

**THE RELIABILITY OF  
THE ACHIEVEMENT OF THERAPEUTIC OBJECTIVES SCALE (ATOS):  
A RESEARCH AND TEACHING TOOL FOR PSYCHOTHERAPY**

Leigh McCullough, Ph.D., Nat Kuhn, Stuart Andrews,  
Harvard Medical School,

Jakob Valen

Norwegian University of Science and Technology

Doxey Hatch,

Montana State University at Billings,

and

Ferruccio Osimo

Psychiatric Clinic-Affori, Università Statale, Milano, Italy

5 June 2003

**THE RELIABILITY OF  
THE ACHIEVEMENT OF THERAPEUTIC OBJECTIVES SCALE (ATOS):  
A RESEARCH AND TEACHING TOOL FOR PSYCHOTHERAPY**

**Abstract**

This study reports on the development and reliability of the Achievement of Therapeutic Objectives Scale (ATOS), a scale designed to assess patients attainment of specific treatment objectives identified as important change mechanisms, both theoretically and clinically in Short-Term Dynamic Psychotherapy (STDP). These items, which also represent common factors across therapies include: patients' recognition of defensive behavior (insight); desire to change the maladaptive responses (motivation); visceral experience of the conflicted feeling (exposure); and adaptive expression of those feelings (new learning). In addition, there is a need to regulate the anxiety or inhibition in the session, and improve the sense of self and others. The ATOS Scale was used to assess patient responses in videotaped STDP in five reliability studies. Three studies were conducted by the originators of the scale at the Psychotherapy Research Program at Harvard Medical School. Two additional studies were conducted at the Psychiatric Clinic-Affori, Università Statale, in Milano, Italy, and at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. Patients in all five studies had received an Axis II diagnosis and were participating in individual outpatient psychotherapy on a weekly basis. These studies demonstrate that the ATOS is a reliable instrument, and show a clear "dose-response" relationship between training on the scales and reliability. At NTNU, there is ongoing work on training, reliability and validity of the ATOS scale.

In the past decade we have entered a new era in the field of psychotherapy. Several previous decades of outcome research have conclusively demonstrated that psychotherapy is effective (Beutler, 1994; Lambert, 2003). Now we are faced with the lengthy and labor-intensive undertaking of process research in order to answer even more challenging questions: Why does psychotherapy work? What are the mechanisms of change? And what instruments will we use to measure them?

This paper describes the historical evolution of an instrument that has been developed over the past 10 years to try to better understand why psychotherapy works by evaluating the degree of in-session change. This instrument, the Achievement of Therapeutic Objectives Scale (ATOS), was designed to assess patients' attainment of specific treatment objectives identified as important change mechanisms, both theoretically and clinically in Short-Term Dynamic Psychotherapy (STDP). The ATOS scale has been subjected to five reliability studies (to be reported in this paper) and several validity studies which are currently underway.

As in any developing field, many new instruments are under development, but to date, only a relatively few process instruments have reliability and validity data published, and, to the best of our knowledge, none of those instruments examine patient response to treatment. The most well-known and well-established are measures of the therapist-treatment focus such as Luborsky's Core Conflictual Relationship Theme (CCRT: Luborsky & Crits-Christoph, 1990), or patient alliance such as the Working Alliance Inventory (WAI; Horvath, 1981). Several instruments have been developed to evaluate therapist interventions; e.g., The Counselor Verbal Response Modes, (Hill, 1974), the Taxonomy of Therapist Response Modes, (Stiles, 1986), The Therapist

Intervention Rating System, Gaston (1990) and the Psychotherapy Q-Sort, (Jones, 1985). Still other instruments evaluate patient characteristics, such as Benjamin's SAS-B (1986).

However, no psychometrically validated scale evaluates the degree to which a patient is adaptively responding to therapy – in other words, the degree of impact of the therapy on the patient. To date, process research has largely overlooked a very important variable -- a patient's in-session response to treatment. As Greenberg notes:

One of the major problems with current clinical trials is that in comparing or evaluating the effects of different treatment interventions on outcome, there is a hidden, intervening variable, which is not accounted for. This variable can be thought of as absorption of the treatment by the client or the activation of the change processes. A therapist may deliver the treatment, but does it take? Does it set the anticipated change processes in motion? ... The link between client change process and outcome needs to be studied and the specific effects of particular processes need to be demonstrated. To do this, rationally, empirically derived client change processes close to the level of performance need to be specified and measures of these processes constructed. ...for a grounded model of how people change in psychotherapy (Greenberg, 1996, p. 451).

One instrument that was specifically designed for measuring a patient's in-session response is the Session Impacts Scale (SIS) (Elliott, 1994). This scale is not theory-driven, but reflects basic research into the patient's report of helpful and hindering events

in treatment immediately after the session. This instrument is an excellent starting point for grounded research into change mechanisms in psychotherapy.

There are also a few instruments that can be co-opted as measures of in-session change. Two examples are: 1) the Experiencing Scale (Klein et al., 1986), which rates the intensity of patient affective response, and 2) the Defense Mechanism Rating Scale (DMRS) (Perry & Cooper, 1986, 1989; Perry, 1990) which hierarchically rates the defenses a patient uses. Neither scale was originally designed to evaluate response to treatment, but the evaluation of patients' shifts from maladaptive to adaptive defensive or affective responses has been found to be a valuable micro-indicator of change in treatment.

Another instrument designed to assess patient in session reactions is the Psychotherapy Interaction Coding System (PIC- System) developed by the first author at Beth Israel Medical Center in New York City (McCullough, 1991). Using videotapes of Short Term Dynamic psychotherapy sessions, the PIC instrument rated eight standard therapist interventions (questions, confrontations, interpretations, etc.) but, in addition included an evaluation of the patients' responses that immediately followed each intervention (affective, defensive or cognitive) for each minute of the therapy session. (See Soldz and McCullough, 2002, for a review of research on the PIC System).

### **THE ACHIEVEMENT OF THERAPEUTIC OBJECTIVES SCALE (ATOS)**

The ATOS scale represents a step beyond the PIC System – which is more of a descriptive analysis of patient response to intervention. In contrast, the ATOS scale reflects the results of a search for the active ingredients in treatment. The ATOS scale

was designed to assess patients' degree of absorption of specific treatment objectives. The ATOS Scale has been developed through the observations of videotapes while, at the same time, attempting to build coherent theory about the mechanisms of change in therapy. Thus, the ATOS scale is an advance beyond the PIC System because it specifically targets theory-derived mechanisms widely recognized to produce change. This study describes five reliability studies, conducted on the ATOS scale, along with the revisions and updates that occurred as a result of each evaluation.

The ATOS Scale was developed from careful observation of videotapes of psychotherapy sessions. This close scrutiny of psychotherapy process was helpful in building coherent theory about the active ingredients and mechanisms of change in treatment (McCullough, 1994). The resulting objectives demonstrate much overlap with other theoretical orientations in regard to well-established change mechanisms or common factors responsible for therapeutic improvement. (This will be discussed in more detail below.) The ATOS scale was designed with the goal of empirically validating these theory-driven and clinically relevant mechanisms.

Short-Term Dynamic Psychotherapy is a highly focused treatment in which the therapist is active and involved, and positive effects on the patient's maladaptive behavior often can be seen following each session. The first author and her colleagues spent thousands of hours observing videotapes, assessing the shifts in patient behavior in response to active interventions. The therapeutic objectives for STDP were generated largely from the combination of patients' reports of helpful events, and the observation of patient changes on videotape. It may not be surprising that the change mechanisms that became evident turned out to be standard change mechanisms, well-known as 'common

factors' (e.g., Orlinsky et al., 2003, in press). As the STDP treatment model was developed and the specific treatment objectives were identified (see McCullough, 1994, McCullough-Vaillant, 1997; and McCullough et al., 2003), there was a need to rate the degree to which treatment objectives were realized by the patient within and across sessions.

The ATOS rating method can best be understood in terms of a biological metaphor. When a physician prescribes medication for a patient, the effect of that medication cannot be determined unless one knows whether the patient actually swallows the pill, the rate that it is absorbed into the blood and/or its effect on target organs. Likewise, when a therapist provides psychotherapy for a patient, the effect of that psychotherapy cannot be determined unless one observes a range of patient responses that indicate that treatment has made an impact on the patient. Analogous to measuring patient blood levels or effect on target organs, the ATOS scale was designed to measure patient behaviors that indicate the degree to which the patient has absorbed or assimilated the therapeutic interventions.

The designated goals of treatment in STDP are to help a patient recognize defensive behavior (insight), desire to change the maladaptive responses (motivation), viscerally experience the conflicted feeling (exposure) and adaptively express those feelings (new learning). In addition, there is a need to regulate the anxiety or inhibition in the session, and improve the sense of self and others. Thus, the variables measured by the ATOS scale represent theory-derived, clinically relevant mechanisms. Research on the PIC system lends preliminary support to the power of these objectives (See the review by McCullough et al, 2003).

As noted earlier, although the objectives rated by the ATOS scale were derived from observation of videotapes of STDP, we found that these objectives overlapped with standard common factors in psychotherapy (see below). Therefore the ATOS scale also can be adapted to be used to evaluate psychotherapies of a variety of theoretical orientations, not only STDP. These seven treatment objectives (stated first in common factor terms, and second as STDP objectives) include the following:

- 1) Awareness or insight of maladaptive behavior: How clearly can a patient recognize his or her maladaptive cognitive schemas or defensive behavior patterns? In STDP, insight is determined by the patient's recognition of maladaptive defensive behavior patterns.
- 2) Motivation to change: How much does the patient want to give up the maladaptive or defensive behavior? In STDP, motivation is determined by the patient's willingness to give up the defensive behavior.
- 3) Affective arousal/exposure: How much bodily arousal of feeling is experienced in the session? In STDP this is a measure of exposure to and desensitization of conflicted feelings.
- 4) New learning: How effectively is the patient able to express these feelings interpersonally outside the session? In STDP this is called adaptive affective expression.
- 5) Degree of Inhibition: How much anxiety, guilt, shame, or emotional pain is elicited in this process? In STDP, anxiety or inhibition must be regulated to be kept within bearable limits to the patient.

- 6) Improvement in the sense of self. In STDP this is referred to as Self Restructuring.
- 7) Improvement in relations with others. In STDP, this is referred to as Other Restructuring.

ATOS ratings are made from videotaped segments of psychotherapy sessions (audiotapes or transcripts may also be used, but videotape provides the most complete view of the session). Each of the above objectives is rated on a 1-100 scale, (from least adaptive to most adaptive responding). Between sessions and within sessions, these ratings capture the degree of change in these objectives.

The ATOS scale was initially constructed in 1995. The reliability studies above were conducted in 1996, 1997 and 1999 with continued refinement of the rating scales between each testing. Most of the revisions in this study involved better operational definitions of the variables – grounding each level of the rating scale in behavioral examples. For instance, ATOS raters had been having difficulty reliably rating the extent to which patient were experiencing feelings of closeness. This was resolved by clearer definitions (the patients had to be verbally stating that they were experiencing a tender, warm, close, affectionate or loving feeling) and reliability improved.

Another change involved the length of the segment to be rated. In Study 1 (1996), the entire therapy session (usually 50 minutes) was given one rating per objective. After discovering that there was too much variation within the 50-minute period, rating periods were cut down to a more manageable time period of ten minutes. Thus, each session has five or six segments that are each ten minutes in length. By graphing the results of these ratings, it can be seen when the patient was more or less insightful, motivated, or

emotionally activated as the session progresses. Using this data, therapy can be analyzed sequentially to evaluate when there are adaptive (or maladaptive) shifts in the levels of these important factors. These shifts or peaks can be correlated to outcome, and can be evaluated regarding their effects on the other objectives. Later, the strongest predictors of outcome can be identified and analyzed to determine what interventions the therapist employed to bring about the change.

Another important change involved the identification of the predominant affect focus for each segment to be rated. Starting with Study 2 (1998), before rating each objective, raters were required to choose the predominant affect focus of the session in question from a list of eight common affects observed in STDP (anger/assertion, grief, closeness, positive feelings about the self, interest/excitement, joy, sexual excitement, healthy fear responses. (See McCullough et al, 2003). This choice is particularly important because the numerical ATOS ratings (e.g., of affect experiencing) depend on the identified focus.

Although we have highlighted the treatment objectives in STDP, these same common factors are also salient in many other therapies, such as Cognitive Behavioral Therapy (CBT), Interpersonal Therapy (IPT) or Dialectical Behavior Therapy (DBT). For example, insight in Cognitive therapy could be the degree of awareness of maladaptive cognitive schemas; in DBT insight might be the awareness or mindfulness. Motivation to change is similar throughout. Exposure to feelings (or lack thereof) as well as degree of patient inhibition can be assessed in any therapy. And patients' report of new learning of adaptive behaviors is a familiar part of any successful treatment.

Many studies have established that these standard common factors predict favorable outcome in psychotherapy (see the review by Orlinsky et al., 2003). Preliminary ATOS studies on STDP in our research laboratory have also shown correlations between the ATOS objectives to improvement (McCullough et al., 2001). When further research firmly establishes the factors that are most strongly associated with improvement, later studies can begin identifying the therapist interventions that have been the most useful in promoting these factors.

The manual describing ATOS ratings for each of the ‘therapeutic objectives’ or common factors listed above can be obtained from the following website:

<http://www.affectphobia.org>

Research has been ongoing for seven years at Harvard Medical School (HMS) to develop and refine the ATOS scale as an effective measure of change mechanisms in successful psychotherapy as well as monitoring how those mechanisms operate. What follows is a summary of three reliability studies on the ATOS scale conducted by the Psychotherapy Research Program at HMS, and two additional studies conducted in Italy and Norway.

### **General Methods**

The ATOS Scale was used to assess patient responses in videotaped short term dynamic psychotherapy. All patients in each of the five studies had received an Axis II diagnosis and were participating in individual outpatient psychotherapy on a weekly basis. All patients gave informed consent to be videotaped and to participate in this research study.

The ATOS scale was developed and adapted collaboratively by the scale's authors and other colleagues since 1995. The raters in Studies 1 and 2 were developers of the scale (and the first three authors of this paper). In Study 3 graduate and undergraduate student raters were added to the reliability testing group and received training as follows: 1) between 4 and 6 months didactic training; 2) practicing rating videotaped short-term psychotherapy segments; and 3) comparing their ratings to "Gold Standard" ratings previously made by the scale's developers.

Two studies were conducted outside of our laboratory: one at the Psichiatria Clinica-Affori, Università Statale, in Milano, Italy, and the other at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. It should be noted that the inhibition scale was not included in the original ATOS scale because inhibition is not, strictly speaking, a treatment objective. However, the process of managing inhibition (anxiety regulation) is central to each of the STDP treatment objectives, so the inhibition scale was added in 2001.

JakobsLadder: A web-based training tool

The Norwegian Study (Study 5) uses a web-based reliability and training tool for psychotherapy process evaluation (developed by Jakob Valen and Leigh McCullough). On this website ([www.jakobs ladder.com](http://www.jakobs ladder.com)) manuals for a variety of instruments (such as the ATOS scale) may be downloaded and transcripts of sessions are available for training.

JakobsLadder also contains an interactive tutorial that helps the trainee to practice learning the selected instrument, in this case, the ATOS Scale. Once the trainee feels proficient, reliability testing may be done: The trainee selects test transcripts, videotaped

or audiotaped sessions, and as ratings are made, they are then entered into the JakobsLadder program. Once the ratings are entered and saved, the trainee is given feedback on the correct answers. This feedback is in the form of ‘Gold Standard’ ratings with accompanying commentary explaining why the ratings were made. This feedback information must be prepared in advance by the originator of the instrument, often a laborious process. However, once the ratings and explanations are generated and entered into the JakobsLadder program, it can provide automated training for years to come.

DROP There are currently three ATOS reliability tests of 18 segments each that are available for ATOS training. This paper reports on the data collected from ATOS reliability test #1. Those students who attain reliability at a level of 0.65 or better receive a ‘Reliability Certificate’ which is automatically printed by the JakobsLadder program. Those students who did not attain reliability on Test #1 must proceed to Tests #2 and #3 until reliability is attained so that they may qualify to rate the clinical trial videotapes in the Norwegian Videotape Archive Library (discussed below).

### **Statistical Analyses**

There are two statistics used in this paper; one for the agreement on categories of affect, and one for agreement between continuous ratings.

The main affect focus for ATOS analysis is a categorical decision, where raters either agree or disagree. Cohen’s Kappa is suitable for analyzing agreement between pairs or groups of raters of categorical data. Therefore, average kappa values will indicate the level of agreement between the raters on the Affect Focus; i.e., the predominant affect experienced by the patient in each ten-minute segment.

ATOS objective ratings are scored from 1-100; the higher the score, the more adaptive the behavior (such as strong insight into maladaptive patterns, or high intensity of feeling being experienced). In each study the same group of raters evaluated the same ten-minute segments. The raters were considered a random selection of a larger group of possible raters. We did not assume lack of any rater bias. Therefore, a two way analysis of variance with a random model (random effects of raters, random effects of subjects) were used to calculate the Intra Class Correlation (Shrout & Fleiss, 1979). Absolute agreement models were used according to McGraw & Wong (1996). In the Norwegian Study, students' ratings were compared to the Gold Standard ratings. Therefore a one-way analysis was used, also with single measures.

According to Shrout's (1998) revision of Landis and Koch's (1977) reliability "labels" or "values," ratings below 0.41 are slight or none, ratings between 0.41-0.60 are only fairly reliable, ratings between 0.61-0.80 are moderately reliable and ratings over 0.81 are substantially reliable. We will use these categories in evaluating our results. It is often difficult to interpret what a correlation statistic denotes in terms of actual scoring agreement. After careful examination of the rater's scores, we noted that ICC correlations roughly corresponded to the following ATOS difference ratings: above 0.80 differed approximately 4-8 ATOS rating points, scores between 0.60 and 0.80 differed approximately 8-12 ATOS points. The problem with trying to link difference ratings to statistical correlations is that often one strong outlier in a group of otherwise consistent ratings will yield a low ICC; e.g. one subject who rates 40 points from the mean will lower the ICC score, even if there is only a 6 point average difference to the other comparison scores. Of course, the higher ICC correlations reflect less such variability,

while the lower correlations might reflect one strong outlier (which happened in approximately 8 % of our data) or simply wider variation across all raters. We offer these examples of concrete difference scores only to give the reader some grounding regarding what it means to be ‘reliable’ on the ATOS scale; i.e., raters must receive an ICC score of above .65 (moderate to substantial), roughly corresponding to scoring within ten ATOS points (or within one ATOS scale level) in comparison to each other or in comparison to the Gold Standard rating.

### **Study 1**

In the first systematic assessment of the ATOS instrument, three raters worked together for 1-2 years to develop the ATOS scale by practicing making ratings of videotapes of short term psychotherapy. During this time the scale was continually developed and revised. For the formal reliability study, the three raters individually viewed new videotapes of STDP therapy sessions and then made ATOS ratings on each of the objectives for each full 50-minute session. Once all the raters had finished rating the full session, and recorded those ratings, they discussed the session and compared each other’s ratings to arrive at a consensus rating. This procedure was carried out for thirty-four different therapy segments for 5 subjects with Axis II, Cluster C diagnoses.

The average kappa value between pairs of judges was .40 (SD= 0.20), which according to Shrout (1998) is slight to fair. ICC (2A,1) correlations were calculated for each rater’s reliability with each of the others raters (See Table 1). Two of these ratings fell at the top of the ‘fair’ range (Motivation - 0.56 and Image of Self – 0.55). The remaining four ratings fell just inside the ‘moderate’ category (Insight into Defenses – 0.61; Exposure to Affect – 0.65; Expression of Feeling – 0.66; Image of Others – 0.64).

The variation of the individual raters scores with the consensus ratings ranged from 0.19 to 0.80 with half of the ratings in the fair category and three considered as moderate.

## **Study 2**

Following Study 1, the ATOS manual and rating scales were revised for almost two years to address areas of ambiguity discovered in the level descriptions used in Study 1. The main innovation introduced at this point was generating ATOS ratings for each ten-minute segment of a therapy session, rather than generating only one set of ratings for the entire session. The scale developers felt that there was far too much variation in a 50 minute session, and a single ATOS rating for each objective did not capture the nuances of the work being done. Therefore, several different time periods were evaluated; five, ten and twenty minute segments of sessions. Twenty-minute segments revealed the same problem as whole sessions: they contained too much variation to obtain clear ATOS ratings. Five minute segments were too brief to provide the context necessary for quality ATOS ratings. However, the ten-minute time frame provided enough material to make clinically satisfactory ratings but was not too long to be confounding. We decided that rating each ten-minute segment of therapy optimally captures the flow of therapy session and patients' response to it. The shorter rating segments were also desirable because they made possible the more rapid generation of data points which would permit time series analyses.

There was one exception to this new practice of restricting ratings to ten-minute segments. The ratings of self and others did not vary across a session as much as the other objectives and thus did not require the more intensive ten-minute segment ratings. Therefore it was decided to rate Self and Other Restructuring only once, at the end of

each full session. In fact, in Study 2 we decided not to include the Self and Other Ratings because of time restrictions; each rating required time-consuming discussion and deliberation among the raters. Therefore, for both Study 2 and Study 3, we concentrated only on the main four objectives (e.g, defense recognition, defense relinquishing, affect experiencing and affect expression). In Study 2, three very well-practiced researchers rated thirty-five segments from seven therapy sessions. These sessions featured five patients, all with Axis II, Cluster C disorders.

The results of Study 2 are presented in Table 2, and are much improved over Study 1. The raters agreed perfectly on the Affect Focus. Pooled judges ratings for the four objectives ranged from .66 to .84. Two of the objectives (Motivation for change and Exposure to Affect) were reliable in the ‘substantial’ category (.84 and .80 respectively). The other two objectives (Insight into Defenses, and Affect Expression) fell in the moderate level of reliability (0.66 for both).

### **Study 3**

Study 3 was undertaken after an additional year of refinement of the ATOS manual and scales. At this point, changes in the ATOS method involved more extensive specification of the discrete behaviors for each 10-point rating on each scale. This has been an ongoing process, with the scale levels becoming more and more distinct each year by being more grounded in specific behavioral indicators. For example, in the recognition of defenses, there is the distinction between the therapist pointing out the defense and the patient merely acknowledging and agreeing. If the therapist points out the maladaptive behavior, the rating would be 39 or below on the ATOS scale. If patients are able to verbally demonstrate insight, on their own –without the therapist’s help – that

they could see themselves falling into a maladaptive or defensive pattern, the Insight/Defense Recognition rating on the ATOS scale is rated above 40. The degree of detail included (i.e., how much insight the patient demonstrates into their maladaptive patterns) would determine how far above 40 the rating would be. These levels have become increasingly behaviorally grounded to assist the ease in making ratings.

Study 3 attempted a more rigorous method than the previous two studies. Six raters were separated into two groups that independently rated the same eighteen therapy segments of ten minutes each. One rater dropped out before completion, so only five raters' scores are used.<sup>1</sup>

Prior to study 3, the raters had always ended their rating sessions by generating consensus ratings from extensive discussion to insure that all viewpoints had been considered, and to generate “Gold Standard Ratings” that all could agree with. We decided to study consensus ratings across two groups because consensus data has been shown to be more reliable than data generated by individual raters. Furthermore, it is both enjoyable and educational to discuss the ratings with colleagues. In Study 3 we wanted to evaluate whether these consensus ratings were reliable from one group to another.

Therefore, in Study 3, the raters split into two groups and met in separate rooms to generate separate consensus ratings for 18 videotaped segments. Within each group, after each researcher had completed his or her ratings for a therapy segment, the ratings were discussed and a consensus rating was obtained for that group. This procedure was followed in both groups, yielding two blind, independent consensus ratings for each

---

<sup>1</sup>We want to thank Cara Lanza-Hurley, Jonathan Wolf, and Amelia Kaplan for their participation in Study 3 conducted at Harvard Medical School.

therapy segment. The groups had no access to the other group's consensus scores until all ratings were made and recorded.

Then both groups met together, compared their respective consensus ratings, and through a spirited discussion, arrived at a final, overall 'gold standard' consensus ratings, taking all of the raters' perspectives into account. Study 3 thus provides a particularly rich harvest of information about the inter-rater reliability of the ATOS. Again the interjudge agreement on the affect focus was excellent. The kappa was substantial in group 1 (Kappa 0.83) and perfect in group 2 (1.00). The agreement on affect focus between the two groups was moderate (Kappa = .64). Group 1's agreement with the overall consensus was .91 and Group 2 was .67.

Table 3A presents the ICC correlations for the ATOS objectives for both Group 1 raters and Group 2 raters. Group 1 fell in the moderate range with scores ranging between 0.56 and 0.77, except for motivation. Group 2, which included two less experienced raters, had greater variation in ratings and the correlations were lower (ranging from 0.43-0.68), falling in the fair to moderate levels.

Table 3B presents ICC figures for the agreement between the two groups' consensus ratings. Even though Group 2 showed more variation across raters, when the overall 'Consensus' score was compared to Group 1's consensus score, the reliability was moderate for Exposure to Feeling and Expression of Affect (0.66 and 0.77 respectively) and 'substantial' for Insight into Defenses and Motivation (0.80 and 0.81 respectively).

#### **Study 4 – The Italian ATOS Study**

As a further test of the general usefulness of the ATOS scale, a study was conducted by the research group<sup>2</sup> at the Affori Psychiatric Clinic of Università Statale, Milano, Italy. Five therapists, one with extensive clinical experience, two with 6 years, one with 2 years, and one with no clinical experience, rated videotapes of STDP sessions with four adult outpatients (2 male, 2 female, all with Axis II disorders). For each patient, ten segments were rated, five from each of two consecutive sessions for a total of 40 segments. These sessions were drawn from different stages in the course of therapy, from initial evaluation to final session.

The therapists treating the patients were also varied in terms of experience, with two beginning therapists and one having over 20 years experience. One rater in the Italian study had 8 months of training on the ATOS, and the additional raters received four months of training from him (2 hours once per week for a total of approximately 32 hours).

In this study, no consensus rating was generated, and gold standard ratings were not available to be used for comparison (the only videotapes with a Gold Standard Rating were in English!). Instead, ratings from each researcher were compared to each other and ICC analyses were performed.

The raters' agreement on the Affect Focus was, like the three previous studies, excellent ( $\kappa = 0.74$ ,  $SD = 0.08$ ). Table 4 shows that the Italian raters did very well in their ability to similarly rate the ATOS objectives among each other. All four objectives fell within the moderate range of reliability with scores ranging from 0.61 to 0.78.

#### **Study 5 – The Norwegian ATOS Study**

---

The final and most rigorous study to date, involved eighteen graduate students in clinical psychology at the Norwegian University of Science and Technology, in Trondheim, Norway. This study tested not only the ATOS scale, but also the new web-based system for data entry discussed above (JakobsLadder) that provides feedback on the correctness of the rating, after the rating is entered. This is an automated way of providing some ‘discussion’ of the “gold standard’ ratings and education on STDP to the rater following each rating entry.

The Norwegian raters participated in this study as an elective research class. These students, who generally have either no clinical experience or no more than one or two years of clinical experience, were given only eight hours of training (2 hours per week for 4 weeks) in the use of the ATOS prior to being tested for reliability. During the class, the students were shown videotapes of STDP therapy sessions (in English with accompanying English transcription) and instructed to rate each designated ten-minute segment on the four main ATOS dimensions. We also asked them to rate “Inhibition” with only a brief introduction, because we were just developing this scale and wanted a preliminary evaluation of the degree of difficulty of this scale.

Due to the limitations of class time, these procedures were considerably more challenging for the raters than all the previous studies. The students were restricted on time permitted to score the tapes. They watched the 10 minute segment and were instructed to take no more than 10 minutes to rate 5 of the subscales. Students were instructed not to discuss their rating sessions with each other because we wanted to evaluate how well each rater could match the Gold Standard Score. The first half of the

---

<sup>2</sup> We thank: Arduini L., Carta I., Fava E., Landra S., Masserini C., Merlo A., and Pazzaglia P., who took part in the Italian ATOS study at the Affori Psychiatric Clinic, Università Statale, Milano, Italy.

reliability test was carefully monitored, the second half of the test was not monitored, but students reported following the guidelines fairly well.

The students entered their ATOS ratings on the ATOS data entry program of the JakobsLadder website and, following their data entry, received the answers for the Gold Standard rating as well as the typed commentary discussing why the gold standard rating was made. ICC correlations were computed by comparing each students' ratings to the "Gold Standard" ratings generated by the members of the Psychotherapy Research Program at Harvard Medical School (See Table5).

The results are as follows: Nine of the 18 students completed rating all 25 segments in time for inclusion in this study. The ICC(1) scores across all the Norwegian students fell in the 'slight' range for Motivation (0.28) and Inhibition (0.39), and the 'fair' range for Insight into Defenses (0.52) and Exposure to affect (0.55) and Expression of Feeling (0.51). The Norwegian students were not tested for the Affect Focus in this study, but instead received that information prior to rating the objectives.

The results are as follows: Nine of the 18 students completed rating all 25 segments in time for inclusion in this study. The average ICC scores across all the Norwegian students fell only in the 'slight' range for Motivation (0.34), the 'fair' range for Insight into Defenses (0.53) and Inhibition (0.57), and the 'moderate' range for Exposure to affect (0.73) and Expression of Feeling (0.62). The Norwegian students were not tested for the Affect Focus in this study, but instead received that information prior to rating the objectives.

## DISCUSSION

Taken together these studies demonstrate that a wide variety of raters can reliably rate the ATOS objectives. This has held true in five studies in three countries (USA, Norway and Italy) over 7 years time. Raters have been both experienced clinicians, as well as graduate students with minimal or no clinical training.

The variations in the results across the five studies reflect the degree of development of the scale as well as the level of training of the raters. For example, Study 1 has lower reliability correlations, as might be expected in the preliminary stages of development of the ATOS Scale. In contrast, Study 2 showed the strongest reliability scores, probably due to several factors; 1) the move to ten-minute ratings, 2) the better operationalization of the objectives, and 3) the years of cooperative work and practice on the ATOS Scale among the three raters. Because these three raters (who were also the developers of the scale) had been working together intensively for three years in refining the scale, it is not surprising that their reliability scores are high.

In Study 3, the reliabilities are somewhat lower than in Study 2, but this also might be expected because graduate student raters without clinical experience were included for the first time. In fact the lower, though acceptable, scores of Study 3's Group 2 are attributed to the variation in scoring of the two graduate students in Group 2 who had received only a few months of training, compared to the years of practice by the three senior raters. Study 1 and the Norwegian student group (Study 5) demonstrate the lowest reliability scores.

All five studies are compared to each other in Figure 1. Scanning this figure, there are four studies in which most scores fell in the moderate to substantial range (above

0.60) (Study 1, Study 2, Study 3 - Group Consensus comparison and the Italian Study. (Study 3, Group 1 alone and Group 2 alone are not included in this figure). There were only two exceptions that fell in the high 'fair' range (Study 1 for Motivation, (0.57), and Image of Self (0.57). Only two ICC score in all five studies fell below a 'fair' level of reliability; the Norwegian study for Motivation (0.34), and Inhibition (.39). In fact, figure 1 demonstrates that the motivation scores receive the lowest reliability of all the objectives for three of the studies (Study 1, Italy, and Norway). Thus, the motivation scale will need further attention and refinement.

In retrospect we can identify reasons why Study 1 and Study 5 attained somewhat lower overall levels of reliability than the other studies. Study 1 occurred during the early development of the ATOS scale before scale levels were well defined and before the shift to 10-minute segments.

In Study 5, the Norwegian students were the least trained in ATOS, had little or no clinical experience, and were given the shortest time to make the ratings. In addition, the videotaped sessions were in English, their second language. In addition, the case that received the poorest reliability was a very difficult case to rate. The case involved restructuring of the sense of self where the therapist and patient discussed how the patient thought the therapist felt about her. This was an unusual focus for the relatively inexperienced students, and the use of personal pronouns was particularly difficult to translate (e. g., the patient saying, "How do I think, YOU feel about ME? I think you might feel good about me, but it is hard for me to believe that). Such examples illustrate the many complex factors that make up reliability testing. In the future we might restrict beginning reliability tests to more standard patient-therapist interactions and relegate

these more complex interactions to tests of ‘advanced levels.’ The majority of the Norwegian students had difficulty with the Motivation scale, so we will also discuss the difficulty and generate clearer rating level descriptions.

Given the constraints upon them, it is surprising that the Norwegian raters did as well as they did. What is even more surprising about Study 5 is that, despite these constraints, 5 of the 9 Norwegian students or 55% of the class were reliable at a ‘fair’ to ‘substantial’ level (range 0.49-99) and 3 of these students achieved above 0.59 on all ratings. We think this is a testimony to the dedication and hard work of these students as well as to the clear descriptions of the levels that have been achieved in the development of the ATOS scale. Such impressive results suggest that the ATOS can be quickly learned and employed if one is interested and highly motivated – as were these high-performing students.

In summary, 8-12 hours of training might be considered the lower limit for training on the ATOS scale. The more probable amount of training needed is probably between 20- 30 hours (e.g., 2 hours per week for 10-15 weeks) as in the Italian study and our Study 3 for the graduate student raters. Further research is currently underway to establish time needed for training.

### **Conclusion and Future Directions**

To the best of our knowledge, the ATOS scale is the first instrument to measure the degree of patient assimilation or absorption of therapy, as manifest in achievement of some specific therapeutic objectives. Identifying objectives, and then rating to what degree the therapist assists the patient in achieving those objectives, offers a new and potentially useful method for assessing therapist competence in applying the treatment

model. At the same time, it serves as an index of patients' in-session response to treatment. The low frequency of marked failures in agreement not only provides evidence for the psychometric soundness of the ATOS, but also suggests that some raters can master use of the ATOS within relatively few training and practice sessions.

As demonstrated in these studies, the ATOS is a reliable instrument which is designed to measure objectives in STDP. Studies are currently underway to explore whether these objectives can be empirically linked to common factors in other therapies (e. g., CBT, DBT, etc.). These studies also show a clear “dose-response” relationship between training on the scales and reliability. Ongoing work on training and reliability at NTNU should provide more information not just about the reliability of the scale, but on methods that can be used to reliably train reliable raters. We plan to incrementally increase the time for training and the time allowed for generating ratings until we establish the minimal time needed to achieve reliability on the ATOS Scale for the largest percentage of trainees.

Of course, the reliability of a scale is not important if the scale is not valid, and a validity study is currently underway comparing the ATOS defense and affect objectives to ratings on the DMRS and the Experiencing Scale (Klein et al., 1969), as well as to therapeutic outcome. If high ATOS Scores are found to correlate well with positive therapeutic outcome, a new realm will be open in psychotherapy research, because we will have a relatively simple way of evaluating and comparing the effectiveness of individual, specific therapeutic techniques by evaluating their effect on ATOS ratings.

In addition, at NTNU we plan to conduct a large process-outcome study using the ATOS scale to examine the change mechanisms listed above across clinical trials of three

forms of therapy. These clinical trials include the comparison of STDP to Cognitive Therapy conducted at NTNU (Svartberg, Stiles and Seltzer, 2003 in press), and a clinical trial of Dialectical Behavior Therapy, contributed by Marsha Linehan from the University of Washington, Seattle. We are particularly fortunate to have a large body of students in a central location in Trondheim, Norway who, after attaining reliability, will intensively rate the thousands of videotapes that are available from these trials. We are also fortunate to have the web-based reliability assessment tool, [jakobsladder.com](http://jakobsladder.com) to provide rigorous training and reliability checks on the ATOS scale (as well as other process instruments). This will make possible an unprecedented volume of data on the intricate workings of psychotherapy.

### **The Value of the ATOS Scale as a Training Tool**

Beyond its' potential importance in research, the ATOS may also offer significant benefits in psychotherapy training. It has the potential to augment the usefulness of videotape review by helping trainees focus on key areas related to therapeutic effectiveness. In this sense, the process of learning to do the ratings by reviewing training tapes can be highly educational. Just as a pilot is required to spend hundreds of hours of practice in the test cockpit before being allowed to taking a commercial plane into the air, it is now possible for therapists-in-training to log hundreds of hours at the videotape machine -- not in passive watching, but in interactive coding, responding and evaluating the interventions, methods, and patient responses that lead to improvement in psychotherapy.

Beyond that, clinicians in private or clinic settings, alone or in groups can use the ATOS scale to hone their clinical skills and help them maintain focus on important

clinical objectives or to upgrade their skills by learning the basics of broad-spectrum STDP.

## References

- Benjamin, L. S., Foster, S. W., Roberto, L. G., & Estroff, S. E. (1986). Breaking the family code: Analysis of videotapes of family interactions by Structural Analysis of Social Behavior (SASB). In L. S. Greenberg & W. M. Pinsof (Eds.), The psychotherapeutic process: A research handbook, New York: Guilford, pp. 391-438.
- Beutler, L., Machado, P. P., & Neufeldt, S. A. (1994). Therapist variables. In A. E. Bergin & S. L. Garfield (Eds.), Handbook of psychotherapy and behavior change. New York: Wiley, pp. 229-269.
- Elliot, R. (1986). *Session Impacts and Intentions Questionnaires*. (Unpublished manuscript, University of Toledo [OH], Department of Psychology)
- Elliot, R., & Wexler, M. M. (1994). Measuring the impact of sessions in process-experiential therapy of depression: The Session Impacts Scale. Journal of Counseling Psychology, *41*, 166-174.
- Gaston, L., (1990). The Therapist Intervention Rating Scale.
- Greenberg, L. S., & Foerster, F. S. (1996). Task analysis exemplified: the process of resolving unfinished business. Journal of Consulting and Clinical Psychology, *64*, 439-446.
- Henry, W. P., Schacht, T. E., & Strupp, H. H. (1986). Structural analysis of social behavior: Application to a study of interpersonal process in differential psychotherapeutic outcome. Journal of Consulting and Clinical Psychology, *54*, 27-31.

- Hill, C. E. (1978). Development of a counselor verbal category system. Journal of Counseling Psychology, 25, 461-468.
- Hill, C. E., & O'Grady, K. E. (1985). List of therapist intentions illustrated in a case study and with therapists of varying theoretical orientations. Journal of Counseling Psychology, 32, 3-22.
- Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. Journal of Counseling Psychology, 36, 223-233.
- Høglend, P., Bøgwald, K. P., Amlo, S. Heyerdahl, O., Sørbye, Ø., Marble, A., Sjaastad, M. C., & Bentsen, H. (2000) Assessment of Change in Dynamic Psychotherapy. Journal of Psychotherapy Practice and Research, 9(4), 190-199.
- Gaston, L. (19xx). The Therapist Intervention Rating Scale.
- Jones, E. E. (1985). *Manual for the Psychotherapy Process Q-set*. Unpublished manuscript, University of California, Berkeley.
- Klein, M. H., Mathieu, P. L., & Kiesler, D. J., Gendlin, E. T. (1969). *The Experiencing Scale*. Madison, WI: Wisconsin Psychiatric Institute.
- Lambert, M. L. (2003). Bergin and Garfield's handbook of psychotherapy and behavior change (5<sup>th</sup> ed.). New York: Wiley.
- Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33 159-174
- Luborsky, L., & Crits-Christoph, P. (1990). Understanding transference: The CCRT method. New York: Basic Books.

- Marmar, C. R., Horowitz, M. J., Weiss, D. S., & Marziali, E. (1986). The development of the Therapeutic Alliance Rating System. In L. S. Greenberg & W. M. Pinsof (Eds.) The psychotherapeutic process: A research handbook. New York: Guilford, pp. 367-390.
- McCullough, L. (1991). The Psychotherapy Interaction Coding System. Unpublished manuscript available from the author; 943 High Street, Dedham, MA. 02026.
- McCullough Vaillant L. (1994). The Next Step in Short-Term Dynamic Psychotherapy: A Clarification of Objectives and Techniques in an Anxiety-Regulating Model. Psychotherapy, 31(4), 642-654.
- McCullough, L., Kuhn, N., Andrews, S., Kaplan, A., Wolf, J., & Lanza, C. (2003). Treating Affect Phobias: A Manual for Short Term Dynamic Psychotherapy, New York: Guilford Publications.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30-46.
- Orlinsky, D. E., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy—noch einmal. In M. Lambert (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change. New York: Wiley, pp. 270-376.
- Osimo F., Merlo A., Arduini L., Landra S., Fava E., Masserini C., Carta I., & Pazzaglia P.(1998). La scala ATOS: Achievement of Therapeutic Objectives Scale. Ricerca in Psicoterapia. Journal of the Italian section of the Society for Psychotherapy Research, 1(2),153-166.

- Perry, J. C., & Cooper, S. H. (1985). Psychodynamics, symptoms, and outcome in borderline and antisocial personality disorders and bipolar type II affective disorder. In T. H. McGlashan (Ed.), The borderline: Current empirical research, Washington, D.C.: American Psychiatric Press, pp. 21-41.
- Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research, 7(3), 301-317.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428.
- Stiles, W.B. (1986). Development of a taxonomy of verbal response modes. In L. S. Greenberg & W. M. Pinsof (Eds.), The psychotherapeutic process: A research handbook, New York: Guilford, pp. 161-199.
- Svartberg, M., Stiles, T., and Seltzer, M. (2003, in press). Archives of General Psychiatry.

Table 1

*Study 1(USA) interrater reliability estimates (intraclass correlations; ICC), with 95% confidence intervals (CI), for single raters, randomly drawn from a group of three raters (N=51).*

Objective	ICC(2A,1)	95% CI
Insight into Defenses	.61	.47-.74
Motivation For Change	.56	.40-.70
Exposure to Affect	.65	.51-.77
Expression of Affect	.66	.52-.77
Image of Self	.55	.26-.79
Image of Others	.64	.37-.84

Table 2

*Study 2(USA) interrater reliability estimates (intraclass correlations; ICC), with 95% confidence intervals (CI), for single raters, randomly drawn from a group of three raters in Study 2. (N=23)*

Objective	ICC(2A,1)	95% CI
Insight into Defenses	.66	.46-.82
Motivation For Change	.84	.73-.92
Exposure to Affect	.80	.68-.90
Expression of Affect	.66	.45-.82

Table 3A

*Study 3(USA) interrater reliability estimates (intraclass correlations; ICC(2A,1) for members of group 1 and 2.*

Objective	Group 1 (k=2)	Group 2 (k=3)
Insight into Defenses	.77	.43
Motivation For Change	Ns	.64
Exposure to Affect	.63	.68
Expression of Affect	.56	.51

Table 3B

*Study 3 (USA) Interrater reliability estimates (intraclass correlations; ICC(2A,2)*

*Of Group 1 and Group 2 Consensus Ratings in Study 3*

Objective	ICC (k=2)
Insight into Defenses	.80
Motivation For Change	.81
Exposure to Affect	.66
Expression of Affect	.77

Table 4

*Italian Study interrater reliability estimates (intraclass correlations; ICC), with 95% confidence intervals (CI), for single raters, randomly drawn from a group of four raters (N=23).*

Objective	ICC(2A,1)	95% CI
Insight into Defenses	.76	.63-.87
Motivation For Change	.61	.44-.77
Exposure to Affect	.75	.60-.87
Expression of Affect	.78	.64-.88

Table 5

*Norwegian Study interrater reliability estimates (intraclass correlations; ICC(1)), for 9 raters compared to the Gold Standard with 95% confidence intervals (CI) .*

Objective	ICC(2A,1)	95% CI
Insight into Defenses	.52	.37-.70
Motivation for Change	.28	.14-.50
Exposure to Affect	.55	.40-.72
Expression of Affect	.51	.34-.71
Inhibition	.39	.25-.58

Figure 1

*ICC reliabilities compared for all five studies.*

