JOURNAL OF PERSONALITY ASSESSMENT, *87*(3), 330–348

A Rorschach Stability Study in a Nonpatient Adult Sample

Serge Sultan
Institute of Psychology
University of Paris–René Descartes

Anne Andronikof
Department of Psychology
University of Paris 10

Christian Réveillère
Department of Psychology
University of Tours

Gilles Lemmel
Department of Psychology
University of Paris 10

The objective of this study was to provide new primary data on Rorschach Comprehensive System stability levels. To achieve this, we tested 75 French nonpatient adults twice on the Rorschach with a 3-month interval between the tests. Interrater reliability was in the excellent range for most of the variables studied. The overall stability level in a selected set of previously studied variables was below expectations (median *r* = .53). Personality, cognitive or self/relational variables yielded higher test–retest correlations than emotional and coping variables. Moderators of stability could be identified: (a) overall level of Task Engagement (TE) in F, m, FM + m, a, FC, Sum C', Sum V, Sum Shd, Fr + rF, INC + FAB, COP, es, Adj es, EGO, and Blends; (b) variations in TE in F, FM, and p; (c) state distress in Zd, m, FM + m, a, C, CF + C, WSumC, FD, and es; (d) variables derived from the number of responses impacted stability in P, Zf, m, FC, CF + C, Sum C', Sum V, MOR, EA, es, and Blends. These results provide further support for the reliability of several measures. Examiner effects as an influence on productivity and TE were identified as an important area for future research.

Stability is of great importance in psychological assessment procedures. First, from a psychometric perspective, large short-term test–retest correlations are necessary if we are to be sure that the measures used are reliable. Second, one should expect reasonable long-term stability on measures that are supposedly related to stable personality characteristics. The aims of this article are to present an empirical study of Rorschach stability and to explore some reasons for instability. We focus on the computation of stability measures for interpretively significant variables that have already been studied in the past and study moderators' effects on these measures by relating stability levels to emotional states and test interaction styles.

In a review article, Viglione and Hilsenroth (2001) reported mean test–retest correlation coefficients ranging from .75 to .82 (based on intervals ranging from 3 weeks to 1 year) for commonly used Comprehensive System (CS) variables. Yet, as stated by Meyer (1997a) and Grønnerød (2003), most arguments on test–retest in the Rorschach literature derive from five sets of data (Exner, 1980; Exner, Armbruster, & Viglione, 1978; Exner, Thomas, & Cohen, 1983, as cited in Exner, 2003b; Haller & Exner, 1985; Thomas, Alinsky, & Exner, 1982, as cited in Exner, 2003b). Two of these reports were published in peer-reviewed journals and provide detailed information on the method used. Exner et al. (1978) tested 100 nonpatient adults with an interval of 3 years. Participants were screened for evidence of personality disorganization. Twenty-six and 22 examiners took part in the test and retest, respectively. None of these retested a participant they had tested originally. The *M r* among the 19 variables studied was .79 (*Mdn* = .80) with a minimum–maximum range of .66 to .90. Haller and Exner (1985) tested 50 newly admitted patients presenting symptoms of depression or helplessness with an interval of 3 to 4 days between tests. Twenty-five patients received standard instructions. The others were instructed to give responses different than they had in the first test. Ten examiners were involved at test and retest, with no patient tested twice by the same examiner. Among the 28 variables studied in the group with standard instructions, the *M r* was .71 (*Mdn* = .74) with a minimum–maximum range of .28 to .87. In people who received modified instructions, the mean *r* was .66 (*Md* = .71) with a minimum–maximum range of .33 to .88.

On the basis of these data, the main arguments developed in the Rorschach literature on the test–retest issue in adults have been that (a) almost all the CS variables that supposedly relate to trait characteristics have exhibited substantial stability in adults both in the short and long term; (b) lower stability is mainly associated with inanimate movement (m) and diffuse shading (Y), which are considered state related (Exner, 2003b, pp. 176–183; Meyer, 1997a, 1997c; Viglione, 1999; Viglione & Hilsenroth, 2001; Weiner, 2001).

Three meta-analyses have directly addressed the issue of the stability of Rorschach measures (Grønnerød, 2003; Parker, 1983; Parker, Hanson, & Hunsley, 1988). The empirical basis for these analyses is broad since they include all Rorschach studies with subsequent follow-up. Most recently, Grønnerød systematically examined the temporal stability of different Rorschach scoring systems in studies published between 1921 and 2002. He reached two conclusions: First, the Rorschach method exhibits an overall high level of stability, with a combined weighted temporal stability level of .68 (main average) for an average retest period of 3 years and 2 months. When one considers CS variables individually, the mean of the retest coefficients was .85 for a 6-month period. Second, stability estimates were higher in CS studies as well as when short retest intervals, large samples, and traitlike variables were used, that is, when variables that are less dependent on Y and m were used. These findings are similar to those obtained from previous analyses in which stability coefficients were estimated at the level of .80 (Parker et al., 1988). On the basis of Rorschach literature, expectations for stability should be in the .70 and .80 range for most interpretively significant variables.

Research into the stability of non-Rorschach personality assessment methods may help set goals and standards for Rorschach research. Watson (2004) reviewed all studies with short-term test–retest intervals published in personality journals over a period of more than 13 years. When 23 studies with intermediate intervals of 2 to 4 months are considered from Table 1 of Watson's article (p. 325), the minimum correlation was .19 and the maximum was .92, with the average of the 23 minimums equaling .63 and the average of the 23 maximums equaling .79. Thus, if the Rorschach is comparable to existing personality assessment procedures, the expected range of test–retest correlations in the Rorschach should be in the .60s and .70s range for intermediate retest intervals.

Although research on stability in personality assessment is necessary, the interpretation of stability results may be complicated for various reasons. In the psychometric tradition, the term *test–retest reliability*, or *dependability*, is used in relation to subsequent testings over shorter periods. Over such periods, one can assume that the construct is fairly stable, and instability can therefore be attributed to the unreliability of the measure. Over longer periods, changes in the underlying construct introduce an additional effect in the reliability data. Traditionally, the Rorschach literature has interpreted low stability levels as reflecting state characteristics, whereas high stability levels have been assumed to indicate traitlike features (Exner, 2003b; Exner et al., 1978). This hypothesis derives from a recurrent pattern of high coefficients for almost all CS scores with the exception of a few on which low scores are obtained, namely, Y and m. Meta-analyses have also used this strategy based on the recurrence of results to interpret differences in consistency over various age groups (Roberts & DelVecchio, 2000; Trzesniewski, Donnellan, & Robins, 2003). In a single empirical study, another rather more concrete strategy aimed at disentangling trait and state is to use external criteria for which stability information is available and to examine whether these criteria may explain a significant proportion of the "error variance" (Kraemer, Gullion, Rush, Franck, & Kupfer, 1994). As for the constructs measured by the Rorschach, it is likely that most of them simultaneously possess both state and traits aspects, each of which might be of interest. It would therefore be interesting to determine whether external markers of state, or trait, are capable of moderating stability levels. So far, no test–retest Rorschach study has used such a strategy.

In another meta-analysis, Grønnerød (2004) studied Rorschach sensitivity to changes in psychotherapy. Variables related to self-concept, self-perception, and interpersonal relations, such as reflections and pairs, were less susceptible to change than affective and coping features, such as the D score or CDI. These results are in line with Grønnerød's (2003) previous meta-analysis, which showed some data to support the hypothesis of a lower stability in affective and coping features, including m and Y. In fact, stability may depend on underlying constructs. Research has suggested that measures which focus primarily on behavioral, cognitive, and affective characteristics generally do not have the same expected stability (e.g. Watson, Hubbard, & Wiese, 2000). This conclusion is based on early observations of the stability of intelligence and cognitive abilities (Conley, 1984) and more recent demonstrations that the temperament and trait type have an effect on trait consistency (Watson, 2004). Hence, both the Rorschach and non-Rorschach literature suggest that the state or trait nature of the variables should act as a potential moderator of stability. To summarize, state-related measures of the Rorschach should undergo greater change than trait-related measures over intermediate and long periods. One strategy used to isolate the state or trait component of the measure is to include an external criterion.

Many factors may impact stability levels including reliability and base rates. For example, in the case of instability, the issue of the preciseness of the instrument is always inextricably intertwined with changes in the measured constructs. This is why interrater reliability may be conceptualized as a suppressor of stability levels when reliability estimates are low (Meyer, 1997a). Rorschach studies have also shown that base rates could strongly influence reliability estimates (Meyer et al., 2002). Examining variables with a base rate lower than .05, Viglione and Taylor (2003) showed that intraclass correlation coefficient (ICC) reliability statistics were lower and variability in correlations higher.

However, one factor that is somewhat specific to the Rorschach and performance-based procedures may be of central importance. Rorschach data, more than data obtained with other personality assessment instruments (see Meyer, 1997b), are dependent on an active interaction with the examiner, with the interpersonal dynamics of both the test person and the examiner being important for the completion of the task. Rorschach scores are, to an extent, dependent on the ability of test persons to spontaneously engage with the task and articulate responses. Consequently, one factor that might moderate stability in numerous Rorschach variables could be the level of engagement in the task, as revealed by longer and more complex records versus shorter and more simplistic records. Although we may expect engagement to moderate stability levels for variables that are related to it, the way it may influence stability is still unclear for a variety of reasons.

First, participants with lower engagement levels may be more prone to change, especially when they approach the task for the second time; this second experience may render them less defensive and facilitate engagement. Second, participants with a high level of engagement may express themselves differently on the two occasions. For instance, one can imagine that strong negative feelings might be expressed by a higher Sum C' at baseline (T1) and a higher Sum V or Sum Y at retest (T2). Finally, in most test–retest studies, the fact that the examiners have administered the test to different participants at T1 and T2 may have generated some examiner-related differences in interpersonal dynamics during the conduct of the test. This emphasizes the need to explore examiner-related effects on stability. In addition, given the pervasive and structural role of engagement (e.g., Lindgren & Carlsson, 2002), one might expect that changes in engagement would affect the stability of individual scores. To summarize, stability is expected to vary across domains of functioning and, among other factors, is thought to depend on the variation of states such as emotional states and the engagement or openness of the person, which may secondarily be related to examiner effects.

OBJECTIVES AND HYPOTHESES

One primary objective of this study was descriptive in nature. Given the lack of recent test–retest CS data, we wanted to provide new primary data on the stability of CS measures. We therefore discuss the appropriateness of the coefficients to be used and devote some time to describing stability in nominal variables, given that this is of relevance for the interpretation process. We wished to study an intermediate time interval (3 months) that permits changes in some of the constructs measured, that is, emotions and coping-related constructs.

Hypothesis 1

We expected moderate to high levels of stability for most Rorschach variables. Given the literature on personality assessment, a high level of stability over intermediate intervals was defined in this study as exceeding .70, and a moderate level as exceeding .50. On the basis of Grønnerød's (2003) meta-analysis (Alternative Models, CS only; Table 2), the expected value for overall stability in a CS study would be .82 for a sample of 75 and a 3-month retest period.[1] However, given the limited number of samples contributing to these estimates, it would be better to use predictions based on all scoring systems. In a reanalysis of the data, Grønnerød (2006) computed new regression models and found overall predicted stability for all scoring systems to be estimated at .74 and .69. Thus, the expected stability for a CS study on a 3-month interval should be no less than .69 and should probably be in the .70s or .80s. No expectations could be formulated concerning categories as defined by cut-points or ratios (e.g., EB introversive) given the very small number of existing analyses.

Hypothesis 2

In relative terms, along with Grønnerød's (2004) observations and the traditional hierarchy of stability expected by personologists (Conley, 1984; Watson, 2004), we expected stability to be higher for personality, cognitive or self/relational construct-related variables (e.g., M, a, EA, EGO) than for emotional, coping, or state-related variables (e.g., m, Y, D, shadings).

Hypothesis 3

We expected that a part of the instability (i.e., discrepancies between T1 and T2) could be attributable to specific factors.

[1] This figure was kindly computed by the author of the meta-analysis.

We hypothesized that changes in distress, as measured by an external criterion, could account for "error variance" in state-related emotional variables, particularly m, Y, es, D, DEPI, and S–CON.

Hypothesis 4

We expected that Task Engagement (TE) would moderate stability, such that controlling for TE would increase stability in Rorschach variables known to be related to this factor, that is, Sum Y, Sum C', all color determinants, m, R, S, FM, Sum V, MOR, M, Lambda, and all the variables that are expected to be more frequent or higher in longer and more complex records like WSum6 or Blends.

Hypothesis 5

We expected that large variations in TE would be related to lower stability levels for state-related variables and negative emotion markers.

Hypothesis 6

Because interaction and relational dynamics with the examiner underpin some crucial aspects of TE, some of the instability could also be due to effects related to the examiner's administration of the test. We expected that pairs of examiners whose results indicated lower stability would comprise at least one examiner who was associated with a lower level of engagement in the participants.

METHOD

Participants

Seventy-five persons were recruited from the ongoing French-language normative project (*M* age = 39.2 years; 28 men, 47 women). They were tested twice between November 2001 and March 2002. They were employed in private businesses, sports clubs, and a charity organization. Participants were included provided that they accepted that the individual data would remain strictly anonymous and no individual feedback would be given to anyone. Each participant signed an informed consent form. They were recruited by means of posters and intranet messages stating that in exchange for their participation we would donate a certain amount of money to a charitable organization of their choice. The same donation was made at test and retest.

The first 100 participants included in the normative project were asked to perform a retest after 3 months. Among these, 94 accepted, but 10 were not retested because they were not considered to be "nonpatients" (discussed next), and 5 could not be seen because of practical difficulties. To implement the nonpatient requirement, we used a post hoc screening based on three open-ended questions and a psychiatric screening questionnaire (General Health Questionnaire [GHQ–12]; Goldberg, 1978). A participant was rejected if he or she had three or more positive items on the GHQ–12 or had two positive items and endorsed symptoms on one of the open-ended questions. Consequently, the 79 participants who were retested were initially selected on this nonpatient criterion. Four more participants were excluded because their baseline Rorschach protocols had one or more response that could not be reasonably scored because the location was poorly reported or the inquiry was defective and did not respect the Workbook guidelines (Exner, 2003b). This resulted in a final retest sample of 75. No T2 protocols were excluded

TABLE 1
Sample Description for 75 Nonpatient Adults Included in the Stability Study

	n	%	*M*	*SD*	*Mdn*	*Min*	*Max*
Gender							
Men	28	37					
Women	47	63					
Age (years)			39.2	12.9	37	20	64
20 to 25	12	16					
26 to 35	22	30					
36 to 45	15	20					
46 to 55	16	21					
56 to 65	10	13					
Education (years)			13.3	3.1	14	5	21
Under 12	16	21					
12	11	15					
13 to 15	30	40					
16 +	18	24					
Retest interval (days)			95.0	8.1	95	79	115
Origin							
Dijon	43	57					
Paris	13	17					
Tours	19	26					
Setting							
Private corporation	43	57					
Charity organization	23	31					
Sports club	9	12					
No. of protocols/examiners[a]							
T1			8.0	—	—	3	14
T2			8.0	—	—	2	17
Yes replies to open-ended questions[b]							
T1 and T2							
Medication[c]	4	5					
Psychotherapy[c]	10	13					
Difficult times[d]	9	12					
T2							
Life event	22	29					
Type of life event							
None	53	71					
Minor	18	24					
Major	4	5					

[a]Nine examiners for T1 and T2. [b]Medication: "Do you take (have you ever taken) medication for your nerves?"; Psychotherapy: "Do you follow (have you ever followed) a psychological treatment?"; Difficult times: "Do you currently feel as usual or do you experience difficult times?"; Life event: "Since last time has anything happened in your life?". [c]Same frequencies for T1 and T2. [d]For T2, *n* = 8 (11%).

for problematic administrations. All the protocols obtained at least 14 responses. Characteristics of the final sample are provided in Table 1. The participants who were not included in the final sample had the following characteristics at T1:

1. The 10 people excluded on the basis of GHQ–12 scores had a mean GHQ–12 of 3.70 (Lambda = .51, R = 26.5).
2. The 5 people who could not be scheduled for a retest had means for the same variables of 1.20, .61, and 26, respectively.
3. The 4 people excluded because of administration quality had means of .79, 1.00, and 20.2. No information is available on the 6 persons who initially declined except that they were all men.

Examiners

Twelve examiners participated in the test–retest study. Among the 9 who participated at T1, 6 participated at T2 and 3 additional examiners took part in the retest. No examiner tested the same person twice. This approach was adopted to minimize any examiner–participant relation biases and memory effects and is consistent with previously published studies. It introduces a maximum error variance attributable to participant–examiner effects. All examiners were clinical psychologists (in France, this requires an MA degree) and had previously been trained in the CS with a training equivalent to Rorschach Workshop Level I or II. In addition, all the examiners attended a 2-day training session focusing on how to establish a rapport with the participant and how to inquire complex responses in an accurate way. The examiners were paid for their work. The mean number of protocols per examiner in the retest study was 8 at T1 and T2 (respective ranges = 3 to 14 and 2 to 17).

Procedures

Baseline test (T1). First, after initial informal contact, the examiner reminded the participant of the objective of the study, namely, scientific research examining how most participants respond to this test. Some general questions were first asked of the participant to complete a sociodemographic questionnaire. Then the examiner asked for previous experience and preconceptions with/about the Rorschach. Answers to questions were given in accordance with the Workbook guidelines (Exner, 2003b). This introduction was terminated with the examiner stating, “It is a widely used test in psychology and we need to know how most people in the community respond to it.” Second, the Rorschach CS was then administered using current standardized practice. This was followed by the examiner saying, “I also need to ask you a few questions about your health,” followed by three open-ended questions (see Table 1). Finally, the GHQ–12 was administered. The participants were then asked if they were prepared to be contacted in 3 months for a retest. If they asked any questions about this, the examiner explained that it is a usual procedure to validate psychological tests. The whole assessment took up to 1½ hr.

Retest (T2). The retest took place on average 95 days after T1 (range = 79 to 115). First, after the initial informal contact, the Rorschach was administered. If the participant asked why he or she was being retested, the examiner replied, “We ask you to retake this test because it is the usual scientific practice for validating tests”; if the participant asked whether she or he should give the same responses, the examiner replied, “Just tell me what you see now”; if other questions were asked, the examiners said, “You can do as you wish.” Overall, we adhered to the standardized administration procedures. This was followed by the same three open-ended questions as at T1, but an additional question was asked: “Since last time, has anything important happened in your life?” Positive responses were classified as minor events (birth of a child, new affective relationship, professional changes, n = 18) and major events (separation, death of a first-degree relative, n = 4).

Selection and Calculation of the Variables

We based our analysis on 47 variables that had already been presented by Exner (2003b, p. 179) and Viglione and Hilsenroth (2001) in their summary article.[2] This is a set of core variables in the interpretation process. Based on Grønnerød’s (2004) psychotherapy change results (Table 3, p. 262), we also distinguished between two types of variables for which expectations of stability were different: more stable personality, cognitive, or self or relational variables versus less stable emotional, coping, or state-influenced variables. Although data on change in treatment are not equivalent to stability data, the findings can cautiously be interpreted as implying different levels of consistency over time. The personality, cognitive, and self or relational variables were R, Zf, F, M, a, p, WSumC, L, EA, EGO, WSum6. The emotional, coping, or m and Y influenced variables were P, FM, m, FC, Sum T, Sum Y, Sum V, Adj es, D, Afr.

Other variables also were considered in our analyses. The total number of positive DEPI and S–CON criteria were included as additional markers of negative emotions (as in Meyer, 1997b, and Lingren & Carlsson, 2002). TE was calculated using a formula originally suggested by Meyer (1992) and then reconfirmed (Meyer, 1997b) and replicated (Lindgren & Carlsson, 2002). TE is the interpretation Meyer gave to the first and largest factor among the Rorschach variables.[3] To facilitate the interpretation of the results, two vari-

[2] Several variables presented in this article also were used in Perry, McDougall, and Viglione’s (1995) study.

[3] Task engagement is a weighted combination of Rorschach variables. The precise formula for calculating the task engagement vari-

ables were computed: an overall TE level (mean of T1 and T2 levels) and TE variations (absolute value of the difference between T1 and T2 values).

To avoid artificial correlations in moderation analyses involving TE, we computed 13 modified versions of the engagement scale that did not include the score to be predicted. For composite criterion scores like D-score, EA, es, and Adj es, defining direct part–whole relationships is complex. For instance, the D score is a global measure that is ultimately a function of all the determinant scores. To correct for part–whole associations between TE and this criterion, one could remove all individual determinant variables that contribute to both (i.e., M, FM, m, etc.) and/or one could remove a global index of determinant use (i.e., Lambda). If all the individual determinant-related variables were removed, the TE scale would be computed from just 4 of its 14 components, which would comprise its ability to accurately measure the TE construct (in fact, the correlation between the original and the modified scale would be as low as .80). Consequently, for the composite criterion variables based on determinant use, TE was computed after excluding Lambda, which itself is a broad index of determinant use. To ensure that the 13 modified scales adequately reflected the first factor, we factor-analyzed them within the full data set. Loadings of each scale on the first principal component ranged from .96 to .99 (M = .99). The mean correlation with the original scale was .99 (Min = .96).

GHQ–12 scores were used as an external criterion for measuring distress (Goldberg, 1978). The GHQ–12 is a 12-item self-report instrument for the detection of mental disorders in the community and in nonpsychiatric clinical settings. It measures aspects of psychological distress and social dysfunction (Kalliath, O'Driscoll, & Brough, 2004). The GHQ–12 asks the respondents to report how they have been feeling over the past few weeks (e.g., "Have you recently felt unhappy or depressed?"), using a 4-point scale from 0 (*not at all*) to 1 (*not more than usual*), to 2 (*more than usual*), to 3 (*much more than usual*). It includes six healthy-functioning items and six unhealthy-functioning items. The wording is reversed for positive items (*more than usual, as usual, less than usual, much less than usual*) so that higher values are associated with unhealthy functioning. Following standard practice focusing on the presence and absence of negative features, items were scored in a dichotomous fashion (0, 0, 1, 1; Goldberg et al., 1997). Yet we also explored the impact of a traditional Likert coding of the items (0, 1, 2, 3). Average values in nonpsychiatric samples vary from 1 to 3 (standard coding), and a high probability of caseness for psychiatric disorder was associated with values of 3+ or 4+ (Goldberg et al., 1997). Although the GHQ was conceptualized as a screening instrument in general population, scores are often interpreted as indicative of levels of distress with values of 0, 1 to 3, and 4+ understood as low, moderate, and high levels of distress, respectively (Goldberg, Oldehinkel, & Ormel, 1998). Analyses revealed that GHQ means did not vary between T1 and T2: Cohen's $d = .17$, $t = –1.35$, $N = 75$, $p = .18$; and the coefficient of stability was $r = .44$, $p < .01$. We used the absolute differences between T2 and T1 as an indicator of state distress variation between T1 and T2. However, the fact that participants were excluded if they had more than two positive items limited our ability to use the GHQ–12 as a moderator index of distress, and thus the results involving this measure should be cautiously considered.

Descriptive statistics for the variables used in the analyses are presented in Table 2. The overall mean and standard deviations of the core variables, such as Lambda and DEPI, were close to other nonpatient samples observed in various countries around the world (Erdberg & Schaffer, 1999). An alpha level of .05 was used for all statistical tests.

Rorschach Interrater Reliability and Quality of Data

To guarantee accurate scores, we adopted a consensus scoring procedure. First, the area coordinator scored all protocols. The protocols were then rescored blind by an independent rater. These scores were compared, and each area coordinator then decided on the adoption of the final scores. All the members of the scoring team were ignorant of the identification of each protocol and could not relate test and retest for the same individual.

Of the 150 protocols for T1 and T2, 25% (40 protocols = 20 tests & 20 retests) were randomly selected and rescored independently by one of three other psychologists who were blind to the initial consensus scoring and had not been involved in the consensus process. The total number of responses in the 40 protocols was 1,027 (M R = 25.7). Interrater agreement was calculated for the variables used in subsequent analyses at the protocol level of summary scores using the exact agreement for a single-rater ICC according to a one-way random effects model (see Table 3). The ICCs had a mean and median of .86 and .89, respectively. The standard deviation was .11 and the 25th and 75th percentiles were .81 and .95 with approximately the same pattern of results for T1 and T2. According to established criteria (Chiccetti, 1994; Chiccetti & Sparrow, 1981; Shrout & Fleiss, 1979), ICC was reasonable for C (ICC = .40 to .59) and good for CF, MOR, and X – % (ICC = .60 to .74). For all other variables, it was excellent (ICC = .75 to 1.00).

To ensure the integrity of the data entry procedures, we systematically reviewed the database. We did this in a number of different ways by (a) generating scatter plots for fixed variables, such as age and years of education, to identify matching errors; (b) identifying five outliers in test–retest correlations for R, Pure F%, and TE, and checking each case to verify that the correct participants were properly

able is (using sample based z-transformed Rorschach scores): .436 (Col Shd Blends) + .372 (FY) + .325 (FC') + .3 (FC) + .3 (CF + C) + .29 (m) + .29 (R) + .27 (S) + .24 (FM) + .22 (FV) + .21 (W) + .19 (MOR) + .18 (M) – .24 (L).

TABLE 2
Description of Variables Used in Subsequent Analyses

	T1					*T2*							
	M	*SD*	*Skew*	*Kurt.*	*Freq.*	*M*	*SD*	*Skew*	*Kurt.*	*Freq.*	*t*	*p*	*d*
Variables													
R	23.93	7.37	2.05	6.65	75	23.60	6.56	1.10	1.33	75	0.575	.57	.05
P	5.75	1.85	0.15	1.13	75	5.75	1.72	–0.46	0.62	74	0.000	1.00	.00
Zf	13.19	5.11	0.95	1.69	75	12.41	5.09	0.54	0.74	75	1.911	.06	.15
Zd	–1.61	4.37	–0.10	–0.23	75	–0.53	3.84	0.00	0.33	75	–2.176	.03*	–.26
F	9.44	5.15	1.12	2.42	75	9.33	5.89	0.88	0.76	73	0.214	.83	.02
M	3.92	2.45	0.60	0.10	70	3.91	2.46	0.61	0.36	70	0.067	.95	.00
FM	3.77	2.32	1.19	1.85	74	3.57	2.37	0.60	–0.21	69	0.720	.47	.09
m	1.71	1.61	1.24	2.32	54	1.59	1.32	0.99	0.98	60	0.682	.50	.08
FM + m	5.48	3.11	1.49	2.97	75	5.16	2.84	0.61	–0.24	74	0.932	.35	.11
a	4.68	3.10	1.13	1.50	73	4.89	2.88	0.41	–0.61	73	–0.697	.49	–.07
p	4.81	3.29	0.95	0.38	74	4.21	2.56	0.58	0.07	73	1.816	.07	.20
FC	2.49	1.92	0.83	0.68	64	2.75	2.19	1.18	1.39	66	–1.198	.23	–.13
CF	1.67	1.46	0.82	0.22	57	1.53	1.65	1.47	2.77	49	0.719	.47	.09
C	0.36	0.58	1.40	1.02	23	0.33	0.55	1.44	1.18	22	0.300	.77	.05
CF + C	2.05	1.72	0.76	0.05	60	1.87	1.77	1.34	2.05	58	0.980	.33	.10
WSumC	3.45	2.29	0.99	1.33	75	3.41	2.13	0.72	0.37	75	0.232	.82	.02
S	2.93	2.04	0.91	0.52	70	2.81	2.14	1.14	1.39	68	0.645	.52	.06
Sum T	0.88	1.04	1.36	2.24	41	0.84	1.03	1.18	0.91	38	0.359	.72	.04
Sum C'	1.60	1.59	1.00	0.07	54	1.93	1.86	0.97	0.34	55	–1.498	.14	–.19
Sum Y	1.33	1.45	1.13	0.85	47	1.04	1.32	1.56	2.57	40	1.424	.16	.21
Sum V	0.79	0.98	0.98	–0.18	36	0.88	1.08	1.44	2.33	40	–0.757	.45	–.09
Sum Shd	4.60	3.17	0.52	–0.55	70	4.69	3.30	1.00	0.92	72	–0.233	.82	–.03
FD	1.04	1.13	1.13	0.65	46	0.83	0.91	0.91	0.05	42	1.793	.08	.20
Fr + rF	0.39	0.70	1.78	2.54	21	0.37	0.82	2.40	5.76	16	0.178	.86	.03
Pairs	6.47	3.52	1.63	4.47	75	6.97	3.52	1.04	1.37	75	–1.853	.07	–.14
DV + DR	1.37	2.02	3.26	14.80	47	0.96	1.10	1.03	0.25	42	1.766	.08	.25
INC + FAB	1.87	2.01	1.47	2.35	52	1.96	2.11	1.58	2.75	55	–0.443	.66	–.04
COP	1.31	1.21	0.99	0.81	54	1.36	1.23	0.58	–0.60	52	–0.341	.73	–.04
AG	0.68	0.99	1.46	1.44	31	0.83	1.08	1.67	3.14	38	–1.169	.25	–.14
MOR	1.97	1.62	0.99	0.94	61	1.81	1.75	1.36	2.10	57	0.942	.35	.09
Ratios and percentages													
L	0.76	0.52	1.61	5.02	75	0.94	1.48	5.94	42.99	75	–1.356	.18	–.16
EA	7.37	3.87	0.75	0.71	75	7.31	3.57	0.36	–0.27	75	0.205	.84	.02
es	10.08	4.95	1.08	1.60	75	9.85	4.86	0.68	0.52	75	0.384	.70	.05
Adj es	8.39	3.71	0.62	0.19	75	8.56	3.88	0.30	–0.15	75	–0.379	.71	–.04
D	–0.72	1.47	–0.69	3.08	75	–0.76	1.42	–0.82	1.56	75	0.208	.84	.03
XA%	0.77	0.12	–0.66	0.91	75	0.79	0.11	–0.44	0.14	75	–1.327	.19	–.17
WDA%	0.81	0.11	–0.98	1.39	75	0.83	0.10	–0.62	0.04	75	–1.472	.15	–.19
X + %	0.54	0.14	–0.28	0.20	75	0.58	0.13	0.46	0.56	75	–2.613	.01*	–.30
X – %	0.22	0.12	0.84	1.34	75	0.20	0.11	0.60	0.41	75	1.251	.21	.17
Xu%	0.23	0.10	0.36	–0.49	75	0.21	0.09	0.05	–0.43	75	1.686	.10	.21
Afr	0.53	0.15	1.01	1.68	75	0.50	0.14	0.61	–0.41	75	2.218	.03*	.21
3r + (2)/R	0.32	0.15	0.96	1.79	75	0.35	0.15	0.54	0.55	75	–2.148	.04*	–.20
Sum6	3.45	3.36	2.57	10.37	75	3.09	2.80	1.44	2.52	75	1.003	.32	.12
WSum6	10.37	11.41	2.60	10.15	75	9.11	9.91	1.62	2.19	75	1.082	.28	.12
Blends	4.48	2.90	0.69	0.01	70	4.35	3.03	1.00	1.05	72	0.455	.65	.04
Intell	3.35	2.46	0.38	–0.64	62	3.05	2.26	1.09	1.23	68	1.113	.27	.13
Isolate/R	0.18	0.12	0.98	0.96	75	0.15	0.11	1.09	1.48	75	2.383	.02*	.26
Additional variables													
DEPI (total)	4.07	1.52	–0.40	–0.54	75	4.09	1.49	–0.16	–0.48	75	–0.128	.90	–.01
S–CON (total)	4.15	1.67	0.38	–0.58	75	3.75	1.61	–0.05	–0.39	75	2.047	.04*	.24
Task-engagement	0.00	2.11	1.14	1.64	75	–0.14	2.28	–0.03	2.13	75	0.642	.52	.06
GHQ–12[a]	0.84	1.15	2.01	5.27	75	1.07	1.52	2.22	6.25	75	–1.347	.18	–.17
GHQ–12[b]	9.05	2.36	.439	1.14	75	9.61	3.03	1.41	2.65	75	–1.544	.13	–.21
TE (*M* T1,T2)	–0.07	1.97	0.46	1.84	75								
TE variation	1.46	1.27	1.54	2.16	75								
GHQ–12 variation[a]	0.95	1.13	1.74	3.54	75								
GHQ–12 variation[b]	2.16	2.34	1.49	2.34	75								

Note. One person had no Popular at T2, 2 had no F at T2, and 5 had no M on both occasions. Values for Cohen's *d* were computed directly from the observed *M*s and *SD*s. T1 = baseline; T2 = retest; Kurt = kurtosis; DEPI (total) and S–CON (total) = the number of positive criteria for these indexes; TE (*M* T1,T2) and TE variation = average value of task-engagement of T1 and T2 and task-engagement variations between T1 and T2, respectively.

[a] Traditional coding (0, 0, 1, 1). [b] Likert coding (0 to 1 to 2 to 3).

*$p < .05$.

TABLE 3
Protocol-Level Interrater Reliability of Summary Scores Using the Exact Agreement ICC According to a One-Way Random Effects Model

	ICC		
	T1[a]	*T2*[a]	*T1 + T2*[b]
Variable			
R	1.00	1.00	1.00
P	0.84	0.86	0.89
Zf	0.96	0.98	0.97
Zd	0.77	0.86	0.81
F	0.99	0.98	0.99
M	0.91	0.96	0.95
FM	0.92	0.94	0.94
m	0.85	0.73	0.80
FM + m	0.95	0.92	0.94
a	0.96	0.94	0.95
p	0.87	0.92	0.91
FC	0.87	0.94	0.90
CF	0.65	0.66	0.65
C	0.42	0.47	0.45
CF + C	0.79	0.85	0.83
WSumC	0.91	0.97	0.94
S	0.88	0.95	0.92
Sum T	0.70	0.90	0.81
Sum C'	0.72	0.92	0.83
Sum Y	0.92	0.67	0.78
Sum V	0.88	0.89	0.88
Sum Shd	0.88	0.90	0.89
FD	0.76	0.74	0.76
Fr + rF	1.00	1.00	1.00
Pairs	0.98	0.98	0.98
DV + DR	0.63	0.58	0.60
INC + FAB	0.70	0.89	0.81
COP	0.89	0.80	0.85
AG	0.92	0.92	0.92
MOR	0.68	0.78	0.74
Ratios and percentages			
L	0.97	0.99	0.99
EA	0.93	0.98	0.96
es	0.95	0.95	0.95
Adj es	0.88	0.96	0.93
D	0.73	0.90	0.83
XA%	0.71	0.89	0.82
WDA%	0.75	0.86	0.81
X + %	0.77	0.90	0.83
X – %	0.75	0.65	0.69
Xu%	0.66	0.85	0.78
Afr	1.00	1.00	1.00
3r + (2)/R	0.93	0.98	0.96
Sum6	0.66	0.91	0.78
WSum6	0.67	0.89	0.78
Blends	0.94	0.94	0.94
Intell	0.84	0.86	0.85
Isolate/R	0.93	0.91	0.92
Additional variables			
DEPI (total)	0.85	0.85	0.85
S–CON (total)	0.77	0.88	0.84
Task engagement[c]	0.95	0.95	0.95

Note. Based on 40 Rorschach protocols; 1,027 responses. ICC = intraclass correlation coefficient; T1 = baseline; T2 = retest.
[a]*n* = 20. [b]*n* = 40. [c]Median ICC for the 13 modified scales of task-engagement = .95 (T1), .94 (T2), .95 (T1 + T2) with a minimum/maximum range of .94 to .97, .94 to .97, and.94 to .97, respectively.

matched; and (c) examining data entry or transcription problems in the Rorschach Calculations Program files by comparing each of the 150 scoring data files with the corresponding paper records.

We studied the differences in examiners' results on basic scores usually considered to be related to TE and the capacity to establish rapport: TE, as defined earlier; R; Pure F%; and EA. Pure F% is obtained by dividing F by R; it was recently introduced as a good alternative to Lambda for research purposes (Meyer, Viglione, & Exner, 2001). Because six examiners contributed fewer than eight protocols to the study, we computed mean ranks for each variable by examiner and tested the equality of mean ranks using a Kruskal–Wallis test. This analysis was performed for all 150 protocols taken together as a single sample (75 tests & 75 retests). Significant differences were found between examiners for TE, $\chi^2(11, N = 150) = 27.92, p < .01$; R, $\chi^2(11, N = 150) = 40.62, p < .001$; and EA, $\chi^2(11, N = 150) = 34.65, p < .001$; but not for Pure F%, $\chi^2(11, N = 150) = 16.54$, *ns*.

We then counted the number of values below or above the median of the whole sample for each examiner. We considered that values would be lower for an examiner if the number of values below the median was more than twice the number of values above the median (and the reverse for higher values). We observed that Examiners 2 and 4 had lower values for TE; Examiners 4, 9, 15, and 16 had lower values for R; and Examiners 2 and 5 had lower values for EA. Overall, this analysis suggested that the quality of administration might have been poorer for the examiners mentioned and especially Examiners 2, 4, and 15, who respectively contributed 26, 5, and 19 protocols to the study. Their protocols were checked for administration errors by members of the working team (AA, CR, GL, and SS), and no systematic errors could be diagnosed. In addition, Examiner 15 administered the test in a very different setting compared to the other examiners (a sports club as opposed to a private company), which prevented us from distinguishing between an examiner or setting effect. We therefore decided to retain all the protocols but to consider potential examiner effects in subsequent moderation analyses.

RESULTS

Preliminary Analyses

The statistical distribution of Rorschach and self-report variables was systematically examined to help us make correct decisions on stability coefficients. For instance, Dunlap, Burke, and Greer (1995) suggested that a high degree of skew can suppress the correlation between two variables. Among 51 variables included in this analysis, 30 showed deviations from a normal distribution as assessed by graphical analyses using a normal probability plot, a significant Kolmogorov–

Smirnov test, and a high ratio of skewness and kurtosis to their standard errors.[4]

These variables could be examined with Spearman rank-order correlations (ρ). However, as will be observed in subsequent analyses (cf. Table 4), the values for linear and rank-order correlation coefficients were close to each other (all differences < .10). For the following variables a difference in favor of rank-order correlation greater than .05 was found, thus suggesting that a low correlation may, in part, be artificially due to distribution issues: FM (*r* = .48, ρ = .54), m (*r* = .47; ρ = .56), FM + m (*r* = .50, ρ = .59), FC (*r* = .61, ρ = .69), C (*r* = .08, ρ = .16), CF + C (*r* = .55, ρ = .62). Conversely, in some cases, *r* could overestimate stability, such as in the case of Sum T, Sum V, and L, for the same reasons.

As can be observed from Table 2, most variables with problematic distributions had positive skew. To decrease problematic positive skew, we applied the procedure recommended by Behrens (1997, pp. 145–150) by subsequently raising variables to increasing negative and then positive exponents. In this process, to prevent the distribution from reversing the order of observations, negative reciprocals were applied to negative values of the initial scores. We used the following exponents and transformations: $–x^{–2}$, $–x^{–1}$, $–x^{–1/2}$, $\log_{10}(x)$, ln(x), $x^{1/2}$, x^1, x^2, x^3. However, all such transformations are not possible when values are negative or equal to zero. In count variables, we added 1.0 prior to transformation. In ratios (such as Intell or XA%), we added .05, a value that one more count would yield in an average 20-response protocol. For Zd, D, and TE we added 13.0, 6.0, and 9.0, respectively, depending on the minimum values of the range. When considering the transformation that was most beneficial, (i.e., that minimized skew and kurtosis) for each particular variable, the *M r* of the optimally transformed scores was .59 (as compared to .53 for the raw variables).[5] Given the limited magnitude of differences between coefficient types and between transformed and nontransformed variables, we decided to rely on *r* in nontransformed variables. This also facilitates a comparison of our results with previously published research (cf. Exner, 2003b).

Stability Coefficients for Dimensional Variables

Mean-level analysis. Overall changes between T1 and T2 could be identified using Student's *t* and Cohen's *d* between T1 and T2 for all the variables described in Table 2. Using Cohen's (1992) thresholds, no difference could be labeled as "large" (i.e., *d* ≈ .80). Differences were small (i.e., *d* ≈ .20) for the following variables, with higher values at T2 for Zd, Sum C', WDA%, X + %, EGO, and GHQ–12, and lower values for p, Sum Y, FD, DV + DR, Xu%, Afr, Isolate/R, and S–CON. For all the other variables, *d* could be considered as negligible. In general the findings show that participants were less puzzled by the task at T2 and became involved in the task more easily (better form quality, lower special scores).[6]

Rank-order analysis. Because correlation coefficients are sensitive to extreme values, we conducted an exploratory analysis to identify outliers in intercorrelations using a casewise diagnostic procedure to identify extreme standardized residuals in simple regression analyses. Although some protocols appeared as outliers in several correlation analyses, no noticeable correlation differences were observed when outliers were excluded from the analyses. We therefore computed *r* and ρ in the overall sample for variables already studied in test–retest adult samples. To permit comparisons with other reliability studies (Meyer et al., 2002), Pearson's correlation coefficients between T1 and T2 for the entire structural summary are presented in the Appendix. For the 47 variables previously studied (Table 4), the mean *r* was .51 (*Mdn* = .53, min = .08, max = .78) and *M* ρ was .51 (*Mdn* = .53, min = .15, max = .76).

In the full structural summary, the most stable variables (*r* > .70) appeared to be W, R, Zf, EA, EGO, D location, M, L, H + A, Hd + Ad, H, S, DQo, DQ+, Cg, Ge, Xy. The least stable (*r* < .20) were Col Shd Bl, C, C', VF, FY, Y, (A), AB, PSV, DV, DV2, DR2, ALOG, (A + Ad), Mnone (see Appendix). Note that a majority of these variables have extreme base rates (< .05).

We then examined differential stability levels in construct-related variables, as defined in the Method section. Emotional, coping, and m and Y influenced variables yielded a mean *r* of .46 (± .13; 10 variables). Personality, cognitive, and self or relational variables yielded a mean *r* of .70 (± .09; 11 variables). When ranking coefficients for these 21 variables, we observed that the mean ranks were 6.10 and 15.45 for the two types of variables, thus indicating greater stability in personality, cognitive, or self/relational variables (U = 6.00, *p* < .01).

Stability coefficients for styles and cut-off scores. Some categories may be of major importance in the interpretation process. They are usually defined by predetermined interpretive cut-points (e.g., Exner, 2003b). Exner et al. (1978) dedicated much of their article to examining the consistency and shifts in direction of widely used ratios. This is a good way of answering simple questions like, "In this sample, what are the chances for an extratensive at T1 to be-

[4] These variables were R, F, FM, m, FM + m, p, FC, CF, C, CF + C, WSum C, Space, Sum T, Sum C', Sum Y, Sum V, FD, Fr + rF, (2), DV + DR, INC + FAB, COP, AG, MOR, L, es, D score, Sum6, WSum6, Isolate/R, GHQ–12.

[5] Increases in *r* of more than .10 were observed for individual transformations in P (r_{-x}^{-2} = .82, *r* = .54), m (r_{-x}^{-1} = .63, *r* = .47), CF + C (r_{-x}^{-2} = .72, *r* = .55), and WDA% (r_{-x}^{-2} = .55, *r* = .45).

[6] A full description for all variables mentioned by Exner (2002, Table 1) can be obtained on request.

TABLE 4
Correlation Coefficients for Several Test–Retest Studies in Adult Nonpatients

	3 Month	*3 Month*	*3 Week*	*1 Year*	*Meta-Analysis*
	(Our Study)[a]	*(Our Study)*[b]	*(Exner, 2003)*[c]	*(Exner, 1999)*[d]	*(Grønnerød,*
	r	ρ	*r*	*r*	*2003)*[e] r_w
Variable					
R	.75	.76	.84	.86	.84
P	.54	.56	.81	.83	.77
Zf	.76	.61	.89	.85	.83
Zd	.46	.48	—	—	—
F	.70	.73	.76	.74	.72
M	.76	.76	.83	.84	.82
FM	.48	.54	.72	.77	.70
m	.47	.56	.34	.26	.53
Fm + m	.50	.59	—	—	.66
a	.61	.56	.87	.83	.82
p	.55	.53	.85	.72	.77
FC	.61	.69	.92	.86	.84
CF	.47	.51	.68	.58	.53
C	.08	.16	.59	.56	.57
CF + C	.55	.62	.83	.81	.76
WSum C	.69	.72	.83	.82	—
S	.70	.68	—	—	—
Sum T	.56	.47	.96	.91	.91
Sum C'	.38	.37	.67	.73	.70
Sum Y	.17	.15	.41	.31	.40
Sum V	.46	.36	.89	.87	.81
Sum Shd	.42	.39	.71	—	.63
FD	.51	.41	.90	.88	.86
Fr + rF	.65	.65	.89	.82	.86
(2)	.77	.74	.83	.81	.82
DV + DR	.26	.16	—	.72	—
INC + FAB	.61	.50	—	.89	—
COP	.38	.43	.88	.81	—
AG	.45	.53	.81	.82	—
MOR	.62	.54	.83	.71	—
Ratios and percentages					
L	.72	.65	.76	.78	.76
EA	.77	.74	.84	.83	.81
es	.46	.41	.59	.64	.68
Adj es	.46	.39	.79	.82	.83
D	.34	.39	.88	.91	.80
XA%	.49	.45	NC	.89	—
WDA%	.45	.31	NC	.92	—
X + %	.55	.58	.87	.86	.84
X – %	.51	.45	.88	.92	.91
Xu%	.32	.29	.89	.85	.87
Afr	.57	.54	.85	.82	.84
3r + (2)/R	.78	.74	.90	.89	.85
Sum6	.50	.33	.81	.81	—
WSum6	.56	.43	.86	.86	—
Blends	.63	.64	.71	.62	.73
Intell	.53	.49	NC	.84	—
Isolate/R	.67	.60	.83	.84	—
Additional variables					
DEPI (total)	.28	.26	—	—	—
S–CON (total)	.47	.44	—	—	—
Task-engagment	.61	.50	—	—	—
GHQ–12	.43	.39	—	—	—

Note. Mean correlations for modified task-engagement scales were .61, with minimum = .57, and maximum = .66. NC = not coded or calculated in the Comprehensive System at the time of the study.

[a] *N* = 75, *Mdn r* = .53. [b] *N* = 75. [c] *N* = 35, *Mdn r* = .83. Sample from Thomas, Alinsky, and Exner (1982), additional calculations provided by Viglione and Hilsenroth (2001). [d] *N* = 50, *Mdn r* = .82. Sample from Exner, Thomas, and Cohen (1983). [e] *N* = 350, *Mdn r* = .81. Weighted average correlations from Comprehensive System only samples (Grønnerød, 2003, Table 4, p. 283).

TABLE 5
Stability Frequencies and Percentages for Interpretively Significant Indexes and Ratios (2 × 2 Tables)

	T2		%		
Indexes at T1	*Positive*	*Negative*	*Consistent From T1 to T2*	ϕ	κ
S–CON					
Positive	0	2	0.0		
Negative	0	73	100.0	.00	.00
HVI					
Positive	4	6	40.0		
Negative	7	58	89.2	.28*	.28*
OBS					
Positive	0	0	NA		
Negative	2	73	97.3	.00	.00
SCZI					
Positive	5	5	50.0		
Negative	9	56	86.2	.31**	.31**
DEPI					
Positive	17	14	54.8		
Negative	15	29	65.9	.21	.21
CDI					
Positive	10	10	50.0		
Negative	11	44	80.0	.30*	.30*
L ≥ 1.00					
Positive	11	11	50.0		
L < 1.00					
Negative	11	42	79.2	.29*	.29*
Fr + rF > 0					
Positive	13	8	61.9		
Fr + rF = 0					
Negative	3	51	94.4	.62***	.61***

Note. T1 = baseline; T2 = retest.
*$p < .05$. **$p < .01$. ***$p < .001$.

come an introversive at T2?” Consequently, we calculated frequencies in 2 × 2 tables for indexes, such as the DEPI, and other dichotomously interpreted variables, such as Fr + rF > 0. The results are reported in Table 5, where the consistency percentages describe the frequency of people who retain their baseline characteristics at T2. Shift percentages can easily be derived from the presented data. For scores typically interpreted according to three categories, we calculated frequencies in 3 × 3 tables, which are presented in Table 6. For each variable, the outer columns and rows in the nine-cell matrix describe the directions of the ratios, whereas the central column and row represent the individuals for whom there was no established ratio direction.

As far as the dichotomous indexes are concerned, the consistency percentages suggest that the absence of a positive index was a fairly stable characteristic, specifically for S–CON, OBS, HVI, SCZI, CDI, and to a lesser extent for DEPI, whereas the presence of a positive index was more likely to change between T1 and T2, with around 50% of the initially positive scores shifting to below the threshold at retest. This is consistent with the usual observation that Positive Predictive Power (or the consistency percentage for positive indexes) is lower than Negative Predictive Power (or the consistency percentage for negative indexes) when a positive index has a low base rate (Streiner, 2003).

Of the consistency percentages in Table 6, the most stable conditions were Fr + rF = 0, EGO > .44, EGO < .33, EB introversive (both for 1978 and 1993 definitions, see note of Table 6), EB ambitent (1993), a > p + 1, T = 0, |Zd| = 0 to 3.0, EGO < .33, FM + m > Sum Shd, and Sum Shd > FM + m, with consistency percentages of 65% and higher. The least stable were CF + C > FC + 1, FC > CF + C + 1, D > 0, EA > es + 1, |EA – es| = 0 or 1, EB extratensive (2003), EB ambitent (1978), p > a + 1, |a – p| = 0 or 1, Zd > + 3.0, T = 1, T > 1, FM + m > Sum Shd + 1, |FM + m – Sum Shd| = 0 or 1, with stability percentages lower than 50%. The ratios that remained fairly stable regardless of the initial direction at T1 were the EGO index and eb (FM + m:Sum Shd). The other conditions that remained stable concerned one of the two directions of the ratio (see Table 6). For example, it is noteworthy that protocols with predominantly active movements at T1 were more stable than protocols with predominantly passive ones; the same is true for EB introversive, as opposed to EB extratensive. However, analyzing stability exclusively through stability percentages can be misleading because this approach does not take shifts into consideration. This is why we indicate ϕ, Cramer’s V, and κ coefficients in Tables 5 and 6. On the basis of these criteria, the most stable variables appear to be Fr + rF, EGO and EB (1993), with ϕ/V coefficients of .62, .56, and .46, respectively.

Moderators of Stability

We computed hierarchical regression models for the primary Rorschach variables. For each, the dependent variable was the T2 score, whereas the independent variables were the T1 score and the proposed moderators. The T1 score was entered in Block 1, and unstandardized residuals were saved for further analysis. The moderators were then entered individually as alternative Block 2s with the residuals as the dependent variable. We also computed an alternative global model with all moderators entered in Block 2. Within this approach, the semipartial correlation associated with each moderator was determined after the contribution of the T1 scores had been fixed. The main moderators under consideration were TE (overall level and T1/T2 variations) and state measure of emotional distress (GHQ–12 variations between T1 and T2).

Do moderators explain additional variance? To answer this question, we tested variance changes (ΔR^2) to estimate whether moderators could explain a significant proportion of the variance in T2 scores beyond what can be predicted on the basis of the T1 scores. (The results are summarized in Table 7, last columns, Model 4.) A positive answer was given for 14 variables out of 49. For es and Adj es, the increase in variance was quite substantial with $\Delta R = .55$

TABLE 6
Stability Frequencies and Percentages for Interpretively Significant Indexes and Ratios (3 × 3 Tables)

	T2					
Indexes at T1	*First Category*	*Second Category*	*Third Category*	*% Consistent From T1 to T2*	*V*	*κ*
EB (1978) Introversive	20	5	1	76.9		
EB (1978) Ambitent	8	14	10	43.8	.43***	.38***
EB (1978) Extratensive	2	5	10	58.8		
EB (1993) Introversive	16	5	1	72.7		
EB (1993) Ambitent	7	25	5	67.6	.46***	.45***
EB (1993) Extratensive	2	6	8	50.0		
EB (2003) Introversive	8	4	1	61.5		
EB (2003) Ambitent	3	11	4	61.1	.38*	.34**
EB (2003) Extratensive	1	5	5	45.5		
EA > es + 1	5	3	7	33.3		
|EA – es| = 0 or 1	2	4	5	36.4	.21	.16
es > EA + 1	4	14	31	63.3		
D score < 0	22	12	1	62.9		
D score = 0	12	16	4	50.0	.26*	.20*
D score > 0	1	5	2	25.0		
FM + m > Sum Shd + 1	16	11	7	47.1		
|FM + m – Sum Shd| = 0 or 1	8	9	6	39.1	.25	.20*
Sum Shd > FM + m + 1	2	6	10	55.6		
FM + m > Sum Shd	30	6	9	66.7		
FM + m = Sum Shd	5	0	2	0.0	.32**	.28**
Sum Shd > FM + m	6	2	15	65.2		
a > p + 1	20	6	2	71.4		
|a – p| = 0 or 1	5	7	8	35.0	.35**	.24**
p > a + 1	6	11	10	37.0		
FC > CF + C + 1	11	8	4	47.8		
|FC – CF + C| = 0 or 1	9	23	4	63.9	.21	.17*
CF + C > FC + 1	3	10	3	18.8		
Zd < –3.0	12	12	0	50.0		
|Zd| = 0 to 3.0	6	29	7	69.0	.37***	.29**
Zd > +3.0	0	5	4	44.4		
T = 0	24	7	3	70.6		
T = 1	9	8	7	33.3	.31**	.27**
T > 1	4	5	8	47.1		
EGO < .33	30	12	1	69.8		
EGO = .33 to .44	5	9	3	52.9	.56***	.48***
EGO > .44	1	2	12	80.0		

Note. EB styles were computed in the 1978 fashion (differences over one point between the two sides), in the 1993 fashion (i.e., differences = 2, when EA = 10, and differences > 2, when EA > 10), and in the 2003 fashion (i.e., same way as 2003 but excluding protocols with L ≥ 1.00). No systematic change between T1 and T2 could be detected in frequency tables using a McNemar test (all *p*s > .18). T1 = baseline; T2 = retest.
$^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

and .51, respectively ($p < .001$). All 14 of these variables had initial test–retest correlations lower than .75 (i.e., $r^2 < .50$), thus indicating that moderators can, to some extent, account for instability. For Sum C', Sum Shd, es, Adj es and D score, we observed that the semipartial correlation with individual moderators or the ΔR with all moderators even exceeded the test–retest correlations (i.e., the r values in the Block 1 column). In these cases, the moderators can be considered to be particularly powerful enhancers of stability. Finally, in certain variables, most of the T2 score variation was accounted for by the predictors with a total R close to .75: F, a, WSumC, Fr + rF, EA, and Blends. These are major variables in the interpretation process, and it should be noted that few other sources of variation are able to explain instability. This signifies that, at least for these variables, the moderation analysis is reasonably adequate and saturated.

Which moderators? We then focused on each moderator in turn as an alternative independent variable in Block 2. The results are summarized in Table 7, columns 2 through 8, Models 1, 2, and 3.[7]

Task engagement, as measured by the different versions of the TE scale, appeared to be a moderator of stability in 15 of the 49 variables studied here. Higher levels of engagement were associated with higher stability in m, FM + m, a, FC, Sum C', Sum V, Sum Shd, Fr + rF, INC + FAB, COP, EA, es, Adj es, EGO, and Blends. Lower levels of engagement were associated with higher stability in F. For Sum C', es, and Adj es, semipartial correlations with this moderator were about

[7] A full table of F, p, and coefficient values can be obtained from the first author on request.

TABLE 7
Summary of Hierarchical Regression Analyses With T2 Scores As the Dependent Variable and T1 Scores and Moderators As Independent Variables

		Block 2 Single Moderators								
	Block 1	*Model 1*		*Model 2*		*Model 3*		*Block 2 All Moderators: Model 4*		
	T1 Score r	*TE (M) ΔR*	*Total R*	*TE (Variation) ΔR*	*Total R*	*Distress (Variation) ΔR*	*Total R*	*ΔR*	*Total R*	*Adj. R*
Variables										
R	.746	–.037	.747	–.251*	.787	.020	.746	.251	.787	.773
P	.544	–.009	.544	–.161	.567	.128	.559	.205	.581	.548
Zf	.764	.200	.790	–.162	.781	.146	.778	.305	.823	.811
Zd	.464	.201	.506	.136	.484	.248*	.526	.319	.563	.527
F	.702	–.226*	.737	–.250*	.745	–.057	.704	.313*	.769	.753
M	.756	.089	.761	–.105	.763	–.012	.756	.153	.771	.756
FM	.475	.161	.502	–.246*	.535	.149	.498	.347*	.588	.556
m	.472	.249*	.534	–.116	.486	.288*	.553	.389**	.612	.582
FM + m	.503	.237*	.556	–.216	.547	.261*	.567	.424**	.658	.633
a	.609	.362**	.708	–.044	.611	.256*	.661	.429**	.745	.728
p	.546	.147	.565	–.234*	.594	.003	.546	.303	.624	.596
FC	.611	.283*	.673	.103	.620	–.056	.614	.306	.683	.661
CF	.473	.102	.484	.010	.473	.098	.483	.133	.491	.445
C	.084	.076	.113	–.113	.141	.335**	.345	.364*	.374	.301
CF + C	.553	.083	.559	–.085	.559	.245*	.605	.268	.615	.585
WSumC	.692	.212	.724	.009	.692	.278*	.746	.329*	.766	.751
S	.703	.135	.716	–.161	.721	.003	.703	.232	.740	.723
Sum T	.564	.092	.571	.027	.565	.112	.575	.137	.580	.547
Sum C'	.384	.406***	.559	–.004	.384	.028	.385	.413**	.564	.528
Sum Y	.170	.189	.254	.003	.170	.091	.193	.202	.264	.129
Sum V	.463	.237*	.520	–.123	.479	.089	.471	.290	.546	.508
Sum Shd	.424	.388**	.575	–.145	.448	.103	.436	.437**	.609	.579
FD	.507	.031	.508	–.019	.507	–.275*	.577	.286	.582	.549
Fr + rF	.645	.302**	.712	–.125	.657	.011	.645	.364*	.741	.723
Pairs	.774	.084	.779	–.073	.777	–.009	.774	.125	.784	.770
DV + DR	.263	.057	.269	–.011	.263	.019	.264	.062	.270	.142
INC + FAB	.608	.246*	.656	–.046	.610	.162	.629	.290	.674	.650
COP	.381	.231*	.446	.197	.429	–.014	.381	.284	.475	.426
AG	.452	.195	.492	.090	.461	.122	.468	.225	.505	.461
MOR	.620	.184	.647	.002	.620	–.055	.622	.207	.654	.628
Ratios and percentages										
L	.718	–.108	.726	–.057	.720	–.082	.723	.129	.729	.711
EA	.771	.220	.802	–.009	.771	.180	.792	.282	.821	.810
es	.457	.464***	.651	–.069	.462	.239*	.516	.552***	.717	.697
Adj es	.455	.458***	.646	–.020	.455	.171	.486	.512***	.685	.662
D	.337	–.212	.398	.175	.380	–.071	.344	.347*	.484	.436
XA%	.494	.026	.495	–.053	.497	.040	.496	.071	.499	.454
WDA%	.446	–.140	.467	.131	.465	–.011	.446	.204	.490	.444
X + %	.553	.023	.553	.084	.559	–.029	.554	.091	.560	.524
X – %	.512	–.025	.513	.042	.514	–.109	.523	.117	.525	.484
Xu%	.316	.024	.317	–.114	.336	.105	.333	.155	.352	.272
Afr	.571	–.093	.579	.001	.571	–.107	.581	.133	.586	.553
3r + (2)/R	.779	.313**	.840	–.045	.780	–.052	.781	.339	.850	.840
Sum6	.503	.205	.543	–.012	.503	.062	.507	.213	.546	.508
WSum6	.556	.178	.584	.004	.556	.061	.559	.184	.586	.553
Blends	.634	.366**	.732	.095	.641	.174	.657	.391**	.745	.728
Intell	.534	.141	.552	–.019	.534	.143	.553	.191	.567	.532
Isolate/R	.666	.056	.668	–.021	.666	–.148	.682	.170	.687	.665
DEPI (total)	.284	.192	.343	–.074	.293	.084	.296	.228	.364	.288
S–CON (total)	.470	.115	.484	–.131	.488	.143	.491	.228	.522	.481

Note. *N* = 75. T1 = baseline; T2 = retest; TE = task engagement.
**p* < .05. ***p* < .01. ****p* < .001.

the same magnitude as the test–retest correlations. This shows that for these variables, TE was a strong moderator of stability. Because R is a significant component of TE, supplemental analyses examined the moderating role of the average number of responses alone. Only two coefficients were significant. Higher levels of *R* were associated with higher stability in FC ($\Delta R = .288, p < .05$) and Sum V ($\Delta R = .244, p < .05$). Changes in TE between test and retest were identified as moderators in 4 variables, with less change being associated with higher stability in R, F, FM, and p. Supplemental analyses showed that variation in R, as defined by the absolute value of $R_{T1} - R_{T2}$, was a moderator in 9 variables with less variation being correlated with higher stability in Populars ($\Delta R = -.236, p < .05$), Zf ($\Delta R = -.372, p < .01$), m ($\Delta R = -.289, p < .05$), CF + C ($\Delta R = -.260, p < .05$), Sum C' ($\Delta R = -.246, p < .05$), MOR ($\Delta R = -.257, p < .05$), EA ($\Delta R = -.297, p < .05$), es ($\Delta \mathrm{R} = -.279, p < .05$), and Blends ($\Delta R = -.347, p < .01$).

We also identified state distress as a moderator for nine variables, with higher variability in distress from T1 to T2 being associated with a lower stability in Zd, m, FM + m, a, C, CF + C, WSumC, and es. Lower distress variations were associated with a higher stability in FD (see Table 7, columns 7 to 8, Model 3). When using the Likert coding of the GHQ–12, we observed the same pattern of results: State distress was identified a moderator for FM + m ($\Delta R = .230, p < .05$), CF + C ($\Delta R = .231, p < .05$), C ($\Delta R = .295, p < .01$), FD ($\Delta R = -.230, p < .05$).

To further explore the impact of distribution issues, we examined regression models using transformed variables. When using the most beneficial transformation for the test–retest correlation (Block 1), few differences were observed in the moderation analyses (Models 1 to 3, Block 2): for m, the average TE level and state distress became nonsignificant ($\Delta R = .221$ and $\Delta R = .199$, respectively). For CF + C, state distress became nonsignificant ($\Delta R = .139$). Other results remained at the same significance level.

Exploring the effect of examiners. To assess potential examiner effects on the stability of certain variables, we compared test–retest rank-order correlation coefficients for different pairs of examiners: A test–retest involved two examiners because no examiner could administer the test twice to the same person. We systematically compared correlations for seven key variables: R, Zf, Pure F%, EA, es, D score, and TE. Rank-order correlation was chosen because of the limited size of the subsamples in the comparison. When subsamples with a size of 5 or higher were considered, we made two kinds of observations: First, there was a high disparity between pairs of examiners on stability coefficients: For example, for es and D score, the range for coefficients was –.03 to .68 and –.03 to .69, respectively. Second, some pairs of examiners were "well below" the level of the total sample in terms of stability: Pair 2–1 (es and D score, $N = 12$) and Pair 4–5 (R, $N = 5$). Although in our study this type of analysis can only be exploratory in nature given the limited subsample sizes, it partly confirmed our preliminary impressions about examiner effects, and involved Examiners 2 and 4 (cf. Method section). Thus, participants who exhibited lower stability in the Rorschach were administered the test at least once by an examiner who generally obtained protocols with lower levels of TE, R, or EA (at T1 and T2). We computed stability coefficients in a subsample of 58 participants, excluding protocols from Examiner Pairs 2–1 and 4–5. The median *r* for the variables studied was .56, as compared to .53 when considering the total sample of 75 participants. As stated earlier, a close examination of protocols collected by these examiners was conducted by members of the team. They could not identify any obvious mistakes. As a result, no clear-cut conclusion can be drawn as to whether this examiner effect was systematic or random. In conclusion, it is probable that some examiner effect played a role in reducing stability. However, this would have had a relatively small impact in our study because it concerns a small number of protocols.

DISCUSSION

Hypothesized Stability

First, our results show that among the 47 variables that have been studied previously, 9 had stability coefficients above the .70 threshold, which is indicative of high stability, and 21 had coefficients above .50, thus indicating moderate stability. However, the overall level of stability across all variables was lower than expected (< .69). A comparison of our findings with those reported by Exner and in Grønnerød's (2003) recent meta-analysis reveals lower coefficients in our study. Using a 3-week and a 1-year interval, Exner reported a mean *r* of .79 (*Mdn* = .83, min = .34, max = .96) and .78 (*Mdn* = .82, min = .26, max = .92), respectively. Among 30 variables that were also examined by Grønnerød, the *M r* was .76 (*Mdn* = .81, min = .40, max = .91), where we found a mean of .54 for the same subset of scores (*Mdn* = .55, min = .08, max = .78). The overall level of stability in our study is much lower than that obtained in previous research, and it applies across variables, including characteristics that we found to be fairly stable: Zf, M, a, p, WSumC, and so on. The correlations for most of these variables are about .10 to .15 lower than those observed by Exner in the cited studies.

Second, the majority of the categories defined by cut-off scores were fairly unstable with two variables showing $\phi > .50$. Also, main indexes that are interpreted dichotomously did not produce evidence of stability when they were initially positive (e.g., DEPI, CDI). Two interesting results emerge from our analyses. First, when comparing stability for various definitions of the EB, we observe that the current definition, which excludes those with an avoidant style (i.e., with Lambda ≥ 1.00; Exner, 2003), was no more stable than previous ones (i.e., defined by the 1978 and 1993 criteria). More specifically, the most stable condition was the EB Introversive (1978) which was more stable than the EB Introversive (2003), with consistency percentages of 77% and 62%, respectively. Second, our results contrast with those previously reported by Exner et al. (1978) and Exner (1999) in their directionality analyses. For example, as far as the EB style (1978 definition) is concerned, of the 43 participants exhibiting a clear direction at T1, 33 had a clear direction at T2, and 3 of these 33 (9%) shifted as compared to 1 out of 40 (3%) reported by Exner (1999) in 50 nonpatient adults (data from Exner et al., 1983, 1-year interval). The same pattern emerges from a comparison of consistency and shifts for the FC:CF + C and a:p ratios. As previously mentioned, consistency and shifts percentages are base rate sensitive statistics. Given that the base rates were higher for each of these variables in Exner's studies (e.g., there were more people with a clear introversive or extratensive EB style), this might partly explain the observed differences. Overall, Rorschach stability was lower than expected in our first hypothesis.

The stability levels observed here are closer to the expectations made on the basis of Watson's (2004) review of the literature, which included other personality assessment procedures. The range of correlations observed in our study overlaps the mean range computed from 23 intermediate interval studies cited in Watson. Our coefficients ranged from .08 to .78 (.17 to .78 if C is excluded); Watson's coefficients ranged from an average minimum across the 23 studies of .63 and an average maximum of .79.

Our results actually reflect a wide range of stability levels for variables within the Rorschach. This observation should lead us to consider variables independently and avoid judgments made on the Rorschach "as a whole." In fact, R, Zf, F, M, S, Pair, Lambda, EA, and EGO exhibited stability levels comparable to those obtained with other instruments as reported in meta-analyses and individual empirical studies. Meyer (2004b) reviewed all the meta-analyses of test–retest reliability studies in the psychological literature. For intervals up to 12 months, weighted *r*s ranged from .38 to .92, with most instruments yielding test–retest correlations in the .70s and .80s. More recently, Jiang and Cillessen (2005) reported a similar aggregated correlation in the sociometric status of children, where a median weighted *r* above .70 was computed when the intervals were shorter than 3 months. For instruments on which a meta-analysis is not yet available, empirical results greatly vary. For example, a recent report by Egloff, Schwerdtfeger, and Schmukle (2005) studied the temporal stability of two scores on an indirect procedure for the assessment of test anxiety (Implicit Association Test). For a 1-month and 1-year interval, these authors found test–retest *r*s of .57/.62 and .49/.47, respectively.

So some of the Rorschach variables had test–retest correlations comparable to other instruments ($r > .70$). Yet for the majority of the variables studied, our results contrast with earlier findings reported in the Rorschach literature.

In line with the expected "hierarchy of consistency" (Conley, 1984) and previous results in the Rorschach and non-Rorschach literature, we observed differences between the stability levels of m, Y, and other state-related variables on one hand and trait-related measures on the other hand. Emotional, coping, and state-related measures were less stable than cognitive, personality, or self/relational measures. This is consistent with our second hypothesis and Grønnerød's (2004) findings on Rorschach changes during psychotherapy.

Finally, our results suggest that some moderators play a major role in the stability of Rorschach variables. In some variables with low test–retest correlations ($r < .50$), a high level of variance could be explained when state emotional distress and TE were considered (e.g., es, Adj es, Blends, Sum Shd, a, p, m). Emotional distress was demonstrated to be a moderator that, if controlled, would lead to increased stability in m, FM + m, a, es, FD, Zd, C, CF + C, WSumC, which is consistent with our third hypothesis. Although Active Movements, Zd, and Color Determinants were not expected to vary as a function of state emotions, the other correlations were in line with previous validation studies (as reviewed by Exner, 2003b). However, contrary to our hypothesis, we could not find evidence of any impact of state distress on Sum Y, DEPI, or S–CON stability. The relations observed here support the validity of m, FM + m, es, and FD because instability, or "error variance," is partly attributable to emotional change during the interval. These results should be cautiously interpreted though, given that participants were also excluded from the study on the basis of their GHQ–12 scores. This is a limitation for interpreting the GHQ–12 as an indicator of distress levels, as our selection procedure produced a truncated range of GHQ–12 scores.

However, if the coefficients of state distress appear modest (ΔR = .24 to .34), these results should not be overlooked for two reasons: First, CS stability data in relation to an external criterion of emotional states have never been reported in the past. Second, the state distress measure comes from a self-report; Rorschach variables and self-reports have repeatedly demonstrated low correlations even when they have been supposed to measure the same constructs.

The role of TE in stability could also be clarified. Individuals with a high overall level of TE exhibited higher stability on some interpretively important variables, such as Sum C', es, Adj es, a, Sum Shd, Fr + rF, EGO, and Blends. In line with our expectations, the results also reveal an effect on stability of variations in TE over testing occasions, although this was limited to 4 variables of the 49. Participants whose level of engagement changed at the retest demonstrated lower stability on variables such as R, F, FM, and p. This is consistent with the definition of task engagement and confirms its role in the expression of personal functioning on the Rorschach (Meyer, 1997b). Nevertheless, we identified no effect of TE, either overall level or variation, on M, WSumC and Color Determinants, Sum Y, DV + DR, Form Quality variables, or

Special Scores, although we had expected to observe an effect according to our fourth and fifth hypotheses.

We also observed a limited role for mean values of R, with this factor moderating stability in FC and Sum V. R's impact was similar to the overall level of TE, with approximately the same ΔR values. We identified the variation of R between test and retest as a more important moderator, as it impacted the stability of some interpretively significant variables, including Zf, EA, and es. This result parallels and extends previous analyses (Exner, 1988), which showed that participants who gave very few responses (R < 14) at test or retest had lower stability on the very same variables as those observed in our study: Populars, Zf, m, CF + C, Sum C', EA, and es.

The impact of engagement immediately raises the question of the determinants of engagement, in particular with reference to context and examiner factors. In effect, as we expected, our results show that each pair of examiners whose results indicated lower stability included one examiner who aroused lesser engagement. This result is compatible with previous analyses. Although no systematic administration error could be detected, it is probable that an examiner effect played a role here. However, this effect appeared to be rather limited, because mean stability levels did not change notably when protocols from the two least stable examiner pairs were excluded.

To summarize, although our results are subject to limitations (see next), they provide some arguments in support of (a) the importance of state negative emotions for the stability of selected CS variables, (b) the role of TE in some important personal features that may be expressed in the Rorschach, (c) the importance of responsivity variation in a number of core variables, and (d) a probable effect of examiner-induced attitude and context on stability. Nevertheless, for each of these aspects, and particularly for the latter, which could not be systematically investigated in our study, our research needs to be replicated in systematic, large-scale studies. However, although the moderation analysis seemed appropriate and somewhat saturated in some cases (e.g., F, WSumC, or Blends), it did not provide any explanation of the "error" variance for many scores (e.g., M, Sum Y, Form Quality variables). This requires us to consider other factors.

Alternative Explanations of Instability

Several factors may be responsible for low stability levels in our study. In our preliminary analyses, we showed that distribution issues may have impacted stability in some variables where r and ρ values diverged, such as FM, m, FM + m, FC, C, and CF + C. In these cases, it is probable that r underestimated the actual stability levels. Also, transformations brought significant increases in stability for the following variables: Populars, m, CF + C, and WDA%. In addition, although overall interrater reliability was excellent for most variables, C and CF had lower levels, with ICCs of .45 and .65. This, too, may account for some part of the error variance between test and retest. To assess this effect, a stability correlation corrected for attenuation was computed by dividing the raw stability correlation by the square root of the interrater reliability estimate. For C and CF, this corrected stability correlation equaled .12 and .58, respectively, as compared to raw stability values of .08 and .47.

Similar reasoning can be applied to the base rates. As already emphasized, infrequent codes may compromise stability coefficients. One way to approach this issue is to consider that extreme base rates are a form of range restriction. The more extreme a score's base rate, the smaller its variance. Although detailed information on base rates is not available from previous stability reports, a comparison with frequencies observed in published normative samples (Exner, 2002) leads us to think that this factor may explain some of the differences observed between our study and previously reported results. In particular, this could be the case for variables such as Sum C', FC, CF, or Sum T. For example, in Exner's last published reference sample, 74% of the respondents had at least one texture in their protocols, whereas in our sample this proportion was 55%. In our study, variables with extreme base rates less than .05 had lower stability coefficients than variables with base rates closer to .50 (*Mdn* = .48 and .56, respectively). These coefficients also exhibited a somewhat greater variability (*SD* = .18 and .13, respectively). These findings parallel the observations made by Viglione and Taylor (2003) on interrater reliability estimates.

State and trait factors beyond those considered in our moderation analyses likely also contribute to instability. For instance, our external state measure (GHQ–12) probably is too coarse to detect certain aspects of state variance. In fact, the GHQ–12 was designed as a screening instrument, not as a fine-tuned measure of distress, anxiety, or other negative emotions. Moreover, we focused on emotional states and not on other kinds of states and processes. Other factors that might contribute to instability are changes in the way participants approach the task the second time. Indeed, participants had a somewhat less defensive and more effective approach to testing at the retest. This could result from a habituation process with a task that is always described as puzzling when first administered. This suggestion needs to be confirmed by further research. In addition, given the 3-month retest interval we chose to study, it is possible that the "state" change was accompanied by "trait" change. However, changes were greater on the emotional or coping measures than the personality, cognitive, and relational variables.

Finally, main differences with existing nonpatient samples should be considered. The normative sample from which the participants in this study were drawn (Sultan et al., 2004) differs from the extant Exner's samples (e.g., Exner, 2003a). Overall, although our sample exhibits rather similar levels of TE, they show greater complexity with more negative emotion markers, lower resources, and lower form quality. This, too, may reduce stability levels, with individuals having a more varied expression of engagement and negative emotions.

Suggestions for Practice and Research

Although some of the results presented here may be sobering for clinical practice, they should not be taken as definite. When reviewing short-term CS stability studies, Meyer (2004a) reported on five adult samples other than this one with fairly complete data collected according to standard administration guidelines. Two points show that this body of research is still weak: Coefficients seem to vary markedly across samples, and the total sample size for these studies is still limited when compared to other personality assessment procedures. Both of these points underline the need for more large-sample studies. Despite this, one positive aspect of our research is to provide fairly complete information on the variables with a saturated moderation analysis.

The consequence for practice and research is that variables such as R, Zf, F, M, a, WSumC, S, Fr + rF, Pairs, Lambda, EA, es, Adj es, EGO, and Blends may be used more easily, knowing that (a) some possess a good test–retest stability over a 3-month period and (b) in the case of the unstable variables, instability may be reduced if we take factors such as TE and state emotional distress or productivity into account. Unlike the more complex first factor task engagement variable, R, could potentially be modified through altered administration guidelines. Our results show that the stability of core variables in the interpretation process, such as Zf, EA, or es, is influenced by variation in productivity. This may help guide decisions in the future about an optimal range for R.

The observed differences in stability between predetermined classes of variables, together with the cross-validation of m, FM + m, es, and FD in the moderation analysis, provides further arguments in support of the existence of different kinds of variables in Rorschach data, some being more stable than others, as Rorschach researchers have traditionally underscored (e.g., Weiner, 2001). The design of this study made it possible to partially isolate state features in the emotional or stress markers of the Rorschach. Some markers, like m, FM + m, es, or FD, may be more prone to change on our measure of state distress than others, like Sum C', Sum V, or even Sum Y. Thus, stability studies may contribute to the validation process before measures are finalized. One way would be to use stability data as a criterion to redefine cutpoints for defining interpretively significant categories, like a > p + 1. Although this is beyond the scope of this study, thresholds could be determined in the future so as to maximize test–retest correlation coefficients using methods such as likelihood ratios and ROC curves (see Streiner, 2003).

ACKNOWLEDGMENTS

This study is dedicated to the memory of John E. Exner, Jr., whose constant encouragement and advice accompanied us all the way. We also are indebted to the Rorschach Research Foundation for financially supporting this research. An earlier version of this research was presented at the Society for Personality Assessment meeting at Miami, FL in March 2004. Some aspects of the methodology and data analysis benefited from fruitful discussions with Gregory Meyer, Cato Grønnerød, and Donald Viglione. We thank three reviewers for their insightful comments on earlier versions of this article. We also thank Thomas Saïas, Damien Fouques, and Hélène Salaün de Kertanguy for their participation in data collection and verification.

REFERENCES

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2,* 131–160.

Chiccetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Chiccetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications of assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127–137.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality, and self-opinion. *Personality and Individual Differences, 5,* 11–25.

Dunlap, W. P., Burke, M. J., & Greer, T. (1995). The effect of skew on the magnitude of product-moment correlations. *Journal of General Psychology, 122,* 365–377.

Egloff, B., Schwerdtfeger, A., & Schumukle, S. C. (2005). Temporal stability of the Implicit Association Test–Anxiety. *Journal of Personality Assessment, 84,* 82–88.

Erdberg, P., & Schaffer, T. W. (1999, July). *International symposium on Rorschach nonpatient data: Findings from around the world 1, 2, 3.* Paper presented at the 26th Congress of the International Rorschach Society, Amsterdam, The Netherlands.

Exner, J. E., Jr. (1980). But it's only an inkblot. *Journal of Personality Assessment, 44,* 562–577.

Exner, J. E., Jr. (1988). Problems with brief Rorschach protocols. *Journal of Personality Assessment, 52,* 640–647.

Exner, J. E., Jr. (1999). The Rorschach: Measurement concepts and issues of validity. In S. B. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 159–183). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Exner, J. E., Jr. (2002). A new nonpatient sample for the Rorschach Comprehensive System: A progress report. *Journal of Personality Assessment, 78,* 391–404.

Exner, J. E., Jr. (2003a, August). *A new nonpatient sample for the Rorschach Comprehensive System (N = 350).* Paper presented at the Rorschach Society Summer Seminars, Spiez, Switzerland.

Exner, J. E., Jr. (2003b). *The Rorschach: A Comprehensive System: Vol. 1. Basic foundations* (4th ed.). Hoboken, NJ: Wiley.

Exner, J. E., Jr., Armbruster, G. L., & Viglione, D. (1978). The temporal stability of some Rorschach features. *Journal of Personality Assessment, 42,* 474–482.

Exner, J. E., Jr., Thomas, E. A., & Cohen, J. B. (1983). *The temporal consistency of test variables for 50 nonpatient adults after 12 to 14 months* (Study No. 281). Unpublished Rorschach Workshop.

Goldberg, D. P. (1978). *Manual of the General Health Questionnaire.* Windsor, England: Nfer-Nelson.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., et al. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general care. *Psychological Medicine, 27,* 191–197.

Goldberg, D. P., Oldehinkel, T., & Ormel, J. (1998). Why GHQ threshold varies from one place to another. *Psychological Medicine, 28,* 915–921.

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

Grønnerød, C. (2004). Rorschach assessment of changes following psychotherapy: A meta-analytic review. *Journal of Personality Assessment, 83,* 256–276.

Grønnerød, C. (2006). Reanalysis of the Grønnerød (2003) Rorschach temporal stability meta-analysis data set. *Journal of Personality Assessment, 86,* 222–225.

Haller, N., & Exner, J. E., Jr. (1985). The reliability of Rorschach variables for inpatients presenting symtoms of depression and/or helplessness. *Journal of Personality Assessment, 49,* 516–521.

Jiang, X. L., & Cillessen, A. H. N. (2005). Stability of continuous measures of sociometric status: A meta-analysis. *Developmental Review, 25,* 1–25.

Kalliath, T. J., O'Driscoll, M. P., & Brough, P. (2004). A confirmatory factor analysis of the General Health Questionnaire–12. *Stress and Health, 20,* 11–20.

Kraemer, H. C., Gullion, C. M., Rush, J., Franck, E., & Kupfer, D. J. (1994). Can state and trait variables be disentangled? A methodological framework for psychiatric disorders. *Psychiatry Research, 52,* 55–69.

Lindgren, T., & Carlsson, A. M. (2002). Relationship between the Rorschach and the MMPI–2 in a Swedish population: A replication study of the effects of first-factor related test-interaction styles. *Journal of Personality Assessment, 79,* 357–370.

Meyer, G. J. (1992). The Rorschach's factor structure: A contemporary investigation and historical review. *Journal of Personality Assessment, 59,* 117–136.

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

Meyer, G. J. (1997b). On the integration of personality assessment methods: The Rorschach and MMPI. *Journal of Personality Assessment, 68,* 297–330.

Meyer, G. J. (1997c). Thinking clearly about reliability: More critical corrections regarding the Rorschach Comprehensive System. *Psychological Assessment, 9,* 495–498.

Meyer, G. J. (2004a, July). *An overview of Rorschach reliability, validity, and utility.* Paper presented at the 8th Congress of the European Rorschach Association, Stockholm, Sweden.

Meyer, G. J. (2004b). The reliability and validity of the Rorschach and Thematic Apperception Test (TAT) compared to other psychological and medical procedures: An analysis of systematically gathered evidence. In M. Hersen, M. Hilsenroth, & D. Segal (Eds.), *Personality assessment. Volume 2. Comprehensive handbook of psychological assessment* (pp. 315–342). Hoboken, NJ: Wiley.

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Piers, C. C., et al. (2002). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment, 78,* 219–274.

Meyer, G. J., Viglione, D. J., & Exner, J. E., Jr. (2001). Superiority of Form% over Lambda for research on the Rorschach Comprehensive System. *Journal of Personality Assessment, 76,* 68–75.

Parker, K. C. H. (1983). A meta-analysis of the reliability and validity of the Rorschach. *Journal of Personality Assessment, 47,* 227–231.

Parker, K. C. H., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103,* 367–373.

Perry, W., McDougall, A., & Viglione, D. (1995). A five-year follow-up on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment, 64,* 112–118.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126,* 3–25.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81,* 209–219.

Sultan, S., Andronikof, A., Fouques, D., Lemmel, G., Mormont, C., Réveillère, C., & Saïas, T. (2004). Vers des normes francophones pour le Rorschach en systéme intégré [Towards French language norms for the Rorschach Comprehensive System]. *Psychologie Française, 49,* 7–24.

Thomas, E .A., Alinsky, D., & Exner, J. E, Jr. (1982). *The stability of some Rorschach variables in 9-year-olds as compared with nonpatient adults* (Study No. 441). Unpublished Rorschach Workshop.

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2003). Stability of self-esteem across the life span. *Journal of Personality and Social Psychology, 84,* 205–220.

Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment, 11,* 251–265.

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions and future. *Psychological Assessment, 13,* 452–471.

Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of Rorschach Comprehensive System coding. *Journal of Clinical Psychology, 59,* 111–121.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 38,* 319–350.

Watson, D., Hubbard, B., & Wiese, D. (2000). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology, 78,* 546–558.

Weiner, I. B. (2001). Advancing the science of psychological assessment: The Rorschach Inkblot Method as exemplar. *Psychological Assessment, 13,* 423–432.

APPENDIX
Stability Coefficients for Structural Summary Variables: Pearson's *r* for 75 Nonpatient Adults

LOCATION			DETERMINANTS			CONTENTS	SPECIAL SCORES	
FEATURES			BLENDS		SINGLE		COGNITIVE SP. SCORES	
						H = .76		
Zf	= .76		Nb Blends = .63		M = .76	(H) = .59	Lv1	Lv2
ZSum	= .76		Col Shd Blends = .14		FM = .48	Hd = .63	DV = .06	–.02
ZEst	= .76				m = .47	(Hd) = .24	INC = .26	.20
					FC = .61	Hx = .35	DR = .31	–.02
W	= .82				CF = .47	A = .70	FAB = .50	.48
D	= .79				C = .08	(A) = .18	ALOG = –.18	
Dd	= .61				Cn = N/A	Ad = .70	CON = N/A	
S	= .70				FC' = .35	(Ad) = .48		
(Wv	= .69)				C'F = .23	An = .51		
					C' = –.01	Art = .52	Raw Sum6 = .50	
	DQ				FT = .46	Ay = .63	Wgtd Sum6 = .56	
		(FQ–)			TF = .48	Bl = .55		
+	=.71	(.55)			T = N/A	Bt = .69	OTHER SPECIAL SCORES	
o	=.73	(.54)			FV = .36	Cg =.72		
v/+	= .36	(–.02)			VF = .11	Cl = .52	AB = .14	
v	=.63	(.37)			V = N/A	Ex =.51	AG = .45	
					FY = .06	Fd = .60	COP = .38	
					YF = .23	Fi = .67	CP = N/A	
	FORM QUALITY				Y = –.03	Ge = .77	GHR = .47	
					Fr = .53	Hh = .53	PHR = .69	
	FQx	FQf	MQual	SQ	rF = .72	Ls = .38	MOR = .62	
+	= .31	–.05	.53	N/A	FD = .51	Na = .33	PER = .34	
o	=.67	.58	.67	.58	F = .70	Sc = .65	PSV = .06	
u	= .38	.46	.09	.26		Sx = .65		
–	= .60	.47	.66	.46		Xy = .80		
none	= .15	—	–.02	N/A		Id = .25		
					(2) = .77			

RATIOS, PERCENTAGES, AND DERIVATIONS

R = .75	L = .72	Pure F% =.68	FC:CF + C	= .61:.55	COP = .38	AG = .45
			Pure C	= .08	GHR:PHR = .55:.70	
EB = .76:.69	EA = .77	EB Per = .45	Sum C':WSumC	= .38:.69	a:p = .61:.55	
eb = .50:.42	es = .46	D = .34	Afr	= .57	Food = .60	
	Adj es = .46	Adj D = .38	S	= .70	Sum T = .56	
			Blends/R	= .66	Hum Cont = .68	
FM = .48	C' = .38	T = .56	CP	= N/A	Pure H = .76	
m = .47	V = .46	Y = .17			PER = .34	
					Iso Indx = .67	
					H (Hd):A (Ad) = .37:.15	
					H + A:Hd + Ad = .73:.73	
a:p =.61:.55	Sum6 = .50		XA% = .49	Zf = .75	3r + (2)/R = .78	
Ma:Mp = .59:.42	Lv2 = .42		WDA% = .45	W:D:Dd = .80:.74:.61	Fr + rF = .65	
2AB + Art + Ay = .53	WSum6 = .57		X – % = .51	W:M = .80:.76	Sum V = .46	
MOR = .62	M– = .66		S – % = .39	Zd = .46	FD = .51	
	Mnone = –.02		P = .54	PSV = .06	An + Xy = .53	
			X + % = .55	DQ+ = .70	MOR = .62	
			Xu% = .32	DQv = .55	H:(H) Hd (Hd) = .67:.58	
			F + % = .30			
PTI = .48	SCZI = .50	DEPI = .28	CDI = .58	S–CON = .47	HVI = .62	OBS = .25

Serge Sultan
Institute of Psychology
University of Paris–René Descartes
71 avenue Edouard Vaillant
92000 Boulogne, France
Email: serge.sultan@univ-paris5.fr

Received July 19, 2005
Revised April 5, 2006