Ridenour Jeremy (Orcid ID: 0000-0003-1521-9075)
Stein Michelle (Orcid ID: 0000-0002-4179-0817)

Longitudinal Stability of SCORS-G Dimensional Ratings, Score Ranges, and Narrative "Blandness" in a Clinical Sample

Jeremy M. Ridenour[1]

Katie C. Lewis[1]

Caleb J. Siefert[2]

Michelle B Stein[3]

Author Note

[1] Erikson Institute for Education and Research, Austen Riggs Center, Stockbridge, Massachusetts. [2]Department of Psychology, University of Michigan – Dearborn. [3]Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School.

Correspondence concerning this article should be addressed to Jeremy Ridenour, Erikson Institute for Education and Research, Austen Riggs Center, 25 Main Street, Stockbridge, MA 01262. E-mail: Jeremy.Ridenour@AustenRiggs.net.

Researchers interested in learning more about the data of this study may contact the corresponding author. The data are not publicly available due to their containing information that could compromise the privacy of research participants.

Keywords: performance-based assessment, TAT, SCORS-G, object relations

Abstract

While understanding how internalized representations of others (i.e., object relations) change over the course of treatment is essential for treatment planning and evaluation of progress, few studies have examined the nature of these changes through repeated psychological assessments. In this study, we applied the SCORS-G to Thematic Apperception Test narratives for 61 adult patients with complex psychiatric problems undergoing residential treatment over an 18-month period. Over the course of treatment, patient representations of self and others became more complex, indicating improvements in mentalization. Further, an analysis of default ratings (i.e., an aspect of story "blandness") suggested that certain shifts in SCORS-G dimensional scores over time were accounted for by greater degree of scorable content across time points, rather than changes specific to the dimensions themselves. Findings from novel exploratory analyses aimed at evaluating the test-retest reliability of both default rating proportion and several emerging alternative scoring approaches (including changes in minimum, maximum, and range of scores on individual dimensions) are reported. While the magnitude of change across mean dimensional scores was found to be lower compared to those previously found in outpatient populations, our results suggest that the SCORS-G remains capable of detecting changes in mentalization capacities in individuals contending with longstanding, severe psychiatric

impairment and personality-based psychopathology. Implications for treatment, assessment, and future research are discussed.

Keywords: performance-based assessment, TAT, SCORS-G, object relations

Key Practitioner Points:

- Complexity of representations of people (COM) changed over the course of long-term residential treatment and shifts in this dimension might be an initial indicator of response to treatment.
- When interpreting SCORS-G profiles, percentages of default ratings may be useful in determining whether changes are due to actual change in the target construct or simply due to changes in presence of thematic content available for scoring.
- Across both the traditional mean dimension ratings and emerging alternative scoring approaches, object relations as assessed by the SCORS-G show limited test-retest reliability within individuals characterized by high clinical severity and personality-based psychopathology.

Longitudinal Stability of SCORS-G Dimensional Ratings, Score Ranges, and Narrative "Blandness" in a Clinical Sample

Social cognitive capacities and object relations (i.e., internalized representations of significant others) are theorized to be relatively stable entities evolving from early life experiences with primary caregivers (e.g., Bowlby, 1989; Siefert & Porecelli, 2015). The supposition that schemas can be revised and reformulated in response to new relational experiences has served as the bedrock of major theories of development and therapeutic treatment (Lukowitsky & Pincus, 2011; Muran, 2002). The degree to which these schemas change versus remain the stable over the course of treatment remains an area in need of further study (Hopwood, Bleidorn, & Wright, 2021).

The Social Cognition and Object Relations Scale – Global Rating Method (SCORS-G; Stein & Slavin-Mulford, 2018; Westen, 1995) is a clinician-rated measure of object relations and

social reasoning. It has been applied to a range of narrative materials, including early memory protocols (Pinsker-Aspen, Stein, & Hilsenroth, 2007), psychotherapy transcripts (Mullin et al., 2017), the Picture Story Exercise (Galtieri et al., 2021), and Thematic Apperception Test stories (TAT [Murray, 1943]; see also Ridenour et al., 2021; Stein et al., 2014). Though the SCORS-G has been investigated as a predictor of symptom remission and behavioral change in response to treatment (see Stein & Slavin-Mulford, 2018 for extensive review), only a handful of studies have examined change in underlying object relations in and of itself over the course of treatment using the SCORS-G (Fowler et al., 2004; Josephs et al., 2004; Mullin & Hilsenroth, 2014; Mullin et al., 2017; Mullin et al., 2018; Porcerelli et al., 2006). The present study builds on prior work by examining test-retest reliability and mean-level change in SCORS-G dimensions in inpatients undergoing intensive treatment.

Change and Stability in Narrative Assessments of Personality Functioning

The Thematic Apperception Test (TAT; Morgan & Murray, 1935; Murray, 1943), an implicit performance-based measure, is frequently used in clinical settings to assess relational schemas and personality functioning (Mihura, Roy, & Graceffo, 2017). The SCORS-G is widely used to rate TAT narratives along eight dimensions, intended to respectively capture the complexity of an individual's representations of people (COM), understanding of social causality (SC), affective quality of representations (AFF), emotional investment in relationships (EIR), emotional investment in values and moral standards (EIM), experience and management of aggressive impulses (AGG), self-esteem (SE) and identity and coherence of self (ICS). Each

dimension is scored on 7-point scale with lower scores representing more pathological responses and higher scores indicating healthier responses. Several of the dimensions (e.g., AFF, EIR, EIM, AGG, SE, and ICS) also have default ratings if the content relevant to the dimension is absent from the narrative, and stories with multiple dimensions with default ratings can lead to narrative "blandness", which limits the overall interpretability of the SCORS-G dimensional profile (Stein et al., 2020).

Most contemporary research on the SCORS-G has utilized all eight dimensions to explore a range of clinical factors, including personality psychopathology, interpersonal functioning and behaviors, and changes in relational schemas over the course of psychotherapy (Stein & Slavin-Mulford, 2018). However, research into the factor structure of the SCORS-G has yielded evidence for both a two-factor (cognitive [COM & SC] and affective-relational [AFF, EIR, EIM, AGG, SE, & ICS], see Lewis et al., 2016) and three-factor model (cognitive [COM & SC], relational tone [EIM & AGG], and self-affective states [SE & ICS], see Stein et al., 2012). We followed the recommendations of Siefert and colleagues (2018) and interpret the SCORS-G dimensions by separating them into two domains: cognitive (COM & SC) and affective-relational (AFF, EIR, EIM, AGG, SE, and ICS).

The SCORS-G has been used to evaluate changes in object relations over the course of treatment, using psychotherapy transcripts (Mullin & Hilesnroth, 2014; Mullin et al., 2017; Mullin et al., 2018), interpersonal narratives from daily life (Clemence & Lewis, 2018), and clinical material gathered from individual case studies (Josephs et al., 2004; Lewis et al., 2021;

Porcerelli et al., 2007). All studies report evidence that object relations improve from treatment; however, specific findings vary. Mullin and colleagues (2018) evaluated changes in the relational schemas of 75 clinical outpatients receiving six-to-nine months of short-term psychodynamic psychotherapy, finding improvements across all eight dimensions with effect sizes ranging from medium (e.g., COM [$d = 0.73$]) to large (e.g., AFF [$d = 1.43$]).

Studies in more severely distressed populations (e.g., those characterized by higher rates of personality pathology) suggest more limited improvement over the course of treatment. Porcerelli and colleagues (2006), for example, utilized an earlier iteration of the SCORS system (Westen, 1995) to assess changes in a clinical sample enrolled in residential treatment over the course of 15 months, finding only small (e.g., Complexity [$d = 0.30$]) to medium (e.g., Affect-Tone [$d = 0.72$]) effects across dimensions. Similarly, Fowler and colleagues (2004) applied the SCORS-G to TAT narratives collected from a separate long-term residential sample, finding medium effect sizes for both cognitive dimensions [COM ($d = 0.46$) & SC ($d = 0.41$)] but less significant changes on the affective-relational dimensions (i.e., only SE showed a small effect size, $d = 0.34$) over the course of 16 months of treatment. Again, the breadth and magnitude of change in this study was more limited in comparison to findings with outpatients.

This pattern may suggest that object relations change more slowly in individuals with more severe psychopathology and that less mature object relations. This aligns with prior studies suggesting that less mature object relations predict poorer response to treatment (Høglend, 1993; Knekt et al., 2017). Single-case studies also lend support to this position. Porcerelli et al. (2007)

examined changes in relational functioning using the SCORS-G (applied to psychotherapy transcripts) in a five-year psychoanalysis of a man diagnosed with avoidant personality disorder, finding considerable gains with large effect sizes across all eight dimensions (e.g., COM [d = 3.00]). Similarly, Josephs and colleagues (2004) reported changes in the eight SCORS-G dimensions for a client with schizoid personality disorder following four years of treatment. Personality disordered clients in these single case designs ultimately achieve results that were similar to those reported for outpatients (e.g., Mullin et al., 2018), but longer periods of time were required. Notably, the SCORS-G dimensions that most consistently demonstrated change across studies regardless of clinical severity or lack of treatment were the "cognitive" dimensions of COM and SC, suggesting that engagement in therapeutic treatment may improve mentalization and social reasoning capacities.

Methodological Differences Across Studies

While clinical severity and degree of personality disturbance may contribute to differences in findings across studies, differences in study methods may also play a role. Studies applying the SCORS-G to narratives of daily life experiences or biographical memories have yielded larger effects than studies using standardized stimulus sets, such as the TAT. For example, Porcerelli et al. (2007) and Mullin and colleagues (2017) used the SCORS-G to rate session narratives and obtained larger effect sizes for change compared to studies that applied the SCORS-G to TAT stories (Fowler et al., 2004; Porcerelli et al., 2006). In comparison to session

narratives featuring spontaneous accounts of daily life situations, detecting changes on structured assessment tasks such as the TAT may be more challenging.

Several factors may contribute to this divergence, including greater task demands and restrictions on narrative content compared to spontaneous speech in a psychotherapy session, interference in the production of new stories based on respondent recollection of previous narratives, or the less interactive and familiar nature of structured assessment tasks in comparison to exchanges occurring in the context of treatment with a known, trusted provider (Slavin-Mulford, Amerson, Hilsenroth, Zodan, Charnas, Cain, & Stein, 2021; Slavin-Mulford, Amerson, Cain, Hilsenroth, Wilcox, & Stein, 2021).

Jenkins (2017a, 2017b) discussed the complexity of measuring reliability and stability for the TAT arguing that certain constructs such as personality style, cognitive capacities, and the cognitive aspects of object relations (see also Hibbard et al., 2001) are more trait-like in terms of their stability over time and across circumstances, while affective-relational dimensions are more likely to show greater variability in response to context. The cognitive dimensions of the SCORS-G evaluate an individual's capacity to define, integrate, and differentiate between the mental states of characters in their narratives (COM) and to logically articulate the flow of social interactions and outcomes (SC). Prior studies have shown that these dimensional scores are rated more consistently across TAT cards, reflecting systematic and enduring characterological approaches to organizing interpersonal content (Ridenour et al., 2021). In contrast, the affective-relational dimensions (AFF, EIR, EIM, AGG, SE, & ICS) have been shown to be influenced by

both characterological tendencies (i.e., ingrained expectations for interpersonal outcomes) as well as the stimulus properties of the TAT cards, with the latter reflecting patterns of responsiveness to the test stimuli itself (Stein et al., 2014; Siefert et al., 2016).

The Current Study

While the SCORS-G has received increasing empirical attention over the past few decades, surprisingly few studies examine the relative degree of change versus stability in its eight dimensions over time. Among studies examining change over the course of treatment, improvements in object relations reflected by the cognitive dimensions of the SCORS-G (i.e., COM and SC) have been consistently found regardless of clinical severity or treatment context. The detection of change across the affective-relational dimensions, however, has varied across study populations (e.g., inpatient/residential and outpatients) and methods (e.g., rating therapy transcripts vs. responses to TAT cards). Furthermore, the most recent study of SCORS-G changes utilizing TAT narratives (excluding the case study of Lewis and colleagues [Lewis et al., 2021]) was conducted 15 years ago. Since then, alternative approaches to scoring the SCORS-G dimensions (e.g., range scores; Clemence & Lewis, 2018) and extra-dimensional factors such as degree of ratable content versus "blandness" within a story narrative (Stein et al., 2020) have been identified as important factors predicting clinical outcomes. The test-retest reliability of these factors over time remains unknown. In addition, few studies have evaluated test-retest reliability across the eight dimensions when applied specifically to TAT narratives. Prior studies suggest the need to consider method variance as an important factor influencing the magnitude of

change in relational schemas over time. The TAT, as a standard stimulus set that requires respondents to draw primarily on internalized relational schemas rather than recent events or memories (which may be more subject to selection bias), remains an invaluable assessment method for evaluating underlying changes in personality dynamics and structure.

The primary purpose of the present study was to evaluate the test-retest reliability and mean-level change across the eight SCORS-G dimensions derived from TAT narratives collected over an 18-month period in a clinical sample marked by complex psychiatric problems enrolled in long-term residential treatment. Given the overall clinical severity of our sample, we hypothesized that participants would show greater consistency across dimensions compared to clinical samples with lower severity of psychopathology (e.g., Knekt et al., 2017; Rush et al., 2006). Moreover, a recent study on patients at this long-term residential facility (Perry & Fowler, 2021) found that recovery can be a long and arduous process that takes years if not decades. As a result, we would expect that changes in mean dimensional scores would be modest over the course of an 18-month period.

We hypothesized that the cognitive dimensions of COM and SC would be the most likely to evidence changes in mean scores over time, based on past research showing that ratings on these dimensions tend to be more consistent across cards and reflective of the mentalization capacities that are typically a focus in psychotherapy treatment (Siefert et al., 2016, Stein et al., 2014; Ridenour et al., 2021). An essential element in most therapies focuses on the patient's ability to take a step back and observe themselves and other people whether it be in the service of

insight, emotion regulation, and/or behavior change (Linehan, 2015). This generally precedes any other type of skill usage (i.e., if one cannot take a step back and observe their and/or other's experiences, then insight, emotion regulation, and behavior change is less likely).

In contrast, we did not anticipate changes in mean scores across ratings for the six affective-relational dimensions. These dimensions are influenced to a greater degree by proximate factors, such as card content, that activate more deeply entrenched interpersonal beliefs and expectations. We also predicted that participants would generally remain consistent across all dimensions in terms of their relative degree of pathological versus adaptive responses – in other words, mean dimension scores on a given dimension would be significantly positively correlated across the two assessment periods, suggesting adequate test-retest reliability. Given the lack of prior studies evaluating the reliability of newer rating indices derived from the SCORS-G system, we considered our analyses of changes in the minimum, maximum, range, and percentage of default scores to be exploratory.

Methods

Participants

The study employed a medical record review to identify participants. Permission to access participant medical records to obtain demographic and psychological testing data was granted by the treatment center's Institutional Review Board. A review of records for adult psychiatric patients enrolled in long-term residential treatment in the northeastern part of the United States between 2015 and 2020 revealed 61 who completed at least two TAT protocols

during their stay. The final sample consisted of 37 female patients (*M* age = 29.73, *SD* = 10.21), 20 male patients (*M* age = 31.65, *SD* = 11.08), and four transgender and gender nonconforming patients (*M* age = 21.75, *SD* = 2.22). The sample identified predominantly as European American (96.7%) and single (88.5%). The most common principal *DSM-5* diagnoses were unspecified/other specified personality disorder (26.2%), depressive disorder (23.0%), borderline personality disorder (19.7%), schizophrenia spectrum and other psychotic disorders (8.2%), bipolar disorder (6.6%), and posttraumatic stress disorder (6.6%). On average, each participant was diagnosed with 3.9 *DSM-5* diagnoses (*SD* = 1.3), with 85% percent of participants having at least one personality disorder diagnosis. Clinical diagnoses were assigned to participants by their psychiatrist and/or psychologist psychotherapist during the first five weeks of residential treatment using the Longitudinal, Expert, All Data (LEAD) diagnostic standard (Pilkonis, Heape, Ruddy, & Serrao, 1991).

Procedures

All patients were administered a battery of psychological tests (including the TAT) by doctoral-level psychologists within the first five weeks of treatment as a part of routine clinical practice. The TAT was administered in accordance with the procedures and guidelines outlined by Murray (1943). Five cards were selected for analysis from a standard nine card protocol administered to all patients: 1, 14, 13MF, 12M, and 2. These cards were selected both because of their prevalence of use in clinical settings, as well as their relative balance of "pull" for positive (Card 2), negative (Card 13MF), and neutral (Cards 1, 12M, and 14) narrative themes according

to recent card pull research within this population (Ridenour et al., 2021). All TAT narratives were recorded and transcribed verbatim and rated according to the SCORS-G training manual (Stein & Slavin-Mulford, 2018). Overall protocol word count at each time point and length of time between assessment periods (in days) were documented for each participant.

Measures

Social Cognition and Object Relations Scale-Global rating method (SCORS-G; Stein & Slavin-Mulford, 2018; Westen, 1995). The SCORS-G is a rating system applied to narrative material that assesses facets of object relations and social cognition. The system is comprised of eight dimensions scored on a 7-point rating scale. Lower scores are assigned in instances of more pathological responses, while higher scores are given to more adaptive responses. Multiple dimensions have a default score (AFF [4], EIR [2], EIM [4], AGG [4], SE [4], and ICS [5]), which may be given if the specific narrative does not contain content relevant to that dimension. While EIR also has a formal default rating of '2' if only one character is depicted (Stein & Slavin-Mulford, 2018), this code is more often given on cards that only feature a single character (e.g., Cards 1 and 14). In a previous study on "blandness", Stein and colleagues (2020) considered an EIR rating of '2' bland for single character cards and a rating of '2' and '3' bland for all multiple character cards administered. To avoid potential confusion between default ratings associated with respondent-generated interpersonal content (e.g., how relationships were described regardless of stimulus properties) versus those ratings associated more with stimulus properties (e.g., single character cards), we elected not to include EIR in our default coding

computations. Furthermore, distinct from the previous study on blandness and default ratings (Stein et al., 2020), we did not set a predetermined threshold for determining a bland narrative (i.e., if all relevant dimensions were given a default rating than a TAT story was determined to be "bland"). Instead, we computed the percentage of default ratings for all five dimensions (i.e., AFF, EIM, AGG, SE, and ICS) to determine changes over time and across dimension. This atheoretical, dimensional approach enabled us to evaluate changes in ratable content over time without being bound to a particular a priori threshold determined by the investigators. For the present study, two psychologists with extensive experience with the SCORS-G co-rated all subject TAT narratives. Raters were blind to all identifying information as well as the administration time point for each protocol during the coding process. Consensus meetings were used throughout the rating process to review scoring discrepancies of ≥2 points on any dimension.

Mean ratings for each of the eight dimensions were calculated by averaging across raters and individual card scores on each dimension. For minimum and maximum scores, the lowest and highest individual card ratings (averaged across raters) were selected for each dimension. Range scores were calculated by subtracting the minimum score from the maximum score on each dimension. Finally, using the criteria defined by Stein and colleagues (Stein et al., 2020), bland percentage scores were calculated for each dimension by dividing the number of actual default scores assigned across cards and dimensions by the number of potential default scores possible (e.g., five cards per protocol x five dimensions including a default code option = 25).

Data analytic plan

Overall shifts in SCORS-G dimensional ratings between time points were evaluated using paired samples *t*-tests, with Cohen's *d* reported for effect size (SPSS.23). Pearson correlation coefficients were calculated to evaluate whether the relative ranking of SCORS-G ratings remained consistent across the two assessment periods (e.g., whether high scores at T1 remained high at T2; Cicchetti, 1994; Kazdin, 1982). Correlations were used to assess test-retest reliability while *t* tests were employed to assess sample level mean changes across SCORS-G dimensions over time.

Results

Intraclass correlation coefficients using a two-way random effects model with absolute agreement ranged from good to excellent for all SCORS-G dimensions (reliability coefficients and overall sample dimension means and standard deviations for both time points are reported in Table 1). Associations between T1 and T2 mean dimension ratings for the eight SCORS-G dimensions are reported in Table 2. All eight SCORS-G dimension means were significantly correlated, with effect sizes ranging from small (AFF $r = .30$) to medium (SE $r = .52$). While the cognitive dimensions (COM and SC) showed similar degrees of associations between T1 and T2 in terms of effect sizes (falling in the medium range), the magnitude of associations for the six affective-relational dimensions (i.e., AFF, EIM, EIR, AGG, SE, ICS) displayed less consistency over time.

Significant changes in dimension means between the two time points were examined using paired samples *t* tests. COM was found to increase significantly over time ($t[60] = -2.55$, $p = .01$, $d = .32$), suggesting an overall greater degree of complexity of mental state representations between the two assessment periods. In contrast, EIM was found to decrease significantly between T1 and T2 ($t[60] = 2.46$, $p = .02$, $d = .31$), indicating that participants included greater narrative content that reflected less investment in moral values and standards.

Changes in prevalence of default dimension ratings.

As with the SCORS-G dimension ratings, participants' tendency to produce default responses to TAT card stimuli at T1 was correlated with their likelihood of doing so at T2 ($r = .44$, $p = <.01$). Mean levels of default rating responses decreased from T1 to T2 ($t[60] = 2.57$, $p = .01$, $d = .33$), suggesting that participants provided a greater amount of scorable content in their narratives over time. In addition, we also calculated the percentage of default ratings for the five dimensions over both time points: AFF T1 (24.91%) and AFF T2 (14.43%), EIM T1 (70.49%) and EIM T2 (67.54%), AGG T1 (74.10%) and AGG T2 (69.83%), SE T1 (64.26%) and SE T2 (61.64%), and ICS T1 (36.07%) and ICS T2 (27.54%). There was a decrease of percentage of default ratings across all five dimensions, especially for AFF and ICS (indicating increases in narrative richness).

Reliability of alternative dimensional scoring approaches.

In addition to the mean SCORS-G dimension ratings across time points, we also evaluated patterns of associations and changes in the lowest (minimum) score that a participant

received on a given dimension within their T1 versus T2 protocol, their highest (maximum) score on a given dimension, and the overall dimension range score (calculated by subtracting their minimum from their maximum score on a given dimension). Across time points, COM was the only dimension to show significant change in minimum score (t[60] = -2.69, p = .01, d = .34). In contrast, both EIM and ICS showed a significant reduction in maximum score at T2 compared to T1, indicating a diminution of adaptive content in responses pertaining to moral dilemmas (t[60] = 2.65, p = .01, d = .34) or identity-related concerns (t[60] = 2.31, p = .03, d = .30). Finally, the relative range of protocol scores for participants across dimensions showed no significant mean-level changes between time points. Minimum scores across all eight dimensions were significantly and positively correlated across time points, while correlations between time points for maximum and range scores showed greater variability, suggesting greater consistency in participants' least adaptive response style compared to their more adaptive capacities or relative range of response styles, though all coefficients fell below acceptable standards for demonstrating adequate test-retest reliability (Nunnally, 1978). Pearson correlation coefficients for each of these alternative scoring methods are shown in Table 2.[1]

[1] We additionally ran a series of partial correlations between alternative scoring dimensions for the SCORS-G, controlling for average protocol word count at T1 and T2 and for length of time (in days) between assessment periods. Patterns of significance across associations for the dimensional mean scores remained constant after controlling for these factors. For the alternative scoring approach indices, we found altered associations on three subscales: COM maximum scores were reduced to trend significance after controlling for T1 word count (r = .24, p = .06) and were no longer significantly associated after controlling for T2 word count (r = .16, p = .22). SC range scores were reduced to trend significance (r = .22, p = .09) when controlling for T1 word count, and lost significance (r = .21, p = .11) when controlling for T2 word count. Finally, ICS minimum scores were reduced to trend significance after controlling for T2 word count (r = .25, p = .053), and ICS range scores associations become non-significant after controlling for T2 word count (r = .18, p = .18).

Discussion

Though research with the SCORS-G has increased notably in the past two decades, there are a dearth of studies examining change in SCORS-G dimensions as a function of treatment. In both long-term residential and outpatient populations changes in the cognitive aspects of object relations have been observed across studies and regardless of differences in method. However, findings are mixed for affective-relational dimensions. In general, long-term residential patients assessed with the TAT evidence smaller changes on average (Porcerelli et al., 2006) and across fewer dimensions (Fowler et al., 2004) relative to outpatients (assessed using psychotherapy transcripts; Josephs, 2004; Mullin et al., 2017). The present study built on prior work with long-term residential patients by utilizing more recently developed methods for scoring TAT stories with the SCORS-G. Additionally, a retrospective chart review was used (rather than an opt-in approach) to rule out the possibility of self-selection bias. This is also the first study of its kind to assess changes in "blandness" over time.

Based on prior work, we anticipated that cognitive dimensions (i.e., COM and SC; Lewis et al., 2016; Stein et al., 2012) would change as a function of long-term residential treatment. While mean ratings on the COM dimension increased significantly but modestly (d = .32*)* over the 18-months of residential treatment, significant differences for SC were not observed. Perhaps changes in COM (i.e., the ability to define and differentiate the mental states of characters) might precede shifts in SC (i.e., the capacity to logically sequence and explain the underlying causes of social interactions). In other words, before someone can realistically interpret social situations

and predict interpersonal outcomes, they first might need to develop the capacity to accurately identify and reflect upon the psychological states of self and other (i.e., mentalization) in a realistic and nuanced manner.

Second, we hypothesized that the affective-relational SCORS-G dimensions would be fairly stable over time, given their tendency to be more influenced by the stimulus properties of the cards as well as their reflecting expectations for relational outcomes that may be more intransigent (Ridenour et al., 2021; Stein et al., 2014). Our findings supported this hypothesis for all dimensions except for EIM. EIM showed a significant decrease in mean score between assessment periods (we discuss this finding in detail below). Our hypothesis that the SCORS-G dimensional ratings would remain positively associated across time points was supported by our findings. Nonetheless, the average test-retest correlation across all eight SCORS-G dimensional means (r = 0.43) fell below acceptable range according to the criteria set by Nunnally (1978), while remaining comparable to retest coefficients previously reported for other performance-based narrative assessments (excepting the Rorschach; see Meyer, 1997; Sultan et al., 2006) and behavioral assessments of socio-emotional competencies (e.g., Boon-Falleur et al., 2022; McAdams et al., 2006). In comparison to self-report measures, behavioral assessments of personality tend to show lower consistency across time points, which has been attributed both to the tendency for greater variability in behavioral performance across task administrations compared to self-appraisals of ability and competency (Boon-Falleur et al., 2022), as well as the influence of variance attributable to low between-subject/high within-subject differences in task

performance over time (Enkavi et al., 2019). Additionally, the significant span of time between assessment periods (on average 18 months) likely contributed to lower test-retest reliability. This decrease is consistent with a previous study by McAdams and colleagues (2006) that found test-retest reliability for a narrative measure of life stories diminished over time. While these factors may provide rationale for the low test-retest reliability found in the present study, further studies are needed to better understand the factors affecting low consistency over time in TAT-rated SCORS-G dimensions. The present findings may suggest that longitudinal studies or assessments that include the SCORS-G may be better served by incorporating more frequent sampling schedules, and that SCORS-G ratings should be used as one of several multimethod indicators of interpersonal and personality functioning in the prediction of long-term clinical outcomes (Enkavi et al., 2019).

In addition to investigating the mean dimensional SCORS-G ratings (which have historically been used in research and clinical practice contexts), we also evaluated newer scoring approaches, including changes in minimum, maximum, range, and default rating scores (Stein et al., 2020). As with the main dimensional scores, default score proportions showed test-retest reliability falling below acceptable levels (Nunnally, 1978); further, overall levels of default score proportions decreased across time points, suggesting that patients produced more in-depth stories containing more scorable content at the end of treatment. Further, blandness proved quite salient for contextualizing results. For example, the decrease in EIM between time points was unexpected. These changes are likely related to the overall decrease in EIM default

ratings. When there is no codable content for EIM, a default score of four is given. The majority of the EIM ratings across all narratives in the present study were default ratings (69.02%), which suggests that the mean ratings were heavily influenced by default coding. For some participants, default coding may falsely elevate EIM scores (i.e., default scores of four may "pull up" the average). Thus, T1 and T2 differences may emerge as scorable content increases. More research is needed to assess how elevated proportions of default ratings influence the overall interpretability of the eight SCORS-G dimensions. In addition to the dimensional mean rating, both the minimum and maximum EIM ratings also showed a decrease between time points (significant only for maximum score), suggesting that changes in mean scores were driven by the inclusion of greater scorable content.

For the minimum score, only the COM dimension showed significant changes, as participants produced more pathological responses at T1 (i.e., a tendency towards concrete thinking and egocentrism) relative to T2. There was also a decrease in maximum scores for EIM, discussed above, and ICS across time points. The decrease in ICS maximum score is especially interesting because it indicates less adaptive content (i.e., scores of 6-7 are given if a character is described as having realistic goals and aspirations) over the 18-month interval. One possibility is that higher maximum scores on this dimension at T1 might indicate a defensive elevation, given that participants completed this assessment at the beginning of residential treatment, which they presumably sought due to feeling lost, struggling with their sense of identity, and grappling with severe symptoms such as suicide, psychosis, and mania. Again, significant decreases (8.55%) in

ICS default rating scores indicate that participants told more evocative and content-rich stories at T2 that touched upon themes relevant to identity, increasing the potential for issues of conflict and ambivalence to emerge. This decreased blandness might suggest that individuals were better able to imaginatively engage the card, identify with the characters, and describe more intricate plots that revealed their difficulties with identity.

Furthermore, it was notable that the greatest decrease in default rating scores across the five dimensions was on AFF (10.48%) across time points. This decrease indicates that the respondents were imbuing their story with more emotionally laden interpersonal themes at T2 (codes for AFF), which were likely negative given the dysphoric pull of the TAT cards. Of note, practice effects might have also partially contributed to the decrease in default ratings across the various dimensions. In other words, when the subject completed the TAT for the second administration, the previous exposure of the stimuli and the familiarity with the instructions, might have allowed them to relax and engage more imaginatively with the task. This said, it is important to note that administration was spaced considerably between T1 and T2, which one would expect to mitigate familiarity and practice effects to at least a degree.

Consistent with past studies that have used the TAT to evaluate stability versus change in SCORS-G dimensions (Fowler et al., 2004; Porcerelli et al., 2006), this study also found less robust changes on the SCORS-G compared to earlier studies that have used psychotherapy transcripts (Josephs, 2004; Mullin et al., 2017). The smaller effect sizes obtained in the current study could be driven by multiple factors. First, the sample in the present study (as in prior

studies using the TAT) included patients in long-term residential treatment who presumably were experiencing more complex, enduring psychopathology (i.e., high rates of personality disorder) than those in an outpatient setting, and thus showing a slower response rate to treatment. Second, TAT narratives and psychotherapy transcripts are distinct in nature and likely illustrate different psychological processes. For instance, over the course of a successful psychotherapy, a patient develops a trusting relationship with their therapist and a greater capacity to understand self and other (COM), invest more in their relationships (EIR), and a better ability to logically understand social interactions (SC). These abilities might be more readily detectable in a psychotherapy transcript when anxiety is lower and trust is higher, allowing the patient to describe examples from daily life that reflect these higher-order capacities. In contrast, TAT stimuli feature unfamiliar, ambiguous, and primarily dysphoric interpersonal scenes, placing greater restrictions on narrative content and affording less of an opportunity to establish comfort and familiarity with the examiner (compared to a therapist). The TAT further often generates anxiety, and respondents can react with various defenses (Cramer, 2017) that prevent them from showcasing their more adaptive capacities. Due to these factors, we suspect that changes on the TAT are likely to lag behind development and growth that might be more clearly evident in a psychotherapy transcript, consistent with the phase model of change that posits that personality and interpersonal functioning (capacities targeted by implicit measures of personality like the TAT and Rorschach) tend to change less quickly than subjective feelings of well-being and overt symptomatology (Fowler et al., 2004).

Limitations

There were four major limitations for this study. First, the sample was racially homogenous (96.7% European American) and characterized by overall high socioeconomic status. While this sample was representative of the residential treatment center, it is not representative of broader populations of persons experiencing mental health difficulties. Second, the study did not evaluate treatment progress or other indicators of clinical change that may have been relevant for contextualizing our findings. For example, we might anticipate that participants who made greater clinical gains in treatment would demonstrate correspondent improvement in their object representations assessed through the SCORS-G, though this may vary across cognitive versus affective-relational dimensions; future studies should include multimethod assessment of interpersonal and clinical functioning to evaluate these possibilities and deepen understanding of the relationship between SCORS-G dimensions and change in response to specific psychotherapy treatments (e.g., Lewis et al., 2021). Third, while the five cards included in our protocol are commonly used in both empirical research and clinical assessment, findings of stability and change across the SCORS-G dimensions may vary according to the specific cards selected for protocol inclusion. While the development of a universal standard card set may aid in generalizing results across studies (e.g., Ridenour et al., 2021), the current findings only apply to the specific cards used in our study. Finally, we would have ideally used the SCORS-G to independently code both TAT responses and psychotherapy transcripts for each patient. Because psychotherapy transcripts were unavailable for this sample, this was impossible in the present

study. We would recommend future investigators researching this topic to collect narratives using both TAT and psychotherapy transcripts. This would assist in assessing alternative explanations (i.e., to what degree are differences between studies due to population and/or method variance).

Conclusion and Future Directions

Our findings provide important information regarding the reliability and stability of the SCORS-G dimensions in clinical populations, enabling clinicians who utilize the system in their standard assessment battery to estimate clinical significance of change over time. While the eight SCORS-G dimensions showed limited general consistency over time, our findings suggest that the dimensions of COM may be most likely to evidence mean level change (albeit modest) over the course of treatment, potentially signaling improved mentalization capacities and reflectiveness. This greater degree of engagement with the task was further evidenced by a significant decrease in default ratings over time. Further, our findings provide novel insight into the reliability of alternative scoring metrics for the SCORS-G system, expanding the range of indices that may be useful for interpretation and pointing to important avenues for future empirical study. Particularly when used in conjunction with idiographic interpretive approaches to TAT narratives that highlight more personalized trajectories of change at the individual level, this improved understanding of longitudinal fluctuations in the SCORS-G dimensions supports greater precision and clarity in clinical interpretation.

Finally, this study is among the first to highlight the importance of considering how percentage of default ratings contributes to story blandness when interpreting therapeutic change. Our initial findings regarding EIM for example were counterintuitive; however, examination of the impact of default ratings provided a reasonable explanation for this unexpected finding. Examining how percentage of default ratings per protocol impacts clinical findings is a developing area of study in SCORS-G research. To continue this momentum, we recommend future researchers using the SCORS-G to assess and report default rating percentages for their samples. Lastly, while this study focuses on the percentage of default ratings per protocol and how this contributes to narrative "blandness," we encourage researchers and clinicians to think about other factors that also contribute to this clinical phenomenon.

References

Boon-Falleur, M., Bouguen, A., Charpentier, A., Algan, Y., Huillery, É., & Chevallier, C. (2022). Simple questionnaires outperform behavioral tasks to measure socio-emotional skills in students. *Scientific Reports*, *12*(1), 1-11.

Bowlby, J. (1989). The role of attachment in personality development and psychopathology. In S. I. Greenspan & G. H. Pollock (Eds.), *The course of life, Vol. 1. Infancy* (pp. 229–270). International Universities Press, Inc. (Reprinted from *American Journal of Psychiatry*, 1987, Vol. 144; and from *American Journal of Orthopsychiatry*, 1982, Vol. 52)

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284.

Clemence, A. J. & Lewis, K. C. (2018). Flexibility and rigidity in object relational functioning: Assessing Change in suicidal ideation and global psychiatric functioning using the SCORS-G. *Journal of Personality Assessment, 100*(2), 135-144. https://doi.org/10.1080/00223891.2017.1418747

Cramer, P. (2017). Defense mechanism card pull in TAT stories. *Journal of Personality Assessment, 99*(1), 15-24.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472-5477.

Fowler, J.C., Ackerman, S. J., Speanburg, S., Bailey, A., Blagys, M., & Conklin, A. C. (2004). Personality and symptom change in treatment-refractory inpatients: Evaluation of the phase model of change of using Rorschach, TAT, and DSM-IV Axis V. *Journal of Personality Assessment, 83*(3), 306-322.

Galtieri, L. R., Cobb, H. R., O'Gorman, E. T., & Kurtz, J. E. (2021). Object relations functioning in a community adult sample: Further normative data for the SCORS-G. *Psychoanalytic Psychology*. Advance online publication. https://doi.org/10.1037/pap0000380

Grønnerød, C. (2003). Temporal stability in the Rorschach method: A meta-analytic review. *Journal of Personality Assessment*, *80*(3), 272-293.

Hibbard, S., Mitchell, D., & Porcerelli, J. (2001). Internal consistency of the object relations and social cognition scales for the thematic apperception test. *Journal of Personality Assessment, 77*(3), 408-419. https://doi.org/10.1207/S15327752JPA7703_03

Høglend, P. (1993). Personality disorders and long-term outcome after brief dynamic psychotherapy. *Journal of Personality Disorders*, *7*(2), 168-181.

Hopwood, C. J., Bleidorn, W., & Wright, A. G. C. (2021). Connecting theory to methods in longitudinal research. *Perspectives on Psychological Science*. https://doi.org/10.31234/osf.io/w5huz

Jenkins, S. R. (2017a). The narrative arc of TATs: Introduction to the JPA special section on thematic apperceptive techniques. *Journal of Personality Assessment, (99)*3, 225-237. https://doi.org/10.1080/00223891.2016.1244066

Jenkins, S. R. (2017b). Not your same old story: New rules for thematic apperceptive techniques (TATs). *Journal of Personality Assessment, (99)*3, 238-253. https://doi.org/10.1080/00223891.2016.1248972

Josephs, L., Anderson, E., Bernard, A., Fatzer, K., & Streich, J. (2004). Assessing progress in analysis interminable. *Journal of the American Psychoanalytic Association, 52*(4), 1185-1214. https://doi.org/10.1177/00030651040520041301

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Knekt, P., Lindfors, O., Keinänen, M., Heinonen, E., Virtala, E., & Härkänen, T. (2017). The prediction of the level of personality organization on reduction of psychiatric symptoms and improvement of work ability in short-versus long-term psychotherapies during a 5-year follow-up. *Psychology and Psychotherapy: Theory, Research and Practice*, *90*(3), 353-376.

Lewis, K. C., Meehan, K. B., Cain, N. M., Wong, P. S., Clemence, A. J., Stevens, J., & Tillman, J. G. (2016). Impairments in object relations and chronicity of suicidal behavior in individuals with borderline personality disorder. *Journal of Personality Disorders, 30*(1), 19-34. https://doi.org/10.1521/pedi_2015_29_178

Lewis, K. C., Ridenour, J. M., Pitman, S., & Roche, M. (2021). Evaluating stable and situational expressions of passive-aggressive personality disorder: A multimethod experience sampling case study. *Journal of Personality Assessment, 103*(4), 558-570. https://doi.org/10.1080/00223891.2020.1818572

Linehan, M.M. (2015). *DBT skills training manual* (2nd ed.). New York: Guilford Press.

Lukowitsky, M. R. & Pincus, A. L. (2011). The pantheoretical nature of mental representation and their ability to predict interpersonal adjustment in a nonclinical sample. *Psychoanalytic Psychology, 28*(1), 48-74. https://doi.org/10.1037/a0020849

McAdams, D. P., Bauer, J. J., Sakaeda, A. R., Anyidoho, N. A., Machado, M. A., Magrino-

Failla, K., ... & Pals, J. L. (2006). Continuity and change in the life story: A longitudinal study of autobiographical memories in emerging adulthood. *Journal of Personality*, *74*(5), 1371-1400.

Meyer, G. J. (1997). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment*, *9*(4), 480-489.

Mihura, J. L., Roy, M., & Graceffo, R. A. (2017). Psychological assessment training in clinical psychology doctoral programs. *Journal of Personality Assessment, 99*(2), 153–164. https://doi.org/10.1080/00223891.2016.1201978

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: the thematic apperception test. *Archives of Neurology & Psychiatry, 34,* 289-306. https://doi.org/10.1001/archneurpsyc.1935.02250200049005

Mullin, A. S. J. & Hilsenroth, M. J. (2014). Relationship between patient pre-treatment object relations functioning and psychodynamic techniques early in treatment. *Clinical Psychology and Psychotherapy, 21*(2), 123-131. https://doi.org/10.1002/cpp.1826

Mullin, A. S. J., Hilsenroth, M. J., Gold, J., & Farber, B. A. (2017). Changes in object relations over the course of psychodynamic psychotherapy. *Clinical Psychology and Psychotherapy, 24*(2), 501-511. https://doi.org/10.1002/cpp.2021

Mullin, A. S. J., Hilsenroth, M. J., Gold, J, & Farber, B. A. (2018). Facets of object representation: Process and outcome over the course of psychodynamic psychotherapy.

Journal of Personality Assessment, 100(2), 145-155. https://doi.org/10.1080/00223891.2016.1215320

Muran, J. C. (2002). A relational approach to understanding change: Plurality and contextualism in a psychotherapy research program. *Psychotherapy Research, 12*(2), 113–138. https://doi.org/10.1093/ptr/12.2.113

Murray, H. A. (1943). *Thematic Apperception Test.* Harvard University Press.

Nunnally, J. C. (1978). *Psychometric Theory*, New York: McGrawHill,Perry, J. C., & Fowler, J. C. (2021). A naturalistic study of time to recovery in adults with treatment-refractory disorders. *Psychiatry*, *84*(3), 260-275.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(1), 46.

Pinsker-Aspen, J. H., Stein, M. B., & Hilsenroth, M. J. (2007). Clinical utility of early memories as a predictor of early therapeutic alliance. *Psychotherapy: Theory, Research, Practice, Training, 44*(1), 96–109. https://doi.org/10.1037/0033-3204.44.1.96

Porcerelli, J. H., Shahar, G., Blatt, S. J., Ford, R. Q., Mezza, J. A., & Greenlee, L. M. (2006). Social cognition and object relations scale: Convergent validity and changes following intensive inpatient treatment. *Personality and Individual Differences, 41*(3), 407-417. https://doi.org/10.1016/j.paid.2005.10.027

Porcerelli, J. H., Dauphin, V. B., Ablon, J. S., Leitman, S., & Bambery, M. (2007). Psychoanalysis with avoidant personality disorder: A systematic case study. *Psychotherapy: Theory, Research, Practice, Training, 44*(1), 1-13. https://doi.org/10.1037/0033-3204.44.1.1

Ridenour, J. M., Lewis, K. C., Siefert, C. J., Pitman, S. R., Knauss, D., & Stein, M. B. (2021). Card pull effects of the Thematic Apperception Test using the Social Cognition and Object Relations-Global rating method on complex psychiatric sample. *Clinical Psychology and Psychotherapy.* https://doi.org/10.1002/cpp.2554

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., ... & Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *American Journal of Psychiatry, 163*(11), 1905-1917.

Siefert, C., & Porcerelli, J. H. (2015). Object relations theory and personality disorders: Internal representations and defense mechanisms. In S. K. Huprich (Ed.), *Personality disorders: Toward theoretical and empirical integration in diagnosis and assessment* (pp. 203–224). American Psychological Association. https://doi.org/10.1037/14549-009

Siefert, C. J., Stein, M., Slavin-Mulford, J., Haggerty, G., Sinclair, S. J., Funke, D., & Blais, M. A. (2018). Exploring the factor structure of the Social Cognition and Object Relations–Global Rating Method: Support for two-and three-factor models. *Journal of Personality Assessment*, *100*(2), 122-134.

Siefert, C. J., Stein, M. B, Slavin-Mulford, J., Sinclair, S. J., Haggerty, G., & Blais, M. A. (2016). Estimating the effects of Thematic Apperception Test card content on SCORS–G ratings: Replication with a nonclinical sample. *Journal of Personality Assessment, 98*(6), 598-607.

Slavin-Mulford, J. M., Amerson, L. R., Cain, L. A., Hilsenroth, M. J., Wilcox, M. M., & Stein, M. B (2021). How narrative source impacts convergence of ratings from the Social Cognition and Object Relations Scale–Global Rating Method with psychotherapy process measures. *Clinical Psychology & Psychotherapy*. https://doi.org/10.1002/cpp.2595

Slavin-Mulford, J. M., Amerson, L. R., Hilsenroth, M. J., Zodan, J., Charnas, J. W., Cain, L. A., & Stein, M. B (2021). Are all narratives the same: Convergent and discriminant validity of the Social Cognition and Object Relations Scale—Global Rating Method across two narrative types. *Clinical Psychology & Psychotherapy, 28*(3), 623-632.

Stein, M. B, Calderon, S., Ruchensky, J., Massey, C., Slavin-Mulford, J., Chung, W. J., Richardson, L. A., & Blias, M. A. (2020). When's a story a story? Determining interpretability of social cognition and object relations scale-global ratings on thematic apperception test narratives. *Clinical Psychology and Psychotherapy, 27*(4), 567-580. https://doi.org/10.1002/cpp.2442

Stein, M., & Slavin-Mulford, J. (2018). *The Social Cognition and Object Relations Scale-Global Rating Method (SCORS-G): A comprehensive guide for clinicians and researchers*. New York, NY: Routledge.

Stein, M. B, Slavin-Mulford, J., Siefert, C. J., Sinclair, S. J., Renna, M., Malone, J., Bello, I. & Blais, M. A. (2014). SCORS-G stimulus characteristics of select thematic apperception test cards. *Journal of Personality Assessment, 96*(3), 339-349. https://doi.org/10.1080/00223891.2013.823440

Stein, M. B, Slavin-Mulford, J., Sinclair, J. S., Siefert, C. J., & Blais, M. A. (2012). Exploring the construct validity of the social cognition and object relations scale in a clinical sample. *Journal of Personality Assessment, 94*(5), 533-540. https://doi.org/10.1080/00223891.2012.668594

Sultan, S., Andronikof, A., Réveillère, C., & Lemmel, G. (2006). A Rorschach stability study in a nonpatient adult sample. *Journal of Personality Assessment*, *87*(3), 330-348.

Westen, Drew (1995); *"Social Cognition and Object Relations Scale: Q-sort for Projective Stories (SCORS-Q)"*; Unpublished manuscript; Department of Psychiatry, The Cambridge Hospital and Harvard Medical School, Cambridge, MA.