

THE NEW YORK PSYCHOANALYTIC SOCIETY AND INSTITUTE

September 14, 2009

Report on the project: *Using Process Notes to Explore Treatment Effectiveness and Changes in Theory and Technique over Five Decades in the Treatment Center of the New York Psychoanalytic Institute*

Principal Investigators: Wilma Bucci, Adelphi University; Leon Hoffman, New York Psychoanalytic Institute

Our major goals were to explore the question of whether therapy notes and case reports could be studied in a systematic manner, combining qualitative and quantitative techniques, so as to obtain information concerning the quality of the treatment; in particular, to differentiate successful from unsuccessful treatments. The second question we wished to address was whether we could use these notes so as to track changes in theory and technique over the five decades at the NYPSI.

The study was funded for the period July 1, 2007 to June 30, 2009. This report outlines our methods and results to date.

Methods

Sample Selection

The initial selection of 10 cases was done by the Treatment Center Director and Staff, randomly selecting a case from each decade (50's through 90's) — with a notation for each case as to whether it had been judged as successful or not by the treating analyst. The TC staff then selected a second case in each decade with the opposite contemporaneously judged outcome. Four additional cases were added to the sample based on current clinical verification procedures as discussed below.

In order to ensure full confidentiality, cases involving patients who were potentially identifiable through professional contacts (even after the notes concerning them had been de-identified) were removed from the above selection process.

Each set of the selected notes was then read by members of the TC staff, and all possible identifying information blacked out. After this de-identifying process, further review of the data was done by senior members of the research team in order to insure complete confidentiality before copies of the notes were given to graduate assistants for processing.

Initial Text Preparation

The printed notes were first scanned into the computer, and then corrected. Where this proved impractical, the notes were retyped into the computer and then proofread. The texts were then prepared for our computer processor, DAAP, which requires some text

preparation procedures, including disambiguation of certain common words, such as "like" and "well"; putting some words, such as "o'clock" and "OK" into special format; and ensuring that de-identified names, such as "Dr. S", or "Ms. A", were appropriately changed so as to not interfere with our dictionaries (among others, "a", "I", "s" and "t" are DAAP dictionary words). This initial text preparation for the first ten cases was accomplished during the first year of this project. The texts for the additional four treatments, all of which could be scanned, were prepared during the second year.

Demographic data

As shown in Table 1, the final set of notes was from 14 cases: 2 from the 1950s; 2 from the 1960s; 3 from the 1970s; 3 from the 1980s; 4 from the 1990s. Nine analysts were male; five were female. Four patients were male; ten were female. One patient was in the late teens; ten were in their 20's; two were in their 30's; and one was in the 40's. Ten patients were single at the start of treatment; four were married, and two of the four had children. Three patients were professionals; four were students; three were homemakers; three were in the arts; and one was in a trade or craft.

Description of Cases and Notes

As shown in Table 2, the length of treatment ranged from 22 to 69 months with a mean of 41.8 months and a standard deviation of 15 months. (This is for 13 treatments; it was not possible to determine the length of treatment for 90T4). The notes were somewhat heterogeneous in their organization; a brief description of each set of notes is given in Table 2. Several of these contained two or more summaries for the same time period; for example, 70T1 had both a full set of very short (about one paragraph) weekly summaries and a full set of moderately sized (2 -3 pages each) 6 month summaries. In general, the summaries for the longer periods of time covered much the same material as those for the shorter periods of time, with somewhat less detail.

For each treatment, we chose a complete set of notes, so that each time period was covered exactly once. Broader time periods were selected whenever possible; no time period was greater than 6 months. After limiting each set of notes to this coverage, the note word counts for the 14 cases ranged widely, from 7,800 to 290,940, with a mean of 43,036 and a standard deviation of 74,584.

We tested the four cases where there was a choice as to summaries for shorter vs. longer periods of time, as in 80T2, where, for the first 30 months of treatment, there were both weekly summaries and 6-month summaries. All our language measures showed only minor and non-significant differences between the two sets of notes.

Segmentation

Most of the notes were already segmented in the form of 3 or 6 month summaries. The others, in the form of daily or weekly notes, were grouped into segments, where the

segmentation points were chosen to match the seasons of the year where possible, and/or to follow the natural breaks in the continuity of the treatment.

The measures and results for the 14 cases will be outlined in three sections

I. Clinical Measures of Success and Linguistic Variables

II. Themes Across the Decades

III. Clinical Analysis of a Successful and an Unsuccessful Case.

I. Clinical Measures of Success and Linguistic Variables

METHODS

Clinical ratings of success

We utilized two measures to verify the original clinical judgments: Global Assessment of Functioning (GAF) and the Psychodynamic Functioning Scales (PFS) (Developed by Per Høglend and colleagues). The GAF is a 100 point single-item scale (with 10 point interval anchors), which has been used in many studies; it was originally derived by Luborsky and colleagues as the Health and Sickness rating scale. The PFS consists of Six Scales; constructed similarly to the GAF with ratings ranging from 1–100 and anchors at each 10-point interval. The six PFS scales are Quality of Family Relations, Quality of Friendships, Romantic/Sexual Relationships, Tolerance for Affects, Insight, and Problem Solving and Adaptive Capacity.

Three analyst judges, Leon Hoffman (LH), Jane Albus (JA), and William Braun (WB) were trained in use of these measures on clinical material other than that used for this study. Satisfactory reliabilities were obtained by the 3 raters on all items except the Quality of Friendships item of the PFS. This scale was then excluded from the analysis of the data. Reliability among the three judges, using Intra-Class Correlation (ICC), after training was .874 for GAF, and ranged from .597 to .843 for the five PFS scales that were used.

After reliability had been achieved on the training materials, 20 clinical segments, consisting of 10 initial and 10 ending segments were extracted by LH from the initial 10 cases. The segments were presented in randomized order and scored independently for the GAF and the PFS scales by JA and WB. Following computation of reliability, discrepancies were discussed and resolved with LH as moderator.

Disagreements between original and current clinical ratings were found for four of the first ten cases. The decision was made to retain these four cases in the study, to increase sample size, since the work of text preparation had already been done, and to add four

more cases to provide at least one success and one nonsuccess case, evaluated by current judgment, for each decade.

We added one case from the 70's, one from the 80's, and two from the 90's. These were chosen based on the original evaluations; then examined for verification purposes using the same procedure of rating beginning and end segments with the GAF and PFS scales.

The PFS measures on the five scales of the PFS utilized (excluding quality of friendships) were averaged into one composite score (this is the procedure which Høglend and colleagues utilize). The ratings of the two current analyst-judges on GAF and the composite PFS score were converted into four summary measures

- (1) GAF at the End of treatment
- (2) The GAF Difference between the beginning and end of treatment
- (3) The PFS at the End of treatment
- (4) The PFS Difference between the beginning and end of treatment.

These four measures were then combined to yield a Composite Clinical Effectiveness measure (CCE), using the following groupings:

For GAF End and PFS End (Mean of the 5 scales)

- Less than 61 = 0
- Between 61 and 70 = 1
- 71 and above = 2

(With a GAF of 61, the patient is “generally functioning pretty well and has some meaningful relationships; and with a GAF of 71, there is “no more than slight impairment in social, occupational, or school functioning”).

For GAF and PFS Change Scores: (End – Beginning)

- 0 or negative = 0
- 1 to 10 = 1
- 10 or greater = 2

For each case, the 4 numbers were then added to give a single CCE score, ranging from 0-8, as shown in Table 3.

Language Style Measures

Referential Activity Intensity Measure (MHW): In general terms, Referential Activity (RA) is a linguistic variable that represents the degree to which language is connected to nonverbal experience, including emotional experience. For purposes of this study RA was assessed using a computerized dictionary, the Weighted Referential Activity Dictionary (WRAD) and a derived measure, the Mean High WRAD (MHW). Most items in the WRAD dictionary are function words rather than content words. MHW is the

average amount by which the WRAD curve is above its neutral value of .5, and is the most direct indicator of emotional immersion in verbal expression.

Reflection (REF): The Reflection Dictionary (REF) contains words that concern how people think and communicate thoughts (words such as: *if, think, feel, avoid*). The REF measure is the proportion of words that match the Reflection Dictionary. High REF language indicates dominance of logical thought and reasoning. In contrast to the WRAD dictionary, which is a language style dictionary, consisting of mainly function words, constructed by a modeling procedure, and which is considered complete, the REF is a content dictionary and is open rather than fixed. Words are added for coverage of new data sets, using the same procedures for development of theme dictionaries, as outlined below.

Mean High RA (MHW) – Reflection (REF): The MHW - REF difference score is an indicator of the degree to which the speaker or writer is immersed emotionally in an experience vs. distancing from it.

Computer Analysis

The basic computer program used for this analysis is the Discourse Attributes Analysis Program (DAAP). Previous versions had the capacity to call for only one text at a time, and had the Referential Process Dictionaries as part of the program. In order to carry out the text analysis for this study, we constructed a new version of DAAP, which permits the analysis of multiple texts using different dictionaries, where copies of the texts are placed in one folder and copies of the dictionaries are placed in another. This new version of DAAP compares the words in all the texts with all the dictionaries, and provides linguistic analysis data for each time period for each text, as well as composite data for all the texts.

Using this new version of DAAP, we computed MHW, REF, and MHW - REF for each time period, as described above, within each treatment; and for each treatment as a whole.

RESULTS

Evaluation of success (See Table 3)

Three cases were deemed successful based on both the original classification and a score of 7 or 8 on the current composite rating (CCE); these were 50T2, 60T1 and 70T3.

Five cases showed mixed results, with CCE ratings of 2 or 3; these were 60T2, 80T3, 90T1, 90T2, 90T3.

Six cases were rated as unsuccessful based on a score of 0 or 1 on the CCE; these were 50T1, 70T1, 70T2, 80T1, 80T2, 90T4.

Comparison of Original and Current Clinical Measures of Success

As shown in Table 3, seven of the 14 cases were originally judged as successful, and seven were judged as unsuccessful. For the seven cases originally labeled as successful, the mean of the current CCE scores was 4.14, compared to a CCE of 1.29 for the 7 originally labeled unsuccessful.

Comparison of Current Clinical Measure and Computerized Linguistic Measure

The major result for this section of the research is the correlation between the CCE and the language measures:

- MHW (Mean High WRAD) vs. CCE = .744
- MWRAD (Mean WRAD) vs. CCE= .710
- MHW – REF vs. CCE= .732

These correlations provide evidence that the degree of emotional involvement with which analysts describe their ongoing analytic work with their patients, as reflected in their language style, is closely correlated with the patient's improvement over the course of treatment as indicated by independent systematic clinical ratings of the analyst's report.

We, therefore, suggest that the style in which therapists/analysts write about their patients relates to the way in which they listen to their patients and integrate their patients' communications to them, and this in turn relates to how well the treatment works.

II: Themes Across the Decades

Development of theme dictionaries

In a preliminary study, the content words occurring in Treatment 50T1 were examined by a group of students and separated into themes. The actual list of themes, 25 in all, was chosen by consensus among them. A 26th theme, defense, was added later.

For this study, a computer program separated all 14 sets of treatment notes into distinct words (types). Non-content words, such as prepositions and pronouns, as well as proper nouns, were removed from the list. These types were then grouped; words with the same stem and having the same basic meaning, such as "chill", "chilled", "chills" and "chilly", were placed in the same group. Each group of words was scored by three graduate assistants for inclusion in the 26 theme dictionaries, where the judges were instructed that, when applicable, they could place a group of words in more than one theme dictionary. Words scored for inclusion in a theme by two or more judges were added to the dictionary for that theme. At the same time, these words were scored for inclusion in our basic Referential Process dictionaries, including REF, described above. Here, words

scored as belonging to a dictionary by all three judges were included, and words so scored by exactly two of these judges were then scored by an additional judge.

Using the 14 sets of notes, a cluster analysis of the 26 themes was performed; four clusters of themes emerged; these were labeled as Sex, Ego Functions, Relational and Symptoms. The themes in each cluster, together with sample words from each theme, and the coverage of the cluster of themes for the entire set of treatment notes are listed in Appendix A. The eight themes (5 of which are indicative of negative emotions) that did not occur in these clusters, along with sample words, are also listed in Appendix A.

Results

The variation in use of words from these four clusters over the decades is shown in Appendix B. As expected, Cluster 1 (Sex) had strikingly greater usage in the 1960's than in the other decades; Cluster 2 (Ego functions) showed greater usage after 1970 than previously; Cluster 3 (Relational) showed less usage in the 1950's than in the later decades; Cluster 4 (Symptoms) showed little change over the decades. At this point, we have found no significant relationship between usage of words from these clusters and success of treatment; this question will be examined in future research.

III: Clinical Analysis of a Successful and an Unsuccessful Case

Method

For each of the 14 treatments, a chart was constructed showing the variations in MHW and REF over the course of the treatment; the chart was constructed using the natural segmentation into time periods described above. It had been noted in previous studies that crossing points of these two curves often point to places of clinical interest. One clearly successful case, 70T3, and one clearly unsuccessful case, 50T1, have been chosen for further analysis. The charts showing MHW and REF for these two sets of notes are reproduced in Appendix C.

Results

For the unsuccessful case, 50T1, there is a clear change in the course of treatment, starting with Spring, 1955, when these two measures are roughly equal; going on to summer, 1955, when they start to diverge, with REF dominating MHW; and going on to the first part of fall, 1955 where REF is very much higher than MHW. (There are two segments for fall, 1955, as there is a six week unexplained break in the treatment in the middle of this period). The notes from these periods were read clinically to look for understanding as to why the treatment went into decline and did not recover.

For the successful case, 70T3, there are two time periods where REF is higher than MHW; these are October, 1964 to March, 1975, and April to September, 1978, with the treatment going on to a successful conclusion shortly after the latter period. The notes from the latter period, as well as the subsequent two periods on to the conclusion, were

read by the same three analysts to look for understanding as to what happened in the critical period ending in September, 1978, so as to turn the treatment into its successful conclusion.

Clinical Analysis

An unsuccessful case

50T1 was a professional man in his mid to late 30's who was considered to be a good analyzable patient at intake. The analyst notes in his first session "The patient is an aggressive, outspoken, neat emphatic speaker." The patient complained of a potency problem and concern about the analyst's inexperience and whether psychoanalysis is any good; early in the analysis there were many references to homosexual fantasies, which, as far as could be seen from the notes, were not addressed analytically.

The three analysts (LH, WB, and JA) independently studied the three segments of the clinical material; that is, Spring, 1955, when MHW and MREF are roughly equal; Summer, 1955, when MREF is somewhat higher than MHW; and the first part of Fall, 1955, when MHW is considerably lower than MREF. The question to be addressed was: Why was there such a dramatic decrease in the MHW-REF after the summer break?

In brief, throughout, transference interventions were never in the here and now and homosexual feelings, whether or not related to vacations, were not addressed. It seems as if the analyst did not appreciate that the patient's attacking him could have served as a defense against missing, needy, even castrated feelings. All three analysts independently asked: Didn't the analyst see the unexpressed feelings? The analyst clearly missed the patient's close bond toward him. When they returned after vacation, the RA went down and continued going down to its nadir in Fall 55a. Subsequently when the analyst finally addresses the homosexual transference (whether too late or not is not clear) the supervisor advised that the analysis be discontinued. (It is not clear what the role of a countertransference in both analyst and supervisor to the patient's potential homosexuality played in this analysis. JA speculated that the supervisor was away during a time when the analyst seemed to be doing better analytic work and addressing directly with the patient the patient's homosexual feelings and fantasies.) It was striking to us all that in the final summary the word 'homosexuality' does not appear.

A successful case

70T3 was a married woman in her early 30's with two children. She complained of recurrent depressions which at times would become so severe that she found herself "curled up in her depression corner", which meant that she spent much of her day lying on a couch in the corner of her living room. She was unhappy and dissatisfied with her life but wasn't sure why. The summer before the analysis was to begin she had found herself on a vacation unable to leave the house for approximately three weeks.

For 70T3, the three analysts independently clinically examined the notes from the period, April to September 1978, when the MHW was at its nadir; this occurred right after a summer vacation and termination was in the air. The patient reported a dream in which she was leaving for college and her mother was asking her to take a later train.

The analyst (who throughout addressed the transference) interpreted to the patient that she had mixed feelings about both her mother and the analyst asking her to stay; she wanted them to ask her to stay but protected herself against hurt feelings by insisting she wanted to leave. Both WB and LH independently judged this intervention to be “best” “most complete” example of the analyst’s interpretations. The patient responded that she was not aware of a wish on her part for the analyst to ask her to stay and then went on to describe how she felt that there was no reason for the analyst to care whether she was leaving or staying. She said that she understood that she was in a Treatment Center analysis and that the analyst was getting an education out of her analysis. Again, it was interpreted how much she had to ward off her own feelings of wishing that the analyst care about her as she had the concern that her mother didn't love her.

In sum, Analyst 50T1 did not appreciate the patient’s connection to him, did not address ambivalence: wish to leave/wish to stay; aggression/love and did not address the transference.

In contrast, analyst 70T3 interpreted the transference, that the patient wanted to be asked to stay and that the patient protected herself from hurt by threatening to leave.

This preliminary clinical reading of the sections of the notes with particular WRAD-REF configurations indicates that the analyst in 70T3 understood that the patient was ambivalent about leaving and that the patient was, in fact, connected to the analyst. The clinical reading also indicates that the analyst's interpretations allowed the patient to re-establish emotional connection with the analyst, and then tolerate the analyst’s subsequent pregnancy. This is in contrast to 50T1, where clinical reading of the sections of the notes with particular WRAD-REF configurations indicates that the analyst in 50T1 had an apparent lack of understanding of the patient interfered with the analytic process and led to failure.

Discussion: Questions and Future Research

One interesting and problematic finding is that of the 7 cases noted as successful by the treating analyst, only 3 were deemed successful by the current clinical evaluation, using the standardized measures of evaluation of functioning, and the others showed relatively low ratings of functional change. This finding needs to be further investigated, as to whether the ratings by the analyst were biased in a favorable way, or alternatively, the standard measures are missing analytic change.

The major findings of the study thus far are the strong correlations (ranging from .710 to .744) between the clinical ratings of treatment effects and the computerized Referential Activity measures of language style, indicating variation in therapist engagement in the

clinical material. This high correlation is particularly striking in that the language measures were applied to the entire corpus of notes and are essentially independent of specific clinical contents, whereas the clinical measures are based on evaluations of patient functioning and were applied only to initial and ending segments.

Another interesting set of results concerned changes in clinical themes across the decades. While interesting changes were found that are compatible with known changes in theory and technique in the field of psychoanalysis in general during the half century covered by the study, we should emphasize here that we only examined from 2 to 4 cases per decade, so that the results of these analyses should be considered anecdotal only.

Some next research steps

1. In this preliminary study, the analysts carrying out clinical analyses of the successful and unsuccessful cases were aware of the language measures. In the next step of this research, these analyses will be done by a group of clinicians who are not familiar with this project, and who are unaware of the linguistic measures for the segments. For this study a range of segments with different WRAD-REF configurations will be selected and the clinical evaluations will then be compared with the WRAD-REF difference scores for those segments.
2. Other measures concerning involvement of the analyst and patient with the analytic process are in the process of development. These measures, which can be thought of as measures of transference and counter-transference, are based on the proportion of sentences containing both self-reference and reference to the other, and the order in which these references occur.
3. Several of the notes contain reports of dreams. We expect to develop an automatic method of identifying these dream report segments. We will then compare our language measures applied to these dream segments with the language measures for the notes as a whole and with our measures of success.
4. Use of these measures in supervision is being explored.

Conclusions: The findings of this study indicate that information derived from process notes, using computerized linguistic analysis, can play a significant role in a psychoanalytic research project, not as a substitute for verbatim transcripts but as a supplement to such data. The notes provide indicators of the analyst's emotional experience of the patient and their interaction, and variation in the analyst's involvement in the treatment that could not be derived from the verbatim material alone. Finally, the analyst's emotional experience reflected in the writing of his or her notes seems to be related to treatment outcome.

If there is any further information you need please contact us.

Leon Hoffman at hoffman.leon@gmail.com

Wilma Bucci at wbucci@optonline.net

Table 1
Demographic Data for all Cases

Treatment	Analyst Gender	Patient Gender	Pt. Beginning Age	Pt. Other Info
50T1	M	M	Mid 30's	S, Professional
50T2	M	F	Late Teens	M, Student
60T1	M	F	Late 20's	M (2 Children) Homemaker
60T2	M	M	Mid 20's	S, Student
70T1	M	M	Late 20's	S, Professional
70T2	M	F	Early 20's	M, Homemaker
70T3	F	F	Early 30's	M (2 Children) Homemaker (also some free lance)
80T1	F	F	Late 20's	S, Trade-Craft
80T2	F	F	Mid 20's	S, Arts
80T3	M	F	Mid 20's	S, Arts
90T1	F	F	Early 20's	S, Student
90T2	F	M	Mid 20's	S, Student
90T3	M	F	Late 20's	S, Professional
90T4	M	F	Early 40's	S, Arts

Table 2
Description of Treatments and Notes
Comparison of Clinical Evaluations and Language Ratings

Treatment	Length in months	Structure of Notes	Number of Words in Scored Text
50T1	22	Mainly daily, some weekly	90,247
50T2	46	Mainly daily, some supervisory notes, some summaries	290,940
60T1	39	Weekly for 1 yr., then monthly; final 6 months in single summary	24,338
60T2	29	Daily reports and weekly summaries (some of both missing)	28,248
70T1	23	Weekly and 6 month summaries	10,988
70T2	30	Weekly summaries, 3 six month summaries and final summary	7,800
70T3	69	very short weekly summaries; detailed 6 month summaries	18,623
80T1	46	6 month summaries plus final summary	13,145
80T2	42	6 month summaries; also weekly summaries for first 30 months	17,557
80T3	66	6 month summaries; also weekly summaries for first 56 months	19,558
90T1	49	6 month summaries	10,394
90T2	52	6 month summaries	47,931
90T3	30	6 month summaries; also final summary	13,380
90T4	Unclear	1 Early report; 1 6-month summary; 1 final report	9,349

Table 3
Comparison of Clinical Evaluations and Language Ratings

Treatment	Original Evaluation	Current Clinical Evaluation
50T1	U	0
50T2	S	7
60T1	S	8
60T2	U	3
70T1	S	0
70T2	U	1
70T3	S	8
80T1	U	0
80T2	S	0
80T3	S	3
90T1	U	3
90T2	S	3
90T3	U	2
90T4	U	0