through the center of the head as well. Then, using a pencil, make a line

extending from above the top of the figure through the two, or three, reference points and on down below the base of the figure to the bottom of the page. This is the midline. This procedure is slightly modified for figures drawn in profile. Because the crotch of the figure is not usually drawn in profile, the midline must be drawn through the midpoint of the hip.

Now you need to make a series of 15 measurements; arm length (right and left), arm width (right and left), body length, head height, head width, overall height, overall width, leg length (right and left), leg width (right and left), midline to left shoulder, and midline to right shoulder. Results of these measurements should be placed in the appropriate sections of the coding sheet. The references of right and left refer to the rater's right- and left-hand sides. All measurements should be rounded to the nearest centimeter. For the benefit of the new rater drawing measurement examples have been provided at the back of the manual. Drawing measurement example A provides illustrations of all of the basic measurements as they would routinely be encountered. Examples B through D provide illustrations of unusual measurement circumstances. Direct references will be made to examples B, C, and D where appropriate.

Each of the measurement procedures is outlined briefly below:

Arm Length: Placing the ruler such that it roughly follows the angle of the arm measure from the midpoint of the shoulder to the furthest reach of the fingers. Sometimes the end of the hand will be obscured because the figure has been drawn with an arm behind the back or with hands placed in pockets. For one hand behind the back an estimate should be made of where the arm might be expected to end. Where both arms are behind the back, arm length should be measured to the point where the arm meets the wall of the body. In the case of a hand placed in the pocket, measurement should be taken to the bottom of the pocket (unless the

pocket is excessively large, in which case the rater should estimate the end point of the arm). At other times their will not be a clearly marked shoulder or the arm will extend directly out of the midpoint of the body. In this case the rater should measure from the beginning of the arm (either at the neck or at the point where the arm leaves the body) to the end of the hand.

Another complication is when one or both arms are bent at a sharp angle (e.g., 45 degrees). In the case of a sharply bent arm the rater should measure from the shoulder to the middle of the elbow region, then measure from that point on the elbow to the finger tips, adding the two measures together to find a value for the overall length of the arm (see Drawing Measurement Example B). One rare variation of the bent arm is the arm bent back on itself (see Drawing Measurement Example C). In this special case measure from the midpoint of the shoulder to the bottom of the arm, then from the bottom of the arm up to the end of the hand. Add the two measurements together to determine the length of the arm.

Arm Width: Placing the ruler at right angles to the arm, move it up and down measuring the distance from the right most aspects of the arm to the left most aspects. The greatest distance between the two outside walls of the arm is the arm width. Do not measure additional clothing items, such as elbow pads.

Body Length: Placing the ruler along the midline measure the distance from the crotch to the chin of the figure. In a figure drawn in profile measure to the point where the legs divide or the line of the buttocks joins the leg (see Drawing Measurement Example D). Where a dress is drawn or some other article obscures the leg region of the feature, the rater should estimate the lower bounds of the body.

Head Height: Place the ruler along the midline or, where the midline does not pass at a reasonable distance through the chin of the figure to the top of the head, place the ruler in a line between the chin and the topmost portion of the head. Now measure the distance between the chin and the top of the head. Where the figure is wearing a hat or the top of the head is otherwise obscured, estimate the height of the head by completing the circle which marks the outside boundaries of the head. The hair is used in estimating the top-most aspect of the head and a beard, when drawn, is used in estimating the bottom-most aspect.

Head Width: While keeping the ruler roughly at right angles to the head, move up and down measuring the distance from the right most aspects of the head to the left most aspects. This measurement should be taken somewhere between the chin and the top of the head. The greatest distance observed is the head width. Occasionally the hair will fall below the chin. The rater should not use hair which has fallen below the chin to measure head width, even if the resulting measure would produce a greater width measurement. As well, a hat or other head gear should not be used in measuring the width of the head.

Height: Place the ruler along the midline. Then measure the distance from the lowest point to the highest point on the figure. Hats, boots, and other pieces of clothing should be used when ascertaining the extremes of the figure. However, other objects, such as fishing poles, should not be used in making this measurement (Prytula and Hiland, 1975).

Leg Length: Placing the ruler such that it roughly follows the center line of the leg, measure from the lowest demarcation line found at the waist to the lowest reach of the foot. Where no demarcation lines are drawn, the rater should estimate where that line might fall (see Drawing Measurement Example B). Another special case is

where the leg is cut-off by the bottom of the paper. In this case leg measurements should be taken to the bottom of the page (see Drawing Measurement Example C).

Leg Width: Placing the ruler at right angles to the leg and move it up and down along the leg measuring the distance from the right most aspects of the leg to the left most aspects. The greatest distance between the two outside walls of the leg is the leg width.

Midline to Shoulder (Right and Left): Place the ruler at a right angle to the midline at shoulder height on the figure. Measure the widest distance from the midline to the outside of the shoulder. Where a sharp turn is not present indicating the outside of the shoulder, the rater should measure from the point where the outside line of the arm begins to travel more down along the figure than out and away from the midline.

Sometimes the end of the shoulder is difficult to determine using the above rules. One case where this is true is where the arm is upraised (see Drawing Measurement Example C). Under such circumstances the rater should mark the end of the shoulder at the point where the arm begins to travel upward. Another case is where the arm extends directly out from either the neck or the body (see Drawing Measurement Example B). In this case the point where the arm leaves the neck or body should be treated as the end of the shoulder. This rule should be employed even if shoulders are drawn on a figure in which the arms extend out of the body.

Width: Place the ruler at a right angle to the midline of the figure. Move the ruler up and down the figure measuring the distance between its most extreme left and right points. The greatest distance observed is the width of the figure. Hats, boots,

and other pieces of clothing should be used when ascertaining the outside bounds of the figure. However, other objects, such as fishing poles, should not be used in making this measurement (see Drawing Measurement Example C).

Once measurements have been made the rater should move on to score the following features using the criteria below. All values found using the following criteria should be recorded on the scoring sheet. Results of formal calculations should be recorded on the record sheets as decimal numbers. Decimals should be taken to the second place. Note for asymmetry ratings where only one value can be determined a value missing should be recorded. Missing values for asymmetry ratings are often found when the human figure is drawn in profile.

Arms (Asymmetry of Length): [Thought Disorder] The asymmetry rating is calculated by first dividing the arm length measures for each side of the body by the measure of head width. The smaller of these quotients is then subtracted from the larger producing the asymmetry rating (John, 1974).

Arms (Asymmetry of Width): [Thought Disorder] The asymmetry rating is calculated by first dividing the arm width measures for each side of the body by the measure of head width. The smaller of these quotients is then subtracted from the larger producing the asymmetry rating (John, 1974).

Arms (Length): [Thought Disorder] This feature is measured as the average arm length for the figure, calculated by averaging the two arm length measurements.

Head (Tiny/Small): [Anger/Hostility; Anxiety] **1)** Score 0-3: A score of 0 is given when the head height is 2.5 cms or greater. A score of 1 is given when the head height is between 1.5 and 2.4 cms. A score of 2 is given when the head height is between 1

and 1.4 cms. A score of 3 is given when the head height is less than 1 cm (Handler, 1967).

2) Score 0-1: First divide head height by height. A score of 0 is given when the resulting product is greater than or equal to .10. A score of 1 is given when the product is less than .10 (Koppitz, 1968), for emotional indicators.

Head:Body Ratio: [Anxiety] Total ratio score: To calculate this ratio divide the head height by the body length (Prytula and Hiland, 1975), for anxiety.

Height (Overall): [Anger/Hostility; Thought Disorder] Total score: This score is the height measurement already taken (Prytula and Hiland, 1975), for anxiety.

Height (Small): [Anxiety; Thought Disorder] Score 0-4: Scores for this rating should be made in reference to the height measure. A score of 0 is given when the figure is 17.1 cms or taller. A score of 1 is given when the figure is between 14.1 and 17 cms in height. A score of 2 is given when the figure is between 11.6 and 14 cms in height. A score of 3 is given when the figure is between 5.1 and 11.5 cms in height. A score of 4 is given when the figure is 5 cms or less in height (Handler, 1967, for scores 0-3, and Koppitz, 1968, for score 4), for anxiety and emotional indicators.

Legs (Asymmetry of Length): [Thought Disorder] The asymmetry rating is calculated by first dividing the leg length measures for each side of the body by the measure of head width. The smaller of these quotients is then subtracted from the larger producing the asymmetry rating (John, 1974).

Legs (Asymmetry of Width): [Thought Disorder] The asymmetry rating is calculated by first dividing the leg width measures for each side of the body by the measure of head width. The smaller of these quotients is then subtracted from the larger producing the asymmetry rating (John, 1974).

Leg (Length): [Thought Disorder] Score 0-1: Divide the leg length measurement by the body length measurement. A score of 0 is given when the length of the leg is less than twice the length of the body (a quotient of less than 2). A score of 1 is given when the length of the leg exceeds twice the length of the body (a quotient of 2 or greater) (Holzberg and Wexler, 1950).

Shoulders (Asymmetry of Width): [Thought Disorder] First calculate shoulder:head width ratios for both the left and right shoulders. These ratios can be obtained by dividing the individual midline to shoulder measurements (right and left) by the head width measurement. To find the shoulder asymmetry rating subtract the smaller of the shoulder:head width ratios from the larger (John, 1974).

Vertical Imbalance: [Anxiety; Thought Disorder] Score 0-90: Using a protractor determine the angle between the midline and the bottom of the paper. Subtract this angle from 90. The resulting value is the number of degrees from perpendicular of the figure. This constitutes the vertical imbalance score.

Width: [Thought Disorder] Total width: This measurement is the width measurement already taken (John, 1974; Prytula and Hiland, 1975).

Line Quality

All of the ratings in this section refer to the characteristics of the lines used by the subject. The lines refereed to are usually the body lines, but may also include those used with features or to add extra detail to the drawing. Body lines are found to be the best point of reference because they are 1) common across the vast majority of drawings and 2) usually the largest and most easily discerned lines on the drawing. The rater should look over the complete drawing before judging the quality of line, as decisions regarding the type of line used by the subject often involve general impressions or are combined with

judgments of the percentage of the drawing exhibiting a given characteristic. As a rule of thumb the rater should do the line pressure ratings for the head, lower body, and upper body before completing the ratings of the overall figure in Line (Heavy) and Line (Light). The rater should also strive not to be influenced by non-line aspects of the drawing (e.g., shading) which might make the figure appear dark, despite the use of relatively light lines.

To increase the reliability of heaviness/lightness ratings a Line Reference Chart has been devised which provides seven line samples. The chart indicates the relative scoring of the lines for each of three rating scales; those for Line (Heavy), Line (Light), and Line Pressure. These reference lines have been photocopied such that they can be directly compared with the copies of the drawings in order to determine the average weight of line used.

A final concern worthy of special emphasis is the drawing with very inconsistently weighted lines. Such drawings appear to make use of a variety of line weights from very light to very heavy. Where the rater observes such variability in a drawing, they should first consider the possibility that the subject has employed extensive line reinforcement. Thus, highly variable line heaviness can be used as a cue that reinforcement is occurring. This does not, however, ease the raters task, as reinforced lines are used in determining the heaviness or lightness of a line. The best course for scoring a drawing with highly variable lines is to estimate the percentage of different weights used for each of the head, upper body, and lower body and then sum these estimates in one's head. That weight of line which appears to be most commonly used should be that chosen for Line (Heavy) and Line (Light).

Line (Discontinuous): [Anxiety] For the purpose of scoring this feature a discontinuity is defined as a break in a boundary line of the figure which makes it possible to move

from the outside of the body wall to the inside of the body wall without crossing a body line. Where lines are sketchy, but no entry into the inside walls is unblocked, a discontinuity is not scored. Further, a break must be clearly evident. There must be clear evidence that the body line has stopped. A line which is simply very faint or poorly reproduced in the photocopy would not be scored as a discontinuity. Similarly, an unfinished limb, e.g., a leg drawn to the knee which stops without connecting the leg lines, is not a discontinuity. Score 0-6: A score of 0 is given when body lines are continuous. A score of 1 is given when one body line is broken. A score of 2 is given when 2-3 discontinuities are observed. A score of 3 is given when 4-5 discontinuities are observed. A score of 4 is given when 6 discontinuities are observed. A score of 5 is given when 7-8 discontinuities are observed. A score of 6 is given when 9 or more discontinuities are observed (Wainwright, 1970), for psychopathology.

Line (Heavy): [Anxiety] Score 0-3: Line heaviness refers to the width and darkness of a line. It does not refer to how hard a person pressed on the paper, but whether the line looks heavy or dark. A given heaviness rating is scored only if that quality of line is used for more than half of the drawing. Heaviness ratings include: 0, a predominantly medium line; 1, a predominantly medium-heavy line; 2, a predominantly heavy line; and 3, a predominantly very heavy line (Handler, 1967). Refer to Scale A of the Line Reference Chart for scoring of each type of line.

Line (Light): [Anxiety] Score 0-3: Line lightness refers to the width and darkness of the line. It does not refer to how hard a person pressed on the paper or the sketching of a line, but whether the line looks light and thin. A given lightness rating is scored only if that quality of line is used for more than half of the drawing. Lightness ratings include: 0, a predominantly medium line; 1, a predominantly

medium-light line; 2, a predominantly light line; 3, a predominantly very light line (Handler, 1967). Refer to Scale B of the Line Reference Chart for scoring of each type of line.

Line Emphasis: [Anxiety] Emphasis involves the use of additional lines, beyond those necessary to provide a basic depiction of the human figure, which either bring a three-dimensional quality to the drawing or extra detail to a feature or body part. Detail in this definition refers to lines which accentuate or elaborate on the body and do not refer to extra work in detailing clothing, such as additional items of clothing or jewelry, accents or patterns on clothing, or supplementary clothing details (e.g., pockets, shirt or pant cuffs, or belt buckles). Examples of additional details which would constitute emphasis include; folds or creases in clothing (particularly where they accentuate body features, such as creases at the inside of the elbow), muscle lines, breasts or chest muscles, dimples, wrinkles or creases in the face. Shading does not constitute emphasis. Where cross-hatching, the side of the pencil, or its point has been used to darken an aspect of the figure shading is said to have occurred. Score 0-7: One point is awarded for each body part having emphasis lines up to the maximum of seven.

Line Pressure (Head): [Anxiety] Score 1-5: To score this feature using the definition for heaviness and lightness of line, rather than that used for line pressure. The scale for this feature is: 1, a very light line; 2, a light line; 3, a medium line; 4, a heavy line; 5, a very heavy line (Exner, 1962). Refer to the line reference chart for examples of each scoring.

Line Pressure (Heavy): [Anxiety] Line pressure is assessed by turning over the sheet on which the drawing has been made and feeling whether the surface is smooth or raised where the lines were drawn.) Score 0-4: A score of 0, is given when the

surface of the back of the sheet is smooth. A score of 1 is given when a moderately raised outline can be felt on up to 3/4 of the drawing. A score of 2 is given when a moderately raised outline can be felt on 3/4 or more of the drawing, or a markedly raised outline is found for 1/2 of the drawing. A score of 3 is given when a markedly raised outline can be felt for 1/2 to, but not including, 3/4 of the drawing. A score of 4 is given when a markedly raised outline can be felt for 3/4 or more of the drawing (Handler, 1967).

Line Pressure (Lower Body): [Anxiety] Score 1-5: As in Line Pressure (Head), but applied to the hips and legs (Exner, 1962). Refer to the line reference chart for examples of each scoring.

Line Pressure (Upper Body): [Anxiety] Score 1-5: As in Line Pressure (Head), but applied to the upper torso and arms (Exner, 1962). Refer to the line reference chart for examples of each scoring.

Line (Reinforced): [Anxiety] Score 0-3: Line reinforcement is indicated whenever extra, overlapping lines are used that darken, broaden, or otherwise buildup a single line such that it is darker and more pronounced than it would be otherwise. Reinforcement can occur on either a body line or a detail line. Reinforcement should be distinguished from shading where repeated lines may also be used but where the intent is to darken the feature and not strictly the line, e.g., a belt. Line reinforcement can also be confused with sketchy lines which also use repeated pencil strokes. In the case of a figure drawn with sketchy lines, reinforcement is deemed to occur only where the amount or darkness of the lines suggests the subject has placed particular effort into accenting the lines. In judging whether reinforcement has occurred first look to see whether any lines appear dark relative to other aspects of the feature. Once dark lines have been observed, these should

be examined to see whether a repeating of lines has been used to render them darker. Sometimes large sections or the whole body may be reinforced. This is indicated by clearly darker lines overlapping initial drawing lines. Such overall reinforcement may not appear noticeably darker relative to other aspects of the drawing. However, on examination it will be clear that an initially lighter line has been darkened by the overlapping line indicating that reinforcement has occurred. Score 0-3: A score of 0 is given when no lines are reinforced. A score of 1 is given when lines from one body part are reinforced. A score of 2 is given when lines from two body parts are reinforced. A score of 3 is given when lines from more than two body parts are reinforced (Handler, 1967).

Omissions

An omission can be defined as the failure to draw the body part, when that part might reasonably be expected to be included in the drawing. Omissions include the absence of a body part in its expected location (e.g., a hand at the end of an arm), the placement of the figure such that certain body parts are cut off by the edge of the paper, and the failure to distinguish body parts by shape and/or demarcation lines. As an example of the latter, if the legs are drawn such that they come to a point, feet are counted as omitted unless toes or shoes are indicated or a line marking the cuff is drawn. As well, hands are considered as omitted if fingers are not indicated or fingers are drawn as if they are attached to the arm. A case where a body part might not be drawn and no omission would be scored is a figure in profile. As well, hands behind the figure's back or other body parts obscured by an object in the drawing would not be counted as omissions.

The above guidelines should be used when scoring the drawing features in this section. All scores should be recorded on the coding sheet in the omissions section.

Omission (Arms): [Anger/Hostility; Thought Disorder] Scored 0-1: A score of 0 is given when the arms of the figure are present. A score of 1 is given when the arms of the figure are omitted (Holzberg and Wexler, 1950).

Omission (Ears): [Thought Disorder] Scored 0-1: A score of 0 is given when the ears of the figure are present. A score of 1 is given when the ears are absent or are drawn where they should not be (Holzberg and Wexler, 1950). Because ears are often obscured by hair, where hair is drawn that might reasonably block the view of the ears no omission is scored.

Omission (Eyes): [Anxiety] Score 0-1: A score of 0 is given when the eyes of the figure are present. A score of 1 is given when the eyes are absent (Koppitz, 1968), for emotional indicators.

Omission (Feet): [Thought Disorder] Scored 0-1: A score of 0 is given when the feet of the figure are present. A score of 1 is given when feet are absent (Holzberg and Wexler, 1950).

Omission (Hands): [Anger/Hostility] Score 0-1: A score of 0 is given when the hands and the fingers of the figure are present. A score of 1 is given when the palm or the fingers or both are absent. Fingers are said to be absent when no finger like projections are observed at the end of the arm, even if the demarcation line between the arm and the hand is marked. A round fist like hand should be judged as a complete hand only if there are clear indications that the hand is being held in a fist. For example, a pencil or other object is seen projecting from either side of the hand. The palm is said to be absent when either fingers are drawn as if they are each independently connected to the outside of the arm or fingers are drawn at the end of the arm, but without any clear demarcation between the arm and the hand.

Hands drawn as mittens are also scored as omissions, unless the figure is clearly drawn as wearing winter apparel. Hands hidden behind the figure or placed in pockets are not scored as missing (Koppitz, 1968), for emotional indicators.

Omission (Legs): [Thought Disorder] Score 0-1: A score of 0 is given when the legs of the figure are present. A score of 1 is given when the legs are absent (Holzberg and Wexler, 1950).

Omission (Mouth): [Anger/Hostility; Anxiety] Score 0-1: A score of 0 is given when the mouth of the figure is present. A score of 1 is given when the mouth is absent (Koppitz, 1968), for emotional indicators.

Omission (Neck): [Thought Disorder] Score 0-1: A score of 0 is given when the neck of the figure is present. A score of 1 is given when the neck is absent (Holzberg and Wexler, 1950).

Omission (Nose): [Anger/Hostility; Anxiety] Score 0-1: A score of 0 is given when the nose of the figure is present. A score of 1 is given when the nose is absent (Koppitz, 1968), for emotional indicators.

Omission (Shoulders): [Thought Disorder] Score 0-1: A score of 0 is given when the line of the arm extends outward from the neck or bottom of the head and then bends so as to indicate a transition to the arm. A score of 1 is given when the line of the arm extends straight out from the neck, head, or body with no discernible bend to indicate the presence of a transition from shoulders to arms. (Holzberg and Wexler, 1950). Where arms are attached to the body around the midline the upper torso must have shoulder like bends otherwise shoulders are classed as omitted. For example, a perfectly round body with arms extending out at the waist would be classed as having shoulders omitted.

Omission (Waist): [Thought Disorder] For the purposes of scoring this feature, the waist is defined as either a demarcation line between the upper and lower body or a clear narrowing of the body between shoulders and hip. Score 0-1: A score of 0 is given when the waist of the figure is present. A score of 1 is given when the waist is absent (Holzberg and Wexler, 1950). In some figures the body is drawn as a complete circle or square with legs independently attached at the lower reaches of the body. Where this is the case, the waist is classed as omitted if there is no demarcation line on the body indicating a waist.

Shading

Shading refers to any consistent pattern using the point or the side of the pencil lead to darken the figure or any aspect thereof. Cross-hatching, scribbling, blackening, or any other action that darkens a feature is considered shading. Scribbled lines indicating the presence of facial hair or hair on the head is considered as shading, as are "freckles", "measles", etc. Darkening of circles, including those of the iris, or on a belt buckle is considered shading, though a simple dot, such as that for the eye, is not. An even light shading of skin areas to represent skin color is not treated as shading. Note that shading is distinguished from line reinforcement in that shading involves the feature as a whole, while reinforcement refers to the lines only.

The above guidelines should be used when scoring the drawing features in this section. All scores should be recorded on the coding sheet in the shading section.

Shading (Arms): [Anger/Hostility; Anxiety] Score 0-1: A score of 0 is given when no shading is observed on the arms. A score of 1 is given when shading is observed on the arms (Koppitz, 1968), for emotional indicators.

Shading (Chest): [Thought Disorder] Score 0-1: A score of 0 is given when the chest area of the figure is unshaded. A score of 1 is given when the chest area is shaded (Holzberg and Wexler, 1950).

Shading (Feet): [Anxiety] Score 0-1: A score 0 is given when no shading is observed on the feet of the figure. A score of 1 is given when shading is observed on the feet (Koppitz, 1968), for emotional indicators. Crossed lines indicating laces should not be scored as shading of the feet.

Shading (Hands): [Anger/Hostility; Anxiety] Score 0-1: A score of 0 is given when no shading is observed on the hands. A score of 1 is given when shading is observed on the hands (Koppitz, 1968), for emotional indicators.

Shading (Head): [Anxiety] Score 0-2: A score of 0 is given when the head of the figure is unshaded. A score of 1 is given when either the hair or the face of the figure is shaded. A score of 2 is given when both the hair and the face are shaded (adapted from Koppitz, 1968), for emotional indicators. The face of the figure includes the eyes, nose, mouth, cheeks, ears, eyebrows, mustache, beard, and chin.

Shading (Legs): [Anxiety] Score 0-1: A score 0 is given when no shading is observed on the legs of the figure. A score of 1 is given when shading is observed on the legs (Koppitz, 1968), for emotional indicators.

Shading (Mouth): [Thought Disorder] Score 0-1: A score of 0 is given when the mouth of the figure is not shaded. A score of 1 is given when the mouth is shaded, but only if there is a center line demarcating the lips. If a demarcation line is absent, then it is assumed the mouth is open and shading is deemed not to have occurred (Holzberg and Wexler, 1950).

Shading (Neck): [Anger/Hostility; Anxiety] Score 0-1: A score of 0 is given when no shading is observed on the heck. A score of 1 is given when shading is observed on the neck (Koppitz, 1968), for emotional indicators.

Shading (Waist): [Thought Disorder] Score 0-1: A score of 0 is given when the waist area is unshaded. A score of 1 is given when the waist area is shaded (Holzberg and Wexler, 1950). If only a demarcation line is present at the waist, or no waist line is present, shading is determined based on whether the body or legs are shaded to the waist. If either legs or the body is shaded to the waist, then shading is said to be present at the waist.

Feature Emphasis

Emphasis refers to the apparent use of greater drawing activity than is necessary to produce the basic feature. There are two definitions of "greater drawing activity" which will be used in the scoring of feature emphasis in this manual. The first definition, borrowed from Holzberg and Wexler (1950), is largely one of line reinforcement. This definition requires that the rater look for evidence that the subject has used repeated strokes on a line of a feature which renders the lines of the feature or body part noticeably darker than other nearby lines. Lines which are simply sketchy or overlapping without a sense of emphasis are not classed as reinforced. Similarly, shading of a feature is not classed as reinforcement because this involves darkening of the feature, rather than darkening of the lines of the feature.

The second definition borrows from that provided for line emphasis (Handler, 1967) in the section on line quality. Emphasis according to this second definition is any additional lines employed over and above those necessary to provide a basic

representation of the feature. The definition of feature emphasis used here differs from that employed with line emphasis in that any detail added to a feature or body part is considered an indication of emphasis. Examples of emphasis according to this second definition include muscle lines, folds and creases in clothing, dimples, etc., as in line emphasis, but also includes details on clothing, such as insignia or patterns, shoelaces, pockets, the presence of jewelry, etc. It does not include other objects, such as footballs, pets, etc. Unusual outlines of a feature which give it greater complexity, but does not use additional lines on the feature, is not classed as emphasis.

Two scorings are required for each feature listed below. The first is for line reinforcement and should be placed on the open line on the coding sheet. The second is for detail and should be placed between the brackets on the coding sheet. Where the feature is omitted or absent a score of 0 should be given.

Arms (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when simple lines are used in drawing the arms. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used in drawing the lines of the arms (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent arms. A score of 1 is given when additional lines are used to represent greater detail in the arms, e.g., muscles, elbows, etc.

Clothing (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when simple lines are used in drawing the clothing. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used in drawing the lines of the clothing (adapted from Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent clothing. A score of 1 is given when

additional lines are used to represent greater detail in the clothing, e.g., pockets, insignia, folds, etc.

Eye or Ear (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when the eyes and ears are drawn with simple line. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the eyes or the ears (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the eyes and ears. A score of 1 is given when additional lines are used to represent greater detail in the eyes and ears. Darkening of the irises is not considered emphasis.

Eyebrow (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when simple lines are used in drawing the eyebrows. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the eyebrows (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent eyebrows. A score of 1 is given when additional lines are used to represent greater detail in the eyebrows. Examples of detail emphasis would be wrinkles above the eyebrow and the use of several parallel curved lines indicating the eyebrow hairs.

Feet (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when simple lines are used in drawing the feet. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the feet (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the feet. A score of 1 is given when additional lines are used to represent greater detail in the feet, e.g., shoelaces, insignia, etc. Shading of the soles is not reinforcement.

Hands (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when simple lines are used to draw the hands. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the hands (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the hands. A score of 1 is given when additional lines are used to represent greater detail in the hands, e.g., fingernails, etc.

Leg (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when a simple line is used to draw the leg. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the legs (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the legs. A score of 1 is given when additional lines are used to represent greater detail in the legs, e.g., pockets, crease lines in the pants, etc.

Midline (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when the midline is represented by a simple line. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the midline (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the midline. A score of 1 is given when additional lines are used to represent greater detail in the midline, e.g., drawing of a belt, a belt buckle, etc.

Mouth (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when a simple line is used to represent the mouth. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the mouth (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the mouth. A score of 1 is given when additional lines are

used to represent greater detail in the mouth, e.g., lines separating the lips, curves around the corners of the mouth, etc.

Neck (Emphasis): [Thought Disorder] Scored 0-1: A score of 0 is given when a simple line is used to represent the neck. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the neck (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the neck. A score of 1 is given when additional lines are used to represent greater detail in the neck, e.g., an Adam's apple, etc. Fancy collars and jewelry placed on the neck is also treated as additional detail.

Shoulders (Emphasized): [Thought Disorder] Scored 0-1: A score of 0 is given when the shoulders are drawn with simple lines. A score of 1 is given when repetitive pencil strokes suggesting reinforcement have been used on the lines of the shoulders (Holzberg and Wexler, 1950). Scored 0-1: A score of 0 is given when basic lines are used to represent the shoulders. A score of 1 is given when additional lines are used to represent greater detail in the shoulders, e.g., muscles, etc. As well, extra detail on clothing at the shoulders, e.g., a shoulder flap, would also result in a score of 1.

Miscellaneous Features

The remaining features do not lend themselves readily to being grouped. As well, the vast majority of the items in this section are well enough defined that they do not require an introduction as was done for previous sections. Unless otherwise indicated a score of 0 is given when a miscellaneous feature cannot be scored. For example, when the feet are omitted a score of 0 is given to Feet (Pointed).

Age of Figure Discrepant From Subject: [Anxiety] (Use the following to determine age categories: 2 = infant, age 1-2; 3 = child, age 3-13; 4 = adolescent, age 14-19; 5 = young adult, 20-26; 6 = adult, age 27-40; 7 = middle age, age 41-59; 8 = old age, age 60+) Score 1-8: Score 1, when testee is 3 categories younger than drawing appears; score 2, when testee 2 categories younger; score 3, when testee is 1 category younger; score 4, when testee in same age group; score 5, when testee is 1 category older than drawing appears; score 6, when testee is 2 categories older; score 7, when testee is 3 categories older; score 8, when age cannot be determined (Wainwright, 1970). Several cues can be used in trying to identify the general age of the figure. Drawings of children have larger heads and few creases. They may also have child like items. Adolescent figures are largely identifiable by their dress and their lack of adult features. Common adolescent dress includes runners, t-shirts, radical hairstyles, and jewelry, such as single earrings. Adult figures are identified by the fullness of their figure (e.g., heavy muscled, large bust, etc.) Fullness of the beard, baldness, and wrinkles differentiate young adult from older adult figures.

Arms (Behind Back): [Thought Disorder] Scored 0-3: A score of 0 is given if the arms are revealed. A score of 1 indicates that the hands or part of the hands are behind the back, but the arms are fully exposed. A score of 2 indicates that at least part of the arms (i.e., more than just the hands) are portrayed as behind the back (adapted from Holzberg and Wexler, 1950). Examples of 1 and 2 scores are found in the Arms (Behind Back) example sheet at the back of the manual. If arms are omitted a score of 3 is given.

Arms (Down): [Anxiety] Score 0-3: A score of 0 is given when at least one arm is at an angle greater than 45 degrees from the body. A score of 1 is given when both arms

are at less than a 45 degree angle from the body, but at least one arm is held away from the body or is bent at the elbow. A score of 1 is also given if the arms are across the chest or behind the back. A score of 2 is given when both arms are held down at the sides of the body or are held in a rigidly vertical position (Fox, Davidson, Lighthall, Waite, and Sarason, 1958), for anxiety. A score of 3 is given when arms are omitted.

Arms (Long): [Anger/Hostility] Score 0-1: A score of 0 is given when the arms are of normal proportion to the body or shorter. A score of 1 is given when the arms of the figure reach below the knee or where the knee should be (Koppitz, 1968), for emotional indicators.

Asymmetry of Limbs: [Anger/Hostility] Score 0-1: A score of 0 is given when the arms and the legs are similar in shape. Arms and legs may differ slightly in size. A score of 1 is given when one arm or leg differs markedly in shape from the other arm or leg (Koppitz, 1968), for emotional indicators.

Bizarreness: [Thought Disorder] Bizarreness is defined as the degree to which the human figure drawing deviates from a reasonable representation of the human body. To be considered as bizarre a figure must either 1) possess one or more parts which are distorted in shape such that they very poorly represent, or fail to represent, those parts, 2) be grossly distorted in size relative to the rest of the figure, or 3) be misplaced. In establishing whether a body part is distorted the rater should not ask themselves " is this a good representation of an arm, a leg, etc.?", but "is this an arm, a leg, etc.?"

1) This first set of criteria come very close to the definition of bizarre given above. To score bizarreness using this criteria the rater should look for deviations in the form of exaggerated, distorted, misplaced, or excessive body parts (Kay,

1978). Score 0-2: A score of 0 is given for a non-bizarre figure. A score of 1 is given to a figure which has some bizarre features, but provides a reasonable representation of a human body. A score of 2 is given when a figure is quite bizarre, deviating markedly from a normal figure (Cauthen, Sandman, Kilpatrick, and Deabler, 1969).

2) Distorted Figure: [Anxiety; Thought Disorder] This second definition includes gross distortions of size and shape. To be distorted a body part must be either disproportionately small or large, or be oddly shaped. Score 0-3: A score of 0 is given if all body parts are well proportioned. A score of 1 is given if 1 or 2 body parts are out of proportion or misshapen to a small extent. A score of 2 is given if half of the drawing is out of proportion or distorted. A score of 3 is given if more than half of the drawing is out of proportion or distorted (Handler, 1967), for anxiety, but also used for thought disorder by Maloney and Glasser (1982).

Body (Simple): [Anxiety] Score 0-3: A score of 0 is given if a) the trunk is well proportioned, b) the waist is narrower than the chest, c) the body has a three-dimensional quality, and d) the arms are appropriately placed on the body. A three-dimensional quality is usually indicated by emphasis lines indicating breasts, muscles, creases, folds in clothing, etc. A profile is deemed to have a three-dimensional quality only if there are clear indications of perspective in the drawing, e.g., the far leg rests higher on the page than the near one, the far arm is partially revealed behind the body of the figure, etc. Simply drawing the near arm in front of the body is not sufficient evidence of three-dimensionality in a figure drawn in profile. A score of 1 is given if the three-dimensional quality is absent and/or proportionality of the figure is not quite good enough for a score of 0. As well, if the shoulders of the figure are well-proportioned (i.e., sloping from the neck,

moving outward, and then rounding down to arms) this can be used to distinguish a 1 from a 2. A score of 2 is given if a) the waist is indistinguishable, but chest, etc. can be discerned or b) the trunk is more like a simple circle or square. A score of 3 is given if a) the trunk is square or round, arms are attached inappropriately, or there are other signs of poor form, or b) the figure is bizarre or grotesque, or amorphous (Handler, 1967), for anxiety, but also used for thought disorder by Maloney and Glasser (1982). For examples of figures at each scoring level see the Body Simple and Head Simple Scoring Examples.

Breasts (Delineated): [Thought Disorder] Scored 0-1: A score of 0 is given when there is no indication of breasts on the figure. A score of 1 is given when lines marking the breasts have been drawn. The figure may be either nude or clothed (Holzberg and Wexler, 1950). A score of 1 can only be given to a female figure.

Buttons (# of): [Thought Disorder] Scored 0-9: This feature is measured by counting the total number of buttons drawn down the shirt front of the figure to a maximum of 9 (Wainwright, 1970).

Clothing (Amount of): [Thought Disorder] Score 0-4: A score of 0 is given when the figure is overdressed (e.g., hats, scarves, gloves, etc.). To be overdressed a figure needs more than simply a hat or gloves. The figure must be dressed for a fall or winter day, or dressed in a suit coat or tux. A score of 1 is given when the figure is fully covered by their clothing (e.g., long sleeve shirt, shoes, and long pants or a long skirt). A score of 2 is given when the figure is dressed either in shorts or a short sleeved shirt, or any clothing which exposes skin which would be covered when the figure was fully dressed (e.g., arms, legs, shoulders, stomach, waist). A score of 3 is given when the figure is partly undressed or reveals much of the body. Examples of styles of dressing which would be scored 3 include not wearing a

shirt, dressed only in underwear, dressed in swim wear, or dressed in a halter top which reveals the stomach. A score of 4 is given when no clothes are drawn and the figure has obvious signs of nudity (e.g., a belly button, nipples, pubic hair, etc.) (adapted from Wainwright, 1970), for thought disorder.

Delineation Line Absent: [Anxiety] A delineation line is any line which divides the body into parts; sleeve cuffs, arm holes, cuffs, belt, pant line, neck line, etc. If a body part is omitted, it is not scored for absence of a delineation line. Delineation lines are expected at six body junctions when the figure is fully clothed; sleeve ends for the right and left arms, pant leg ends for the right and left leg, the neck line, and the waist line. In a complete figure one should look for these lines first. Additional lines may be missing from other pieces of clothing, e.g., socks or shoes, when the figure wears less than full clothing. These are not marked unless they are the primary delineation line for a body part. For example, shoes on an otherwise nude figure. It is assumed that any figure that does not show signs of clothing is clothed, unless there are obvious signs of nudity. Score 0-3: A score of 0 is given when no delineation lines are absent. A score of 1 is given when one delineation line is absent. A score of 2 is given when two delineation lines are absent. A score of 3 is given when more than two such lines are absent (Handler, 1967).

Erasures (General): [Thought Disorder] **1)** Score 0-3: A score of 0 is given when no erasures are observed. A score of 1 is given when one or more erasures are observed <u>in</u> any one body part. A score of 2 is given when one or more erasures are observed <u>in</u> any two body parts. A score of 3 is given when one or more erasures are observed <u>in</u> more than two body parts (Handler, 1967).

2) Score 0-9: This alternate measurement of erasures is a count of independent erasures on the drawing, where independent erasures is defined as the

erasure of separate lines or collections of lines. The score for this measurement is equal to the number of erasures observed on the figure up to a maximum of 9 (Handler, 1967; Wainwright, 1970), for anxiety.

Eyes (Circles): [Thought Disorder] Score 0-1: A score of 0 is given when the eyes are complete (with an outside circle and an iris with or without a pupil), or are represented by dots, or are drawn in some unusual manner other than an empty circle. A score of 1 is given when the eyes are represented by simple curved lines (e.g., a circle) with no iris, such as that drawn with the cartoon figure Annie (Holzberg and Wexler, 1950).

Eyes (Dots): [Thought Disorder] Score 0-1: A score of 0 is given when the eyes are not drawn as dots. A score of 1 is given when the eyes are represented by simple dots with no surrounding white space (Holzberg and Wexler, 1950).

Feet (Pointed): [Anger/Hostility] Scored 0-1: A score of 0 is given when the feet are drawn with normal looking blunted toes. A score of 1 is given when the tip of the foot is drawn in a point (Holzberg and Wexler, 1950).

Fingers (Unshapely): [Thought Disorder] Score 0-3: A score of 0 is given when the fingers of the figure are drawn in correct proportion. A score of 0 is also given for the drawing of a closed fist, provided that the hand is obviously a fist and not just short fingers. A score of 1 is given when the fingers are poorly proportioned, but with there length greater than their width. A score of 2 is given when the fingers of one or both hands are drawn with their width greater than their length, such that the fingers appear to be bumps rather than fingers, or look like sticks (adapted from Holzberg and Wexler, 1950). If only one or two fingers are unshapely, don't score as 1 or 2 unless grossly distorted. A score of 3 is given when fingers are absent.

Genitals: [Anger/Hostility] Score 0-1: A score of 0 is given when the genitals are not drawn. A score of 1 is drawn when realistic or unmistakably symbolic representations of genitals are drawn (Koppitz, 1968), for emotional indicators.

Hands (Big): [Anger/Hostility] Score 0-1: A score of 0 is given when the hands of the figure are drawn smaller than the face. A score of 1 is given when the hands are drawn as big as or bigger than face of the figure (Koppitz, 1968), for emotional indicators.

Hands (Hidden): [Thought Disorder] Score 0-1: A score of 0 is given when hands of the figure are apparent. A score of 1 is given when one or both hands are represented as hidden behind the back of the figure, in pockets, or in some other space where they cannot be seen (Holzberg and Wexler, 1950).

Head (Simple): [Anxiety; Thought Disorder] (This index is a developmental-like score assessing the degree to which the head is drawn in an oblong shape with sufficiently detailed features.) Score 0-3: A score of 0 is given if 1) the head is drawn as an oblong, 2) its features have a three dimensional quality, and 3) the features are appropriately placed on the head. Three dimensional qualities are those which give the figure depth, such as creases, wrinkles, a nose drawn in partial profile, etc. A score of 1 is given if 1) the head is somewhat oblong and 2) its features are reasonably placed, but lack three-dimensionality or substance. A score of 2 is given if 1) the head is more circular than oblong and 2) its features are represented by simple lines or dots. A score of 3 is given if the head is extremely circular with very simple features which poorly represent the face (Handler, 1967), for anxiety, but also used for thought disorder by Maloney and Glasser (1982). For examples of figures at each scoring level see the Body and Head Simple Scoring Examples.

High Heels: [Thought Disorder] Score 0-1: A score of 0 is given when high heels are not drawn. A score of 1 is given when high heels are drawn on the figure (Holzberg and Wexler, 1950).

Knee Joint: [Thought Disorder] Score 0-1: A score of 0 is given when the knee joint is not drawn. A score of 1 is given when the knee joint is indicated on the figure, either by a bent leg or by the drawing of the knee cap (Holzberg and Wexler, 1950).

Monster: [Anxiety] Score 0-1: A score of 0 is given when the figure appears to be human. A score of 1 is given when the figure represents a non-human creature, or a degraded or ridiculous person; the grotesqueness of the figure must appear deliberate on the part of the subject and not the result of a lack of maturity or drawing skill (Koppitz, 1968), for emotional indicators.

Mouth (Expression): [Anger/Hostility; Anxiety; Thought Disorder] Score 0-3: A score of 0 is given when the figure is smiling. A smile is most reliably indicated by both corners of the mouth curled up. A score of 1 is given when there is a neutral expression on the mouth. Such an expression is usually represented by a straight line, one side of the mouth curled up, or an open mouth. A score of 2 is given when the mouth is shown as sneering, shouting, or angry in expression. Such expressions are often indicated by both corners of the mouth curled down (adapted from Holzberg and Wexler's (1950) Smile and the insufficiently defined Mouth (Straight)). A score of 3 is given when the mouth is omitted.

Objects: [Anxiety; Thought Disorder] Scored 0-1: A score of 0 is given when no objects other than the figure are drawn. A score of 1 is given when an object other than the figure is drawn. The object may be on the figures person (e.g., in their hand) (Holzberg and Wexler, 1950) or it may be clouds, rain, snow, flying birds, etc. (Koppitz, 1968), for emotional indicators.

Objects (Weapons): [Anger/Hostility] Scored 0-1: A score of 0 is given when no objects other than the figure are drawn. A score of 1 is given when a weapon is drawn along with the figure.

Placement: [Anger/Hostility; Thought Disorder] Score 0-3: A transparent sheet equally divided into four quadrants is used to judge placement. The sheet should be placed over the figure such that the outside edges of the transparency align with the outside edges of three of the four sides of the drawing paper. The writing "top left" and the arrow should be in the top left corner when the base of the figure is along the 8 1/2 inch side of the paper. When the subject has used the 11 inch side of the paper as a base, then the transparency should be set on its side with the "top left" indicator in the bottom left hand corner. Scoring is as follows. A score of 0 is given when the vertical axis of the overlay is within one half inch of the vertical axis of the figure and the horizontal axis of the overlay falls between the knees and shoulders of the figure. A score of 1 is given when a) the vertical axis of the overlay intersects with the figure at a distance of more than 1/2 inch from the figure's midline, the figure is not in any one quadrant, and the intersection is through a body part other than the hands or feet or b) the vertical axis of the overlay is within 1/2 inch of the figure's midline, but the horizontal axis of the transparency falls above the shoulders or below the knees of the figure and figure is not in any one quadrant. A score of 2 is given when a) the figure is on the left hand side of the page with only the hands or feet extending over the vertical axis of the overlay or b) the figure is in any one quadrant but the upper left, only hands or feet may extend over the horizontal or vertical axes of the overlay. A score of 3 is given when the figure is completely in the upper left-hand quadrant. Only the hands and the feet may extend over the vertical and the horizontal axes of the

overlay (Handler, 1967), for anxiety. These rules apply even if a figure is at a considerable angle.

Playful Figure: [Anxiety] Score 0-1: A score of 0 is given when neither the details of the figure (e.g., facial expression) nor expressive stance reflect humor or playfulness. A score of 1 is given when either the details or the expressive stance of the figure communicate a kind of playful, humorous mood (Fox, Davidson, Lighthall, Waite, Sarason, 1958), for anxiety. A smile is usually found on such figures, but to be playful the expression on the face must be more expansive, e.g., a broad smile, dimples, happy eyes, etc. An adequate test of this quality is to cover up the smile. If the face stills looks happy, then give the figure a 1.

Profile Drawn: [Anxiety] Score 0-5: A score of 0 is given when the head and body faces front. A score of 1 is given when either 1) the head is at an oblique angle and the body is full front, or 2) the body is at an oblique angle and the head is full front. A score of 2 is given when the body and head are at an oblique angle. A score of 3 is given when either 1) the head is at a right angle and the body is either full front or at an oblique angle, or 2) the body is at a right angle and the head is either full front or at an oblique angle. A score of 4 is given when both the head and body are at a right angle to the front of the paper. A score of 5 is given when the figure is facing fully away from the rater (i.e., the figure is backwards) (adapted from Wainwright, 1970).

Rigid Figure: [Anxiety] Score 0-1: A score of 0 is given when the figure is in a relaxed and comfortable pose. A score of 1 is given when the figure appears rigid, with body parts appearing inflexible or unable to move. The key to a judgment of rigidity is in the appearance of the arms and legs. If all of the arms and legs are drawn straight, with no curve to them or bend at the joints, the figure will appear rigid. Rigidity

can also be partially determined by the stance of the figure. Rigid figures will usually have the arms and legs nearly vertical or at a slight angle off of the vertical. Another aspect of the figure that is helpful in identifying rigidity is the appearance that the figure would topple if pushed over or would be unable to recover if pushed over because the limbs are not flexible enough (Fox, Davidson, Lightfall, Waite, and Sarason, 1958), for anxiety.

Sex of First Drawn Figure: [Anxiety] Score 0-2: A score of 0 is given when a figure of the same gender as the subject is drawn. A score of 1 is given when the subject draws a figure of the opposite gender first. A score of 2 is given if the gender of the figure cannot be determined (Wainwright, 1970). Long hair on a figure is not enough to determine sex of the figure. Useful features include secondary sexual characteristics, dress, and features such as eyelashes. Determination of the sex of the drawing should be made prior to comparison with the sex of the subject.

Shoulders (Wide): [Thought Disorder] Score 0-1: A score of 0 is given when the shoulders appear to be in proportion to the rest of the body. A score of 1 is given when the shoulders appear to be disproportionately broad for the rest of the body (Holzberg and Wexler, 1950). Disproportionately large is defined as shoulders which appear to be more than twice as large as the hips and the face of the figure, or three times as wide as the head where the hips are also wide. Such a figure is usually square shouldered. That is, the shoulders appear to travel straight out and then bend sharply to the vertical. Further, the arms extending down from the shoulders usually fall some distance away from the body, even though the body is in proportion to the hips and head.

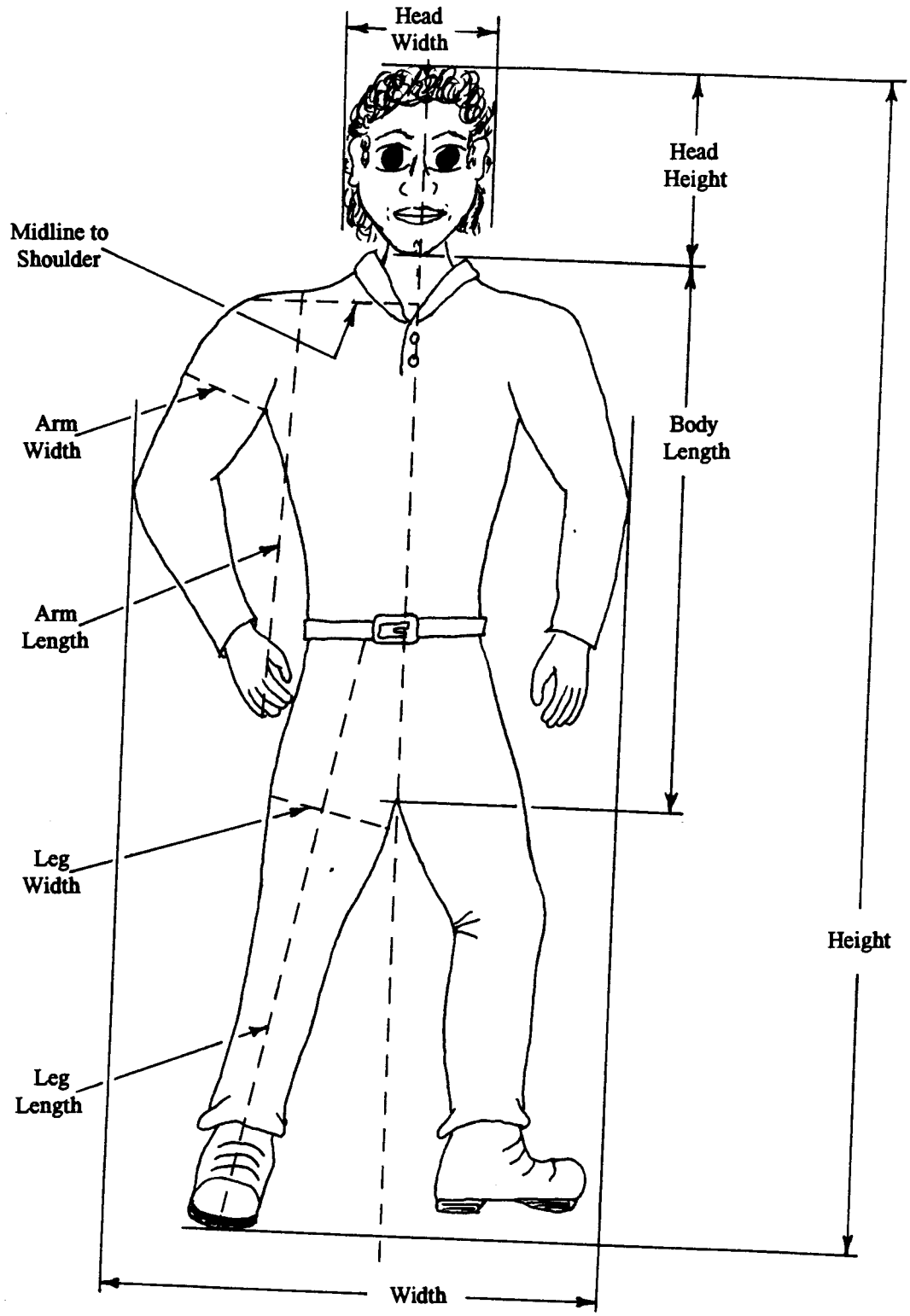
Skirt (Wide): [Thought Disorder] Score 0-1: A score of 0 is given when the skirt is narrow to medium in width or absent. A score of 1 is given when the skirt is wide and billowy (Holzberg and Wexler, 1950).

Teeth: [Anger/Hostility] Scored 0-1: A score of 0 is given when no teeth are indicated on the figure. A score of 1 is given when at least one tooth was evident in the drawing (Koppitz, 1968), for emotional indicators.

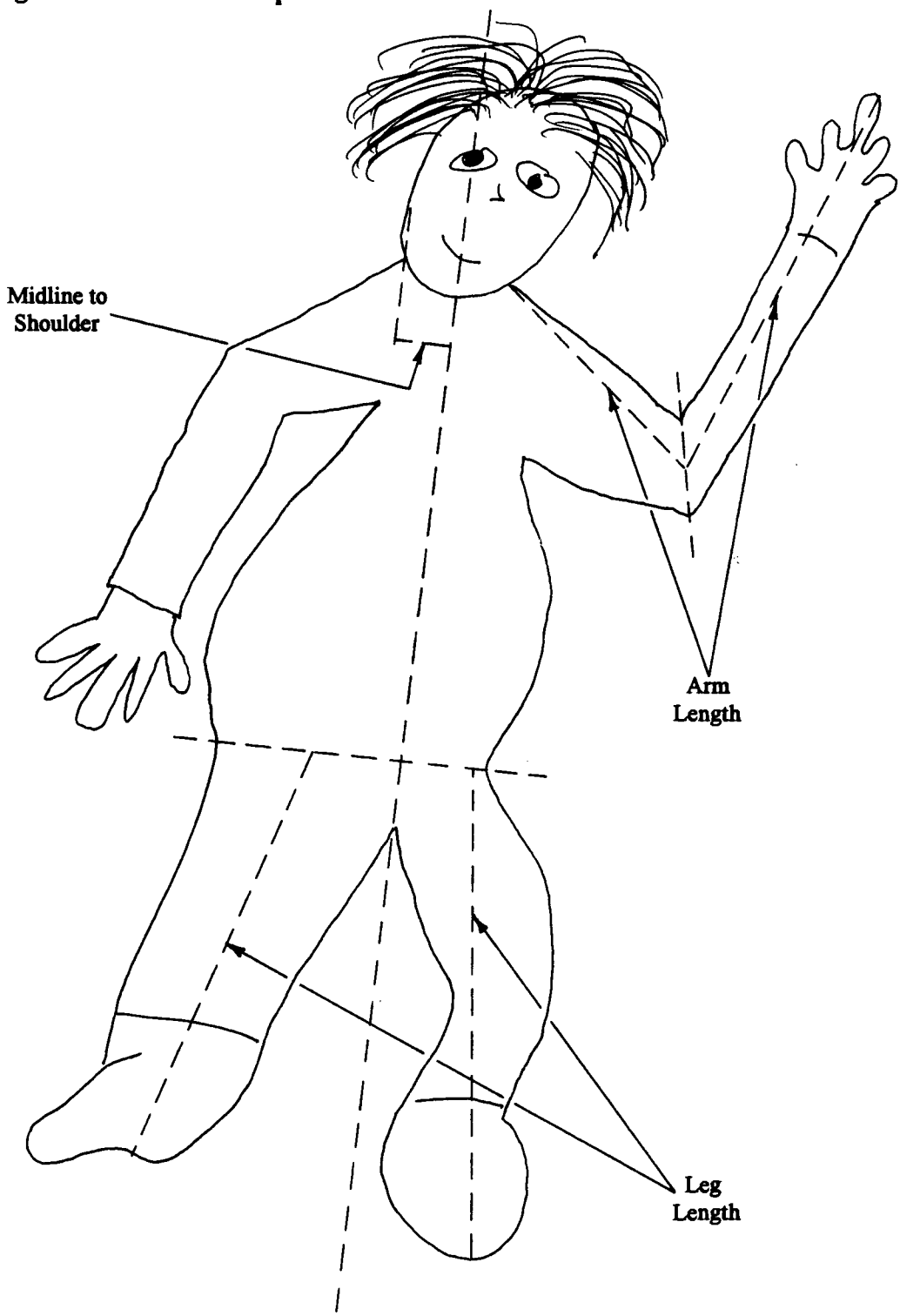
Transparency: [Thought Disorder] A transparency is defined as the inappropriate extension of a body line into clothing at garment borders of the neck, either arm, or either leg, as if the figures apparel were made of glass. An extended line has to exceed .2 cm in length. Body lines which extend into other body lines are also deemed to be transparencies. This is particularly important in figures which do not have clear delineation lines. Score 0-5: For each transparency observed, up to a maximum of 5, score 1 point (Prytula and Hiland, 1975), for anxiety. Note that for a simple figure drawn as a circle with arms and/or legs attached to the body, each point where the body line extends across the joint between the limb and the body is a transparency.

Wrinkles (# of): [Anger/Hostility] Score 0-9: This feature is scored by counting the total number of wrinkles observed on the figure to a maximum of 9 (Wainwright, 1970).

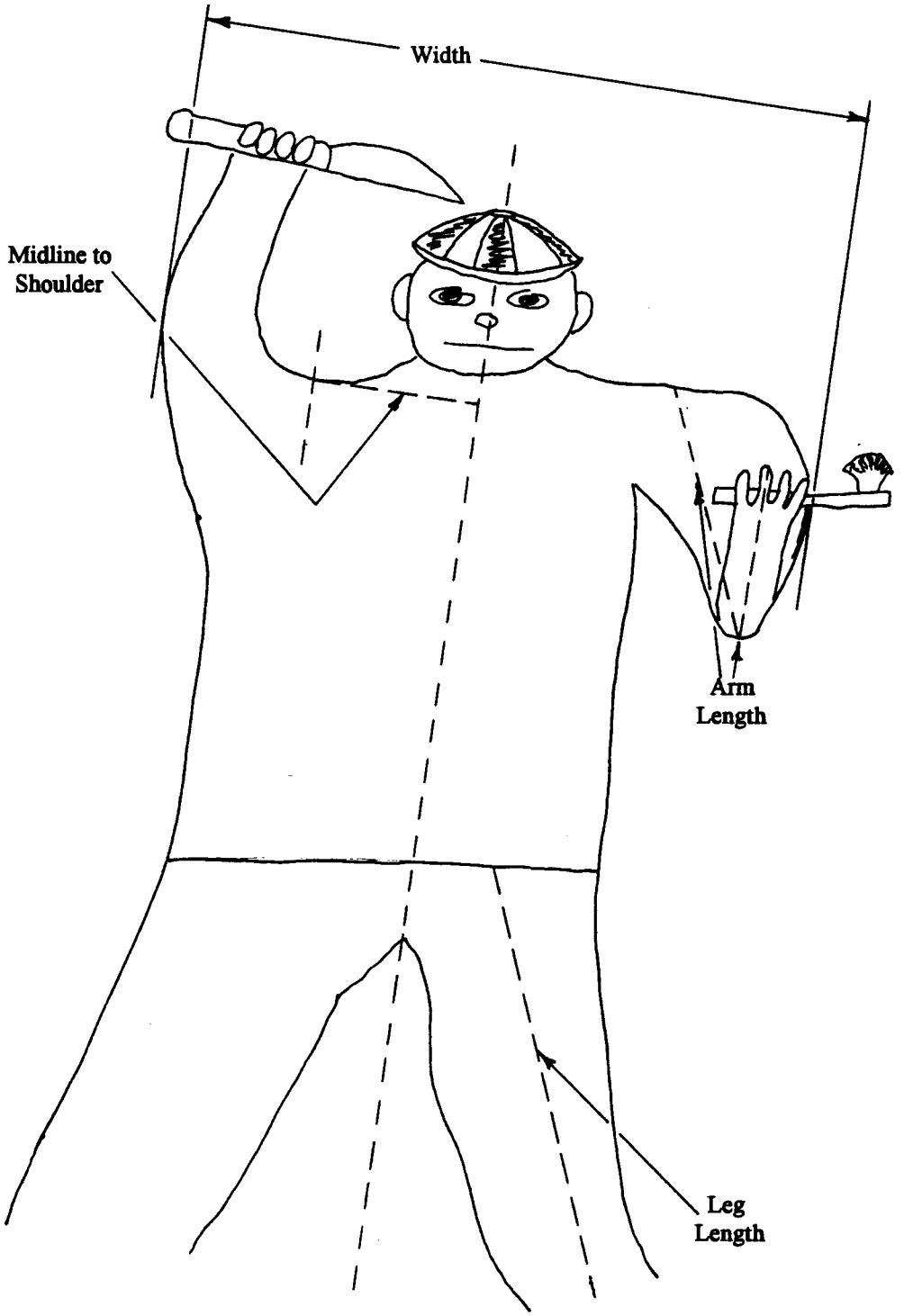
Drawing Measurement Example A



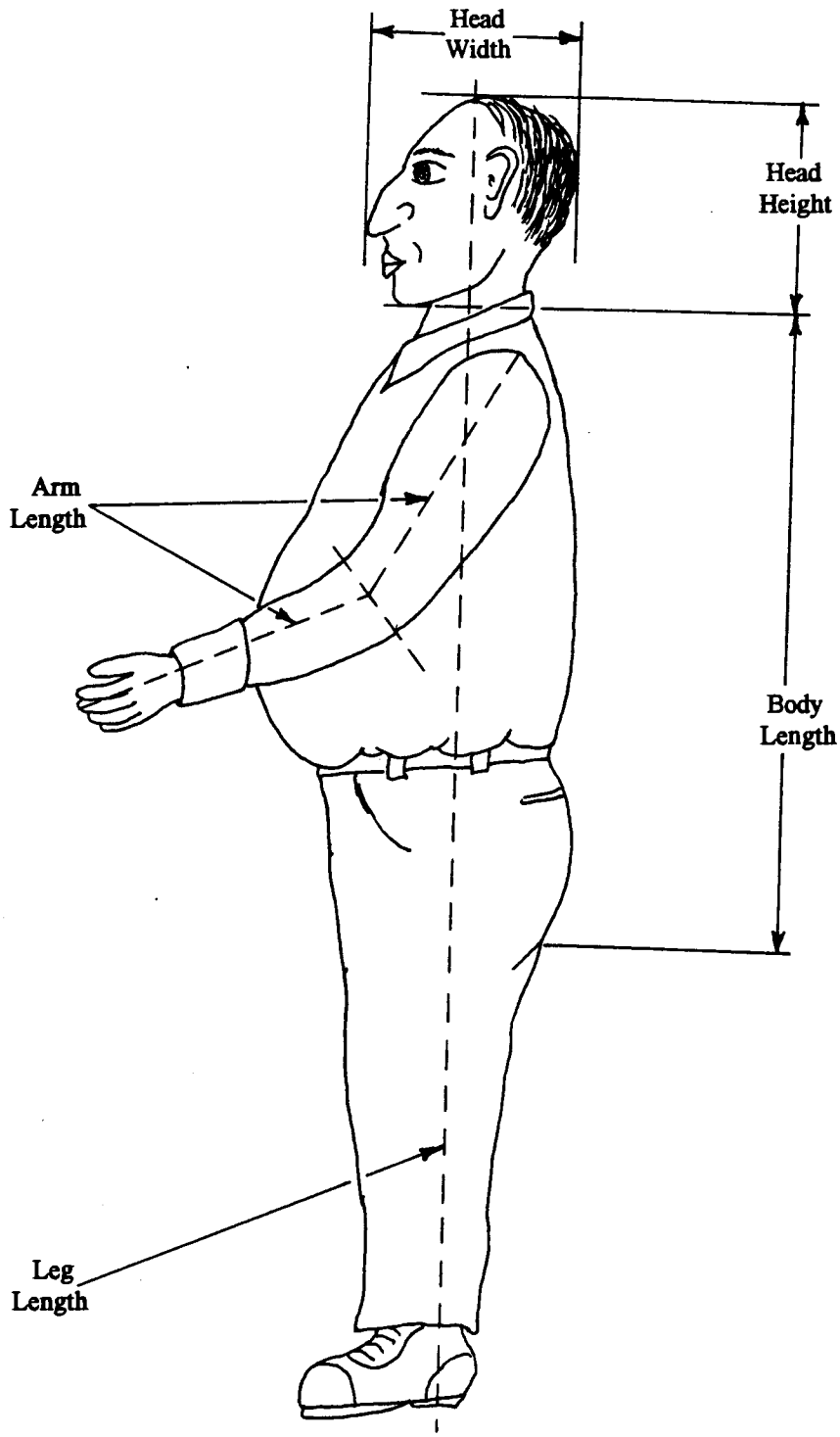
Drawing Measurement Example B



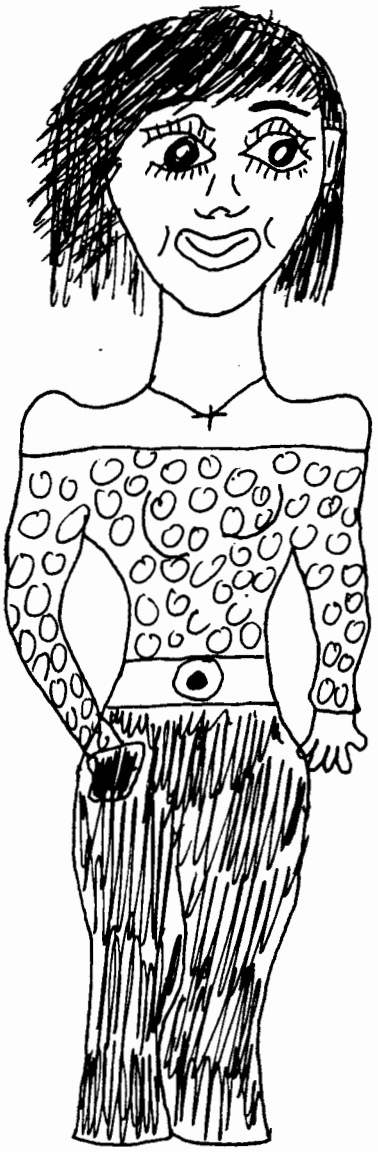
Drawing Measurement Example C



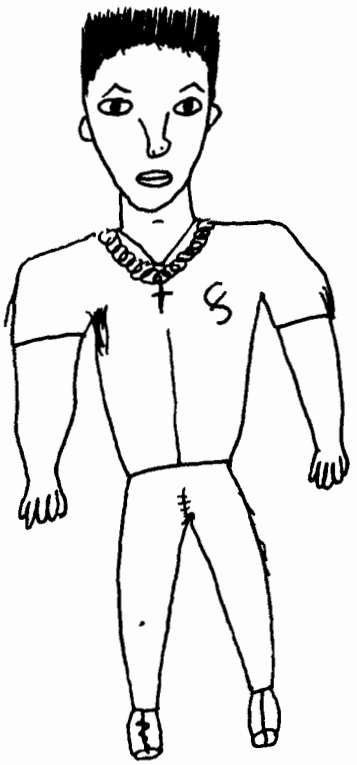
Drawing Measurement Example D



Body Simple and Head Simple Examples: Score 0



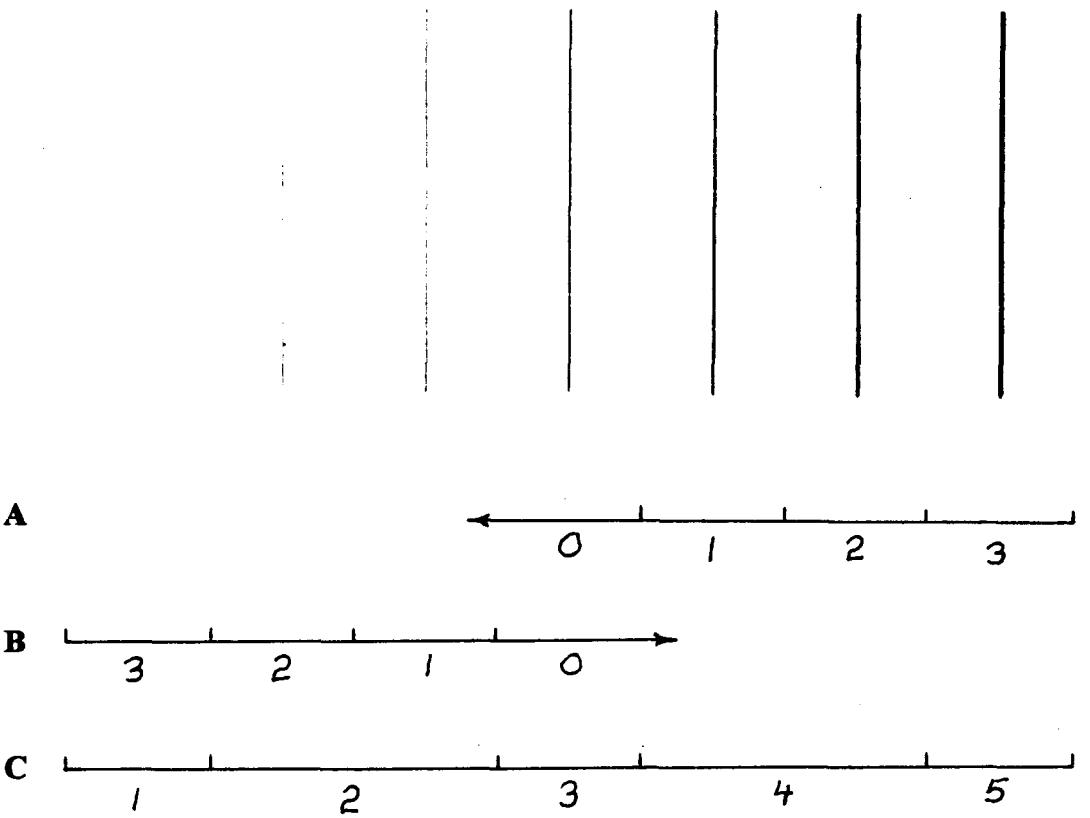
Body Simple and Head Simple Examples: Score 1



Body Simple and Head Simple Examples: Score 2

Body Simple and Head Simple Examples: Score 3

Line Reference Chart

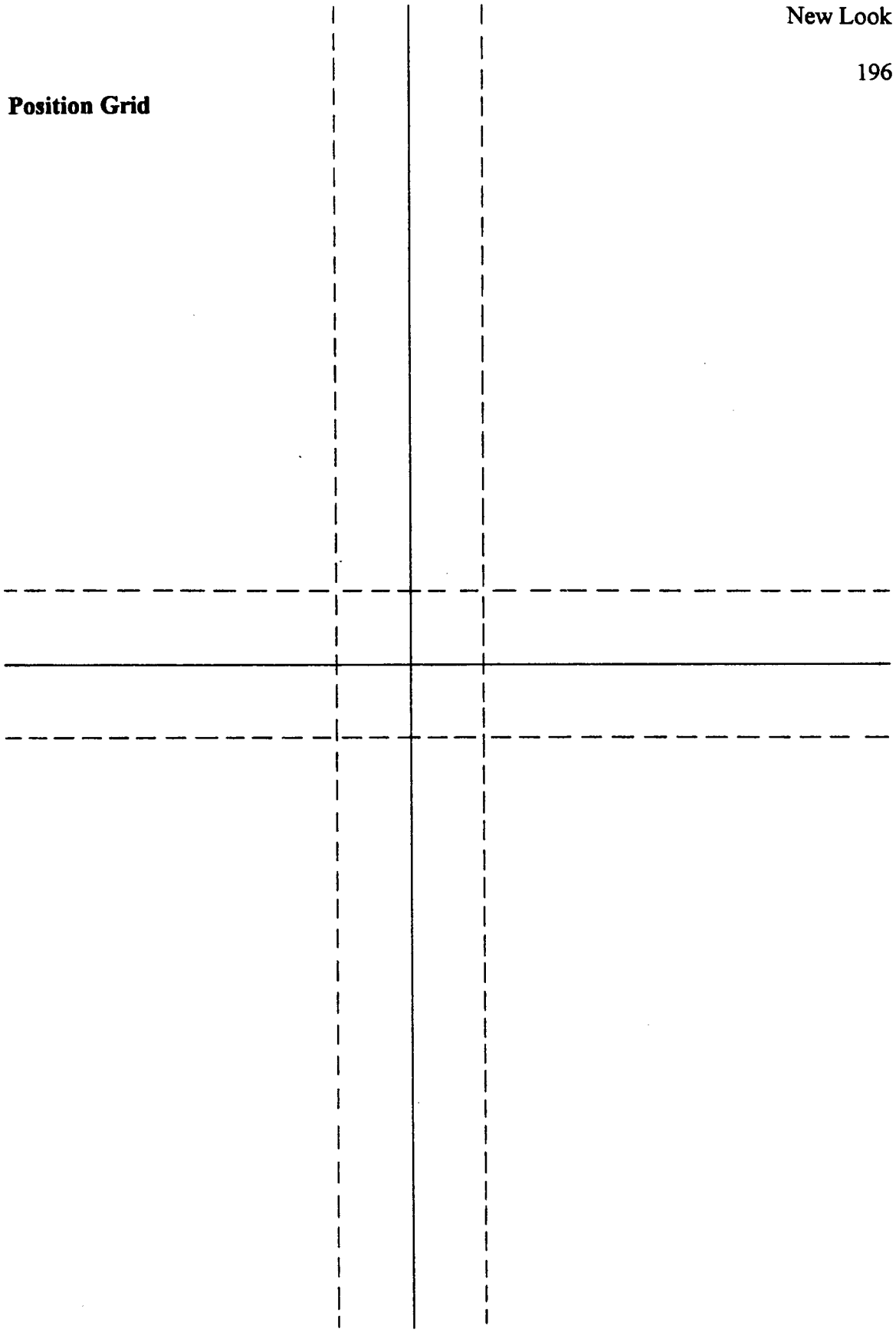


A: Rating scale A is for scoring Line (Heavy). Note that the three darkest lines have the numbers 1, 2, and 3 written below them. These numbers refer to the line heaviness scores, with 1 being medium-heavy, 2 being heavy, and 3 being very heavy. The arrow indicates that any lines from the midpoint to the left are scored as 0.

B: Rating scale B is for scoring Line (Light). Note that the three lightest lines have the numbers 1, 2, and 3 written below them. These numbers refer to the line lightness scores, with 1 being medium-light, 2 being light, and 3 being very light. The arrow indicates that any lines from the midpoint to the right are scored as 0.

C: Rating scale C is for the scoring of Line Pressure (Head), Line Pressure (Lower Body), and Line Pressure (Upper Body). The seven lines at the top of the page are broken into five divisions corresponding to the five levels of line pressure: very light, light, medium, heavy, and very heavy. Numbers below each division refer to line pressure scores. Please note that two lines are encompassed by ratings of "light" and "heavy" lines.

Position Grid



Human Figure Drawing Scoring Manual: Quick Reference Sheet

Head (Tiny/Small):
0 - 2.5 cms or greater
1 - 1.5 to 2.4 cms
2 - 1 to 1.4 cms
3 - 1 cm or less

Line (Discontinuous):
0 - 0 line breaks
1 - 1 line break
2 - 2 to 3 line breaks
3 - 4 to 5 line breaks
4 - 6 line breaks
5 - 7 to 8 line breaks
6 - 9 or more line breaks

Shading (Head):
0 - no shading
1 - shading hair or face
2 - shading hair and face

Arms (Down):
0 - 1 arm greater than 45 degrees
1 - both less than 45, 1 arm away from body
2 - both arms vertical
3 - no arms

Body (Simple):
0 - proportioned; waist narrow; 3-D; arms good
1 - no 3-D or proportion
2 - only chest; like simple square or circle
3 - arms inappr. attached or grotesque

Delineation Line Absent:
0 - no lines absent
1 - 1 line absent
2 - 2 lines absent
3 - > 3 lines absent

Fingers (Unshapely):
0 - correct proportion
1 - poor, but length , greater than width
2 - bumps or sticks
3 - no fingers

Mouth (Expression):
0 - smile; corners up
1 - neutral; straight line
2 - frown: sneer: angry expression
3 - mouth omitted

Placement:
0 - centered
1 - intersect > 1/2" or <1/2 and below knees or above shoulders
2 - left side or 1 quadrant, but not upper left
3 - upper left quadrant

Height (Small):
0 - 17.1 cms or greater
1 - 14.1 to 17 cms
2 - 11.6 to 14 cms
3 - 5.1 to 11.5 cms
4 - 5 cms or less

Line (Emphasis):
1 per body part

Line (Reinforced):
0 - no reinforcement
1 - 1 body part reinforced
2 - 2 parts
3 - 3 or more parts

Arms Behind Back
0 - arms shown
1 - hands behind back
2 - arms behind back

Bizarreness:
1) 0- non-bizarre
1 -some bizarreness
2 - quite bizarre
2) 0 - non-bizarre
1 - 1 to 2 parts bizarre
2 - 1/2 figure bizarre
3 - > 1/2 figure bizarre

Clothing (Amount of):
0 - overdressed
1 - fully clothed
2 - shorts or short sleeves
3 - underwear, swim wear, no shirt, or halter top
4 - obviously nude

Erasures:
1) 0 - no erasures
1 - 1 body part
2 - 2 body parts
3 - > 2 body parts
2) absolute number

Head (Simple):
0 - oblong; 3-D; features well placed
1 - < oblong; no 3-D
2 - circular; lines and dots
3 - circle and simple, poor features

Profile Drawn:
0 - head and body front
1 - head or body oblique
2 - head and body oblique
3 - head or body right angles
4 - head and body right angles
5 - back of figure drawn

Appendix D

Table of Inter-Rater Reliability Results for Scoring of Individual Drawing Features

Table D.1

Pearson Correlations, Intra-class Correlations, Kappa, and Number of Subjects in the Reliability Sample by Individual Drawing Features

Drawing Features	Statistics			
	Pearson Correlation	Intra-Class Correlation	Kappa	N
Arms (Asymmetry Length)	.88	.88		40
Arms (Asymmetry Width)	.89	.90		40
Arms (Length)	.97	.97		40
Head (Tiny/Small, def. 1)	.95		.80	20
Head (Tiny/Small, def. 2)	1.00		1.00	20
Head:Body Ratio	1.00	.99		40
Height (Overall)	1.00	1.00		40
Height (Small)	.99		.93	40
Legs (Asymmetry Length)	.73	.68		40
Legs (Asymmetry Width)	.84	.84		40
Leg (Length)	1.00		1.00	20
Shoulders (Asym. Width)	.48	.57		40
Vertical Imbalance	.96	.96		40
Width	.99	.99		40
Line (Discontinuous)	.85	.85		40

Table D.1 Cont'd

Drawing Features	Statistics			
	Pearson Correlation	Intra-Class Correlation	Kappa	N
Line (Heavy)	.93		.73	20
Line (Light)	.84		.67	20
Line Emphasis	.86	.86		40
Line Pressure (Head)	.96		.80	25
Line Pressure (Lower Body)	.91		.70	25
Line Pressure (Upper Body)	.91		.55	25
Line (Reinforced)	.65		.40	20
Omission (Arms)	1.00		1.00	20
Omission (Ears)	1.00		1.00	20
Omission (Eyes)	.79		.77	20
Omission (Feet)	.91		.90	20
Omission (Hands)	.80		.80	20
Omission (Legs)	.55		.46	20
Omission (Mouth)	.73		.69	20
Omission (Neck)	1.00		1.00	20

Table D.1 Cont'd

Drawing Features	Statistics			
	Pearson Correlation	Intra-Class Correlation	Kappa	N
Omission (Nose)	.80		.80	20
Omission (Shoulders)	.70		.70	20
Omission (Waist)	1.00		1.00	20
Shading (Arms)	.80		.80	20
Shading (Chest)	.80		.80	20
Shading (Feet)	1.00		1.00	20
Shading (Hands)	.60		.60	20
Shading (Head)	.84		.70	30
Shading (Legs)	1.00		1.00	20
Shading (Mouth)	.62		.60	20
Shading (Neck)	1.00		1.00	20
Shading (Waist)	.80		.80	20
Arms (Reinforced)	.80		.80	20
Clothing (Reinforced)	.80		.80	20
Eye or Ear (Reinforced)	1.00		1.00	20
Eyebrow (Reinforced)	.49		.49	20
Feet (Reinforced)	.80		.80	20

Table D.1 Cont'd

Drawing Features	Statistics			
	Pearson Correlation	Intra-Class Correlation	Kappa	N
Hands (Reinforced)	.60		.60	20
Leg (Reinforced)	.70		.70	20
Midline (Reinforced)	.60		.60	20
Mouth (Reinforced)	.40		.40	20
Neck (Reinforced)	.80		.80	20
Shoulders (Reinforced)	.62		.60	20
Arms (Detail)	.70		.70	20
Clothing (Detail)	.82		.80	20
Eye or Ear (Detail)	.80		.80	20
Eyebrow (Detail)	.56		.56	20
Feet (Detail)	1.00		1.00	20
Hands (Detail)	.80		.80	20
Leg (Detail)	.91		.90	20
Midline (Detail)	.80		.80	20
Mouth (Detail)	.80		.80	20
Neck (Detail)	.66		.60	20
Shoulders (Detail)	.80		.80	20

Table D.1 Cont'd

Drawing Features	Statistics			
	Pearson Correlation	Intra-Class Correlation	Kappa	N
Age of Figure Discrepant From Subject	.90	.90		40
Arms (Behind Back)	.90		.80	30
Arms (Down)	.78		.46	30
Arms (Long)	.80		.80	20
Asymmetry of Limbs	.80		.80	20
Bizarreness (Def. 1)	.85		.70	30
Bizarreness (Def. 2)	.83		.40	20
Body Simple	.80		.53	20
Breasts Delineated	.80		.80	20
Buttons (# of)	.99	.73		40
Clothing (Amount of)	.98		.90	25
Delineation Line Absent	.72		.60	20
Erasures (Def. 1)	.96		.87	20
Erasures (Def. 2)	.64	.64		40
Eyes (Circles)	1.00		1.00	20
Eyes (Dots)	.70		.70	20
Feet (Pointed)	.80		.80	20

Table D.1 Cont'd

Drawing Features	Statistics			
	Pearson Correlation	Intra-Class Correlation	Kappa	N
Fingers (Unshapely)	.65		.60	30
Genitals	1.00		1.00	20
Hands (Big)	1.00		1.00	20
Hands (Hidden)	1.00		1.00	20
Head (Simple)	.80		.53	20
High Heels	1.00		1.00	20
Knee Joint	.80		.80	20
Monster	.30		.30	20
Mouth (Expression)	.89		.90	30
Objects	.80		.80	20
Objects (Weapons)	1.00		1.00	20
Placement	1.00		1.00	20
Playful Figure	.70		.70	20
Profile Drawn	1.00	1.00		40
Rigid Figure	.60		.60	20
Sex of First Drawn Figure	.72		.70	30
Shoulders (Wide)	.73		.70	20

Table D.1 Cont'd

Drawing Features	Statistics			
	Pearson Correlation	Intra-Class Correlation	Kappa	N
Skirt (Wide)	.49		.48	20
Teeth	1.00		1.00	20
Transparency	.79		.77	40
Wrinkles (# of)	1.00	1.00		40