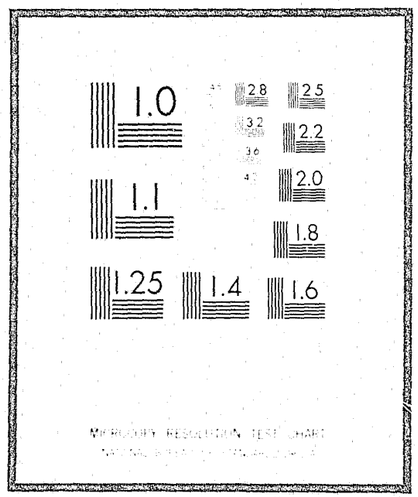

4

A REVIEW OF I-LEVEL RELIABILITY AND ACCURACY IN THE CALIFORNIA COMMUNITY TREATMENT PROJECT

by

Ted Palmer and Eric Werner

Sponsors:

CALIFORNIA YOUTH AUTHORITY

and

NATIONAL INSTITUTE OF MENTAL HEALTH

Community Treatment Project Report Series: 1972, No. 2. Fall, 1972.

STATE OF CALIFORNIA

Ronald Reagan
Governor

Earl W. Brian, M.D.
Secretary
Health and Welfare Agency

DEPARTMENT OF THE YOUTH AUTHORITY

Allen F. Breed
Director

George R. Roberts
Chief Deputy Director

Lyle Egan
Chief, Division of Rehabilitation

Keith S. Griffiths, Ph.D.
Chief, Division of Research
and Development

Youth Authority Board Members

Allen F. Breed, Chairman
Julio Gonzales, Vice Chairman
Ed Bowe
Richard Calvin
Rudolph Castro
William L. Richey

COMMUNITY TREATMENT PROJECT* STAFF

Ted Palmer, Ph.D.
Principal Investigator

James K. Turner
Research Associate

Eric J. Werner, M.A.
Co-Investigator

John Helm
Research Associate

David D. Sams, ACSW
Research Associate

*Official title: An Evaluation of Differential Treatment for Delinquents.
This study is supported by PHS Research Grant No. MH 14734, NIMH,
(Center for Studies of Crime and Delinquency).

TABLE OF CONTENTS

This is a somewhat technical report of updated information on the reliability and accuracy of I-level classifications. These classifications ('diagnoses') relate to a large number of delinquent adolescents, ages 13 through 19, who had been committed to the California Youth Authority (CYA) from local Juvenile Courts. The basic data of this study were I-level diagnoses of youths who comprised the study sample of the CYA's Community Treatment Project (CTP). All diagnoses were made by CTP staff.

Previously reported findings related to the period which extended from 1961 through late 1965.[1] They had reference to the Sacramento and Stockton areas alone; and, they did not differentiate between males and females. The present information extends from 1961 through 1969, thereby covering the entire CTP Phase I and Phase II operation.[2] It relates to all three study areas--Sacramento, Stockton-Modesto, San Francisco--and is broken down separately for males and females.[3]

I Interrater-Reliability At A Single Point In Time

Here, the data in question relate to the situation in which two different research raters each classified--at virtually the same point in time--the tape-recorded intake interview which was conducted with each youth.[4] This situation applied to a total of 364 males, i.e., 45% of all 802 Phase I and Phase II males.[5] (Interrater-reliability for females will be taken up in later pages.)

The overall results are given separately for subtype and I-level classifications.

> Any given youth may receive 1 of 9 subtype classifications: Aa, Ap, Cfm, Cfc, Mp, Na, Nx, Se or Ci. Simultaneously, he may receive 1 of 3 I-level classifications: I_2, I_3, or I_4.[6] Logically speaking, the rater must decide upon the youth's I-level classification prior to determining the subtype classification. In actual practice, the two judgments, or decisions, often take place almost simultaneously.

In the case of males, the first and second research raters agreed with one another as to the youth's subtype 62% of the time. They agreed with one another regarding the youth's I-level 81% of the time.

The percentage of agreement between the first and second research raters was as follows for the separate subtypes. (These figures are shown in relation to the final--i.e., 'true'--subtype-classification which was determined for each given individual): Aa - 33%; Ap - 81%; Cfm - 75%; Cfc - 74%; Mp - 34%; Na - 49%; Nx - 71%; Se - 79%; Ci - 67%. (The subtype sample-sizes were: 3, 16, 51, 38, 47, 82, 78, 19, and 30, respectively.)

The percentage of agreement between the first and second research raters was as follows for the separate I-levels: I_2 - 79%; I_3 - 79%; I_4 - 83%. These figures refer to interrater-agreement in relation to the I-level which was determined to be the youth's true I-level. (As to I-level agreement per se--irrespective of whether the raters had agreed with each other regarding the true I-level--the figures were: I_2 - 84%; I_3 - 79%; I_4 - 85%.) The sample-sizes were: 19, 136, and 209 for the I_2, I_3, and I_4 levels, respectively. Only one I_5 was included within the present analysis. The first and second research raters agreed on his I-level...though not on his subtype. One called him an Na, and the other an Nx.

We will now break this down in various other ways, once again, separately for subtype and I-level.

A. First to the subtype classifications. 48.9% of the 1st-2nd research rater disagreements were 1 subtype-classification apart (e.g., diagnosis by first research rater = Cfm; diagnosis by second research rater = Cfc). 18.7% of the disagreements were 2 subtype-classifications apart (e.g., first research rater's dx = Cfm; second research rater's dx = Mp). 20.9% were 3 categories apart (e.g.,...Cfm vs. Na). The remaining figures were 5.8%, 3.6%, and 2.2% for 4-, 5- and 6- subtype-classifications apart, respectively. Diagnostic disagreements between the first and second research raters were 2.09 subtype-classifications apart, on the average; these same disagreements were separated by a median of 1.56 subtype-classifications. These results rather clearly support the idea that interrater-disagreements were more likely to involve adjacent categories (or, relatively similar classifications), instead of those which were widely or even randomly separated (or, relatively dissimilar classifications). (The results appear to be meaningful irrespective of our belief that there exists no single underlying continuum of subtypes within I-level. See pg. 4, paragraph 2, for further comment.)

Three examples will further illustrate this point:

(1) Among males who were finally diagnosed as Mp's (i.e., the 'Mp' label represented the youth's 'true' diagnosis insofar as CTP was concerned), there were a total of 31 1st-2nd research-rater disagreements during 1961-1969. Of these 31, 7 involved an Mp-Cfc combination (i.e., the first research rater classified the youth as an Mp, whereas the second research rater called him a Cfc...or vice versa); 6 disagreements involved an Mp-Na combination; 4 involved an Mp-Cfm combination; 3 involved an Mp-Ci combination. The remaining disagreements (i.e., subtype-combinations) each had a frequency of fewer than 3.

(2) In the case of youths whose final diagnosis was Na, there were a total of 42 1st-2nd research-rater disagreements during 1961-1969. Of these 42, the most common 'disagreement-combinations' were as follows (together with their frequencies): Na-Mp = 8 disagreements; Na-Nx = 7; Na-Cfm = 5; Na-Cfc = 5. The remaining subtype-combinations each had a frequency of 3 or less.

(3) As to the Nx's (23 disagreements in all): Nx-Na = 12 disagreements; Nx-Se = 4; Nx-Mp = 3. The remaining disagreements each had a frequency of 2 or less.

We performed some statistical tests, and derived a number of readily interpretable indices, with respect to the above sample of 364 males. Included were: Chi Square (X^2); Cramér's Statistic (ϕ'); Goodman and Kruskal's Index of Predictive Association - i.e., lambda (λ); Percentage of Interrater Agreement ("Rater agreement"); Percentage of Interrater Agreement minus the Percentage of Agreement expected on the basis of Chance Guessing alone ("Rater agreement minus chance"). Much of what follows relates to these tests and indices.

Relative to the tests and findings next reported, the 9 subtype-classifications were first reduced to a total of 7. This was done by combining the Aa's with Ap's, on the one hand, and the Se's with Ci's on the other. This seemed to be a worthwhile move in view of the relatively small numerical representation of these particular subtypes--especially the Aa's, Ap's and Se's.

Relative to the classification of youths by subtype (N = 364 males; 7 categories), the interrater-reliability results were:

X^2 = 795.1 (d.f. = 36; $p < .001$)

ϕ' = .60

λ (symmetrical) = .55[7]

Rater agreement (7 categories) = 65%

Rater agreement minus chance[8] = 51%

The obtained ϕ' and λ appear quite encouraging, in relation to the present sample-size. This applies to the 'rater-agreement-minus-chance' figure, as well. (As compared with the latter figure, λ can be considered a more useful measure of reliability with regard to the present analyses. It takes account of CTP's subtype--and, where appropriate, I-level--base rates...whereas 'rater agreement minus chance' does not. Whether at CTP or not, many raters are in possession of at least some information concerning population base rates; and, in certain instances, their information may be rather accurate and complete. If they choose to utilize and/or 'fall back upon' information of this type, individual raters may, under given conditions, make substantially better classifications than if they were merely to guess. As a result, raters could also be more likely to agree with one another than would be the case in relation to 'chance guessing', alone. The present issue would be somewhat less pertinent with respect to differing sets of base rates. Within CTP, however, Na's and Nx's each comprised more than 25% of the male sample. Collectively, the I_4 group comprised approximately 65% of the male sample.)

A Pearson r was computed largely for the purpose of making a general comparison with the other tests and indices (e.g., the ϕ'). It was computed despite the fact that--except at certain steps along the way--the subtype-classification 'series' (moving from Aa to Ap, then on to Cfm... and so on through Ci) does not fully represent a maturational progression. (Nx's are not necessarily "more mature" than Na's; similarly, Mp's are not necessarily more mature than Cfc's.) Because of this, it does not comprise the type of measurement scale which would really satisfy one of the key statistical assumptions involved in the interpretation of a Pearson r. This is apart from the fact that the series of I-level classifications (i.e., $I_2 \rightarrow I_3 \rightarrow I_4$) does, on the other hand, represent a relatively unbroken and coherent maturational progression; on at least this score it comprises a sufficiently suitable type of scale relative to the Pearson r. In any event, the Pearson r was +.72 with respect to the subtype-classifications.

8. The following relates to the I-level classifications (3 categories in all). In any given instance the 1st and 2nd research raters may have disagreed with one another as to the youth's subtype, while still agreeing with each other as to his I-level. Thus, with reference to instances of subtype-disagreement (between 1st and 2nd research raters) in connection with youths whose true I-level was I_2 (total of 5 subtype-disagreements), the raters' joint I-level classifications were: I_2-I_2 - 20%; I_2-I_3 - 40%; I_2-I_4 - 20%; I_3-I_3 - 20%. Comparable figures for youths whose true diagnosis was I_3 (54 instances of subtype-disagreement) were: I_2-I_2 - 0%; I_2-I_3 - 11%; I_2-I_4 - 2%; I_3-I_3 - 46%; I_3-I_4 - 39%; I_4-I_4 - 2%. The figures for I_4's (79 subtype-disagreements) were: I_2-I_2 - 0%; I_2-I_3 - 0%; I_2-I_4 - 1%; I_3-I_3 - 5%; I_3-I_4 - 39%; I_4-I_4 - 54%.

Still within I-level, the most common interrater subtype-disagreements were as follows. (The youths are shown in terms of their 'true' I-level):

I_2 youths (total of 5 1st-2nd research-rater disagreements): The most common type of disagreement involved the Aa-Cfm combination. (N = 2 disagreements of this type.)

I_3 youths (total of 54 disagreements): The most common disagreements were Mp-Na (N = 9 such disagreements); Cfc-Mp (N = 7); Cfm-Mp (N = 7). The remaining subtype-combinations each had a frequency of 5 or fewer.

I_4 youths (total of 79 disagreements): The most common disagreements were Na-Nx (N = 20); Na-Ci (N = 9); Na-Mp (N = 8). The remaining subtype-combinations each had a frequency of 5 or fewer.

Relative to the classifications of youths by I-level (N = 364 males; 3 categories), the above-mentioned tests and indices are summarized as follows:

X^2 = 321.3 (d.f. = 4; $p < .001$)

ϕ' = .66

λ (symmetrical) = .56

Rater agreement (3 categories) = 81%

Rater agreement minus chance = 48%[9]

Pearson r = +.69[10] (r_{max} = +.99)*

*"r_{max}" refers to the numerically highest Pearson r that can be obtained with respect to given joint frequency distributions. (See: Carroll, John B., The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26, No.4. 1961. 347-372.) In line with the issues reviewed on pg. 4 paragraph 3, and in fn.10 as well, r_{max} was computed only in relation to the 3-category (I_2, I_3, I_4) and 4-category (I_2, I_3, I_4 Neurotic, I_4 Non-Neurotic) breakdowns.

The following applies to <u>females</u>:

The 1st and 2nd research raters agreed with one another as to the youth's <u>subtype</u> 70% of the time. They agreed with one another regarding the youth's <u>I-level</u> 85% of the time. The subtype figure has reference to a total of 71 females - everyone whose intake tape had been rated by two different researchers. (This represents 33% of the 212 Phase I and Phase II females.)

The percentage of agreement between the first and second research raters was as follows for the separate <u>subtypes</u>. (These figures are shown in relation to the final--i.e., 'true'--subtype-classification which was determined for each given individual): Aa - 0%; Ap - 100%; Cfm - 100%; Cfc - 67%; Mp - 55%; Na - 57%; Nx - 87%; Se - 100%; Ci - 100%. (The subtype sample-sizes were: 1, 2, 4, 3, 11, 21, 23, 3, and 2, respectively. There was one I_5.) Because of the extremely small sample-sizes, essentially no confidence should be placed in the percentage-of-agreement which is reported for each subtype other than the Na, the Nx, and, possibly, the Mp subtype.

The percentage of agreement between the 1st and 2nd research raters was as follows for the separate <u>I-levels</u>. (Figures are shown in terms of the individual's 'true' I-level): I_2 - 67%; I_3 - 67%; I_4 - 76%. The sample sizes were as follows: 3, 18, and 49 for the I_2, I_3, and I_4 levels. respectively. (As to the I_5, the first and second research raters disagreed on her I-level.)

We will now break this down further, separately for subtype and I-level.

1. First to the <u>subtype</u> classifications. 61.9% of the 1st-2nd research rater disagreements were 1 subtype-classification apart. 23.8% of the disagreements were 2 subtype-classifications apart. The remaining figures were 4.8% and 9.5% for 3- and 4- subtype-classifications apart, respectively. The diagnostic disagreements between the first and second research raters were 1.62 subtype-classifications apart on the average; the disagreements were separated by a median of 1.31 subtype-classifications.

   Relative to the tests and findings next reported, the 9 subtype-classifications were reduced to a total of 7. This was done by

combining the Aa's with Ap's, on the one hand, and the Se's with Ci's on the other.

The results were:

χ^2 = 217.2 (d.f. = 36; $p < .001$)

ϕ' = .72

λ (symmetrical) = .64

Rater agreement (7 categories) = 76%

Rater agreement minus chance = 61%

Pearson <u>r</u> = +.86

For the purpose of further comparison--and particularly in view of the very small sample-sizes which were present relative to several subtypes--the 9 (and/or 7) subtype-classifications were reduced to a total of 4. (Included were: I_2, I_3, I_4 Neurotic, and I_4 Non-neurotic.)

The results were:

χ^2 = 122.9 (d.f. = 9; $p < .001$)

ϕ' = .77

λ (symmetrical) = .66

Rater agreement = 87%

Rater agreement minus chance = 62%

Pearson <u>r</u> = +.78 (r_{max} = +.97)

These findings are not very different than those which resulted from the 7-category analysis.

2. Relative to the classification of youths by <u>I-level</u> (N = 70 females), [11] the results were:

χ^2 = 87.0 (d.f. = 4; $p < .001$)

ϕ' = .79

λ (symmetrical) = .67

Rater agreement = 85%

Rater agreement minus chance = 52%

Pearson <u>r</u> = +.78 (r_{max} = +.98)

II Rater-Reliability Through Time

Thus far we have been speaking of interrater-reliability at a single point in time--viz., intake. A different set of figures was obtained when we analyzed rater-reliability through time. The latter figures were found to be somewhat higher than the former, though not markedly so.

> The 'through time' analysis refers to the 'research rating at point of intake' as compared with the 'research rating based upon a routine followup interview'...or, in a few cases, a 'revocation' or 'discharge' type of followup interview. The time-interval between intake and followup ratings was usually around 8 to 12 months (estimated). The analysis related to all subject-groupings and all locations combined; in addition, it covered the entire Phase I and II operation, 1961-1969.

Results are shown in Table 1--separately for (a) males and females, (b) subtype and I-level, and (c) 'single research rater' (i.e., researcher 'X' classified the youth at intake as well as at followup) as distinct from 'different research raters' (i.e., researcher 'X' classified the youth at intake, whereas researcher 'Y' classified him at followup).

Table 1

CTP Interrater-Agreement Through Time (Intake vs. Followup)

	MALES				FEMALES			
	Subtype		I-Level		Subtype		I-Level	
Type of Rater	No. of Youths	% of Agreemt.	No. of Youths	% of Agreemt.	No. of Youths	% of Agreemt.	No. of Youths	% of Agreemt.
Single Rater	256	75.8	256	91.4	45	80.0	45	93.3
Different Raters	170	74.7	170	91.2	45	75.6	45	84.4
Total	426	75.4	426	91.3	90	77.8	90	88.9

It might be of interest to note that the above figures (the subtype figures in particular) are considerably higher than the 'through time' results reported by Jesness for a sample of 525 males who were institutionalized at the CYA's Preston School of Industry, during 1966-1968.[12] (It is likely that approximately 25 to 35 of these particular youths had been CTP study subjects as well, at some prior point in time.) The analysis related to the following: Subtype diagnoses based upon the Jesness Inventory were determined on a 'before-after' basis--i.e., at point of admission to Preston, and, afterwards, shortly prior to departure. As in the case of CTP data, the usual before-after time-interval was 8 to 12 months (estimated). (In passing, it might be mentioned that the Inventory-based results involved a substantially higher proportion of I_3 classifications than were found within CTP's own study sample over the years. The CTP classifications were based almost exclusively upon the interview method of diagnosis.) The main results were:

Subtype agreement-through-time - 39%

I-level agreement-through-time - 67%.

If one includes only the sub-sample of 270 individuals who had a 'high pre-inventory probability' (.50 or greater) with regard to their highest subtype classification, the figures would then rise to 49% and 71%, respectively.

Getting back to the CTP data on rater-reliability through time, the subtype results were as follows for all males combined (N = 426; 7 categories):

X^2 = 1357.3 (d.f. = 36; $p < .001$)

ϕ' = .73

λ (symmetrical) = .69

Rater agreement = 77%

Rater agreement minus chance = 62%

Pearson r = +.86

Relative to the classification of youths by I-level, the results were as follows for all males combined (N = 426; 3 categories):

X^2 = 560.4 (d.f. = 4; $p < .001$)

ϕ' = .81

λ (symmetrical) = .80

Rater agreement = 91%

Rater agreement minus chance = 58%

Pearson r = +.85 (r_{max} = +.94)

Separate breakdowns for 'single raters' and 'different raters' appear in fn. 13.

With regard to the classification of youths by subtype, the results were as follows for all females combined (N = 90; 7 categories):

χ^2 = 206.8 (d.f. = 36; $p < .001$)

ϕ' = .62

λ (symmetrical) = .64

Rater agreement = 77%

Rater agreement minus chance = 62%

Pearson r = +.78

The results, for females, were as follows, when the subtype-classifications were reduced to a total of 4 (I_2, I_3, I_4 Neurotic, I_4 Non-Neurotic):

χ^2 = 100.4 (d.f. = 9; $p < .001$)

ϕ' = .61

λ (symmetrical) = .67

Rater agreement = 87%

Rater agreement minus chance = 62%

Pearson r = +.93 (r_{max} = +.95)

As to the classification of youths by I-level the results were as follows for all females combined (N = 90; 3 categories):

χ^2 = 57.4 (d.f. = 4; $p < .001$)

ϕ' = .56

λ (symmetrical) = .68

Rater agreement = 90%

Rater agreement minus chance = 57%

Pearson r = +.79 (r_{max} = +.91)

Separate breakdowns for 'single raters' and 'different raters' appear in fn. 14.

III Accuracy Of I-Level Diagnoses

The 'accuracy' analyses will now be reviewed. First, we took the following to be the youth's 'true' diagnosis: the classification which was finally agreed upon on the basis of all available information. The information in question consisted chiefly of interviews. In the case of Experimental subjects it also included behavioral observations, together with various verbal interactions between staff and youth. It would be useful to say a little more about this. In actual practice, a given youth's final classification could have been--and was--arrived at via one of several routes. A 'basic route' occurred in all cases; other routes represented elaborations of the basic route. (It should be remembered that CTP was never particularly well set up to make a systematic and/or methodologically 'clean' assessment of diagnostic reliability and accuracy. At any rate, the basic, or fundamental 'route' was that in which the first research rater classified the youth on the basis of the intake tape. This particular classification occurred without exception.) The routes or conditions in question included one or more of the following--either singly or in various combinations with one another:[15]

A. Intake interview only. (1) Initial research rating of intake interview ("first research rater"); (2) second research rating of intake interview ("second research rater"); (3) third or fourth research rating of intake interview ("third research rater", etc.); (4) rating(s) of intake interview by operations personnel--i.e., by parole agent and/or treatment supervisor. (In this latter case, the operations staff member had also conducted the intake interview. In cases '1', '2', and '3' above, the researcher did the intake interview. In all four cases, at least one researcher rated--i.e., classified--the intake interview.)

B. Followup interview(s) by one or more researchers. There may have been anywhere from one to six or more such followup's. (The median figure is approximately two, in the case of youths who had at least one followup interview.) Collectively, these interviews may have extended over a period of several months, or, more often, a couple of years. The first research rater, and/or any other research rater(s), may have been involved in any one or more of these followup interviews...plus the classifications which resulted from them. Operations personnel were not involved.

C. Discharge interview.[16]

D. Observations/Other interactions. These refer to direct and sometimes rather frequent behavioral observations of--and/or non-interview-centered verbal interactions with--the youths. This applied to Experimental subjects only.

The findings presented below are the results of analyses which involved two of the most common and/or possibly most meaningful "routes" with respect to arriving at the final CTP diagnosis. These are referred to as Case A and Case B, respectively:

Case A:* Here, the first requirement for inclusion within the analysis was that at least two different researchers must have diagnosed the given youth. The ratings in question may have been made at approximately the same point in time (e.g., first research rater--at intake; second research rater--also at intake), or, they may have occurred at substantially different points in time (e.g., first research rater--at intake; a different research rater--at followup). As always, the first researcher's classification was in response to the intake tape. "Level of accuracy" was defined as the percentage of youths for whom the first research rater's classification agreed with the 'true' diagnosis--i.e., the final CTP diagnosis.[17] In 97.6% of all Case A ratings, the true diagnosis was 'verified' by the ratings of at least one researcher other than the first research rater. In itself, the first researcher's rating could not 'verify' the youth's true diagnosis--at least not in relation to the manner in which we conceived of 'verification'. (We will mention at this point that there was virtually no measurable difference in the accuracy findings in connection with our having either included or excluded the remaining 2.4%--this being 12 cases in all.) The researcher who verified the true diagnosis was not necessarily the second research rater.

Main results are given below, following the brief description of "Case B".

Case B: All individuals who fell within Case A were also included under Case B. However, three additional categories of youth were included in Case B:

(1) Individuals who--instead of having been rated by at least one researcher in addition to the first research rater--were later re-rated only by the first research rater himself, in connection with a followup and/or discharge interview. 67 males and 14 females fell within this subgrouping.

(2) Individuals whose only rating--other than the ever-present first research rater's intake rating--was one which had been done by an operations person. 122 males and 25 females fell within this subgrouping. (These youths had no followup or discharge interviews, etc.)

(3) Experimental subjects whose only actual rating was the one which had been completed by the first research rater, yet whose diagnosis had, in a meaningful sense, been 'verified' on the basis of several months or--in most instances--one or more years of direct observation by operations and/or research personnel. 41 males and 10 females fell within this subgrouping.

In sum, Case B involved a total of 765 males and 168 females. Level of accuracy was defined exactly as in Case A, above.

*Case A was one in which the first research rater's diagnosis of an individual was invariably supplemented by that of at least one other research rater. This particular requirement was not always present in relation to Case B diagnoses. Because of this and related reasons (see definition of 'Case B', in the text), a somewhat greater degree of confidence may be placed in the accuracy of those diagnoses which were arrived at via the Case A 'route' only, as vs. those which related to Case B.

Separately for Cases A and B, the main results were as follows, with respect to diagnostic accuracy:

A. As to the classification of CTP youths by subtype (9 categories), the results for males were:

Case A (N = 535): Level of accuracy = 74%
Level of accuracy minus chance = 63%

Case B (N = 765): Level of accuracy = 81%
Level of accuracy minus chance = 70%.[18]

Various supplementary analyses were carried out. Here, as in the case of interrater-reliability, the Aa and Ap categories were combined, as were the Se and Ci categories. For males, the results of these 7-category analyses were as follows:

Case A (N = 535): χ^2 = 1630.0 (d.f. = 36; $p < .001$)
ϕ' = .71
λ (asymmetrical) = .66
Level of accuracy (7 categories) = 75%
Level of accuracy minus chance = 61%
Pearson r = +.89*

Case B (N = 765): χ^2 = 2656.0 (d.f. = 36; $p < .001$)
ϕ' = .76
λ (asymmetrical) = .73
Level of accuracy (7 categories) = 81%
Level of accuracy minus chance = 67%
Pearson r = +.89*

See footnote 19 for comparable figures with regard to females. (In connection with all male as well as female 'accuracy analyses', it should be noted that the asymmetrical λ was used relative to the question

*As mentioned earlier with reference to subtype analyses, the Pearson r was computed for the purpose of general comparison only.

of predicting the 'true' diagnosis on the basis of the rater's diagnosis--not vice versa. The latter approach would have been less pertinent to the issues under consideration and would, of course, have yielded somewhat different asymmetrical λ values.)

B. As to the classification of youths by I-level (3 categories), the results for males were:

Case A (N = 535): X^2 = 685.7 (d.f. = 4; p < .001)
ϕ' = .80
λ (asymmetrical) = .73
Level of accuracy = 89%
Level of accuracy minus chance = 55%
Pearson r = +.89 (r_{max} = +.97)

Case B (N = 765): X^2 = 1041.9 (d.f. = 4; p < .001)
ϕ' = .83
λ (asymmetrical) = .78
Level of accuracy = 92%
Level of accuracy minus chance = 59%
Pearson r = +.80[20] (r_{max} = +.96)

It might be of interest to compare the CTP accuracy results with those from the Center for Training in Differential Treatment (Rita Warren, George Howard, et al) located here in Sacramento. Using raw data supplied by CTDT, an analysis was made of the ratings of trainees who had completed CTDT's nine-weeks course which focused upon differential diagnosis, together with general principles of differential treatment. The analysis related to all 39 individuals who had completed both their 'in-training diagnostic ratings' and their 'one year followup diagnostic ratings' as of several months ago.[21] Most of the 'in-training ratings' took place during the fourth-through seventh-weeks of the nine-weeks course.

As to the in-training ratings, a total of 9 diagnostic categories were present in relation to the subtype-classifications--10, if one considered the I_5's as well.[22] However, very few ratings related to Aa's, Se's, and I_5's.[23] Not all trainees made the exact same number of ratings; nevertheless, they did make approximately the same number of ratings (...8 apiece, on the average). There were 313 ratings in all--the vast majority involving male subjects. Results for the in-training ratings were:

51% of the ratings were accurate at the subtype level.[24] Giving equal weight to each rater (and in this sense holding constant the number of ratings per rater), it was found that the average or typical rater correctly classified the youths 53% of the time as to subtype. Comparable figures for the I-level classifications were: ratings = 74%;[25] average rater = 79%. (Some raters were particularly accurate in their I-level judgments, thereby raising the overall level of rater accuracy.)

As to the one year followup ratings:[26] All 39 trainees rated a set of four standard tapes.[27] Included in this set were a Cfm, an Mp, an Na, and a Ci tape--each of which involved a male subject. Relative to the one year followup subtype classifications, 46% of the 155 ratings were accurate.[28] The average rater was also accurate 46% of the time.[29] Comparable figures for the I-level classifications were: ratings = 66%;[30] average rater = 68%. These figures were somewhat lower than those for the in-training ratings. We doubt that this particular drop would have more than possibly a modest amount to do with, say, a factor such as 'gradual reduction in the rater's overall level of diagnostic skill'. (CTDT provided the trainees with a certain amount of feedback and consultation during the year subsequent to their participation in the nine-weeks course.)

The breakdown by subtype may shed some light on this moderate drop in accuracy from 'in training' to 'one year followup' (see fn. 29).

All in all, the level of accuracy which was associated with the CTDT trainees was noticeably lower than that obtained by experienced CTP research staff.

IV Concluding Remarks

Reliability figures obtained for the period 1961-1969 were approximately the same as those reported for the period 1961-1965. For males, the 'updated' interrater-agreement at point of intake was 62% for subtype and 81% for I-level. Comparable figures for females were 70% and 85%, respectively. Interrater-disagreements usually involved immediately adjacent or nearly adjacent subtype categories. This was in contrast to subtype categories which were widely separated or, for that matter, randomly distributed. Taken together with the Cramér ϕ', the Goodman and Kruskal lambda, and so on, these results would appear to be more than satisfactory by most standards--at

least with reference to the number of differentiations in question (9 for subtype, 3 for I-level). (Even so, see pg. 16, paragraph 3, regarding one particular factor whose influence would reduce the strength of these findings to a moderate degree.) This would apply to the diagnostic accuracy results, as well.

In terms of CTP's own standards, however, much improvement is still in order. These standards relate very much to CTP's need for rather highly individualized treatment planning, beginning at point of intake. Thus, while recognizing the rather substantial conceptual and operational achievements which may be reflected in the findings reported above, we are not at all satisfied with having 'only' 62% - 70% interrater-agreement at the subtype level--even granting that such figures include a 'somewhat-difficult-to-rate' (yet rather sizable) subsample, in addition to several called-for differentiations. The 74% - 81% subtype-accuracy figures for males are a little more encouraging.* While recognizing the difficulties involved, we feel a need to strive for levels of interrater-agreement which would be in the neighborhood of 85% - 90%. With this in mind, it would seem as if CTP's only apparent, current source of optimism might relate to the fact that such levels were achieved at least with reference to subtype-accuracy, in those cases which were rated and then discussed by at least two different raters (viz., two researchers) prior to their having arrived at what we would call the 'operational diagnosis'. (In the case of Experimentals, it was the operational diagnosis which the individualized treatment plans most closely reflected.)

Apart from CTP's particular standards and/or operational needs, it will be noted that the obtained percentages-of-agreement, the lambda's, the Pearson r's, etc., do indicate the presence of a sizable amount of predictive ability with reference to the subtype as well as I-level classifications. In other words, the Phase I and II results do not reflect the presence of a level or type of statistical significance which, in itself, is little other than an expression of low or moderately positive correlations within the context of large sample sizes.

The following should be kept in mind. We estimate that, at the subtype level, most CTP figures for interrater-agreement are perhaps 15% (not 15 percentage points) higher than they would have been in the event that the 2nd research rater had had absolutely no information regarding the 1st research rater's general--and, at times, rather specific-assessment of the youth. (This issue is less germane to the question of diagnostic accuracy.) This same factor would probably have resulted in a 5% - 10% difference in the case of I-level agreement. To quote from a 1970 CTP report: "Among research staff, second raters often received information as to the one, two, or perhaps three possible subtype-diagnoses with which a first rater may have been wrestling....Possession of this information eliminated the second rater's ability to reach a technically independent or literally uncompounded judgment.

*Level of accuracy was as follows for 'Case A' (the corresponding figures for 'Case B' are shown within parentheses) - Males: subtype - 74% (81%); I-level - 89% (92%). Females: subtype - 80% (86%); I-level - 92% (94%).

However, it did not, ipso facto, eliminate the latter's ability to reach a relatively sound judgment--one which was based upon his personal review and integration of the taped interview [plus any other available information]. In this sense, it represented no more and no less than a semi-independent judgment".[31]

> Related to this: The Phase I and II diagnostic accuracy figures were higher than those which involved interrater-reliability. Close inspection of this situation suggests that the first research rater's classification of the youth probably had a stronger influence upon (a) the diagnosis which was ultimately arrived at (viz., the true diagnosis) than upon (b) the diagnosis which was made by the 2nd rater.[32] This might help account for the fact that the accuracy results were moderately yet consistently higher than the interrater-reliability results--a situation which is not often found in connection with studies of psychiatrically/psychologically oriented systems of personality classification.

For males, rater agreement through time (i.e., intake vs. followup--estimated to be 10 months on the average) was 75% for subtype and 91% for I-level. Comparable figures for females were 78% and 89%, respectively. Broadly speaking, this level of agreement suggests the presence of at least moderate--or, quite possibly, sizable--amounts of stability with respect to personality dimensions upon which the raters' attention would ordinarily be focused.

Stability and interactional context aside, the I-level system would doubtlessly profit from continued conceptual and operational sharpening-up with regard to the Na vs. Nx distinction, in particular. (Some progress has been reported along this line, at least at the conceptual level.[33]) This distinction has consistently remained the principal contributor to rater-disagreement--at point of intake, and through time as well. Beyond this, it would be of benefit--particularly to correctional workers outside of CTP--if CTP were to pin down and spell out, at least more comprehensively than has been done to date, the features which operationally distinguish most Mp's from most Na's. (It may be noted that the Mp and Na subtypes represent 'adjacent categories' with respect to the I-level classification schema. They also share with one another a number of readily apparent, as well as underlying, attributes. Seen in this light, it is interesting to note that each such subtype had a noticeably lower-than-average level of interrater-agreement.)

In sum, it is accurate and probably fair to say that CTP's Phase I and Phase II reliability and accuracy results would compare favorably or quite favorably with those obtained in connection with other clinically oriented--and, especially, interview-based--personality typologies. However, very much improvement is needed within the conceptual and operational areas alike. On the latter score, e.g., increased consideration should definitely be given to the idea of almost routinely calling for second ratings, at point of intake.[34]

Footnotes

1. The 1961-1965 data were re-analyzed in 1969 in connection with CTP's reassessment of the reliability index which it had previously utilized. However, no new data was involved.

2. The period 1961-1969 includes all ward-intake during Phases I and II. However, numerous followup interviews took place after 1969 with regard to Phase II youths.

3. Separate analyses were carried out for each of the following time-periods: 1961-1963; 1964-1966; 1967-1969;...also included were 1964-1969 and 1961-1969. (The present analysis relates to the entire Phase I and II period--viz., 1961-1969.) Similarly, for each time-period, separate analyses were made with regard to each of the following areas: Sacramento; Stockton-Modesto; and, San Francisco. (The present analysis relates to all three locations, combined.) Cutting across each such analysis, the data was also looked at separately for: Experimentals; Controls; Ineligibles; and, the San Francisco Guided Group Interaction subjects. Collectively, the latter three subject-groupings are referred to as <u>non-Experimentals</u>. (The present analysis combines all four of these subject-groupings.) These analyses were carried out in order to determine whether any substantial trends or differences were involved in connection with time-period, location, and/or subject-grouping. By and large, reliability and accuracy (as defined in the text) remained unchanged through time, across locations, and with reference to the differing subject-groupings.

4. The researcher who first rated the youth's intake tape is referred to as the "first research rater". The researcher who next rated the youth's intake tape (generally upon request of the first research rater) is referred to as the "second research rater". The latter researcher was never the individual who had conducted the intake interview. During Phases I and II, the first research rater conducted the intake interview in some 87% of the cases (males). The remaining 13% were conducted by operations personnel (mainly during the years 1966-1969).

5. In the remaining 55% of the cases, the first research rater did not consider it necessary to request a second research rating of the intake interview. Most, though not all such tapes were considered relatively "easy" from a diagnostic standpoint, whether rated by an <u>operations</u> person or not--and particularly if they had been rated by an operations person with whom the first research rater agreed. (As indicated in fn. 4, 13% of the 802 males had been interviewed by an operations staff member. This individual--and/or his treatment supervisor--then rated the tape. The operations rating was separate and apart from--and,

temporally speaking, it almost always preceded--that which was invariably done by the first research rater. If the first researcher's classification concurred with that of the operations staff member, the former would usually feel less reason than would otherwise be the case to request a second researcher's rating of the intake tape.) In most such cases, the diagnosis appeared to be relatively clear-cut--at least to the first research rater (and, in many cases, to the operations rater). Yet, the present data suggest that the first research raters were not sufficiently 'conservative' in this regard: That is to say, it would have been better if they had asked for a second researcher's rating more often than they did. For example, the percentage of agreement between the first research rater's classification and the classification which was ultimately arrived at (based upon all contacts and/or interviews with some 427 males) was 79% in the case of subtype classifications and 92% in the case of I-level classifications. These figures refer to Experimental subjects only. These were individuals whom it was possible to observe far more closely than Controls (and GGI subjects as well), and whose original classification had had the greatest opportunity of being modified as the result of post-intake observations and/or interviews. (All instances of what may be described as 'substantial growth' within the youths themselves--e.g., movement from one I-level to the next higher I-level--were excluded.) Comparable figures for Experimental females were 85% in the case of subtype classifications and 91% with reference to I-level classifications (N = 94 females).

6. Theoretically, he may receive a classification of I_5 as well. However, I_5's comprise a negligible quantity within the present sample of youths--less than 1%. As a result, they are not differentiated from I_4's of comparable subtype relative to the present analysis, unless otherwise specified.

7. Thus, the probability of error was reduced quite a bit--viz., 55%--as a result of knowing the row and column categories of the 7 x 7 joint-probability distribution in question.

8. Without knowing either the theoretical or obtained subtype distributions, an individual who possessed no information about the I-level system, or about the CTP ward-sample, etc., would still have had a 14.3% chance of simply guessing the true diagnosis if faced with 7 categories from which to choose, and if given a single choice.

9. Chance equals .33, in the case of 3 categories--and, a single choice.

10. Use of the Pearson r would appear to be justified in this instance, at least as far as the underlying level-of-maturity continuum is concerned. ($I_4 > I_3 > I_2$). - The results of this 3-category analysis were rather similar to those which were based upon the following 4-category breakdown (N = 364 males): I_2, I_3, I_4 Neurotics, I_4 Non-neurotics. Figures for the latter analysis were:

χ^2 = 494.8 (d.f. = 9; $p < .001$)

ϕ' = .67

λ (symmetrical) = .57

Rater agreement = 76%

Rater agreement minus chance = 51%

Pearson r = +.69 (r_{max} = +.98).

(Recent CTP data support the view that Non-neurotic I_4's are, on the whole, somewhat more mature than Neurotic I_4's.)

11. The one I_5 was excluded from this analysis.

12. Jesness, C. The Preston Typology Study Final Report. California Youth Authority and the American Justice Institute. July, 1969. pg. A-125.

13. For male subjects, the <u>subtype</u> breakdown (7 categories) on rater-reliability through time is as follows with respect to 'single research raters' only (N = 256); the figures for 'different research raters' only (N = 170) are shown in parentheses:

χ^2 = 759.3 (500.4) (d.f. = 36 in both cases; $p < .001$)

ϕ' = .70 (.70)

λ (symmetrical) = .71 (.65)

Rater agreement = 78% (75%)

Rater agreement minus chance = 63% (61%)

Pearson r = +.77 (+.87)

For males, the <u>I-level</u> breakdown (3 categories) on rater-reliability through time is as follows with respect to 'single raters' only (N = 256); the figures for 'different raters' only (N = 170) are shown in parentheses:

χ^2 = 354.9 (209.8) (d.f. = 4 in both cases; $p < .001$)

ϕ' = .83 (.79)

λ (symmetrical) = .82 (.77)

Rater agreement = 92% (91%)

Rater agreement minus chance = 58% (57%)

Pearson r = +.87 (+.83) (r_{max} = +.96 (+.91))

14. For female subjects, the <u>subtype</u> breakdown (7 categories) on rater-reliability through time is as follows with respect to 'single research raters' only (N = 45); the figures for 'different research raters' only (N = 45) are shown in parentheses:

χ^2 = 112.6 (98.6) (d.f. = 36 in both cases; $p < .001$)

ϕ' = .65 (.60)

λ (symmetrical) = .71 (.54)

Rater agreement = 80% (73%)

Rater agreement minus chance = 66% (59%)

Pearson r = +.83 (+.73)

For females, the 4-category (I_2, I_3, I_4N, I_4Non-N) breakdown on rater-reliability through time is as follows for 'single raters' (N = 45); figures for 'different raters' (N = 45) are shown in parentheses:

χ^2 = 53.1 (48.4) (d.f. = 9 in both cases; $p < .001$)

ϕ' = .63 (.60)

λ (symmetrical) = .70 (.64)

Rater agreement = 87% (87%)

Rater agreement minus chance = 62% (62%)

Pearson r = +.84 (+.71) (r_{max} = +.90 (+.94))

For females, the <u>I-level</u> breakdown (3 categories) on rater-reliability through time is as follows with respect to 'single raters' (N = 45); the figures for 'different raters' (N = 45) are shown in parentheses:

χ^2 = 25.1 (24.7) (d.f. = 4 in both cases; $p < .001$)

ϕ' = .53 (.52)

λ (symmetrical) = .79 (.56)

Rater agreement = 93% (87%)

Rater agreement minus chance = 60% (53%)

Pearson r = +.86 (+.73) (r_{max} = +.91 (+.92))

15. Once again, all youths had an intake interview. Moreover, this interview was always rated by at least one researcher--viz., the "first research rater".

16. In this case, a researcher would interview the youth and make a diagnostic classification, on the occasion of the latter's favorable discharge from the California Youth Authority.

17. As will be indicated in the text with reference to Case A, subtype accuracy was found to be 74% while I-level accuracy turned out to be 89%. However, subtype accuracy rose to 83% and I-level accuracy became 93% once we dropped the requirement that there be a second research rating. [The issue of this particular requirement is separable from that which relates to the presence or absence of 'verification' of the true diagnosis per se. Thus, in the present instance (...83% and 93% accuracy)--i.e., in contrast to Case A--considerably fewer than 97.6% of all true diagnoses turned out to have been verified..that is, verified by the second research rater or else by some other researcher who classified the youth at a later point in time. Although--still with reference to the present instance--the second rating had indeed been dropped relative to its being an essential requirement, many of the youths in question nevertheless did receive a second, third, or subsequent research rating. At least one of these ratings may have verified--and in most cases actually did verify--the true diagnosis. As a result, a fairly high percentage of the ratings in question not only did receive a second or later rating, but did in fact turn out to have been verified in the sense described within the text.] (These figures all apply to males. Without the 'second research rating' requirement, the comparable figures for females were 91% and 96%, respectively.) We feel that the latter set of figures are not the best one's to use in light of the absence of any substantial external and/or post-intake check on the first rater's diagnosis, particularly in the case of non-Experimentals subjects. Only partially aside from the latter point, it is useful to note that once the 'intake classification-sequence'

had been completed (whether by 1, 2, 3 or 4 research raters...with or without an operations rater in addition) the subtype diagnosis which was settled upon at that point was accurate in 96% of the cases (Experimentals = 95%; non-Experimentals = 97%) as judged in terms of the final, i.e., 'true' diagnosis. By the same token, I-level accuracy was 99% (Experimentals = 99%; non-Experimentals = 99%). This excludes all instances of 'substantial growth' within the youths themselves. For females the comparable subtype figure was 96% (Experimentals = 92%; non-Experimentals = 98%) whereas the I-level figure was 98% (Experimentals = 95%; non-Experimentals = 100%).

18. Level of accuracy was as follows for the separate subtypes. Case A: Aa - 67%; Ap - 83%; Cfm - 82%; Cfc - 79%; Mp - 59%; Na - 69%; Nx - 82%; Se - 78%; Ci - 63%. (The subtype sample-sizes were: 6, 24, 72, 47, 76, 124, 136, 18, and 32, respectively.) Case B: Aa - 67%; Ap - 84%; Cfm - 85%; Cfc - 85%; Mp - 62%; Na - 77%; Nx - 89%; Se - 81%; Ci - 67%. (The subtype sample-sizes were: 6, 25, 101, 68, 85, 174, 245, 21, and 39, respectively.)

Level of accuracy was as follows for the separate I-levels. Case A: I_2 - 87%; I_3 - 88%; I_4 - 90%. (The sample-sizes were: 30, 195, and 310, respectively. Only one I_5 was included within the present analysis. He had been classified accurately by the first research rater.) Case B: I_2 - 87%; I_3 - 90%; I_4 - 93%. (The sample-sizes were: 31, 253, and 480, respectively. Only two I_5's were included within the present analysis. Both had been classified accurately by the first research rater.)

As to subtype classification, the comparable figures for females were:

Case A (N = 119): Level of accuracy = 80%
Level of accuracy minus chance = 69%.

Case B (N = 168): Level of accuracy = 86%
Level of accuracy minus chance = 75%.

Level of accuracy was as follows for the separate subtypes. Case A: Aa - (no cases); Ap - 100%; Cfm - 90%; Cfc - 50%; Mp - 76%; Na - 73%; Nx - 89%; Se - 75%; Ci - 100%. (The subtype sample-sizes were: 0, 2, 10, 4, 17, 45, 36, 4, and 2, respectively.) Case B: Aa - (no cases); Ap - 100%; Cfm - 92%; Cfc - 50%; Mp - 82%; Na - 77%; Nx - 95%; Se - 80%; Ci - 100%. (The subtype sample-sizes were: 0, 3, 13, 4, 22, 56, 64, 5, and 2, respectively.)

Level of accuracy was as follows for the separate I-levels. Case A: I_2 = 100%; I_3 = 86%; I_4 = 92%. (The sample-sizes were: 2, 31, and 87, respectively. No I_5's were included within the present analysis.) Case B: I_2 = 100%; I_3 = 90%; I_4 = 95%. (The sample-sizes were: 3, 39, and 127, respectively. No I_5's were included.)

19. In the case of females the comparable figures were:

Case A (N = 119): X^2 = 429.2 (d.f. = 36; $p < .001$)
ϕ' = .78
λ (asymmetrical) = .69
Level of accuracy (7 categories) = 81%
Level of accuracy minus chance = 66%
Pearson r = +.87.

Case B (N = 168): X^2 = 657.9 (d.f. = 36; $p < .001$)
ϕ' = .81
λ (asymmetrical) = .78
Level of accuracy (7 categories) = 86%
Level of accuracy minus chance = 72%
Pearson r = +.90.

20. As to I-level classification, the comparable figures for females were:

Case A (N = 119): X^2 = 192.5 (d.f. = 4; $p < .001$)
ϕ' = .90
λ (asymmetrical) = .70
Level of accuracy (3 categories) = 92%
Level of accuracy minus chance = 58%
Pearson r = +.78 (r_{max} = +.96).

Case B (N = 168): X^2 = 285.5 (d.f. = 4; $p < .001$)
ϕ' = .92
λ (asymmetrical) = .76

Level of accuracy (3 categories) = 94%
Level of accuracy minus chance = 61%
Pearson r = +.90 (r_{max} = +.98).

21. No selection was involved here, relative to the present analysis. Although more than 39 individuals received extensive training from CTDT during recent years, many did not receive their followup tapes because a full year's period had not yet elapsed subsequent to the completion of their training.

22. The I_5 ratings were analyzed as simply an additional subtype.

23. The figures shown for Aa's, Se's, and I_5's should be considered highly tentative since no more than 1 Aa, 4 Se's, and 5 I_5's had been rated.

24. Broken down by subtype, the in-training levels of rating-accuracy were: Aa = 0%; Ap = 54%; Cfm = 56%; Cfc = 45%; Mp = 56%; Na = 48%; Nx = 50%; Se = 50%; Ci = 42%; I_5 = 60%.

25. Broken down by I-level, the in-training levels of rating-accuracy were: I_2 = 64%; I_3 = 67%; I_4 = 87%; I_5 = 40%.

26. Once again, 9 diagnostic categories were present in relation to the subtype-classifications--10, if one considered the I_5's as well.

27. A fifth, I_2 tape, was later found to be largely inappropriate. Thus, no one year followup figure is presented relative to the I_2 level as such.

28. One rating had to be eliminated due to the rater's chance personal knowledge of the youth in question.

29. Broken down by subtype, the one year followup levels of rating-accuracy were: Cfm = 72%; Mp = 74%; Na = 24%; Ci = 13%. While it is unknown whether the Cfm and Mp tapes were of what one might call "average difficulty" or, perhaps, were even on the "slightly easy side", it does seem highly likely that the Na and Ci test-tapes were on the rather difficult side. Unfortunately, it was difficult to expect the trainees, as a group, to uniformly rate more than a limited number of followup tapes (whether within or across subtypes).

30. Broken down by I-level, the one year followup levels of rating-accuracy were: I_3 = 81%; I_4 = 54%.

Footnotes, Concluded

31. Palmer, T. Reply to Eight Questions Commonly Addressed to California's Community Treatment Project. California Youth Authority. CTP Report Series: 1970, No. 2 pg. 19.

32. Furthermore, the 1st rater's influence upon the true diagnosis was almost certainly stronger in those cases in which there was an absence of any 2nd, 3rd, etc., research rating--i.e., stronger than when any of these latter ratings were present. This would help account for the fact that the 'Case B' figures were higher (moderately yet consistently higher) than those for 'Case A'.

    Additional analysis showed that the 1st research rater's degree of influence upon the true diagnosis was identical to that of the 2nd research rater's in the case of Experimentals. In the case of non-Experimentals, it was slightly but almost negligibly greater--3 percentage points in the case of subtype as well as I-level, for males and females alike.

33. Palmer, T. California's Community Treatment Project - Research Report #11. California Youth Authority. July, 1971. pp. 13-14.

34. This is of particular relevance to the need for high levels of diagnostic accuracy, as one of the first steps in the direction of individualized treatment planning.

END