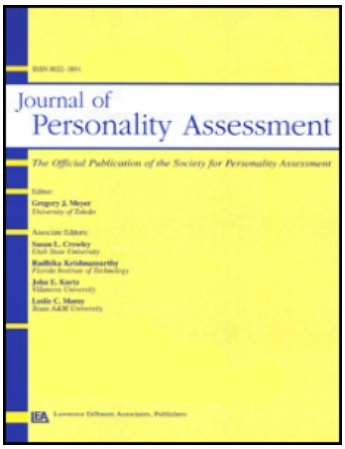

Journal of Personality Assessment
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653663

Does Productivity Impact the Stability of Rorschach Scores?

Serge Sultan [a]; Gregory J. Meyer [b]
[a] Institut de Psychologie, Université Paris Descartes, France [b] Department of Psychology, University of Toledo,

Online Publication Date: 01 September 2009

To cite this Article Sultan, Serge and Meyer, Gregory J.(2009)'Does Productivity Impact the Stability of Rorschach Scores?',Journal of Personality Assessment,91:5,480 — 493
To link to this Article: DOI: 10.1080/00223890903088693
URL: http://dx.doi.org/10.1080/00223890903088693

Journal of Personality Assessment, *91*(5), 480–493, 2009

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223890903088693

Routledge
Taylor & Francis Group

Does Productivity Impact the Stability of Rorschach Scores?

SERGE SULTAN[1] AND GREGORY J. MEYER[2]

[1]*Institut de Psychologie, Université Paris Descartes, France*
[2]*Department of Psychology, University of Toledo*

Research suggests that productivity could impact the stability of Rorschach scores. To explore for this effect, we conducted secondary analyses of test–retest data gathered using the Rorschach Comprehensive System (Exner, 2003) and available for 75 French, nonpatient adults (Sultan, Andronikof, Réveillère, & Lemmel, 2006). We examined how response frequency (R) impacted stability using hierarchical regression models. Results on 83 variables from the lower part of the structural summary showed that stability was impacted by the mean level of productivity in 12 variables with medium to large effects (including Zf, HVI, and W location). Stability was also impacted by variations of productivity in 9 variables with medium to large effects (including Passive Movement, D Location, or Human Contents). Higher mean R and variability of R impacted stability levels negatively. Transforming scores into proportions (i.e., dividing scores by R) was beneficial for some important variables (including FM+m, Zf, DQ+). Procedures should be developed to limit productivity and control for R variations across time if one wishes to derive more reliable descriptions of individuals from the Rorschach.

Recent criticisms of the Rorschach (Exner, 2003) have prompted many different types of research into this procedure, ranging from investigations of temporal stability (Grønnerød, 2003) to interpretation reliability (Meyer, Mihura, & Smith, 2005). However, much remains to be done in other respects such as the impact of productivity on the scores used for interpretation (see Hunsley & Bailey, 2001; Lilienfeld, Wood, & Garb, 2000, p. 34). Verbal and response productivity can be considered to be a core feature in most of the widely used storytelling techniques and inkblots tasks. Some authors have argued that productivity, or the quantity of verbal content produced during these tests, could be confounded with interpretation (Lilienfeld et al., 2000). The idea that response frequency (or R) could have an impact on Rorschach scores goes back many years (cf. Cronbach, 1949; Fiske & Baughman, 1953). However, the role of R has seldom been systematically examined in studies including a specific criterion such as the prediction of external features or the stability of records. In this article, we present secondary analyses on Rorschach stability data recently gathered in a French nonpatient sample (Sultan, Andronikof, Réveillère, & Lemmel, 2006). The primary report focused on stability levels and the moderating role of complexity on the stability of records. The aim of this study was to determine the extent to which R may impact stability levels. To this end, we examined all the variables in the lower section of the Comprehensive System Structural Summary (Exner, 2003), that is, the variables that have the highest impact on the interpretation process. It is important to explore the role of response frequency because R can be controlled by the examiner, whereas complexity cannot. Research into response frequency may also lead to practical recommendations or adaptations of the system (cf. Dean, Viglione, Perry & Meyer, 2007).

Engagement, or record complexity, has long been considered to be a robust first factor across most factor analytical studies of the Rorschach. This first factor consists of R and many variables that relate to productivity (such as D and Dd Locations, Zf or Shadings; see Meyer, 1992, 1997; and Wood, Krishnamurthy, & Archer, 2003). Although there is no clear consensus about whether correlations of R with Rorschach scores are large (Exner, 1992, Kinder, 1992), the average proportion of total variance among Rorschach scores accounted for by engagement is no less than 20% to 25% (about 50% of common variance; Table 2 in Meyer, 1992). This suggests that productivity, as it is related to engagement, may be responsible for individual differences in some scores. This phenomenon would be partly due to a part–whole relationship between R and other scores. For example, in the Comprehensive System, R = W + D + Dd Locations. Despite a stated need for studies of the effects of test-taking styles, little research is available on this topic (see Meyer & Archer, 2001, for a review).

However, response frequency may impact validity. If one client gives 17 responses and a second 34, the second has twice as many opportunities to report aggressive contents or morbid imagery. In addition, some research has demonstrated that the level of R impacts constellations that are central to interpretation (Meyer, 1993). The high correlations described previously mean that some people might exhibit positive indices (DEPI > 5, e.g.) or significant scores (MOR > 2, e.g.) simply because they have a higher response frequency. This is a major concern for clinical interpretation and internal validity. When examining correlations with external criteria, Meyer (1993) showed that the level of R directly impacted the external validity of constellations (SCZI, S-Con, HVI). A similar result was found in the correlation between Blends and IQ (Wagner, Young, & Wagner, 1992). Given the high proportion of variance shared by R and Rorschach scores, a legitimate question that researchers need to answer is whether Rorschach scores provide specific information beyond a mere productivity report. Of course, most contemporary research contends with R in some fashion to ensure that validity findings are due to the variable under consideration rather than R itself by using percentages, ratios, partial

Received March 13, 2009; Revised May 12, 2009.
Editor's Note: Bill Kinder served as the final Editor for this manuscript.
Address correspondence to Serge Sultan, Institut de Psychologie, Université Paris Descartes, 71 avenue Edouard Vaillant, 92000 Boulogne, France; Email: serge.sultan@parisdescartes.fr

correlations, multistep regressions and so on. But few researches have precisely studied the incremental validity of Rorschach variables over R (see Dawes, 1999, 2001; Hunsley & Meyer, 2003; Perry, Moore, & Braff, 1995; Perry, 2001, 2003).

As mentioned previously, the problem of productivity impacting validity also concerns other assessment procedures such as the Thematic Apperception Test (TAT; Murray, 1943). For example, word count (i.e., the total number of words produced by a respondent in a TAT story) has been found to be positively correlated with some dimensions of object relations scored in the TAT (e.g., Ordnuff, Freedenfeld, Kelsey, & Critelli, 1994).

Only one study has investigated the impact of R on the stability of Rorschach scores (Exner, 1988). In this study, Exner explored the moderating role of R on short-term stability (4–30 days) in child and adult records. Two groups of 36 pairs of test–retest protocols were compared. In the first group, one protocol in each pair had fewer than 14 responses and the second more than 14. In the second group, both protocols in each pair had more than 14 responses. A comparison of the stability coefficients for the two groups revealed that the first group exhibited lower overall stability (in 20 of the 26 variables observed), especially when Lambda was high (>.99), which was the case in 28 of the 36 pairs in this group. Exner's (1988) study merits a number of comments. First, the $R < 14$ threshold was determined intuitively because such cases have been found to be responsible for lowering result significance in various validation studies (Exner, 1988, p. 641). Although this threshold may be appropriate, there is no clear evidence to suggest that it is optimal, that is, it is not derived from stability data. Second, the two groups that were compared actually differed on two aspects, namely, productivity itself and variations in productivity. In fact, there was a considerable likelihood of productivity being lower in the first group given the usual test–retest correlation of R and the fact that one response frequency was set to be less than 14. In addition, the size of the R variations between the test and retest was very likely to be larger in the first group given the constraints imposed on R differences between the test and retest ($|R_{T1} - R_{T2}| \geq 2$). As a consequence, Exner's (1988) results may mean that people who exhibit a low level of productivity and who provide a different number of responses at test and retest have less stable protocols. Thus, Exner's (1988) analyses in fact involved a complex interaction between productivity and productivity variations among protocols that were also quite constricted (i.e., high Lambda).

One important aspect of Exner's (1988) study is that the procedure employed very probably created a severe range restriction in the low-R sample (first group).[1] Although Exner's (1988) article does not provide any indication of central tendency or variability in his two samples, one can surmise that most of the protocols in the low-R sample had an $R = 8$ to 13. In contrast, the comparison sample was constructed so that the protocols probably had an $R = 15$ to 30+. Thus, the range of R was probably three to five times larger in Exner's (1988) comparison sample than in his low-R sample. Because, as explained earlier, many Rorschach scores are correlated with R, their range could also be restricted in the low-R sample compared with the comparison sample. Because range restriction usually reduces reliability coefficients, it is probable that Exner's (1988) sampling procedure ensured that test–retest correlations would be lower in the low-R group than in the comparison sample. For these reasons, the impact of Exner's (1988) findings on the possible moderating role of R remains unclear.

Consequently, there are two ways response frequency may impact stability. First, given the correlations between R and Rorschach scores, it would be logical that when R varies, other scores will vary as well. Although this has been demonstrated many times across different individuals, no data are available on the impact that intraindividual variations in R have on other scores except for those provided by Exner (1988). To the extent that scores are linearly related with R, differences in R from one occasion to the next should produce instability in the scores over time. Second, one needs to ask whether the overall level of productivity impacts stability. There is no reason why R should not impact stability when it is above the threshold of 14 responses. Currently, protocols with fewer than 14 responses are generally considered to be invalid. It is probable that the overall level of productivity also moderates stability in valid records.

Although it is difficult to formulate clear expectations on the basis of previous research (see observations made about Exner, 1988), two possible competing hypotheses about the impact of mean R may be formulated. First, higher stability could be observed as a byproduct of better "internal consistency" reliability such that giving more responses in a protocol should provide more opportunities to assess the latent construct for any given score (e.g., M, FD, SumT, WSum6). Thus, having a larger number of responses to work with on each occasion provides a more accurate measure of the latent construct. From this perspective, one could expect higher single-occasion reliability with higher average R and as a consequence higher short-term stability across occasions. Yet, there is an important competing hypothesis that works in the opposite direction. Like many phenomena, the mean of R across two occasions is probably positively correlated with its variability. This can be exemplified by considering the possible range around an average R value. Given that R must be 14 or higher, a mean R value of 20 can only be obtained with an R difference of 12 points or less (i.e., a range from 14–26). However, a mean R of 30 could be associated with an R difference of up to 32 points (i.e., a range from 14–46). As average R increases, there is more opportunity to see R values at test and retest diverge more substantially. As a result, higher levels of mean R may be linked to lower stability. If this hypothesis is correct, we should observe a negative impact of productivity on stability. This would occur not because higher R is in itself associated with lower stability but because it favors variability in R and in the scores associated with R.

On the basis of previous research, we formulated the following three objectives and expectations. First, we evaluated whether the overall mean level of R influenced stability in Rorschach scores. Given the various competing mechanisms, no clearer expectations were formulated. However, one would expect that in cases in which higher mean levels of R negatively influence stability, then the variation of R also would negatively influence stability because higher mean R levels would be associated with more R variability. Second, variations of R should impact stability negatively, with increases or decreases in R over time relating with lower stability levels, especially in variables showing large correlations with R. Finally, because one habitual way of controlling for the impact of R is to divide scores by

[1] We thank an anonymous reviewer for drawing our attention to this phenomenon.

TABLE 1.—Description of 10 ratios of the structural summary ($N = 75$).

	T1					T2							
Ratios	*M*	*SD*	*Mdn*	*Min*	*Max*	*M*	*SD*	*Mdn*	*Min*	*Max*	*r*	*t*	ICC
EB ratio	1.64	1.66	1.20	.00	10.00	1.76	2.50	1.20	.00	18.00	.53	−0.49	.97
eb ratio	2.19	2.70	1.40	.22	16.00	1.84	2.31	1.20	.00	14.00	.21	0.94	.93
FC:CF+C ratio	2.05	2.58	1.00	.00	12.00	2.69	3.89	1.00	.00	20.00	.52	−1.64	.62
SumC':WSumC ratio	0.65	0.95	0.36	.00	6.00	0.62	0.61	0.50	.00	2.00	.08	0.28	.80
Blends/R	0.19	0.13	0.17	.00	.67	0.20	0.14	0.18	.00	.68	.66	−0.24	.94
GHR:PHR ratio	1.86	2.12	1.00	.00	10.00	1.80	1.94	1.20	.00	10.00	.46	0.22	.66
a:p ratio	1.71	1.78	1.00	.00	8.00	1.91	2.37	1.14	.00	16.00	.34	−0.71	.90
Ma:Mp ratio	1.39	1.53	1.00	.00	8.00	1.90	2.39	1.00	.00	12.00	.23	−1.76	.73
W:M ratio	3.37	2.99	2.50	.50	18.00	3.52	3.68	2.43	.00	20.00	.62	−0.46	.97
H:(H)+Hd+(Hd) ratio	0.77	1.14	0.50	.00	7.00	0.91	1.34	0.50	.00	10.00	.76	−1.35	.96

Note. T1 = Time 1; T2 = Time 2; ICC = intraclass correlation coefficient; ICCs indicate interrater scoring reliability for absolute agreement, one-way random model, computed on 40 randomly chosen protocols.

R, we examined whether this procedure was beneficial in cases in which an impact of mean R or variability in R was initially demonstrated.

METHODS

Participants and Examiners

Authors of the primary report recruited 75 persons from the ongoing French-language normative project and tested them twice at an interval of 95 days (range = 79–115 days between Time 1 [T1] and T2). The sample consisted of 28 men and 47 women (mean age = 39.2 years) who were employed in private businesses, sports clubs, and a charity organization. We included the participants provided that they accepted that the individual data would remain strictly anonymous and no individual feedback would be given to anyone. We asked the first 100 participants included in the French normative project to perform a retest after 3 months (Sultan et al., 2004). After taking account of the nonpatient criteria used for selection, the quality of administration, and attrition, there were 75 participants in the final test–retest sample (see additional details on methods in the primary report in Sultan et al., 2006). Twelve properly trained examiners participated in the test–retest study. No examiner tested the same person twice.

Procedure and Measures

The Rorschach Comprehensive System was administered twice using current standardized practice (Exner, 2003). Codes were recorded using the Rorschach Calculation Program, and the records were then transferred to SPSS (Statistical Package for the Social Sciences Version 14.0). We selected 83 variables that are central in the interpretation process. These scores constitute the lower part of the Structural Summary (Exner, 2003; Meyer et al., 2002) and refer to various categories such as Locations, Determinants, Contents or Perceptual Organization. Because, for psychometric reasons, the reliability of ratios tends to be substantially lower than the reliability of their constituent scores (Cronbach, 1949), usual ratios—such as EB, eb, or GHR:PHR—were treated as fractions. When the denominator equaled zero, it was replaced by .5. Thus if W:M = 3/0, it was converted to 3/.5 = 6.[2] To measure the overall level of productivity, we averaged the numbers of responses at test and at retest and labeled this variable *Mean* R ($M = 23.7$, $SD = 6.5$, *Minimum* [*Min*] = 14.5, *Maximum* [*Max*] = 47.5, 25th, 50th, and 75th centiles: 19.5, 22.5, and 27.0, respectively). To assess for variations in R, we used the raw difference of the two values of R in which R_{T1} was subtracted from R_{T2}. This was labeled as *Var* R ($M = -.33$, $SD = 5.02$, $Min = -23$, $Max = 11$, 25th, 50th, and 75th centiles: −3.0, 0.0, and 2.0, respectively). *Mean* R and *Var* R were strongly associated but in a nonlinear fashion according to a U-shaped curve, with low negative *Var* R and high positive *Var* R values related to higher *Mean* R and midrange close-to-zero *Var* R values associated with lower *Mean* R. This association is reflected in the correlation of *Mean* R, with the squared value of *Var* R being .55 (this correlation was .57 with the absolute value of *Var* R). Thus, mean elevations in R are intrinsically associated with variations in R. Table 1 provides basic information for 10 ratios. The same information regarding the other Rorschach variables can be found in the previous report (Sultan et al., 2006).

Interrater Reliability and Quality of Data

We adopted a number of procedures to guarantee a high quality of data and accurate scores (consensus scoring, then blind rescoring, followed by an interrater reliability study on 40 randomly chosen protocols). We calculated interrater agreement for the variables used in the subsequent analyses on the basis of the intraclass correlation coefficients (ICCs) protocol-level, interrater reliability on summary scores. The distribution of ICCs among the Rorschach variables was characterized by a mean and median of .87 and .89, respectively. The standard deviation was .11, and the 25th and 75th centiles were .82 and .95, respectively, with approximately the same pattern of results for T1 and T2. According to established criteria (Chiccetti, 1994), ICC was reasonable for Pure C and PSV (ICC = .40–.59) and good for A+(Ad), MOR, Level2, X − %, FC:CF + C ratio, GHR:PHR ratio, and Ma:Mp ratio (ICC = .60–.74). For all the other variables, it was excellent (ICC = .75–1.00).

[2] We also computed proportions in which the numerator was the same as in the preceding fraction, but the denominator was the sum of both elements. This did not yield any significant difference in subsequent analyses. The following results are based on the common adjustment referred to in the body of the article.

Statistical Analyses

To explore the impact of R on stability, we computed hierarchical regression models with the dependent variable being the Rorschach score at T2 and the independent variables being the Rorschach score at T1, the moderator (*Mean* R or *Var* R), and the interaction term between the T1 score and the moderator. We computed the interaction in the traditional manner as the product of the T1 score and the moderator. We constructed the regression equations in three blocks. Block 1 consisted of just the T1 score. Following this step, the residuals were saved to serve as the dependent variable for Blocks 2 and 3. Using these isolated residuals ensured that the moderator and interaction terms would not be allocated variance they shared with the T1 score that would then make them appear to have a larger impact on the T2 scores than was strictly warranted from their unique contribution over and above the T1 score. We then entered the moderator in Block 2 and the interaction between the T1 score and the moderator in Block 3. Although Block 1 is informative, our focus is on Blocks 2 and 3. We performed the analyses twice, first with the productivity level (*Mean* R) moderator and then with the variation of productivity (*Var* R) moderator. As noted, the criterion for Blocks 2 and 3 were the residuals computed by subtracting from the T2 score the value predicted on the basis of the T1 score. This indicates the share of variability in T2 that is not accounted for by T1. Conceptually, it is the instability of the score. The analyses allow an estimate of the proportion of instability that can be accounted for by productivity or productivity variations in each Rorschach score. They also then indicate to what extent controlling these factors could improve stability levels.

In the case of *Mean* R, we sought to understand to what extent it would predict the general degree of instability in scores and therefore considered the absolute value of the residuals as criteria. That is, we were interested in the overall degree of instability regardless of whether it was due to increases or decreases in the scores over time. However, the situation is very different in the case of *Var* R for which it is essential to know if scores increase or decrease in tandem with R changes. Thus, for these analyses, we kept the raw signed residuals as criteria. For a given score, these residuals are positive when the T2 value exceeds the T1 value. Because *Var* R is computed in the same fashion, that is, it is positive when R_{T2} exceeds R_{T1}, predicting signed residuals maximizes the match between the moderator and the criterion and tests the hypothesis that R variations will predict instability especially in scores that are linearly related to R.

Because of the forgoing, the interpretation of correlations with residuals will be very different according to the moderator under consideration. When predicting the absolute value of the residuals, the criterion is on a scale from zero at the low end to some positive maximum at the high end. Here, a value of zero indicates stability (i.e., no difference between the T2 score and the value predicted based on the T1 score), and increasingly large residual values indicate increasing instability (i.e., increasing differences between the T2 score and the value predicted based on the T1 score). Given this, a positive association of *Mean* R with the absolute value of the residuals indicates that higher levels of productivity are associated with higher levels of disagreement between the T1 and T2 scores (i.e., higher instability). Conversely, a negative association would indicate that lower levels of productivity are associated with score instability.

The considerations differ when predicting the signed residuals. As before, a residual value of zero indicates the T2 score is exactly what was predicted based on the T1 score. Hence a residual of zero still indicates score stability from T1 to T2. However, the criterion can now consist of large negative values and large positive values on either side of zero. Because the residuals are computed as the T2 score minus the value predicted from the T1 score, large negative residuals are obtained when the T2 score is notably lower than what would be predicted based on the T1 score. Large positive residuals are obtained when the T2 score is notably higher than what would be predicted based on the T1 score. Because the *Var* R difference score is computed as R_{T2} minus R_{T1}, a positive correlation between *Var* R and the signed residuals indicates that both sets of differences are aligned such that the signed differences in R are helping to explain or account for the signed differences in the target score residual. In other words, to the extent that a score is positively correlated with R, it will track differences in R over time and produce a positive coefficient in the regression equation. Thus, when R is higher at T2 than at T1, the target score at T2 will be higher than at T1; when R decreases from T1 to the retest at T2, the target score will also decrease in a linear fashion from test to retest. Within these positively correlated variables, when *Var* R is nearer to zero, the signed residual is also nearer to zero (i.e., the score is more stable). The latter principle remains true when there is a negative correlation between *Var* R and a residual. What differs in such cases is the direction of alignment among the two sets of difference scores. Larger positive T2 minus T1 differences in R are associated with larger negative differences in the target score residual and vice versa. In other words, when R increases from T1 to T2, the target score decreases; when R decreases from T1 to T2, the target score increases. This somewhat counterintuitive finding should occur to the extent that a score is negatively correlated with R; that is, when R goes up, it goes down and vice versa.

We computed the interaction terms on the basis of the recommendations of Aiken and West (1991), and we performed all analyses on standard scores. For presentation purposes, in Table 2, the semipartial correlation of each block together with the sign of the association is mentioned in columns. We also computed a total correlation for each full model. Given the fact that we performed analyses for 83 variables and to limit spurious positive interpretations due to a high number of results, we underscore results involving medium-large effect sizes in Blocks 2 and 3 ($\Delta R \geq .30$).

The approach taken here differs in two ways from the common approach to moderator analyses following in the tradition of Baron and Kenny (1986). First, Baron and Kenny reserved the term *moderator* to apply to instances when the product of the moderator and the predictor produces a statistically significant interaction term. A significant interaction indicates that the moderator has a multiplicative effect on prediction such that the magnitude of the association (i.e., the regression slope) between the predictor and the criterion changes depending on the moderator value. Graphically, it also means that the regression plane in three-dimensional space is no longer flat but instead is symmetrically curved, with two opposing corners stretched up and the other two corners pushed down (e.g., Cohen, Cohen, West, & Aiken, 2003). In this context, a significant interaction will mean the ability to predict retest values from baseline values increases or decreases as a function of *Mean* R or *Var* R. However, for the analyses we present here, the term *moderator* is also used to

TABLE 2.—Summary of hierarchical regression analyses with T2 scores as dependent variables and T1 scores, moderators, and their interactions as independent variables ($N = 75$).

	Correlations		Block 1	Model 1: *Mean* R (Absolute Value of Residuals Predicted)			Model 2: *Var* R (Signed Residuals Predicted)		
	T1 With R_{T1}	T2 with R_{T2}	Test–Retest *r*	Block 2 ΔR	Block 3 ΔR	Total R	Block 2 ΔR	Block 3 ΔR	Total R
Core Section									
R	1.000	1.000	.746						
Lambda	.096	.127	.720	.089	.045	.723	.056	−.**346****	.760
M	.235*	.239*	.756	.217	.055	.770	.265*	.071	.777
WSumC	.383**	.143	.692	.000	.045	.693	.089	.134	.702
EB ratio	−.109	.055	.528	.000	−.063	.531	.076	.190	.556
EA	.376**	.250*	.771	.084	.077	.774	.268*	.134	.794
EBPer	−.128	.003	.603	.084	−.152	.619	.088	.118	.614
FM+m	.646***	.276*	.503	**.437*****	−.045	.630	.244*	.100	.552
SumShd	.370**	.201	.424	−.063	.130	.444	.216	.239*	.515
eb ratio	.031	−.112	.211	−.118	−.089	.256	−.149	.077	.267
es	.642***	.297**	.457	**.321***	.032	.540	.286*	.148	.539
D score	−.419***	−.177	.337	**.330***	−.045	.460	−.231*	.071	.407
Adjes	.515***	.349**	.455	.217	−.138	.509	.272*	.032	.516
AdjD	−.164	−.150	.376	.100	.000	.387	−.189	−.032	.416
FM	.477***	.295	.475	.297	−.141	.556	.216	.000	.512
SumC’	.124	−.054	.384	−.045	−.045	.388	.045	.095	.396
[SumT]	.064	.313**	.564	.176	.105	.589	.165	**.349****	.648
m	.559***	.064	.472	.173	.105	.505	.099	.205	.513
[SumV]	.272*	.408***	.463	.182	.173	.514	.260*	.286*	.576
[SumY]	.444***	.000	.170	.071	.045	.189	.044	.184	.253
Affect									
FC	.417***	.410***	.611	**.308***	.187	.674	.112	.126	.625
CF+C	.246*	−.062	.553	.089	.164	.575	−.011	−.071	.556
[Pure C]	.159	−.127	.084	−.071	.000	.110	−.028	.126	.154
FC:CF+C ratio	.131	.384**	.519	**.344****	.077	.600	−.047	−.192	.546
SumC’:WSumC ratio	−.029	−.220	.080	−.228*	.000	.241	.051	−.089	.130
Afr	.133	.163	.571	−.055	.000	.573	.074	−.045	.575
S	.457***	.351**	.703	.214	.095	.723	**.386****	.173	.765
Blends/R	−.125	−.295*	.658	−.105	.063	.664	−.031	−.045	.659
Interpersonal									
COP	.092	−.037	.381	.138	.114	.415	−.182	−.118	.431
[AG]	.188	.182	.452	.000	.000	.452	.082	.055	.460
GHR	.489***	.288*	.555	**.333****	.000	.620	.293*	.105	.612
PHR	.291*	.520***	.700	.114	.000	.705	.257*	−.032	.724
GHR:PHR ratio	−.049	−.284*	.462	−.190	−.084	.497	−.067	.063	.469
active	.337**	.200	.609	.063	.000	.611	.107	.055	.616
passive	.469***	.306**	.546	.152	.138	.572	**.345****	.077	.621
a:p ratio	−.133	−.242*	.337	−.224	−.114	.412	−.082	−.032	.347
[Food]	.465***	.262*	.604	.077	.000	.607	.087	.000	.608
Hum Cont	.488***	.523***	.684	.285	.000	.715	**.429*****	.032	.753
Pure H	.105	.055	.761	.200	.000	.772	−.013	.071	.762
[PER]	.194	.122	.344	.145	−.055	.373	−.002	.000	.344
Isolate/R	.060	−.203	.666	−.055	.032	.668	−.238*	.055	.690
(H)+(Hd)	.452***	.279*	.367	−.063	.138	.393	**.391****	−.071	.521
[(A)+(Ad)]	.081	.207	.146	.045	−.138	.205	.139	.000	.201
H+A	.588***	.496***	.775	**.437*****	.192	.832	**.616*****	.063	.868
Hd+Ad	.415***	.663***	.702	.247*	−.114	.728	.290*	−.063	.733
Ideation									
Ma	.093	.151	.591	.126	−.055	.601	.018	.000	.591
Mp	.245*	.187	.423	.077	.045	.431	.225	.071	.474
Ma:Mp ratio	−.210	−.205	.232	−.184	−.126	.318	−.175	−.158	.326
Intell	.315**	.265*	.534	−.032	.032	.535	.244*	.145	.585
MOR	.146	.074	.620	.290*	.000	.660	.011	.228	.645
Sum6	.301**	.248*	.503	.187	.000	.528	.111	.232*	.550
Lv2	.050	.098	.444	.155	.110	.475	.006	.197	.478
WSum6	.238*	.200	.556	.095	−.232*	.594	.090	.148	.574
[M–]	.062	.341*	.656	.000	.110	.661	.079	.100	.663
[Mnone]	−.126	−.130	−.019	−.114	.055	.116	−.037	.000	.042
Mediation									
XA%	−.047	−.203	.494	−.265*	.045	.546	−.101	−.077	.506
WDA%	−.014	−.057	.446	.095	.055	.457	−.033	−.100	.456
X – %	.046	.244*	.512	−.190	−.055	.539	.127	−.071	.527
S–	.146	.307**	.390	.138	−.055	.413	.124	.100	.417
Pops	.341**	.003	.544	.000	.155	.559	.150	.118	.567

TABLE 2.—Summary of hierarchical regression analyses with T2 scores as dependent variables and T1 scores, moderators, and their interactions as independent variables ($N = 75$). *(Continued)*

	Correlations		Block 1	Model 1: *Mean* R (Absolute Value of Residuals Predicted)			Model 2: *Var* R (Signed Residuals Predicted)		
	T1 With R_{T1}	T2 with R_{T2}	Test–Retest *r*	Block 2 ΔR	Block 3 ΔR	Total R	Block 2 ΔR	Block 3 ΔR	Total R
X+%	−.281*	−.351**	.553	−.148	−.192	.589	**−.350****	−.084	.629
Xu%	.320**	.262*	.316	−.063	.055	.326	.267*	.000	.405
Processing									
Zf	.443***	.183	.764	**.472*****	.089	.825	.274*	.272*	.804
W	.228*	−.075	.816	**.420*****	.293**	.868	.193	.224	.834
D	.763***	.741***	.785	.272*	.089	.805	**.640*****	−.276**	.896
Dd	.535***	.583***	.608	**.300****	.077	.656	**.438*****	.084	.704
W:M ratio	−.073	−.168	.616	.000	.184	.633	−.190	**−.541*****	.764
Zd	−.015	−.015	.464	.071	.000	.468	−.167	.077	.492
[PSV]	−.080	−.024	.063	.000	−.192	.202	−.047	.032	.085
DQ+	.469***	.238*	.714	**.535*****	.032	.807	.220	.179	.741
[DQv]	.389***	−.066	.627	.000	.055	.628	.039	.077	.631
Self-Perception									
EGO	−.062	−.080	.779	.063	.000	.780	−.180	−.205	.798
[Fr+rF]	.079	.043	.645	.164	.089	.661	.034	.202	.664
[FD]	.162	−.035	.507	−.105	−.055	.517	.096	−.063	.517
An+Xy	.291*	.240*	.528	.000	.045	.529	.207	.110	.564
(H)+Hd+(Hd)	.561***	.595***	.579	.167	−.055	.597	**.463*****	−.130	.699
H:(H)+Hd+(Hd) ratio	−.123	−.287*	.761	−.239*	.233*	.791	−.217	**−.542*****	.850
Indexes									
PTI	.049	.248*	.484	.089	.045	.492	.002	−.122	.496
SCZI	.140	.196	.498	.071	.000	.502	.105	−.089	.512
DEPI	.101	.135	.284	.032	.000	.286	.241*	.045	.369
CDI	.014	.080	.576	.045	−.118	.585	.095	−.095	.586
S-CON	.251*	−.019	.470	−.158	.126	.503	.126	.105	.492
HVI	.448***	.374**	.621	**.345****	.134	.685	.269*	.249*	.684
OBS	.379**	.271*	.246	.000	.197	.312	.197	−.134	.337

Note. T1 = Time 1; T2 = Time 2. Model 1 explores the role of the average productivity level (*Mean* R). Model 2 explores the role of productivity variations (*Var* R). Block 1 consists of the T1 variable. Block 2 consists of the moderator and Block 3 the interaction term of T1 variable × moderator. Blocks 2 and 3 predict residuals of Block 1. The absolute value of the residuals are predicted in Model 1. Signed residuals are predicted in Model 2. Full tables of β and t values may be sent on request. Variables in brackets indicate a base rate lower than .05. Bold cases indicate a medium-large effect (ΔR ≥ .30).

*$p < .05$. **$p < .01$. ***$p < .001$.

designate instances when the moderator variable contributes to prediction by changing the regression intercept without changing the regression slope. This occurs when the moderator itself makes a significant additive contribution to predicting the criterion such that the intercept of the predicted value changes as a function of the moderator. In this context, the default value or benchmark for the slope of the line predicting T2 scores from T1 scores is adjusted up or down depending on the values for *Mean* R or *Var* R. Using the term *moderator* when either the slope or the intercept of the T1 to T2 prediction equation changes reflects the goals of this research in which it is as important to identify and correct for instances when scores are systematically overpredicted or underpredicted by incorporating the additive effects of the moderator (i.e., higher or lower intercepts) as it is to identify and correct for instances when scores are more or less easy to predict by incorporating the multiplicative effects of the moderator (i.e., steeper or flatter slopes). Moreover, for *Var* R, we formulated specific directional hypotheses for the linear and additive effect of the moderator. There were not theoretical reasons or previous empirical findings to justify predictions associated with the nonlinear and multiplicative effect of the moderator.

The second way the analyses presented here differs from the traditional Baron and Kenny (1986) approach is by virtue of the fact that Blocks 2 and 3 of the regression equation are predicting the residuals from Block 1. We did this deliberately to model the unique contribution of the proposed moderator on the retest coefficient to obtain a better estimate of how these findings would generalize to protocols obtained in clinical practice under altered administration conditions.

RESULTS

In preliminary analyses, we explored the correlations of the scores with R and observed that out of 83 variables, 42 were related to R at T1 and 37 at T2. The correlations were large ($r \geq .50$) on FM+m, es, Adjes, m, H+A, D Location, Dd, and H+Hd+(Hd) (at T1) and PHR, H+(H)+Hd+(Hd), H+A, Hd+Ad, D, Dd, and (H)+Hd+(Hd) (at T2) (see Table 2). A somewhat closer relation of R with Rorschach scores was observed at T1 than at T2 as shown in Figure 1 by the scatterplot displaying the correlations of scores with R at T2 as a function of the correlations of scores with R at T1. As shown in Figure 1, 3 scores are more closely associated with R at T1 (DQv, SumY, and m), and 5 scores are more closely associated with R at T2 (SumT, M–, FC:CF+C ratio, PHR, and Hd+Ad). Overall, however, the magnitude of the correlations between scores and R were consistent with what had previously been observed in other samples (see review in Meyer, 1992, 1993).

When examining the results of regression analyses as described in the Method section, the average level of productivity was associated with residuals in 18 scores (see Model 1, Table 2), and 12 medium-large correlations were found in

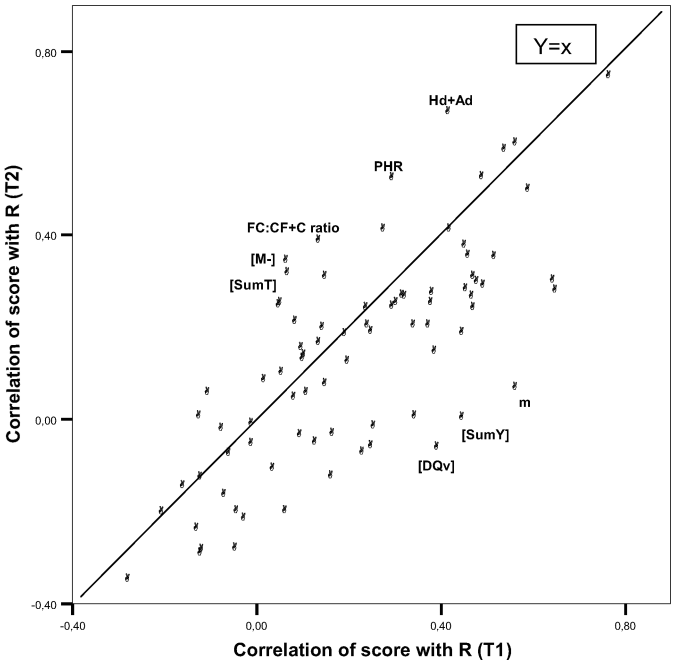

FIGURE 1.—Scatterplot of correlations of 83 scores with R at T1 and T2. Note that labeled variables are variables with residuals over 1.5 *SD*s in the simple regression equation.

FM+m, es, D score, FC, FC:CF+C ratio, GHR, H+A, Zf, W, Dd, DQ+, and HVI. *Mean* R was positively correlated to residuals in all 12 cases, and thus higher average productivity was associated with more instability in these variables.

When examining the effect of the interactions of T1 scores with productivity, three significant predictors were found, but none of them were medium-large in size. To interpret interactions, we used graphical procedures to visualize the regression plane in three dimensions, including the ITALASSI interaction viewer (Provalis Research, 2008) and local linear regression (LLR) smoothing in SPSS.[3] Although the latter allows corners of the regression plane to bend independently, which is not true for an actual interaction in which the bending is symmetrical in opposing corners, the LLR display allows one to visualize the three-dimensional regression surface in the context of the actual data points. For the significant interactions, it was clear that the regression planes bended to meet one or two extreme data points. Additional analyses showed that the cases with extreme data points on the regression surface also had excessive leverage values or Cook's distance values in the regression models and thus were very influential in the final results (Tabachnik & Fidell, 2006). So, regression models were rerun for cases with leverage $< .20$. This resulted in ΔRs associated with interactions dropping to no more than .20. Of importance, omitting these cases had no impact on the magnitude of the main effects. Hence, the significant interactions appeared unstable and due to individual cases with large leverage values. Overall, these analyses suggest that the interactions should not be emphasized here (e.g., Babyak, 2004; Garson, 2008).

[3]This information is presented here in some detail because there was discussion during the review process around the relative importance of main effects versus interaction effects in these analyses.

Interpretations based on the regression analyses were in line with additional analyses in which we examined the relationship of raw differences (absolute stability estimates) to the main effect moderators. For each score, we calculated the difference (T2 – T1), and we correlated this with *Mean* R as summarized in the Appendix.[4] The size and direction of associations reported in the second data column of the Appendix are in line with the results observed for the main effects in the regression analyses.

To summarize, and if we focus on medium-large effects, lower mean scores on R were associated with higher stability in FM+m, es, D score, FC, FC:CF+C ratio, GHR, H+A, Zf, W, Dd, DQ+, and HVI. However, even in those cases with $\Delta R \geq .30$ in Blocks 2 and 3, one must bear in mind that the true additional variance brought by the main effects (and interactions) was rather limited given the fact that Blocks 2 and 3 predicted residuals in which the effect of T1 was initially partialled out. For example, in H+A in which the semipartial correlation of *Mean* R with the residuals was .437 ($p < .001$), the actual additional variance in the T2 score that could be accounted for by productivity and its interaction with T1 was approximately 9% (this figure can easily be computed from Model 1 in Table 2 by subtracting the squared test–retest r in column 4 [$.773^2 = .5975$] from the squared Total R in column 7 [$.832^2 = .6922$; $.6922 - .5975 = .0947$).

When considering variations in the number of responses (*Var* R; see Model 2, Table 2), this was associated with signed residuals in 25 scores, with 9 scores showing medium-large correlations: S, passive, Human Contents, (H)+(Hd), H+A, X+%, D, Dd, and (H)+Hd+(Hd). Higher variations, either increases or decreases from T1 to T2, were associated with greater instability. A negative association was observed in X+% indicating the following pattern: to the extent that R_{T2} was higher than R_{T1}, then $X+\%_{T2}$ was lower than $X+\%_{T1}$. Similarly, to the extent that R_{T2} was lower than R_{T1}, then $X+\%_{T2}$ was higher than $X+\%_{T1}$. The reason for this negative association is due to the fact that X+% is negatively correlated with R at both T1 and T2. As people give more responses, they give more FQu and FQ– relative to FQo or FQ+; so when R increases from test to retest, X+% decreases, and when R decreases from test to retest, X+% increases. As with the positively correlated variables, X+% is more stable the closer *Var* R is to zero.

When examining the effect of interactions of T1 with productivity variations, this factor moderated stability in 12 variables, with medium-large effects affecting four: Lambda, SumT, W:M ratio and H:(H)+Hd+(Hd) ratio. We used the same graphical and statistical procedure as detailed previously to investigate the effect of the interactions. We made the same observations when we reran regressions for cases with nonextreme leverage or Cook's distance values. Excluding these outliers did not affect the results of the main effect. So again the interactions appeared unstable.

Figure 2 illustrates both the instability of the interaction terms and the difference between the main and interaction effects. It

[4]As detailed in Sultan et al. (2006), mean-level changes were very limited in this sample. As a consequence, for a given score, absolute stability and relative stability estimates were close to each other. In fact, within the set of 83 variables considered here, a correlation of .60 was found between the test–retest rs (indicative of relative stability) and the coefficient of absolute variation |T2 – T1|/T1 (indicative of absolute stability).

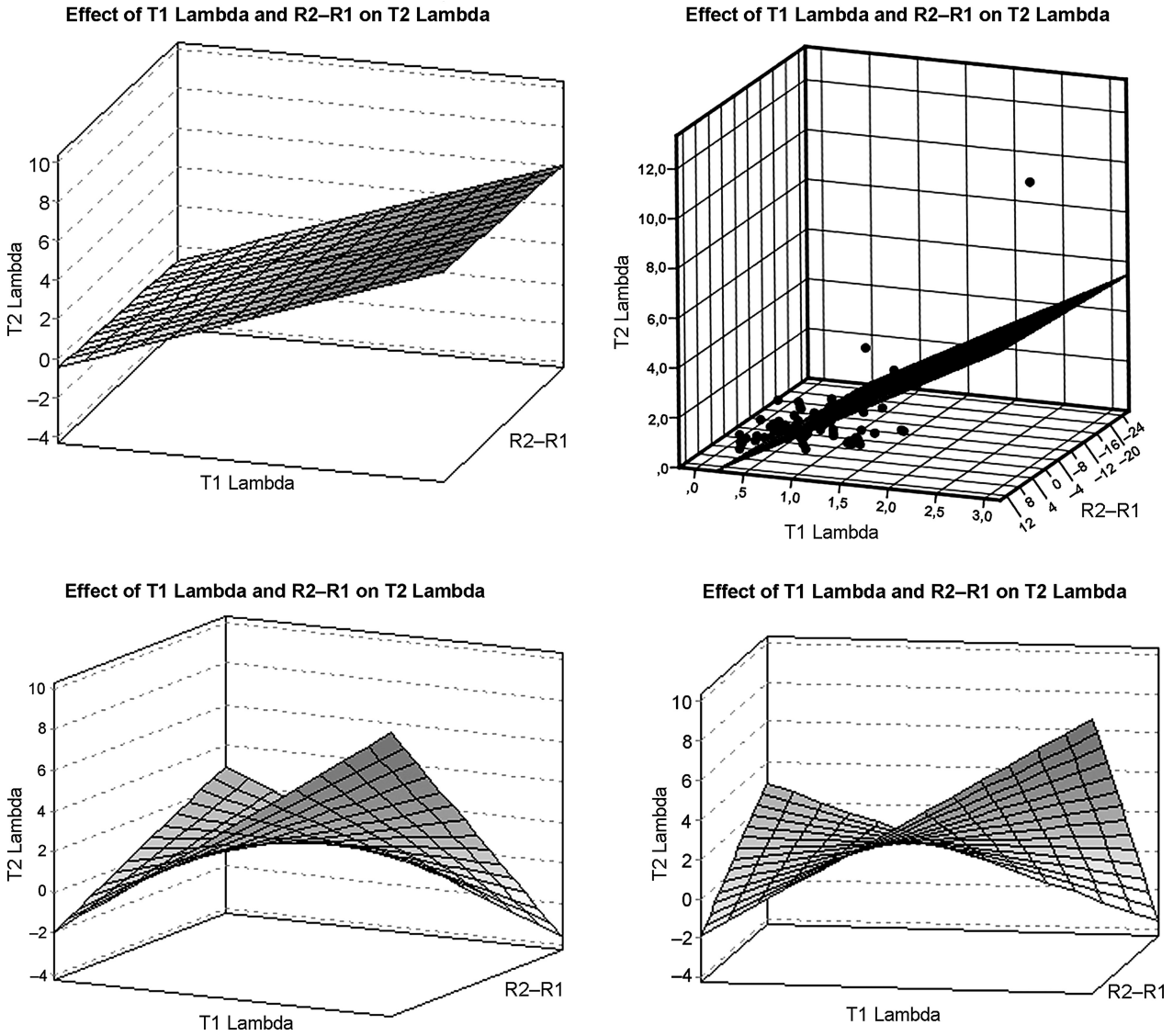

FIGURE 2.—The regression surface for predicting Lambda at T2 from Lambda at T1 and *Var* R with and without the interaction term.

depicts the regression plane for predicting Lambda at T2 (the vertical *y*-axis) from Lambda at T1 (the left to right *x*-axis) and *Var* R (the front to back *z*-axis). For these illustrations, we used the raw variables because identical surfaces are always obtained with raw or centered data (just the placement of the surface differs). The top two panels depict the regression surface without the interaction. The upper panel on the left was produced by the ITALASSI program, which does not depict the actual data points. The upper panel on the right was produced by SPSS and it shows the regression surface in the context of the actual data points. From either graph it is clear that the regression plane is flat and that the slope rises fairly steeply on the *x*-axis but almost not at all on the *z*-axis, consistent with the data in Table 2 showing that the Lambda retest coefficient is .72 (the *x*-axis slope), but the contribution from *Var* R is just .06 (the *z*-axis slope), which in turn is consistent with the first and second data columns showing that Lambda is generally uncorrelated with R at T1 or T2. If *Var* R was a significant positive predictor, the regression surface would remain flat, but it would tilt up when going from back to front,[5] with this slope documenting the changing intercept as a function of differences in R at T1 and T2. An example of this kind of graph is provided in Figure 3, which shows the prediction of all Human Content at T2 from Human Content at T1 and from *Var* R.

Returning to Figure 2, the bottom two panels depict two views of the regression surface with the actual interaction plotted. An interaction is the degree of symmetrical departure from a flat surface; here it can be seen with the corners being elevated when both *Var* R and Lambda at T1 are high or when they are both low. Countering this upward pull, the two opposing corners are pushed down in equal measure (i.e., when *Var* R is low and Lambda at T1 is high or when *Var* R is high and Lambda

[5]The z-axis in these figures is plotted with larger positive values toward the front and increasingly negative values toward the back.

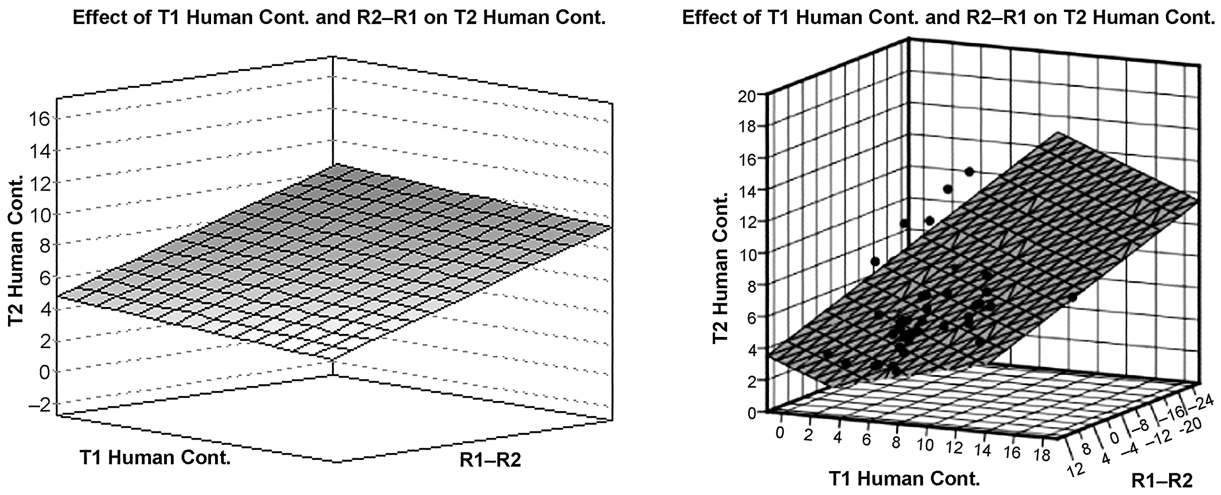

FIGURE 3.—The regression surface for predicting Human Content (Hum Cont) at T2 from Human Content at T1 and *Var* R without the interaction term.

at T1 is low). What is important to note, however, is why this nonlinear bending occurs. It is because a single outlying data point is exerting a strong influence on the regression model. At T1, this man had a Lambda value of 3.2; but it jumped to 12.0 at T2. At T1, R was 21; and at T2, it was 26, producing a *Var* R value of +5.0. The regression model seeks to minimize the squared deviations from the regression surface, and when an interaction term is considered for the model, the surface is pulled up to meet this outlying data point and reduce its degree of poor fit. However, by necessity, the regression surface is symmetrically pulled and pushed in the three other corners to a similar degree. Thus, even though there are no cases above the regression plane in the corner where Lambda at T1 is low and *Var* R is negative, that corner is pulled up by necessity given that a product term interaction is being tested for significance. This is also why the regression surface drops down to produce impossible predicted values of about −4.0 for Lambda at T2 in the back right corner of the graph; it must do this by virtue of the fact that it has risen in the right front corner of the graph to meet the outlying data point. Finally, consistent with the regression diagnostics described previously, the Cook's distance value for the influential outlying case was 4.61, and its leverage value was .30. Because cases with Cook's distance values above 1.0 are considered problematic and because the observed value was more than 13 times larger than the next highest value (.34), it was clear that this case alone was producing a seeming interaction effect that did not generalize to the rest of the data set.

The observations based on the main effects of the regression analyses (i.e., the flat surface) were further confirmed by examining the simple correlations of *Var* R with differences in each score as detailed in the Appendix (columns 4 and 8). All medium-large correlations in the main regression analyses were in the direction predicted by the absolute stability analysis provided in the Appendix.

To summarize, and if we focus on medium-large effects, lower stability was associated with higher response frequency variations in S, Passive Movements, Human Contents, (H)+(Hd), H+A, D Locations, Dd Locations, (H)+Hd+(Hd), and X+%. When taking into consideration graphical and case wise diagnostic procedures, interaction effects were marginal or clearly spurious. Again, however, the actual additional variance of the T2 score residual that was accounted for by productivity variation in Blocks 2 and 3 was limited. For example, for D Location, whereas the ΔRs for the main and interaction effects were .64 and −.11, respectively, the actual additional variance of T2 accounted for by productivity variation and its interaction with T1 was only approximately 16% (this figure can be computed from Model 2 in Table 2 by subtracting the squared r in column 4 [$.785^2 = .616$] from the squared Total R in column 10 [$.882^2 = .778$; $.778 - .616 = .162$]).

The effect of response variation on score stability was in line with the extent to which the scores were correlated with R. Within the 83-variable data set, we observed a correlation of .78 between the mean of the correlation of R with each individual score at T1 and T2 on one hand and the correlation of *Var* R with the T2 – T1 difference in the score on the other hand. So, the correlations of the score differences with *Var* R (reported in column 4 of the Appendix table) are virtually identical to the average of the correlation between the scores and R at T1 and T2 (reported in columns 2 and 3 of Table 2). In other words, to the extent that scores are correlated with R, variability in R over time produces instability in the score. We also observed, although it is not clear why, that the correlation between the score differences and *Var* R was more closely associated with the correlation of each score with R at T1 ($r = .84$) than at T2 ($r = .54$).

Finally, we transformed 19 variables for which an impact of R on stability had been demonstrated by dividing the scores by R. At this step, only variables for which at least one medium-large main or interaction effect was previously observed were included. We excluded variables that already control for R through the use of percentages (X + %) or ratio values (FC:CF+C, W:M, and H:nonPureH). For Lambda, we used the alternative score PureF%, which is obtained by dividing F by R and is superior to Lambda for research purposes (Meyer, Viglione, & Exner, 2001). We then computed the same regression analyses as in Table 2 on the transformed variables.

As shown in Table 3, this transformation was beneficial for 14 of the 19 variables: PureF%, FM+m%, D score%, SumT%, FC%, S%, passive%, Human contents%, (H)+(Hd)%, Zf%, Dd%, DQ+%, (H)+Hd+(Hd)%, and HVI%. In fact, when these

TABLE 3.—Summary of hierarchical regression analyses with T2 transformed variables as dependent variables and T1 transformed variables, moderators, and their interactions as independent variables ($N = 75$).

	Correlations		Block 1	Model 1: *Mean* R (Absolute Value of Residuals Predicted)			Model 2 : *Var* R (Signed Residuals Predicted)		
	T1 With R_{T1}	T2 With R_{T2}	Test–Retest *r*	Block 2 ΔR	Block 3 ΔR	Total R	Block 2 ΔR	Block 3 ΔR	Total R
Core Section									
PureF %	.137	.264*	.672	−.055	−.164	.684	.055	.000	.673
FM+m %	.045	−.236*	.574	.100	.055	.582	−.063	−.145	.588
es%	−.039	−.258*	.439	.000	.158	.461	.063	−.**318****	.527
D score%	−.160	−.005	.269	−.032	.000	.271	−.100	−.045	.289
[SumT]%	−.166	.056	.516	−.141	.000	.530	.053	.195	.519
Affect									
FC%	.042	.050	.575	−.032	−.071	.578	.045	.184	.596
S%	.025	.021	.652	−.173	−.141	.674	.042	.077	.653
Interpersonal									
GHR%	−.136	−.263*	.569	−.134	−.**315****	.635	−.032	−.032	.570
passive%	−.051	−.107	.612	−.184	−.202	.649	.049	−.063	.613
H+(H)+Hd+(Hd)%	−.173	−.035	.664	−.089	−.283*	.700	.033	.045	.664
(H)+(Hd)%	.007	−.077	.349	−.210	.000	.401	.225	−.126	.408
H+A%	−.271*	−.386**	.805	.045	.173	.812	−.205	−.**327****	.817
Processing									
Zf%	−.312**	−.431***	.766	−.164	−.118	.777	−.235	−.221	.794
W%	−.324**	−.459***	.817	−.138	.000	.821	−.220	−.**338****	.829
D%	.245*	.333**	.749	−.045	−.055	.750	.078	−.**392****	.758
Dd%	.209	.348**	.536	.045	−.055	.539	.221	.000	.568
DQ+%	−.062	−.220	.727	.167	−.071	.738	−.122	−.173	.741
Self-Perception									
(H)+Hd+(Hd)%	−.056	.174	.486	.217	−.241	.562	.186	−.077	.512
Indexes									
HVI%	−.064	−.120	.623	−.071	−.126	.633	.063	.170	.639

Note. T1 = Time 1; T2 = Time 2. Model 1 explores the role of the average productivity level (*Mean* R). Model 2 explores the role of productivity variations (*Var* R). Block 1 consists of the T1 variable. Block 2 consists of the moderator and Block 3 the interaction term of T1 variable × moderator. Blocks 2 and 3 predict residuals of Block 1. The absolute value of the residuals are predicted in Model 1. Signed residuals are predicted in Model 2. Full tables of β and *t* values may be sent on request. Variables in brackets indicate a base rate lower than .05. Bold cases indicate a medium-large association with residuals ($\Delta R \geq .30$).

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

variables were divided by R, no significant main or interaction effects could be observed with either of the moderators. In 5 cases, the transformation reduced the impact of the previously found moderator but also introduced a new moderator: es%, GHR%, H+A%, W% and D%. In each case, what had been a main effect was now an interaction effect. In summary, transforming scores into ratios by dividing them by R was beneficial in important categories of the interpersonal sections and some Location scores (Dd and Space). Importantly, applying this transformation did not have a negative impact on test–retest stability coefficients. The stability of the transformed variables remained in the same interpretative range as the original scores. In 6 variables, the transformation even yielded slightly higher test–retest correlations: FM+m% (.57 vs. .50), GHR% (.57 vs. .56), H+A% (.81 vs. .78), Passive% (.61 vs. .55), Zf% (.77 vs. .76), and DQ+% (.73 vs. .71).

DISCUSSION

Our analysis first demonstrated that the stability of 18 variables out of 83 (22%) was associated with the overall mean level of productivity. In 12 cases, effects were medium-large, with the stability of variables such as es, D score, FC:CF+C ratio or Zf being lowered when the average R was larger. So, the impact of productivity appeared important, with high productivity negatively impacting the stability of Rorschach scores. This result was somewhat unexpected because previous research has suggested a lower R would favor instability. However, this was in line with expectations formulated on a psychometric basis in which by definition a higher average of R allows for greater variability in R across time.

Second, when we examined the impact of productivity variations between the two tests, main effects analyses revealed that the stability of 25 variables out of 83 (30%) was associated with variations in R, with medium-large effects observed for 9 variables including Passive Movements, Human Contents, and D Location. Higher variations (such as an important increase in R) were associated with lower stability. These results are very much in line with our expectations because variations in R led to instability in those variables most correlated with R. Examination of the variable set showed that to the extent a score is correlated with R variability in R over time produces instability in the score. For example, the semipartial correlation associated with productivity variations in Space responses was .39 ($p < .01$; Table 2, column 8) and this score correlated .46 and .35 with R at each test (both ps $< .01$; Table 2, data columns 2 and 3). In this score, response frequency variations were associated with the score variation ($r = .47$, $p < .001$; Appendix, column 4). Overall, these results suggest that if we could control for variations in R, we should find benefits in approximately 30% of the variables from the lower part of the structural summary. So, in addition to average productivity, productivity variation had a clear impact on the stability of scores. The examination of interactions between T1 scores and productivity was somewhat disappointing though because the few interactions with medium-large effects were revealed to be unstable. They largely disappeared once

cases with problematic Cook's distance or leverage values were excluded.

However, the correlations of each individual score with response frequency were lower at T2 than they were at T1. The magnitude of the associations at T1 looked more typical and expected than the values at T2. Two hypotheses may account for this. First, at T2, two examiners had to leave the study and were replaced by others who were more experienced. The effect of more experienced examiners might have had an impact on the effectiveness of inquiry or rapport with the examinee. Thus, even records with low R*s* may include variability in Determinants or Contents such as Colors, Textures, or Human and Animal details (as shown in Figure 1). Recent research has shown that the examiners' experience could impact Rorschach data (Lis, Parolin, Calvo, Zennaro & Meyer, 2007). Second, as a result of habituation to the task, people may have been less guarded or anxious at retest and thus expressed relative richness despite low response frequencies. At T1, higher response frequency was indeed accompanied by a relative high frequency of m, SumY, and DQv as Figure 1 suggests. This may reflect higher guardedness and anxiety at T1 than at T2.

In summary, our findings were contrary to the widespread belief that short Rorschach records have lower temporal stability, as cases with more responses were less stable on 12 variables with medium-large effects. Several reasons may account for this phenomenon. First, from a psychometric viewpoint, higher means should be associated with increased variability. This is reflected in the large correlation between the two moderators. Some scores whose stability is impacted by the average R are also impacted by *Var* R: H+A and Dd. However, for some important variables, the impact of the average R was not accompanied by an impact of the R differences, which suggests that productivity in itself may have a negative impact on stability. Reasons specific to the Rorschach task may play a role to explain this phenomenon. In fact, longer records increase the risk that nonreplicable examiner effects might influence the record such as inquiry for Locations and Determinants. This may take place even if R does not vary from one test to the next. In the results, the variables affected by *Mean* R but not so much by *Var* R were FM+m, D score, FC, FC:CF+C, Zf, W, and DQ+. For these variables, differences in inquiry style, including the number and type of inquiry questions posed, could have some influence; and longer records may allow for these kind of undesirable effects to be manifest. To the extent this supposition is true, it would suggest the need for more cohesive inquiry guidelines. Another possible reason may relate to the personality traits of people who give longer records. These people may tend to be more creative, spontaneous, or flexible. As such, they may be more prone to change their responses from one testing occasion to the next. This hypothesis would certainly deserve to be researched in the future.

The results also show that when the response frequency varies between test and retest in one and the same individual, some of the features of the Structural Summary will probably vary at the same time, and the overall description of the individual will be different (all 9 medium-large correlations in the expected direction). However, the impact of productivity is not homogeneous across domains of psychological functioning. The impact of R was clear on the Processing (6 variables out of 9), Interpersonal (5 variables out of 17), and Core sections (4 variables out of 19). It was limited in the Affect, Mediation, and Self-perception sections, and the Constellation Indexes (1–3 variables). No impact was observed within the Ideation section.

As indicated in the introduction, comparisons with Exner's (1988) findings are difficult. Our results, however, do not seem to be easily compatible with Exner's (1988). Of the 20 variables whose stability was moderated by R in Exner's (1988) study and that were also examined here (Exner, 1988, Table 1, p. 643), 6 exhibited modified stability in response to mean R or variations in R with medium-large effects: Zf, passive, FC, SumT, es, and D score. Among these, for SumT, the interaction effect was due to extreme data points in two cases with excessive leverage. In contrast, two of the variables whose stability levels were not impacted in Exner's (1988) study (X+% and D score) were impacted by variations in R in our analyses (medium effect sizes). However, in contrast with Exner's (1988) study suggesting that higher productivity is desirable, our results show that higher productivity may be detrimental for stability.

Our results complement Exner's (1988) in that the impact of R is now clearer and independent of any threshold. Effects were demonstrated here in a sample in which all the records had a frequency higher than 14. The results also lead to a very important and useful insight: The problem is not one of increasing the number of responses because this can also impair stability, but to limit productivity, at least in so-called valid records ($R > 14$). In addition, results suggest it could be beneficial to limit the variation in R to minimize the impact of response frequency on stability. As we showed in the results, stability is impacted by response variations especially for scores that are correlated with R.

Finally, the results obtained after the transformation of scores into percentages were positive for 14 variables out of 19. This transformation was particularly beneficial for the Interpersonal section (e.g., Human Contents), the Affect section, and the Core section. One important observation is that this transformation had no negative impact on test–retest stability. We also note that in a large majority of cases (with the exception of X+%), variables that already control for R, such as Afr and XA%, were not influenced by the mean level of R or differences in R. This also suggests that using ratios and percentages may be beneficial for some important variables that are correlated with R. Although the stability of these transformed variables will not be impacted by productivity or productivity variations, their stability levels remain comparable to the original variables and, in some cases, may be enhanced (e.g., FM+m%, passive%, or H+A%).

This study has some limitations. The main limitation deals with sample size: 75 cases is usually considered a small sample for studying test–retest reliability, especially if we consider the high number of variables in the structural summary. To partly deal with this problem, we focused on medium-large associations. Yet it is probable that the limited *N* also limited the power of the analyses. In addition, some of the results should be interpreted with care given that the base rates were low for many scores. It has been shown that low base rates could influence correlations and reliability estimates (Meyer et al., 2002; Viglione & Taylor, 2003). The variable for which we have significant results and that may be affected by low base rates in this data set was SumT.

Our results therefore indicate the need to find ways of controlling for the impact of R to enhance stability. This observation is in line with previous formulations in favor of treating the

"problem of R" for psychometric purposes (Cronbach, 1949). Three strategies have been suggested by researchers and clinicians to control for the effects of R. First, we could extend the use of ratios and percentages when evidence shows that this is beneficial. Ratios and percentages are already used for some scores, and they are easy to calculate. This ratio method can be studied in new empirical research and also applied retrospectively to the vast body of available archival data to determine its utility. However, this transformation was not beneficial in all cases in our study and did not systematically improve test–retest correlations of Rorschach variables. In addition, psychometricians have objected it could not be optimal in variables in which the association with R is not linear (e.g., Fiske & Baughman, 1953; Kalter & Marsden, 1970).

Another option calls for modified administration guidelines. A recent study suggested that modified administration guidelines to limit the range of R actually increased validity for assessing psychosis (Dean et al., 2007). Dean et al. examined the effect of a nonintrusive method for constraining responses by prompting for an extra response when only one is offered to a card and by removing the card after four responses are given. Among patients, this procedure demonstrated improved Comprehensive System validity in assessing external criteria of thought disorder. However, the question is raised as to what range of R should be tolerated, that is, what degree of variability in R would have a negligible impact on reliability and stability, validity, or utility. Had the Dean et al. procedure been used in this study, it should have limited the range of both moderators and, as a consequence, it is probable that stability would be less impacted by average R or differences in R. Yet, one cannot be sure of the precise effect these changes would have on stability data before they are evaluated.

Finally, another possibility would be to restrict the number of responses to a fixed number. Meyer (1992) suggested requiring precisely two responses by card. Although it is supported by psychometric experts who have underlined the pervasive influence of R on Rorschach scores, we are not aware of any empirical research using this strategy on the Rorschach. This option was adopted in Holtzman's inkblot test (Holtzman, Thorpe, Swartz, & Herron, 1961) and yielded satisfactory psychometric properties (Lilienfled et al., 2000). As for the Rorschach, this proposition would solve the problem of R radically. Yet, it would imply ignoring information clinicians usually find meaningful such as productivity variations between the 10 cards or the Affective ratio. Moreover, like any important adaptation to current guidelines, it would require collecting new normative samples around the world.

To deal with the "problem of R," one should be cautious and perhaps differentiate short- and long-term issues. In the short term, and on the basis of stability data included in this work, one could recommend the use of ratios and percentages whenever empirical evidence shows it is beneficial. In that respect, reanalyses of existing databases would be very useful to ascertain that ratios have equal validity and other important properties as the original raw variables. Once confirmed, clinical practice should be modified by replacing raw count variables in the Structural Summary with ratios and percentages. In the longer term, exploratory work should address the issue of modifying administration guidelines including the possibility of optimizing the number of responses. Given the effects of R shown in this study, we urge researchers to compare these different alternatives in empirical studies, as they may well have important effects on the quality of the information derived from the Rorschach. Of course, before commencing this research agenda, it would be wise to replicate these findings in other samples to confirm the impact of productivity on stability.

Finally, we note that the same pattern of results might equally well be observed with other procedures such as storytelling techniques in which verbal productivity may play a big role. In such cases, productivity should be routinely examined as a covariate as it is done in recent Rorschach research and be the focus of applied research to improve validity, utility, and reliability of psychological assessment.

ACKNOWLEDGMENTS

A preliminary version of this article was presented at the 8th Congress of the European Rorschach Association in Padua, Italy (August 2006). We thank other contributors whose input was of immense value at an earlier stage of the study: Anne Andronikof, Christian Réveillère, Gilles Lemmel, Damien Fouques, and Thomas Saïas. This project received support from the Rorschach Research Foundation. Aspects of the analyses benefited from fruitful discussions with Cato Grønnerød, Donald Viglione, and three anonymous reviewers.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions* (3rd ed.). Newbury Park, CA: Sage.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66,* 411–421.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1949). Statistical methods applied to Rorschach scores: A review. *Psychological Bulletin, 46,* 393–429.

Dawes, R. (1999). Two methods for studying the incremental validity of a Rorschach variable. *Psychological Assessment, 11,* 297–302.

Dawes, R. M. (2001). Incremental validity of the Ego Impairment Index: It's fine when it's there. *Psychological Assessment, 13*, 408–409.

Dean, K. L., Viglione, D. J., Perry, W., & Meyer, G. J. (2007). A method to optimize the response range while maintaining Rorschach Comprehensive System Validity. *Journal of Personality Assessment, 89,* 149–161.

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

Exner, J. E., Jr. (1992). R in Rorschach research: A ghost revisited. *Journal of Personality Assessment, 58,* 245–251.

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

Fiske, D. W., & Baughman, E. E. (1953). Relations between Rorschach scoring categories and the total number of responses. *Journal of Abnormal and Social Psychology, 48,* 25–32.

Garson, G. D. (2008). "Regression analysis," from *Statnotes: Topics in multivariate analysis.* Retrieved July 18, 2008, from http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm

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

Holtzman, W. H., Thorpe, J. S., Swartz, J. D., & Herron, E. W. (1961). *Inkblot perception and personality.* Austin, TX: University of Texas Press.

Hunsley, J., & Bailey, J. M. (2001). Whither the Rorschach? An analysis of evidence. *Psychological Assessment, 13,* 472–485.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15,* 446–455.

Kalter, N., & Marsden, G. (1970). Response productivity in Rorschach research: A caution on method. *Journal of Projective Techniques and Personality Assessment, 34,* 10–15.

Kinder, B. N. (1992). The problems of R in clinical settings and in research: Suggestions for the future. *Journal of Personality Assessment, 58,* 252–259.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Lis, A., Parolin, L., Calvo, V., Zennaro, A., & Meyer, G. J. (2007). The impact of administration and inquiry on Rorschach Comprehensive System protocols in a national reference sample. *Journal of Personality Assessment, 89,* S193–S200.

Meyer, G. J. (1992). Response frequency problems in the Rorschach: Clinical and research implications with suggestions for the future. *Journal of Personality Assessment, 58,* 231–244.

Meyer, G. J. (1993). The impact of response frequency on the Rorschach constellation indices and on their validity with diagnostic and MMPI-2 criteria. *Journal of Personality Assessment, 60,* 153–180.

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

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13,* 486–502.

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., Mihura, J. L., & Smith, B. L. (2005). The interclinician reliability of Rorschach interpretation in four data sets. *Journal of Personality Assessment, 84,* 296–314.

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.

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Ordnuff, S. R., Freedenfeld, R., Kelsey, R. M., & Critelli, J. (1994). Object relations of sexually abused female subjects: A TAT analysis. *Journal of Personality Assessment, 63,* 223–228.

Perry, W. (2001). Incremental validity of the Ego Impairment Index: A reexamination of Dawes (1999). *Psychological Assessment, 13*, 403–407.

Perry, W. (2003). Let's call the whole thing off: A response to Dawes (2001). *Psychological Assessment, 15*, 582–585.

Perry, W., Moore, D., & Braff, D. (1995). Gender differences on thought disturbance: Measures among schizophrenic patients. *American Journal of Psychiatry, 152,* 45–67.

Provalis Research. (2008). Italassi v1.1 interaction viewer. Retrieved July 10, 2008, from http://www.provalisresearch.com/ITALASSI/ITALASSI.php

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

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

Tabachnik, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (4th ed.). Boston, MA: Pearson Education.

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

Wagner, E. E., Young, G. R., & Wagner, C. F. (1992). Rorschach blends, IQ, and the effect of R. *Journal of Personality Assessment, 59,* 185–188.

Wood, J. M., Krishnamurthy, R., & Archer, R. P. (2003). Three factors of the Comprehensive System for the Rorschach and their relationship to Wechsler IQ scores in an adolescent sample. *Assessment, 10,* 259–265.

APPENDIX

Correlation of score differences between T1 and T2 with *Mean* R and *Var* R for each Rorschach score.

Scores	*Mean* R With Raw Differences T2 – T1	*Mean* R With Absolute Value of Differences |T2 – T1|	*Var* R With Raw Differences T2 – T1	Scores	*Mean* R With Raw Differences T2 – T1	*Mean* R With Absolute Value of Differences |T2 – T1|	*Var* R With Raw Differences T2 – T1
Lambda	.017	.002	.118	H+A	**−.310****	**.396*****	**.658*****
M	−.069	.198	.270*	Hd+Ad	**.301****	**.349****	**.299****
WSumC	−.123	.047	.248*	Ma	.085	.102	.029
EB ratio	−.068	−.026	.017	Mp	−.122	.176	.223
EA	−.132	.084	**.353****	Ma:Mp ratio	−.010	−.129	−.186
EBPer	−.066	.014	−.010	Intell	−.082	−.014	**.306****
FM+m	**−.384*****	**.598*****	**.387*****	MOR	−.094	.289*	.013
SumShd	−.141	.062	.307**	Sum6	−.139	.275*	.128
eb ratio	−.071	−.067	−.107	Lv2	.004	.075	−.025
es	**−.319****	**.397*****	**.434*****	WSum6	−.130	.244*	.078
D score	.259*	.260*	−.290*	[M–]	.235*	−.064	.037
Adjes	−.160	.207	.360**	[Mnone]	−.029	−.161	−.076
AdjD	.070	−.079	−.188	XA%	−.071	−.207	−.048
FM	−.200	**.498*****	.278*	WDA%	−.002	−.147	.000
SumC’	−.058	−.081	.165	X – %	.094	−.151	.060
[SumT]	.096	.142	.067	S–	.172	.105	.149
m	**−.432*****	**.363****	**.316****	Pops	−.279*	.111	.282*
[SumV]	.099	.163	.278*	X+%	.019	−.167	**−.362****
[SumY]	**−.322****	.265*	.217	Xu%	−.090	.116	**.358****
FC	.160	**.334****	.199	Zf	−.284*	**.491*****	**.355****
CF+C	−.181	.001	.115	W	**−.360****	**.463*****	.238*
[Pure C]	−.150	.037	.137	D	.056	**.335****	**.725*****
FC:CF+C ratio	**.371****	**.354****	−.040	Dd	−.011	.265*	**.433*****
SumC’:WSumC ratio	−.107	−.084	.030	W:M ratio	−.049	−.064	−.182
Afr	.061	−.121	.133	Zd	.104	.202	−.078
S	−.050	.232*	**.472*****	[PSV]	.048	.056	−.073
Blends/R	−.152	−.210	−.009	DQ+	−.284*	**.588*****	**.317****
COP	−.012	.151	−.089	[DQv]	**−.336****	.283*	**.332****
[AG]	−.083	.106	.034	EGO	.086	.065	−.162
GHR	−.142	.202	**.455*****	[Fr+rF]	.086	.171	.082
PHR	.046	.108	.193	[FD]	−.225	−.007	.112
GHR:PHR ratio	−.061	−.213	.078	An+Xy	−.088	.035	.270*
Active	−.189	.212	.145	(H)+Hd+(Hd)	−.046	.164	**.457*****
Passive	−.280*	**.384*****	**.420*****	H:(H)+Hd+(Hd) ratio	−.102	−.263*	−.204
a:p ratio	−.089	−.213	−.065	PTI	.133	.044	−.054
[Food]	−.146	.137	.174	SCZI	−.044	.028	.047
Human Contents	−.061	.204	**.431*****	DEPI	.008	−.031	.287*
Pure H	−.040	.186	.011	CDI	.019	−.009	.077
[PER]	−.054	.282	.057	S-CON	−.269*	−.022	.175
Isolate/R	−.071	.046	−.092	HVI	−.122	**.377*****	**.311****
(H)+(Hd)	−.238*	.144	**.402*****	OBS	−.086	.028	.296**
[(A)+(Ad)]	.052	−.118	.103				

Note. T1 = Time 1; T2 = Time 2. *Mean* R has a pattern of stronger correlations, with the absolute value of score difference in columns 3 and 7 than with raw differences in columns 2 and 6. Positive correlations in columns 3 and 7 indicate higher *Mean* R is associated with more instability. Given the way *Var* R is computed, raw variations in each score maximize the match between the moderator and the differences in individual scores and are reported in columns 4 and 8. Bold cases indicate a medium to large effect ($r \geq .30$).

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.