Development and criterion validity of a computerized text analysis measure of reflective functioning

Eric A. Fertuck [a] [g] , Erhard Mergenthaler [b] , Mary Target [c] , Kenneth N. Levy [d] [f] & John F. Clarkin [e]

[a] Department of Psychology, City University of New York of CUNY, New York

[b] Clinic for Psychosomatic Medicine and Psychotherapy, University of Ulm, Germany

[c] Research Department of Clinical, Educational and Health Psychology, University College London, UK

[d] Department of Psychology, Penn State University, PA

[e] Psychiatry, Weill Cornell Medical College, NY, USA

[f] Weill Cornell Medical College, New York, USA

[g] Department of Psychiatry, Columbia University Medical College, New York, USA

Available online: 09 Feb 2012

PLEASE SCROLL DOWN FOR ARTICLE

Routledge
Taylor & Francis Group

Development and criterion validity of a computerized text analysis measure of reflective functioning

ERIC A. FERTUCK[1], ERHARD MERGENTHALER[2], MARY TARGET[3], KENNETH N. LEVY[4], & JOHN F. CLARKIN[5]

[1]Department of Psychology, City University of New York of CUNY, New York, and Department of Psychiatry, Columbia University Medical College, New York, USA; [2]Clinic for Psychosomatic Medicine and Psychotherapy, University of Ulm, Germany; [3]Research Department of Clinical, Educational and Health Psychology, University College London, UK; [4]Department of Psychology, Penn State University, PA, and Weill Cornell Medical College, New York, USA & [5]Psychiatry, Weill Cornell Medical College, NY, USA

Abstract
The Reflective Functioning scale (RF) is a narrative-based assessment of the capacity to coherently conceptualize one's own and others' subjective motivations, emotions, beliefs, and desires. We report the preliminary results of an effort to develop a computerized text analysis version (CRF) of the RF assessment system. A sample of 113 clinical and non-clinical Adult Attachment Interviews (AAI) were utilized to develop the CRF measure. Using the Marker Approach (Mergenthaler & Bucci, 1999), 54 linguistic markers of high RF language were identified. The associations between CRF and RF were significant in both a clinical sample of patients diagnosed with Borderline Personality Disorder (BPD) (Spearman rho = .57, $p < .0001$) and a non-clinical sample of adults (Spearman rho = .57, $p = .002$). These results suggest that a CRF rating scale is feasible, has preliminary criterion validity, and, therefore, has potential to facilitate the efficient assessment of RF.

Keywords: attachment; personality disorders; process research; psychoanalytic/psychodynamic therapy; test development; technology in psychotherapy research and training; language and psychotherapy; computer text analysis

The Reflective Function (RF) scale (Fonagy, Target, Steele, & Steele, 1998) is a narrative based assessment that measures an individual's capacity to "mentalize." RF is defined as the ability to coherently conceptualize in narrative form one's own and others' subjective motivations, emotions, beliefs, and desires, particularly as they are related to, and activated within, relationships with significant others. "RF is defined as the capacity to envision and think about mental states, in oneself and in others, in the service of building realistic models of why they behave, think, and feel as they do" (Bouchard et al., 2008, p. 47).

Controversies in mentalization theory and research have now advanced to questions of whether it is a unitary or multidimensional construct, the relative contributions of cognitive and emotional determinants, the transaction between a self and other focused appraisals, and the role of neurobiological systems (Gullestad & Wilberg, 2011; Jurist, 2005, 2010; Pronin, 2008; Semerari, Carcione, Dimaggio, Nicol,

& Procacci, 2007; Van Overwalle, 2009). Additionally, the concordance between RF and other indices of mentalization has been elaborated at both theoretical (Choi-Kain & Gunderson, 2008; Jurist, 2005, 2010; Semerari et al., 2007) and empirical levels (Bouchard et al., 2008; Gullestad & Wilberg, 2011; Lecours & Bouchard, 2011; Vermote et al., 2009). Among the several emerging measures of mentalization, RF is unique in that it is rooted in a framework of developmental, attachment, and psychoanalytic theories, is assessed from linguistic narratives, and focuses on the coherence of verbalizations about self in relation to significant others. RF is scored on a one-dimensional scale, but can be dissociated into sub-dimensions (Gullestad & Wilberg, 2011).

Psychotherapy and reflective function. RF is a psychological capacity at the core of many psychotherapies and forms of psychopathology. Certain

forms of psychotherapy have the potential to foster mentalization among patients as part of their mechanisms of therapeutic change. The patient-therapist dyad provides a rich interpersonal milieu in which therapists can encourage and promote patients' RF capacities. Assessing constructs such as RF from the linguistic discourse within this relationship can provide an index of improvements in mentalizing capacities. Several methods in psychotherapy research have begun to elucidate the role of RF and mentalization more broadly in the mechanism of therapeutic change. This research has begun to identify the techniques that foster mentalization among patients, and link these processes to improvements in behavior, symptoms, and emotional resiliency (Bateman & Fonagy, 1999; Dimaggio et al., 2005, 2007, 2009; Fertuck, Bucci, Blatt, & Ford, 2004; Gullestad & Wilberg, 2011; Karlsson & Kermott, 2006; Levy, Beeney, Wasserman, & Clarkin, 2010; Levy, Meehan, et al., 2006; Lysaker, Buck, & Ringer, 2007; Semerari et al., 2003). Specific to the RF scale, psychotherapy process research has begun to explore the complex relationships between RF, forms of psychopathology, symptom change, and psychotherapeutic processes (Gullestad & Wilberg, 2011; Lysaker et al., 2007). Particularly compelling support for the importance of RF in psychotherapy outcome has been reported from a clinical trial of three forms of manualized psychotherapy in the treatment of Borderline Personality Disorder (BPD) (Levy et al., 2006). One form of psychoanalytic psychotherapy, Transference Focused Psychotherapy (TFP), resulted in significantly greater increases in RF over one year of treatment compared to supportive psychodynamic therapy and Dialectical Behavior Therapy. The TFP therapist helps patients to appreciate and verbalize emotionally imbued mental images of self and other, which may uniquely promote RF compared to other forms of therapy.

Rationale for the development of additional RF instrument. Currently, RF is assessed by extensively trained raters who code the narratives elicited during the Adult Attachment Interview (AAI) (Main & Goldwyn, 1991). The established RF rating system has been shown to be reliable and to possess clinical and research utility (cf., Bouchard et al., 2008; Fischer-Kern et al., 2010; Fonagy, Steele, Steele, Moran, & Higgitt, 1991; Grienenberger, Kelly, & Slade, 2005; Levy, Clarkin, et al., 2006; Slade, Grienenberger, Bernbach, Levy, & Locker, 2005). However, there are aspects of the current RF scoring method that justify the development of an alternate assessment system. (1) RF scoring requires trained judges who must undergo time-intensive, supervised reliability training. Consequently, the measurement of RF is an expensive and time-consuming process. There is a need for a less labor intensive and more efficient system to rate RF in these types of studies. (2) The current RF rating system requires the use of the AAI for its language data. An alternate scoring system could pave the way for the scoring of RF using alternative data sources, such as psychotherapy sessions, interviews, or other language data. (3) An alternate scoring method could add to our understanding of the linguistic markers of RF speech. For instance, the Referential Activity (RA) measure has been computerized as computerized RA (CRA) (Mergenthaler & Bucci, 1999), leading to increased understanding of the linguistic features of RA. Moreover, as validation of the RF construct and scales is an ongoing process, the CRF scale can advance the scope and number of psychometric validation studies of RF. (4) Of interest to psychotherapy researchers, an alternate RF measure could facilitate the empirical study of how some forms of psychotherapy promote the development of RF. A computerized system may be a component of a valid empirical methodology to assess mechanisms of symptom and personality change.

Computerized text analysis. Content analysis in general utilizes many methods that systematically identify content or stylistic features from language samples. There are two basic assumptions inherent in content analysis: (1) that there is meaning in the quantity and frequency of word usage; and (2) inferences can be made about individuals or groups of individuals from the language they produce. The speed and accuracy of computer technology allows for reliable and robust analysis of large amounts of text. The most common form of computer-assisted content analysis employs "dictionaries" of words that have been sorted into categories, or, "tags." For example, a hypothetical category named "fruits" might search a text for such words such as "apple" and "banana." The quantity and frequency of categories in a particular text sample can be tabulated by the computer program for later statistical analysis.

This study investigates the possibility that computer aided text analysis can be used to assess Reflective Functioning in an efficient manner, while simultaneously identifying the linguistic markers of RF speech. A computerized text analysis system would require no training of raters, and it could, therefore, be more portable and easily disseminated into the research community. For example, a computer program to rate RF from text samples could be downloaded from the Internet and used by researchers relatively easily once they format the texts

appropriately for analysis. From this point forward, we label the current RF system RF, short for Rater RF. We will label the computerized RF we are developing, CRF.

This study, then, aims to evaluate the feasibility and criterion validity of developing a computerized measure of RF. We hypothesize that there are unique linguistic indicators of high and low RF speech that can be identified and employed to evaluate narratives for levels of RF. We evaluate these hypotheses utilizing two samples of RF rated Adult Attachment Interview (AAIs) to derive linguistic indicators of RF, then, using AAIs that were not used to develop the indicators, correlate the collective frequency of the indicators with judge scored RF on those AAIs.

Method and Materials

Participants

Two samples of participants completed AAI interviews, which were scored for RF by reliable raters. AAIs from both clinical and non-clinical samples were employed to prevent developing a dictionary that might simply differentiate the speech and word usage of clinical vs. non-clinical populations. All of these AAIs were transcribed and formatted for computer text analysis. A total of 113 RF Scored AAIs were utilized for this study.

Sample 1 consisted of 40 non-clinical adults residing in the United Kingdom. The interviews were drawn from a sample of adults between the ages of 30 and 52 years who were recruited as a comparison group for other individuals who had received psychotherapy. Seventy-eight percent were female, 58% married or cohabiting, 61% were college graduates and 78% were White. None had ever received psychiatric or psychological treatment, and none met current criteria for any DSM-IV Axis I or II disorder.

Sample 2 consisted of 73 individuals diagnosed with Borderline Personality Disorder (BPD) residing in the New York metropolitan area. This clinical sample has been previously characterized in Levy, Meehan et al. (2006). Ninety-two percent of this clinical sample were female, 7.7% married, 51% were college graduates, and 68% were White. All met criteria for BPD, and, additionally 77% had a lifetime history of at least one DSM-IV Axis I mood disorder, 48% an anxiety disorder, 44% an eating disorder, and 39% a substance use disorder.

Design and Procedures

The Marker Approach. The Marker Approach first introduced by Mergenthaler (1996) and further

extended by Mergenthaler and Bucci (1999) is a procedure for transforming an established, manualized coding system for verbal data into a computerized text scoring method. It was developed to model *stylistic* aspects of speech, in contrast to themes or contents within speech. This approach treats word markers as operational indicators of a psychological state or capacity. The Marker Approach applies an inductive method to derive a Characteristic Vocabulary, or set of verbal indicators of a particular rater-judged construct. These verbal indicators tend to be relatively common words and markers, and can occasionally seem counterintuitive in their association with the construct under investigation. However, while a Characteristic Vocabulary is often composed of common words that would be difficult to predict a priori, their relative frequency, when assessed collectively, can be a powerful index of many stylistic aspects of speech. For example, the Marker Approach was utilized to computerize Bucci's Referential Activity (RA) construct, which, similar to RF, is scored by trained raters assessing narrative text samples. The Marker Approach as applied to RA identified that the frequency of just 40 relatively common words correlates between .5 and .8 with rater judged RA, even in relatively brief texts (Mergenthaler & Bucci, 1999).

The Marker Approach to dictionary development can be summarized in several steps (see Figure 1). (1) First, text samples that have been reliably rated with the manualized scoring system (RF in this case) are transcribed and formatted for computer text analysis. (2) One-third of the texts are then chosen randomly, and are designated to be part of the Dictionary Development Corpus. (3) Within this text corpus, extreme cases from both the High and Low ends of the constructs' measurement are identified. This is usually achieved by selecting those cases at the top and bottom terciles of the
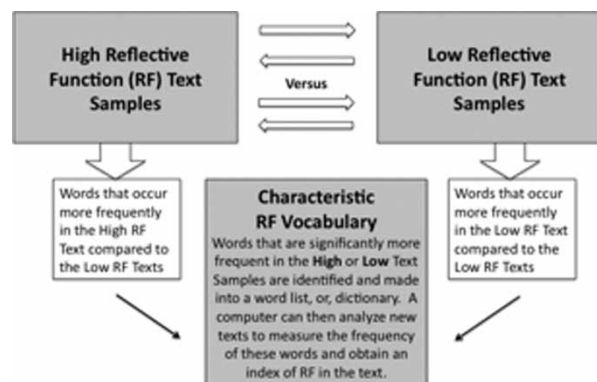

Figure 1. The CRF Characteristic Vocabulary development: The Marker Approach.

distribution. These two terciles constitute the Dictionary Development Corpus. (4) The next step involves the identification of Characteristic Vocabularies for each of the two High and Low corpora in reference to each other. That is, words that are significantly more frequent in one text versus the other are identified. Low frequency and overly specific words are eliminated, as well as words that are usually not transcribed in a reliable way (e.g., uhm, hm, mm, m, . . .). The result is two dictionary word lists, High and Low. (5) Once these word lists are identified, texts can be analyzed by a computer that calculates the collective frequency of these markers, in this study, of High and Low CRF words. (6) The initial criterion-related validity of the Dictionary can be assessed by analyzing all original texts, called a Validation Corpus, that were *not* used to derive the dictionary. That is, the frequencies of the words in the Development Corpus are assessed in this Validation Corpus and this number is correlated with the RF scores in the Validation Corpus. A significant correlation between RF and CRF for this Validation Corpus, can establish the initial criterion related validity of the computer measure.

Assessment

The Adult Attachment Interview (AAI) (Main & Goldwyn, 1991). This semi-structured interview is designed to elicit thoughts, feelings, and memories about early attachment experiences, and to assess a participant's state of mind with regard to these early attachment experiences. The AAI consists of a set series of 20 questions, and requires the interviewees to reflect on their caregivers' styles of parenting and how their childhood experiences with their caregivers impacted their development and adult lives.

The Reflective Function scale. The AAI was scored with the RF scale. The RF scale has an 11-point range that captures the quality of mentalization in the context of attachment relationships. The lowest score is − 1 and considered "negative" RF. A − 1 score is characterized by verbalizations that are very concrete, where reflectiveness is entirely absent, or the mental states of others are described in extremely distorted fashion by the interviewee. On the other end of the scale, a score of 9 represents exceptional RF, in which the interviewee demonstrates unusually multifaceted, sophisticated, or novel inferences regarding the mental states of themselves and others.

RF coding for the non-clinical sample was carried out by one coder (author MT) who is one of the developers of the coding manual, is the main trainer

of RF coding, and runs the reliability training and certification system for the measure. MT is a PhD clinical psychologist and psychoanalyst. The clinical RF coders were one of the authors (KNL), a licensed PhD in clinical psychology, and two advanced doctoral PhD students in clinical psychology. The PhD student coders were in their mid-twenties and had master's degrees in psychology. KNL was trained to code RF by one of the developers of the scale (MT) and completed reliability training with this person. KNL then trained both PhD students, who completed the same reliability training (see Levy, Meehan, et al., 2006). Reliability was obtained between the PhD student coders and one of the developers of the coding manual on practice sets. After training was completed and reliability was established, the two PhD student coders coded a subset of each other's transcripts to a high level of reliability ($n = 28$, $ICC = .86$). Raters for both samples were blind to the participants' clinical status, and other demographic variables.

Results

Descriptive Statistics

Across the two samples, overall RF scores ranged from 1 to 8 ($M = 3.38$, $SD = 1.76$, $n = 113$). There were significant differences between mean RF scores for the two samples with the clinical sample lower in RF than the non-clinical sample (Non-clinical: $M = 4.16$, $SD = 2.40$, $n = 40$; Clinical: $M = 2.95$, $SD = 1.09$, $n = 73$; $df = 111$, $t = 3.70$, $p < .001$).

RF Linguistic Markers

All the AAI transcripts were transcribed verbatim from the audiotapes of the interviews. They were then formatted to meet the standards for computerized text analysis as described in Mergenthaler and Stinson (1992). Following the Marker Approach, from 113 RF scored AAIs two groups were formed: (a) all AAIs with RF greater than 5 ($n = 18$); (b) all AAIs with RF less than 3 ($n = 33$). The AAIs were drawn from the lowest 4 (−1, 0, 1, 2) and the highest 4 (6, 7, 8, 9) points of the RF rating scale rather than choosing an approximately equal distance from the mean. This ensured we would utilize only the extreme ratings, rather then ratings that fell in the middle of the scale. From each group, nine AAIs were randomly selected, yielding nine High and nine Low RF AAIs, and utilized as the Dictionary Development Corpus (see Figure 1).

Based on these 18 High and Low AAIs, a characteristic High and Low CRF word list was derived. These two groups represent a High RF and

a Low RF Text corpus. Three criteria were used to select the High and Low dictionary word lists. First, a $p < .01$ difference is needed between the high and low corpora in the frequency of the word or marker to be considered for the dictionary. Second, a relative frequency $\geq 0.1\%$ of the candidate marker relative to the total number of markers in the text sample must be present in the corpus. Third, from the resulting list 27 words or particles were deleted that were (a) content carrying (e.g., brother, mother), (b) interjections (e.g., er, uh, uhm), or (c) ambiguous terms (e.g., like, mean, well). These three criteria yielded a dictionary of $n = 54$ High CRF markers and $n = 49$ Low CRF markers from the High and Low RF corpora. Table I lists the 20 most frequent High and Low linguistic markers identified with this procedure.

Criterion Validity of CRF

Next, using these word lists we calculated the frequency of the word markers in the High and Low CRF dictionaries on the Validation Corpus, using the remaining randomly selected 95 AAIs. With this Validation Corpus, then, we assessed the initial criterion validity of the CRF measure. As RF was significantly correlated with AAI length (number of words), the CRF measure (relative frequency of CRF markers) was multiplied by the natural logarithm of the respective text length $(\ln(n))$ to correct

Table I. The 20 most frequent High (left side) and Low (right side) CRF word markers (in decreasing order)

and	I
was	to
that	you
the	my
of	t
a	s
it	me
she	so
know	they
he	just
in	would
but	when
think	don
her	no
with	m
had	not
very	then
sort	can
about	one
because	go

Note. The single-letter words (except "I") stem from word contractions: "t" from "not" such as in "can't"; "s" from "is" such as in "it's"; "m" from "am" such as in "I'm."

for AAI length (for the remainder of the article we refer to this corrected measure when using the term CRF). Additionally, due to the non-normal distribution of the data, we utilized the non-parametric Spearman rank-order correlation coefficients (rho) to assess the association between RF and CRF. Table II reports the correlations between RF and CRF for the AAIs from dictionary development corpus, the dictionary validation corpus, and from the non-clinical and BPD sub-samples within the dictionary validation corpus. The High CRF dictionary, made up of just 54 word markers, provided the most robust and consistent correlations with RF; including the Low CRF dictionary by subtracting the proportions of Hi CRF and Low CRF words did not consistently improve upon the correlation between High CRF only and RF (see Table II). The High CRF dictionary demonstrated moderate to high and consistent correlations with RF in both the non-clinical and in the clinical samples.

Discussion

Utilizing a Marker Approach, this study indicates that a CRF dictionary is feasible and correlates significantly with RF ratings in both clinical and non-clinical samples, indicating initial criterion validity for the CRF measure. The most robust associations were between the High CRF dictionary and RF. Adding the Low RF dictionary did not improve the correlations between CRF and RF. The High CRF dictionary contains only 54 linguistic markers, composed of relatively common, high-frequency words, suggesting that CRF is a relatively powerful dictionary that could be reliably utilized with even smaller text samples (as few as 50 words). Moreover, the strength of the High CRF dictionary is consistent with what was found in the development of computerized Referential Activity (CRA) (Mergenthaler & Bucci, 1999) in that the High CRA dictionary in this previous study also provided the highest correlation with manual scored RA.

To illustrate the way in which items from the characteristic vocabulary can be identified, below is an example of a text from an interviewee talking about her mother as part of the AAI. This interview rated high for both RF and high for CRF (the High CRF Dictionary items are bolded):

> **She** liked people round **her** constantly, er, **she was**, well my father stood **for** parliament all **of** my childhood **and**, actually I **suppose that** goes **back** to [be the cause]. Although **he was the** politician, **the** one **that was** standing up at **the** front, **she was** chair **of the** local party, **she was the** one **who** organized everything. **He was the** figurehead **and**

Table II. Spearman rank order correlations (rho) between RF and Computerized Reflective Functioning (CRF)

Sample	Variable	By variable	Spearman rho	Significance p value
Dictionary development corpus ($n = 18$)				
	RF	High CRF	.84	<.0001
		High-Low CRF	.77	<.001
Dictionary validation corpus ($n = 95$)				
	RF	High CRF	.57	<.0001
		High–Low CRF	.47	<.0001
Dictionary validation sub-samples ($n = 95$)				
Non-clinical ($n = 27$)	RF	High CRF	.57	.002
		High–Low CRF	.65	<.001
Clinical ($n = 68$)	RF	High CRF	.57	<.0001
		High–Low CRF	.43	<.001

he believed desperately **in the** [deleted] party. **But it was actually her who** packaged him and **a** lot **of the**, sort **of,** what would now be **the** sort **of** spin **and the** way everything **was** put together **was** my mother. **And the** way everything **was** run **and** thinking **of** just [how] sociable our house **was,** [it was] always **the** headquarters **of** the local [deleted] party and **there were** always people **in** there. **And it** might have **been** partly, I **think it probably was** partly because **she was** aware **that the** relationship **was** flawed, **and was** filling **it with** other people.

While some of the words in the High CRF dictionary would be expected in high RF speech, others would have been difficult to predict. This is an advantage of using the empirically driven Marker Approach. In the High CRF Dictionary, words like "because," "think," or "and" are abstract and elaborative indicators that would seem in keeping with high RF speech. Many of the other CRF words as presented in Table I have no obvious theoretical connection to the RF construct. However, two aspects must be considered. Due to the marker approach we utilized, these words can be understood as markers for higher RF narratives (e.g., a coherent narrative may have to use conjunctions like the word "and"). Moreover, reflective processes may tend to include references to the past as expressed by words such as "had" or "was." Since the technique for finding CRF words was inductive, this approach provides the possibility of identifying new, undiscovered verbal aspects of high RF discourse. As a result, the identification of a CRF dictionary helps to advance the theoretical understanding of high RF speech.

At this stage, criterion validity is preliminary due to a number of issues. The non-clinical sample in this study was relatively small and the clinical sample did not show much variability in the RF scale. Most of the ratings in this clinical sample were in the lower middle range of the scale. Additionally, there may have been subtle rater differences among the sub-samples. Relatedly, there were differences in the quality of the AAIs in terms of either transcription or administration among the sub-samples. Also, the AAI was developed for use with non-clinical samples, and may not be asking questions most relevant to assessing RF for clinical samples. Finally, since the approach presented here is based on lexical items, the findings may be confounded by individual differences in verbal fluency or verbal IQ. There is evidence for an association between better verbal memory and metacognitive awareness in individuals with schizophrenia (Lysaker et al., 2005). This, however, is also a problem for judge-rated RF. Researchers may have to routinely incorporate measures of verbal IQ and verbal memory to clarify whether the construct of RF is partially dependent on higher verbal capacities in general. Finally, the correlations between RF and CRF may be inflated due to similarities in clinical and demographic profiles between the development and validation corpuses. However, the two samples that were used for both corpuses were derived from two different countries (USA and UK) and contained both clinical and non-clinical participants. Given these limitations, at this point, the CRF dictionary should not replace the RF scale. Further, caution is warranted in using the CRF measure until it can be further validated in other samples and the impact of potential covariates can be further clarified.

To conclude, we have identified significant verbal markers of RF speech using an empirically driven, computerized, text analytic approach. At this point we have a prototype CRF dictionary that can be efficiently and readily utilized by other researchers interested in further developing this alternate method of RF rating, investigating the validity of the RF construct, and eventually assessing CRF's utility in a wider variety of linguistic contexts and populations. In subsequent stages of this research effort, we plan to investigate whether RF is best modeled as a one-dimensional or a multidimensional construct that might be better captured by more than one

characteristic vocabulary (cf., Bouchard et al., 2008; Gullestad & Wilberg, 2011; Lecours & Bouchard, 2011; Semerari et al., 2007). This consideration of multidimensionality may be particularly important in clinical samples, where, thus far, Low CRF markers have not yielded consistent associations with low RF speech.

Acknowledgements

References

Bateman, A., & Fonagy, P. (1999). Effectiveness of partial hospitalization in the treatment of borderline personality disorder: a randomized controlled trial. *American Journal of Psychiatry, 156*(10), 1563–1569.

Bouchard, M., Target, M., Lecours, S., Fonagy, P., Tremblay, L., Schachter, A., & Stein, H. (2008). Mentalization in adult attachment narrative: Reflective functioning, mental states, and affect elaboration compared. *Psychoanalytic Psychology, 25*(1), 47–66.

Choi-Kain, L.W., & Gunderson, J.G. (2008). Mentalization: Ontogeny, assessment, and application in the treatment of borderline personality disorder. *American Journal of Psychiatry, 165*(9), 1127–1135.

Dimaggio, G., Carcione, A., Nicolò, G., Conti, L., Fiore, D., Pedone, R., & Semerari, A. (2009). Impaired decentration in personality disorder: A series of single cases analysed with the metacognition assessment scale. *Clinical Psychology & Psychotherapy, 16*(5), 450–462. doi: 10.1002/cpp.619.

Dimaggio, G., Carcione, A., Petrilli, D., Procacci, M., Semerari, A., & Nicolò, G. (2005). State of mind organization in personality disorders. Typical states and the triggering of inter-state shifts. *Clinical Psychology & Psychotherapy, 12*(5), 346–359. doi: 10.1002/cpp.456.

Dimaggio, G., Procacci, M., Nicolò, G., Popolo, R., Semerari, A., Carcione, A., & Lysaker, P.H. (2007). Poor metacognition in Narcissistic and Avoidant Personality Disorders: four psychotherapy patients analysed using the Metacognition Assessment Scale. *Clinical Psychology & Psychotherapy, 14*(5), 386–401. doi: 10.1002/cpp.541.

Fertuck, E.A., Bucci, W., Blatt, S.J., & Ford, R.Q. (2004). Verbal representation and therapeutic change among anaclitic and introjective inpatients. *Psychotherapy: Theory, Research, Practice, Training, 41*(1), 13–25.

Fischer-Kern, M., Schuster, P., Kapusta, N.D., Tmej, A., Buchheim, A., Rentrop, M., ... Fonagy, P. (2010). The relationship between personality organization, reflective functioning, and psychiatric classification in borderline personality disorder. [doi: DOI: 10.1037/a0020862]. *Psychoanalytic Psychology, 27*(4), 395–409.

Fonagy, P., Steele, M., Steele, H., Moran, G.S., & Higgitt, A.C. (1991). The capacity for understanding mental states: The reflective self in parent and child and its significance for security of attachment. *Infant Mental Health Journal, 12*(3), 201–218. doi: 10.1002/1097–0355(199123)12:3 <201::aid-imhj2280120307> 3.0.co;2–7.

Fonagy, P., Target, M., Steele, M., & Steele, H. (1998). *Reflective-Functioning manual (Version 5) for application to adult attachment interviews*. Unpublished manual. Psychoanalysis Unit, Sub-Department of Clinical Health Psychology, University College London.

Grienenberger, J.F., Kelly, K., & Slade, A. (2005). Maternal reflective functioning, mother, infant affective communication, and infant attachment: Exploring the link between mental states and observed caregiving behavior in the intergenerational transmission of attachment. *Attachment & Human Development, 7*(3), 299–311.

Gullestad, F.S., & Wilberg, T. (2011). Change in reflective functioning during psychotherapy: A single-case study. *Psychotherapy Research, 21*(1), 97–111. doi: 10.1080/10503307.2010.525759.

Jurist, E.L. (2005). Mentalized affectivity. [doi:10.1037/0736–9735.22.3.426]. *Psychoanalytic Psychology, 22*(3), 426–444. doi: 10.1037/0736–9735.22.3.426.

Jurist, E.L. (2010). Mentalizing minds. [doi: 10.1080/07351690903206496]. *Psychoanalytic Inquiry, 30*(4), 289–300. doi: 10.1080/07351690903206496.

Karlsson, R., & Kermott, A. (2006). Reflective-functioning during the process in brief psychotherapies. [doi:10.1037/0033–3204.43.1.65]. *Psychotherapy: Theory, Research, Practice, Training, 43*(1), 65–84. doi: 10.1037/0033–3204.43.1.65

Lecours, S., & Bouchard, M.-A. (2011). Verbal elaboration of distinct affect categories and BPD symptoms. *Psychology and Psychotherapy: Theory, Research and Practice, 84*(1), 26–41. doi: 10.1111/j.2044–8341.2010.02006.x

Levy, K.N., Beeney, J.E., Wasserman, R.H., & Clarkin, J.F. (2010). Conflict begets conflict: Executive control, mental state vacillations, and the therapeutic alliance in treatment of borderline personality disorder. [doi: 10.1080/10503301003636696]. *Psychotherapy Research, 20*(4), 413–422. doi: 10.1080/10503301003636696.

Levy, K.N., Clarkin, J.F., Yeomans, F.E., Scott, L.N., Wasserman, R.H., & Kernberg, O.F. (2006). The mechanisms of change in the treatment of borderline personality disorder with transference focused psychotherapy. *Journal of Clinical Psychology, 62*(4), 481–501.

Levy, K.N., Meehan, K.B., Kelly, K.M., Reynoso, J.S., Weber, M., Clarkin, J.F., & Kernberg, O.F. (2006). Change in attachment patterns and reflective function in a randomized control trial of transference-focused psychotherapy for borderline personality disorder. *Journal of Consulting and Clinical Psychology, 74*(6), 1027–1040.

Lysaker, P.H., Buck, K.D., & Ringer, J. (2007). The recovery of metacognitive capacity in schizophrenia across 32 months of individual psychotherapy: A case study. [Article]. *Psychotherapy Research, 17*(6), 713–720. doi: 10.1080/10503300701255932.

Lysaker, P.H., Carcione, A., Dimaggio, G., Johannesen, J.K., Nicolò, G., Procacci, M., & Semerari, A. (2005). Metacognition amidst narratives of self and illness in schizophrenia: associations with neurocognition, symptoms, insight and quality of life. *Acta Psychiatrica Scandinavica, 112*(1), 64–71. doi: 10.1111/j.1600-0447.2005.00514.x.

Main, M., & Goldwyn, R. (1991). *Adult Attachment Classification System, Version 5*. Unpublished manuscript. University of California, Berkeley.

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: a new way of describing psychotherapeutic

processes. *Journal of Consulting and Clinical Psychology, 64*(6), 1306–1315.

Mergenthaler, E., & Bucci, W. (1999). Linking verbal and non-verbal representations: computer analysis of referential activity. *British Journal of Medical Psychology, 72*(3), 339–354.

Mergenthaler, E. & Stinson, C. (1992). Psychotherapy transcription standards. [doi: 10.1080/10503309212331332904]. *Psychotherapy Research, 2*(2), 125–142. doi: 10.1080/10503309212331332904.

Pronin, E. (2008). How we see ourselves and how we see others. *Science, 320*(5880), 1177–1180.

Semerari, A., Carcione, A., Dimaggio, G., Falcone, M., Nicolò, G., Procacci, M. & Alleva, G. (2003). How to evaluate metacognitive functioning in psychotherapy? The metacognition assessment scale and its applications. *Clinical Psychology & Psychotherapy, 10*(4), 238–261. doi: 10.1002/cpp.362.

Semerari, A., Carcione, A., Dimaggio, G., Nicol, G. & Procacci, M. (2007). Understanding minds: Different functions and different disorders? The contribution of psychotherapy research. [Article]. *Psychotherapy Research, 17*(1), 106–119. doi: 10.1080/10503300500536953.

Slade, A., Grienenberger, J., Bernbach, E., Levy, D., & Locker, A. (2005). Maternal reflective functioning, attachment, and the transmission gap: A preliminary study. *Attachment & Human Development, 7*(3), 283–298.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

Vermote, R., Fonagy, P., Vertommen, H., Verhaest, Y., Stroobants, R., Vandeneede, B., … Peuskens, J. (2009). Outcome and outcome trajectories of personality disordered patients during and after a psychoanalytic hospitalization-based treatment. *Journal of Personality Disorders, 23*(3), 294–307.