DOI: 10.26387/bpa.289.2

The impact of extrapolation and examiner's judgment on Rorschach form quality coding: Interrater reliability and clinical implications

Davide Ghirardello[1], Francesca Ales[1], Valeria Raimondi[1], Donald J. Viglione[2], Alessandro Zennaro[1], Luciano Giromini[1]

[1] Department of Psychology, University of Turin, Italy
[2] Alliant International University, San Diego, California, U.S.

francesca.ales@unito.it

ABSTRACT. Nel test di Rorschach, la Qualità formale (*Form Quality*, FQ) descrive il grado di somiglianza tra la risposta e la corrispondente localizzazione nella macchia, ed è derivata dalla frequenza con cui la risposta stessa è identificata e dal giudizio degli esaminatori (rater) riguardo l'aderenza della sua forma ai contorni della macchia. Un ampio numero di ricerche ha dimostrato che la FQ ha un'eccellente validità come misura dell'esame di realtà e di gravità della psicopatologia. Tuttavia, alcuni studi hanno riportato valori di *interrater reliability* (IRR) non ottimali. Nel presente articolo abbiamo esaminato 1588 risposte raccolte in 60 protocolli Rorschach d'archivio. Abbiamo esaminato la frequenza con cui FQ è stata ricavata dalle Tabelle (T), Estrapolata (E) o stabilita sulla base del Giudizio dell'esaminatore (*Judged,* J), e testato la forza dell'associazione tra il processo di siglatura della FQ e (a) i punteggi delle variabili FQ, e (b) la IRR. I risultati hanno mostrato che, quando confrontate alle risposte T, le risposte E e J erano caratterizzate da FQ progressivamente più scadente e IRR progressivamente meno ottimale. Nel complesso, questi risultati confermano che il processo di siglatura della FQ ha un notevole impatto sull'accuratezza della siglatura e sulla IRR della FQ. Al fine di ridurre le incoerenze riscontrate nella codifica della FQ, gli autori suggeriscono che gli sviluppi futuri dell'R-PAS possano provare a incorporare algoritmi computerizzati in grado di aiutare gli esaminatori nell'attribuzione della codifica FQ.

SUMMARY. Rorschach Form Quality (FQ) describes how well a response fits a given inkblot location and is derived from how frequently it is identified and whether raters judge it to be a good fit. A large body of research has established that FQ has excellent validity as a measure of reality testing and severity of psychological disturbance. However, some studies have reported sub-optimal interrater reliability (IRR). In this article we inspected 1588 responses from 60 archival Rorschach protocols. We examined the frequency with which FQ was Tabled (T), Extrapolated (E) or Judged (J), and tested the strength of the association of FQ determination path to FQ scores and IRR. Results showed that when compared to T responses, E and J responses were characterized by increasingly poorer FQ and less optimal IRR. Taken together, these results confirm that the determination path used to code FQ has a notable impact on the scoring accuracy and IRR of FQ. In order to reduce the FQ coding inconsistencies, the authors suggest that future R-PAS developments might try to incorporate computer algorithms to help with the attribution of FQ codes.

Keywords: *Rorschach, Form quality, Interrater*

INTRODUCTION

Rorschach Form Quality (FQ) measures how well a Rorschach response fits a particular inkblot location, and how frequently it is seen in that location (Meyer, Viglione, Mihura, Erard & Erdberg, 2011). To score FQ one determines whether the chosen inkblot area looks like the objects or object that the respondent sees. This is done by comparing the respondent's perceptions of the inkblot to other respondents' perception of the same inkblot. Thus, FQ is an essential measure of perceptual accuracy and reality testing and one of the key variables of the Rorschach test since its inception (Meyer et al., 2011; Mihura & Meyer, 2018).

Hermann Rorschach himself noted the relationship between the accuracy of response objects offered by the examinee in terms of whether their form matches the shape of the blots and the person's ability to perceive the world in a realistic way (Rorschach, 1921). Although Rorschach created a list of objects to help determine the quality of the forms perceived by the examinees, his premature death interrupted his work and his preliminary interpretations left much to debate (Exner, 1969). In the following years, several Rorschach systems were developed that varied in administration, coding, and interpretation (e.g., Beck, Beck, Levitt & Molish, 1961; Klopfer, Ainsworth, Klopfer & Holt, 1954). Nevertheless, every major Rorschach system included FQ coding, and research established it as a core variable when evaluating psychotic processes, regardless of the Rorschach system being used (e.g., Berkowitz & Levine, 1953; Dao, Prevatt & Home, 2008; Goldfried, 1962; Harder & Ritzler, 1979; Kimhy et al., 2007). These FQ scoring systems incorporated some version of fit and frequency ratings even if they were based on examiner judgment rather than by comparing the given response to tabled lists derived from previously collected quantitative data. In other Rorschach systems, the person scoring FQ looks up the verbalized response object(s) in a list organized by card number and location areas within each card (i.e., FQ tables).

The Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011) was introduced about ten years ago to overcome some of the known psychometric and validity limitations of the comprehensive system (CS; Exner, 2003; Mihura & Meyer, 2018). Like previous systems, it defines FQ as a function of two components of perceptual accuracy: (1) *fit*, i.e., whether the inkblot location looks like the object described, and (2) *frequency*, i.e., how common it is to see that object at that particular location. It improved on other systems by including much more fit and frequency data in its empirical foundation of its tables (Su et al., 2015). When participants respond to what the inkblot might be, the FQ of their visual percepts is categorized as either ordinary (FQo), unusual (FQu), minus (FQ–), or none (FQn). FQo responses are accurate, relatively common, and thus quickly and easily seen (e.g., "a butterfly" to the whole of Card I). FQu responses are less accurate and typically less common. However, they are not extremely inconsistent with stimuli contours (e.g., "bones" to the D7 of Card III). FQ– responses are inaccurate, infrequent, and difficult to see (e.g., "a face" to the D1 of Card X). Therefore, FQ–, FQu and FQo lie on a continuum of increasing accuracy and frequency (Meyer et al., 2011). Finally, FQn responses are typically impressions of the blot based on the color or shading of the ink without any reference to form or shape (e.g., "Blood, it's all red there, there's no particular shape" to the whole of Card II). Unlike the other FQ codes, FQn responses are not coded based on their degree of fit to the stimuli.

It is worth mentioning that these criteria are theoretically and empirically grounded in the Exner's notions of distal properties and critical bits of the blots. Distal properties are defined as true components of the inkblots, while critical bits are powerful visual features of the blots that contribute to the perceptual organization of many responses (Exner, 1996). As such, drawing from the distal properties of the stimuli and recognizing the critical bits in the inkblot can lead to conventional responses, which are currently scored ordinary. Similarly, those percepts that exceed the distal properties of a certain stimulus, may results in non-conventional responses, and, consequently, they are likely to be coded with poorer formal quality (i.e., FQu or FQ–).

Whereas a large body of research has established that FQ codes possess excellent validity as a measure of reality testing abilities and psychopathology (Meyer et al., 2011; Mihura, Meyer, Dumitrascu & Bombel, 2013; Su et al., 2015), some recent studies have reported sub-optimal results with regard to interrater reliability (IRR), as it had been the case with the CS (Viglione & Meyer, 2008). More specifically, four studies were designed to address IRR of Rorschach variables, including FQ: two were conducted at protocol-level, with IRR evaluated via intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979); the other two examined IRR at response-level via Cohen's *k* (Cohen, 1960).

In the first report of R-PAS IRR at protocol-level, Viglione and colleagues (Viglione, Blume-Marcovici, Miller, Giromini

& Meyer, 2012) found that FQo and FQ– were characterized by an excellent IRR, with ICC values of .84 and .81 respectively; these values were comparable to the average ICC of .88 across all variables. FQu, instead, was characterized by good IRR (ICC = .64) which was still satisfactory, but less optimal. More recently, Pignolo and colleagues (2017) provided the first account of R-PAS IRR at protocol-level in a non-American context, basing on raw data and complexity adjusted scores. Concerning raw data, the average IRR for all the 60 variables was excellent, with an ICC of .78; FQo reached an excellent IRR, with an ICC of .82, whereas less satisfactory findings emerged for FQ– and FQu, with fair ICC values of .53 and .59, respectively (the results did not change significantly with complexity adjusted scores).

The two recent studies assessing IRR at response-level yielded comparable results. Kivisalu and colleagues (Kivisalu, Lewey, Shaffer & Canfield, 2016) reported an average κ across 50 variables of .66, reflecting good IRR; while they found that FQo was characterized by excellent agreement (k = .77), FQ– and FQu showed barely good agreement (k = .62 and .59, respectively). The IRR was re-assessed on the same protocols by different raters in a subsequent study by Lewey and collaborators (Lewey, Kivisalu & Giromini, 2018); in this newer study the authors reported excellent agreement for FQo (k = .73), whereas FQu and FQ– were characterized by fair agreement (k = .53 and k = .52, respectively).

Taken together, the results of these four IRR studies indicate that when compared to other R-PAS variables, FQ codes (especially FQu and FQ–) yield relatively poorer IRR, both at the protocol- and at the response-level. From an applied, clinical perspective, only protocol-level IRR results are crucial to ensure that FQ–based clinical interpretations are made reliably. This is because ultimately clinicians only interpret scale level data and do not overly focus on item level results. However, we argue that response-level IRR data are important too, for at least two reasons. Firstly, consistent with our years-long teaching experience, empirical evidence (Viglione, Meyer, Resende & Pignolo, 2017) indicates that learning how to reliably code FQ at the response level is particularly challenging, which potentially contributes to discouraging new learners from wanting – or feeling confident enough – to adopt the Rorschach in their clinical practice. Secondly, response-level uncertainties and disagreements may give to both novel and more experienced Rorschach users an uneasy feeling that their coding may be inaccurate or arbitrary. As a result, they might take some extra-time to score FQ codes and ultimately their FQ based clinical interpretations may be under-weighted or considered with more skepticism that they probably should. By saying this, we do not intend to dramatize FQ as a critical code, but merely to acknowledge that all of these weigh on the cost side of the cost-benefit ratio and thus diminish test utility so that improving both protocol-level and response-level IRR of FQ codes would be beneficial.

In this article, we hypothesize that a possible explanation for the sub-optimal IRR of FQu and FQ– codes is that these codes are at times coded based on the examiner's subjective judgment of the degree of fit between the form of the response object and the contour of the blot where it was seen. The section in the R-PAS manual addressing these procedures (Meyer et al., 2011) is an extension of Exner's CS approach (1974, 2003) and largely derived from refinements to the procedure (Viglione, 2002, 2010). A few years after publishing the manual, the authors identified some limitations to the procedure in the R-PAS manual and uploaded a document (Viglione et al., 2016) on the R-PAS website (www.r-pas.org), which specifies three distinct FQ determination paths: Tabled, Extrapolated, and Judged. Tabled FQ determination occurs when the important response objects are found in the FQ tables. For example, in Card I, W, "The face of a witch" would be coded FQu and would consist of a Tabled determination because in the FQ tables, "Face, Witch" is listed as FQu. At times, however, the response object is not found in the FQ tables and an extrapolation process is required. Typically, Extrapolated FQ determination occurs when FQ is derived from similarly shaped tabled item, e.g., when extrapolating from a rat to a mouse or a hat to a bonnet. For example, in Card V, upside-down, W, one might say "A flower". In the FQ tables, W(v), no objects resemble a flower. However, in the standard position, flower is FQu. Thus, by extrapolation, "A flower" seen upside-down also is coded FQu. This would be called an obvious extrapolation. Extrapolation may also be less obvious and occur when, based on examination of multiple, tabled items, the preponderance of the evidence clearly favors one FQ score over another – or a more reasonable middle way. For example, in Card VIII, D3, "Skull of Bigfoot". By looking at the FQ tables, "Skull (Animal)" is coded FQo, while "Skull (Human)" is coded FQ–. Since Bigfoot has both some animal and human features, and given that there is equal evidence for FQo and FQ–, a reasonable coding would be FQu. It is important to specify that the R-PAS extrapolation procedure, similarly to the CS, is based on the fact that the degree of fit

is relative to the shape of the percept and not to the content per se. Lastly, Judged determination requires the examiner to look at the response in the location where the response was seen, so to establish FQ by answering the question: "Can I see that object in this location quickly and easily?". Coders may resort to Judgment in two situations. First, when FQ tables do not provide comparable responses for extrapolation; second, when FQ tables provides support for both FQ– and FQu (or for FQu and FQo), without a clear basis for preferring one over the other. For example, in Card IX, W, "The hand of a person, kinda like making the sign of peace… like with the two fingers up, you know what I mean?" would be coded using examiner judgment. In the FQ tables, "Hand" or "Fingers" are not listed with reference to W and looking for a rationale among similar or near-W locations also does not help. There is no location at the bottom half of the card to look for the palm of the hand; D3 would likely be the two fingers, but there is nothing similar in shape to fingers there. Thus, there are no guidance or comparable responses for extrapolation in the FQ tables.

Moreover, it is worth mentioning that multiple-object responses represent a tricky element that could generate inconsistencies in FQ coding. Basically, three cases should be taken into account here. Firstly, the FQ tables contain entries that refer to overarching percepts, such as landscape or anatomy. These are superordinate categories that could be used as tabled entries for multiple-objects responses composed by multiple objects or components (the FQ determination path would be Tabled). Secondly, the examiner should search for the most common multiple-object responses that are already listed in the FQ tables (also in this case, the FQ determination path would be Tabled). When the overarching category cannot be used, and the multiple-object response is not listed in the appropriate location area of the FQ tables, the guideline is to determine the FQ code for each important object following the procedure outlined above for single-object responses, and then to use the code down principle by choosing the least accurate (or lowest) FQ code and apply it to the overall response. Here, the attribution of the FQ determination paths follows the same rules described above (Viglione et al., 2016): if FQ is determined based on FQ, the path will be Tabled; if extrapolation is needed, the path will be Extrapolated and, finally, if the FQ is determined via judgment of fit, the path will be Judged. It should be noted that, when coding FQ (and its determination path) for multiple objects responses, it might be difficult to distinguish between important and unimportant objects.

AIM

Because no research has yet reported on the frequency with which FQ is coded based on Tabled (T) *versus* Extrapolated (E) *versus* Judged (J) determination paths, we inspected FQ codes from 60 archival Rorschach protocols and examined the percentage of cases in which FQ was determined based on each of those three paths. Next, as we anticipated that the more a response object is likely to be seen in a specific location of a given inkblot, the higher the likelihood that such a response object would appear also on the FQ tables, we tested the extent to which non-tabled, i.e., E and J paths, associated with poorer FQ outcomes. Lastly, and most importantly, we aimed at quantifying the extent to which the greater the use of some judgment (i.e., E and J paths) in the determination of FQ, the lower the IRR of the resultant FQ codes.

MATERIALS AND METHODS

Rorschach data

Rorschach protocols. For this study, we randomly selected 60 protocols from a broader data set we had access to, consisting of 96 Rorschachs from healthy, undergraduate volunteers with no previous neurological/psychiatric disorders. As further detailed in the journal article describing that data set (Burin et al., 2019), participants' recruitment was undertaken in Turin, in the north of Italy, either at the University of Turin or via snowball sampling, and Rorschach administrations were carried on using standard R-PAS guidelines. Most of the protocols analyzed for the current paper were from women (83.3%), and our sample mean age was 21.48 years (SD = 2.69). The total number of responses was 1588 with an average of 26.47 responses per protocol (SD = 2.77). Six out of the 1588 responses received the code FQn by rater 2, so the number of responses on which the analyses are based is 1582 (i.e., the total number of responses having a form demand).

Rorschach coders. Two of the authors of the current article (i.e., Ghirardello - DG - and Ales - FA) coded the great majority of the protocols originally analyzed in Burin et al. (2019) and all of the 60 protocols selected for the current study. Additionally, together with a third rater (Raimondi - VR), Ghirardello and Ales also independently

re-coded all responses of a selected number of protocols, so that the second coders were blind to any previous coding. Thus, all 60 protocols were eventually coded twice by two different and independent raters. To prevent the same protocol from being coded twice by the same rater, half of the protocols initially coded by Ales were randomly assigned to Ghirardello for the second coding; the other half were assigned to Raimondi. Similarly, half of the protocols originally coded by Ghirardello were randomly assigned for a second coding to Ales; the other half to Raimondi. All three raters were graduate students who had been trained by a member of the R-PAS Research and Development Group (last author).

Procedure

As noted above, the 60 protocols examined for the current study were coded twice, by two different and independent judges. More specifically, at t_1, coding was performed with the purpose of conducting Burin et al.'s (2019) study; at t_2, coding was performed to examine the frequency with which FQ was coded based on Tabled (T), Extrapolated (E), and Judged (J) determination paths, and to test the IRR of FQ codes. As such, in addition to coding FQ, t_1 raters also reported, for each response, what determination path was used to code FQ; t1 occurred in 2016, t_2 occurred in 2017. At both times, when coding FQ, all coders relied on both the coding guidelines reported on the R-PAS manual (Meyer et al., 2011) and the online document elaborated by Viglione et al. (2016) and uploaded in the R-PAS website (www.r-pas.org).

To test the IRR of the FQ determination path classifications, a subsample of 16 protocols from t_1 (8 protocols coded by Ghirardello and 8 coded by Ales) was randomly extracted, and the same raters who had coded FQ at t_1 were asked to re-examine the same responses a second time, to indicate what FQ determination path characterized the attribution of their FQ codes. For these 16 protocols comprising a total of 436 responses (27.5% of the total sample), the FQ determination paths were thus assigned twice (i.e., at t_1 and at t_2), by two independent judges (the 8 records coded by Ghirardello at t_1 were independently re-corded by Ales at t_2, and the 8 records coded by Ales at t_1 were independently re-corded by Ghirardello at t_2). Analyses of the IRR of the FQ determination path yielded a highly satisfactory Cohen's k of .79 (Cicchetti, 1994). Two out of the 436 responses received an FQn code, so the number of responses on which these analyses are based is 434 (i.e., the total number of responses having a form demand).

It should be noted that all judges were blind to the chief hypotheses of the study at t_1 and at t_2. Also, at t_2 each rater was blind to the other rater's codes provided at t_1.

Statistical analysis

Statistical analyses mainly focused on descriptive statistics and χ^2 analyses to determine the frequency with which Tabled, Extrapolated and Judged determination paths were used to code FQ across the ten inkblots. Next, FQ IRR was assessed both at response-level (using Cohen's k) and protocol-level (using ICC). IRR classification are based on Cicchetti (1994) and Shrout and Fleiss (1979): k or ICC values lower than .40 indicate poor IRR, between .40 and .59 fair IRR, between .60 and .74 good IRR, and values at or above .75 suggest excellent IRR. In many studies focusing on Rorschach variables, IRR evaluated via ICC was computed using the two-way random effect model (e.g., Acklin, McDowell, Verschell & Chan, 2000; Viglione et al., 2012), which assumes that the same pair of raters have rated each protocol. In our study, the pair of raters was not the same for all protocols, thus we used a one-way random effects model (for details, see Meyer et al., 2002; Shrout & Fleiss, 1979).

RESULTS

Tabled, Extrapolated, and Judgment determination paths and Form Quality

Table 1 shows the percentage of responses in which FQ was coded based on Tabled, Extrapolated, or Judged determination paths, divided by card. In total, about 60% of the responses were found in the R-PAS FQ tables (T), extrapolation (E) was required in about 30% of the cases, and judgment (J) was required in about 10%.

The distribution of T, E, and J, however, varied across all ten cards, $\chi^2(18) = 59.0$, $p<.001$. More specifically, when compared to all other cards, Card IV was characterized by a significantly higher proportion of E responses ($z = 2.1$), Card

Table 1 – Total and card by card FQ determination path

		T	E	J	Total
Card I	R	118	41	11	170
	% in Card	*69.4%*	*24.1%*	*6.5%*	
	Std. Residuals	1.7	−1.7	−1.2	
Card II	R	99	52	13	164
	% in Card	*60.4%*	*31.7%*	*7.9%*	
	Std. Residuals	.2	.1	−.6	
Card III	R	108	43	14	165
	% in Card	*65.4%*	*26.1%*	*8.5%*	
	Std. Residuals	1.0	−1.2	−.3	
Card IV	R	72	**59**	11	142
	% in Card	*50.7%*	***41.5%***	*7.7%*	
	Std. Residuals	1.3	**2.1**	−.6	
Card V	R	100	**29**	12	141
	% in Card	*70.9%*	***20.6%***	*8.5%*	
	Std. Residuals	1.8	**−2.3**	−.3	
Card VI	R	91	51	16	158
	% in Card	57.6%	*32.3%*	*10.1%*	
	Std. Residuals	−.3	.2	.3	
Card VII	R	107	39	12	158
	% in Card	*67.7%*	*24.7%*	*7.6%*	
	Std. Residuals	1.4	−1.5	−.7	
Card VIII	R	92	51	16	159
	% in Card	*57.9%*	*32.1%*	10.1%	
	Std. Residuals	−.2	.1	.3	
Card IX	R	**60**	**69**	**22**	151
	% in Card	***39.7%***	***45.7%***	***14.6%***	
	Std. Residuals	**−3.1**	**3.1**	**2.1**	
Card X	R	90	64	20	174
	% in Card	*51.7%*	36.8%	*11.5%*	
	Std. Residuals	−1.3	1.2	1.0	
Total	R	937	498	147	1582
	% in Card	*59.2%*	*31.5%*	*9.3%*	

Note. Bolded values represent standardized residuals greater than $|z| = 1.96$.

Legenda. T = Tabled; E = Extrapolated; J = Judged.

V was characterized by a lower proportion of E responses ($z = -2.3$), and Card IX was characterized by a lower number of T responses ($z = -3.1$), and by a higher number of E ($z = 3.1$) and J ($z = 2.1$) responses.

Also noteworthy, different FQ determination paths were associated with different FQ coding outcomes, $\chi^2_{(4)} = 391.1$, $p<.001$. Indeed, Table 2 shows that T responses were positively associated with FQo ($z = 8.1$), and negatively associated with FQu ($z = -7.7$) and FQ– ($z = -3.7$). Conversely, E responses associated positively with FQu ($z = 6.7$) and negatively with FQo ($z = -6.3$), and J responses associated positively with FQu ($z = 7.0$) and FQ– ($z = 6.4$) and negatively with FQo ($z = -8.8$). That is, in line with our hypotheses, compared to T determination path, E and J paths associated with increasingly poorer FQ.

Form Quality interrater reliability at response and protocol levels

The third step of our analyses entailed the evaluation of the impact of the determination path, i.e. Tabled (T), Extrapolated (E) and Judged (J), on FQ IRR at response and protocol level (see Table 3). Focusing on response-level IRR, when disregarding the type of determination path used to code FQ, a general good agreement was found, with $k = .68$. Comparing Cohen's k separately for T, E and J responses, however, revealed that IRR was excellent ($k = .77$) for T, but dropped to fair ($k = .48$) and to poor ($k = .37$) for E and J, respectively.

We next focused on protocol-level IRR (see Table 4). Overall, a good to excellent IRR was observed in all cases

Table 2 – Response-level percentage of Tabled (T), Extrapolated (E) and Judged (J) responses along with their FQ codes

FQ determination path	FQ–	FQu	FQo	Total
T	72	194	671	937
% in T	7.7%	20.7%	71.6%	
Std. Res.	**–3.7**	**–7.7**	**8.1**	
E	71	267	160	498
% in E	14.3%	53.6%	32.1%	
Std. Res.	1.6	**6.7**	**–6.3**	
J	44	103	0	147
% in J	29.9%	70.1%	0%	
Std. Res.	**6.4**	**7.0**	**–8.8**	
Total	187	564	831	1582
%	11.8%	35.7%	52.5%	

Note. Bolded values represent standardized residuals greater than $|z| = 1.96$.

Table 3 – Response-level IRR based on FQ determination path

FQ determination path	N	Cohen's *k*	Classification
Tabled	937	.77	Excellent
Extrapolated	498	.48	Fair
Judged	147	.37	Poor
Total	1582	.68	Good

Note. Cohen's *k* classification based on Cicchetti (1994) and Shrout & Fleiss (1979).

Table 4 – Protocol-level IRR

FQ determination path	FQ	ICC	Classification
All responses			
	FQo%	.77	Excellent
	FQu%	.66	Good
	FQ–%	.68	Good
T & E only			
	FQo%	.75	Excellent
	FQu%	.65	Good
	FQ–%	.64	Good
T only			
	FQo%	.79	Excellent
	FQu%	.75	Excellent
	FQ–%	.77	Excellent

Note. ICC classification based on Cicchetti (1994) and Shrout & Fleiss (1979).

(ICCs were comprised between .66 and .77). However, in line with our expectations, ICCs was notably higher when T determined responses only were examined (ICCs were comprised between .75 and .79).

Additional analyses

Sub-optimal agreement for tabled responses. It is surprising that FQ ratings were inconsistent between raters when the path for determining FQ, as reported by Rater 2 at t_2, was T (k = .77, see Table 3). Obviously, raters were using different approaches to derive their FQ, but what is the nature of these differences? To answer this question, we inspected the 16 protocols with 434 responses for which both independent raters identified the FQ determination paths, in addition to the FQ codes themselves (see Procedure). Confirming this hypothesis, we found 20 out of the 266 responses that had been classified as T by Rater 2, had been classified as E or J by Rater 1 (see Table 5), and that these inconsistencies typically resulted in FQ coding inconsistencies too.

To our surprise, inconsistencies on FQ coding occurred also for 29 of 246 (11.8%) responses classified as T by both raters. That is, it did happen – albeit relatively infrequently – that both raters considered the FQ determination to be Tabled, yet disagreed on FQ. We thought that raters were likely using different tabled entries to derive their FQ, so we examined the verbatim responses and location documentation to better understand this puzzling outcome.

This review revealed a number of sources for these Tabled FQ coding inconsistencies. The first involved multi-object responses and whether or not a given, tabled, response object should be considered to be an "important object". For example, the response "a flower and a bush"

Table 5 – Contingency table for the IRR of the FQ determination path

		Rater 1 FQ determination path			
		T	*E*	*J*	*Total*
Rater 2 FQ determination path	*T*	246	19	1	266
	E	13	114	8	135
	J	2	5	26	33
Total		261	138	35	434

could have one or two important objects depending on how they are elaborated. If both are tabled and have different FQ, disagreement on which objects are important would lead to different FQ. A second source of disagreement is whether or not a multi-object response would qualify as an overarching table entry. For example, the response "these look like lungs [tabled], these like bones [tabled]" could have a different FQ, if one looks up lungs and bones in the FQ Table but a second rater found "anatomy" tabled for the entire response location. Additional sources of inconsistencies included raters' misunderstandings related to the location of the response objects, particularly in the case of quasi-W or quasi-D responses and linguistic ambiguities in the description of a response and/or FQ tables entry (e.g., can "a cockroach" be automatically coded based on an FQ tables' entry such as "bug" or does one only use the "bug" entry as Tabled FQ determination when that exact word used by the examinee?).

A tentative approach to reduce judgment in FQ determination. As noted above, at the response-level, the characterization of Cohen's *k* was excellent for Tabled (T) responses, but fair and poor for Extrapolated (E) and Judged (J) responses respectively (see Table 3). At the protocol-level, when only T responses were analyzed, the characterization of ICC was excellent for all three FQ codes, whereas it decreased to good, for FQu% and FQ–%, when considering also the E and J responses (Table 4). Overall, Non-T responses (i.e., E and J responses) thus appeared to be characterized by relatively poorer IRR.

Given that, we wondered whether one could possibly predict the FQ data obtained when scoring the entire protocol basing on the FQ codes assigned in the T responses only. Ideally, such procedure could notably simplify the coding procedures of FQ, while increasing IRR. Indeed, as noted above, teaching how to code FQ is particularly challenging (Viglione et al., 2017) and FQ coding difficulties potentially discourage practitioners from using the Rorschach in their practice as they require a lot of time and effort. Consequently, difficulties in learning and uncertainties about the accuracy of the coding are time-consuming, impacting the cost-benefit ratio associated with using the Rorschach. We thus ran three hierarchical regression models to predict the three key protocol-level scores of FQ, i.e., FQo%, FQu% and FQ–%.

Because when compared to T responses, Non-T responses associated with poorer FQ, in each model we considered two predictors. One predictor (step 1) was the percentage of the target FQ code found in the T responses only. For example, the predictor of FQo% at the protocol-level was represented by the proportion of the FQo responses given to T responses divided by the total number of T responses in that protocol. The second predictor, entered at step 2, consisted of the number of responses whose FQ determination was not T, divided by the total number of responses in the protocol (i.e., the % of Non-T responses in the protocol).

The results of these three models are reported in Table 6. Their adjusted R^2 values were comprised between .50 (for FQ–%) and .70 (for FQo%), thus indicating that at least half of the variance of the overall score of each FQ variable could theoretically be estimated basing on two predictors only. Besides, ΔR^2 decreased from FQo to FQu and FQ–, which suggests that adding the % of Non-T responses to the models impacted more notably the prediction of FQo% than that of FQu% or FQ–%.

DISCUSSION

This study aimed at shedding some light on why IRR of FQ is sometimes less optimal than that of other R-PAS variables, despite its well-established validity. To this aim, we coded the percentage of different FQ coding paths, namely Tabled (T), Extrapolated (E) and Judged (J), and tested some hypotheses concerning FQ and its IRR across judges. In line with our hypotheses, we found that E and J responses were characterized by increasingly poorer FQ and less optimal IRR compared to T responses. Noteworthy, using the % of E and J responses (i.e., Non-T) and the FQ assigned to T responses, we were able to predict 50% to 70% of the variance of the FQ values found when coding FQ for the entire protocol. Taken together, these results confirm that the FQ determination path used to code FQ may have a notable impact on IRR.

An interesting result is that, as shown in Table 1, in approximately 60% of the cases, the percepts to be considered to code FQ were found in the FQ tables, without the necessity to make any extrapolations or judgments. This may be the reason why, even though subject to a certain degree of variability, the IRR of FQ is usually satisfactory across studies, albeit at times lower than optimal (Kivisalu et al., 2016; Lewey et al., 2018; Pignolo et al., 2017; Viglione et al., 2012). Moreover, extrapolation and judgment were required in about 30% and 10% of the cases, respectively.

Table 6 – Hierarchical regression models

Criterion/predictors	β_1	β_2	R	R^2	Adj. R^2	ΔR^2
FQo%						
(step 1) FQo% (T only)	.69**	.72**	.69	.48	.47	–
(step 2) % of Non-T	–	−.48**	.84	.71	.70	.23**
FQu%						
(step 1) FQu% (T only)	.71**	.69**	.71	.50	.50	–
(step 2) % of Non-T	–	.37**	.80	.64	.62	.13**
FQ−%						
(step 1) FQ−% (T only)	.65**	.72**	.65	.42	.41	
(step 2) % of Non-T	–	.33**	.72	.52	.50	.10**

* $p<.05$, ** $p<.01$

Since this is the first study to document the use of T, E, and J coding paths, we have no reference parameters to evaluate these frequencies in the context of a non-clinical sample. Nonetheless, these percentages represent an evidence of the unique contribution that each person can bring to the Rorschach task. When inspecting the percentages of T, E, and J responses across cards, however, we found that Card IX produced a notably greater number of Non-T responses. As such, it might be useful, for future R-PAS developments, to try to extend the FQ tables' list of percepts especially for that specific inkblot. It should be noted that Card IX could be considered one of the most difficult ones in the test, as it is typically characterized by fewer responses, and its Popular response is not so common or obvious (Berry & Meyer, 2019; Pianowski, Meyer & de Villemor-Amaral, 2016).

A second interesting result is the strong association between J responses and FQ– and, more generally, the decline in FQ when moving from T to E to J responses. This result was somehow expected based on technical and theoretical grounds; nonetheless, this is the first study to provide evidence on this matter. On the technical side, the criteria to code FQo when judgement of fit is required are quite strict, since the only case when FQo can be assigned is when the FQ tables provide conflicting support for both FQu and FQo, without clear guidance to help the decision (Viglione et al., 2016). To code FQo rather than FQu, the answer to the question "Can I see that object in this location quickly and easily?" is closer to "Yes. I can see that. It matches the blot

pretty well", whilst to code FQu rather than FQ– or FQo the answer is closer to "A little. If I work at it, I can sort of see that". When FQ tables do not provide comparable responses for extrapolation, the two possible codes are FQu or FQ–. On the theoretical side, the negative association between E (and J) responses and FQo has a basis on the critical bits concept (Exner, 1996), as implemented in the extrapolation for FQo decisions (Viglione et al., 2016), that is: to extrapolate FQo (vs FQu) it is required that the response includes critical bits matching those included in the FQ tables. By definition, when the rater has to resort to judgement, there could be no match between critical bits of the response objects and the critical bits of the tabled ones.

A third interesting finding obtained from this investigation is that response-level IRR tended to decrease when moving from T to E to J responses. While this result was largely expected, to date no study had yet empirically documented the existence of this phenomenon. In this regard, two main considerations may be drawn. First, Rorschach trainers should try to make some extra efforts when teaching trainees how to code FQ if the relevant percepts are not in the FQ tables, and therefore the examiner has to rely on E or J determination paths. Second, if possible, it would be useful to try to further extend the list of percepts included in the FQ tables, so to minimize the need to use E or J to code FQ. It should be noted, however, that when inspected at the protocol-level, the IRR values of FQ codes were always highly satisfactory, even when including Non-T responses. As such, these recommendations for future improvements may be considered to be 'desirable' but certainly not 'mandatory.'

Indeed, a possible source of interrater disagreement could be the weight of local coding conventions (see Meyer, Shaffer, Erdberg P. & Horn, 2015). The Rorschach coders in this study strictly followed the coding guidelines provided by the R-PAS manual (Meyer et al., 2011), along with the guidance provided by Viglione and colleagues (2016), and this should have avoided the IRR being affected by local coding conventions. Nonetheless, our results help to pinpoint two important aspects about FQ coding. First, a disagreement on the FQ determination path could end up in a disagreement concerning the FQ coding. Secondly, the fact that two raters found a response in the FQ tables does not guarantee agreement on the resulting FQ. Indeed, the FQ coding procedure is complex and it is often much more difficult than just "Look it up in the FQ table".

When closely examining possible sources of FQ coding inconsistencies, we found that differences in the determination of which objects are important in a multi-object response often leads to inconsistent FQ scoring. In addition, examiners sometimes disagree on whether or whether not to use overarching category entries such as anatomy or landscape. For instance, in a response such as "these are lungs and these are bones", to what degree one can safely code the FQ of the response by relying on an overarching category such as "anatomy"? For some locations, the FQ tables clarify whether the potentially overarching category "anatomy" may or may not be used to code a specific anatomic part of the body such as the lungs. For instance, on Card VIII, W, the FQ tables present different entries for "anatomy (unspecified)" versus "anatomy (specific)." However, this distinction is not made explicit for other locations (e.g., on D2 in Card I, or W in Card III, the FQ tables only report "anatomy," with no distinction between unspecified vs specific), so that different examiners could treat the same anatomy-related response differently, for those areas.

Some other sources of FQ coding inconsistency identified in our Additional analyses section involved possible uncertainties or misunderstandings about the location of the important response objects and the use of potentially ambiguous categories and synonyms in the description of the response in the FQ tables itself. These were all cases where, despite the existence of seemingly clear rules, a minimum degree of judgment was still somehow required. In fact, Pignolo et al. (2021) stated that FQ judgments made by individual examiners are not always reliable. Therefore, when scoring FQ, one should carefully scrutinize the empirically supported FQ tables and base the FQ score on these rather than personal judgments (Pignolo et al., 2021). We believe that future developments of the R-PAS should therefore make an effort to address each and every one of those issues, so to further improve interrater reliability. Indeed similar issues led to the publication of more thorough coding procedural instructions for the CS in 2002 (Viglione, 2002), many of which were adopted into R-PAS (Meyer et al., 2011).

From a broader perspective, we believe that many of the FQ coding inconsistencies result from failures to search the FQ tables thoroughly, forgetfulness about complex coding guidelines, and the need for subjective examiner judgment. To reduce the resulting, observed inconsistencies, one could add details, distinctions, and clarifications to the FQ guidelines and tables. However, doing so would make it more and more

difficult for the Rorschach examiner to remember all specific FQ coding procedures at the right times. To avoid this from happening, it would be best if FQ coding were delegated to computers as much as possible. We argue that advances in computer technology should be applied to increase reliability and decrease FQ coding time and effort and thus increase utility in terms of the cost-benefit analyses given the unique contributions to assessment offered by the Rorschach in general and FQ in particular (Meyer et al., 2011; Mihura et al., 2013). To be clear, we are not stating that all Rorschach problems could (nor should) be solved by exclusively relying on computer algorithms. Yet, if the administration and coding processes were more automated, the examiner could dedicate more attentional resources to other interpretively meaningful, subtle behavioral manifestations put in place by the examinee while taking the Rorschach. In this direction, it is perhaps noteworthy that the R-PAS team is trying to develop a new feature that will allow an advanced, speech-to-text function, which will likely simplify the examiner's task during the administration phase.

Because IRR was lower for Non-T than for T responses and learning how to code FQ based on E or J determination paths is challenging and intricate (Viglione et al., 2017), we investigated if one could avoid coding the Non-T FQ, by estimating the FQ scores at protocol level on the basis of two predictors: T FQ%, and % of Non-T responses. Results showed that the information generated by using these data alone was sufficient to estimate, with relative accuracy, what FQ values one would obtain if FQ was coded across the entire protocol. Given that (1) Non-T responses represented almost 40% of the total number of responses, and (2) extrapolating FQ for non-tabled objects has been rated by R-PAS new learners as challenging or difficult and time-consuming (Viglione et al., 2017), this approach could potentially notably simplify the learning and practical usage of the test while increasing IRR. This notwithstanding, presently this approach is going to lose some important clinical information, mainly because the accuracy of the estimation is less satisfactory particularly for FQ– %, a key variable for reality testing interpretation. Thus, future studies should replicate our findings by including some validity criteria, so to test the extent to which the supposedly increased IRR would have any influence on FQ validity.

On the basis of the points discussed above, for the time being we recommend using the online R-PAS document authored by Viglione et al. (2016) to solve any extrapolation and judgment issues/doubts. We also suggest that it might be useful, in the future, to code the path used to determine FQ, i.e., T, E or J, as it might add context to the interpretation. Given their higher IRR, T FQ scores will be the ones on which to ground the interpretation. In turn, E and J responses will be treated more tentatively because of their lower IRR, while at the same time potentially providing a more nuanced interpretation. In fact, J responses appear to document a stronger deviation from what is commonly seen in the card (as documented in the FQ tables), since they are generally characterized by a higher percentage of FQ– compared to E responses (30% vs 14%, respectively).

A few limitations of this study should be kept in mind, while reading this article. Firstly, the study was conducted on a non-clinical sample, comprising undergraduate volunteers. As such, the generalizability of our findings may be questioned. Thus, future studies should inspect both clinical samples and controls composed of subjects pertaining to other professional areas and with different ages. This is important because the prevalence of T over Non-T responses, and of FQ– and FQu over FQo, may significantly change in clinical samples. In fact, one would expect clinical protocols to include a higher number of percepts that are not listed in the FQ tables, thus impacting IRR. Moreover, FQ– is interpreted as a perceptual lapse or distortion, and high FQ– % is strongly associated with reality testing problems and psychopathology. Therefore, the conclusions we drew from our results might be questionable in a clinical sample with a higher proportion of FQ–. Somewhat related to this point, one cannot rule out that possible examiners' disagreements on coding FQ– could in fact associate with (and thereby possibly even indicate) the presence of severe problems in the examinee's psychological functioning. To investigate this possibility, one should test the association between the presence of psychopathology and the amount and possibly type of J responses (e.g., using external criteria such as psychiatric diagnosis). Secondly, the study was focused on IRR, so validity was not evaluated. Thus, criterion measures to assess validity should be included in future research. Despite these limitations, this study is the first to analyze FQ scores with respect to the FQ determination paths, contributing to a deeper understanding of both the FQ variability and the issues regarding the IRR of FQ codes.

Conflicts of interest. Donald Viglione (fourth author) owns a share in the corporate (LLC) that possesses rights to Rorschach Performance Assessment System.

References

ACKLIN, M.W., McDOWELL, C.J., VERSCHELL, M.S. & CHAN, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment*, *74*, 15-47. doi.org/10.1207/S15327752JPA740103

BECK, S.J., BECK, A.G., LEVITT, E.E. & MOLISH, H.B. (1961). *Rorschach's test: I. Basic processes (3rd ed.)*. Grune & Stratton.

BERKOWITZ, M. & LEVINE, J. (1953). Rorschach scoring categories as diagnostic "signs". *Journal of Consulting Psychology, 17*, 110-112. doi.org/10.1037/h0062113

BERRY, B.A. & MEYER, G.J. (2019). Contemporary data on the location of response objects in Rorschach's inkblots. *Journal of Personality Assessment*, *101* (4), 402-413. doi.org/10.1080/00223891.2017.1408016

BURIN, D., PIGNOLO, C., ALES, F., GIROMINI, L., PYASIK, M., GHIRARDELLO, D., ZENNARO, A., ANGILETTA, M., CASTELLINO, L. & PIA, L. (2019). Relationships between personality features and rubber hand illusion: An explorative study. *Frontiers in Psychology*. doi.org/10.3389/fpsyg.2019.02762

CICCHETTI, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290. doi.org/10.1037/1040-3590.6.4.284

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, *XX* (1), 37-46. doi.org/10.1177%2F001316446002000104

DAO, T.K., PREVATT, F. & HOME, H.L. (2008). Differentiating psychotic patients from nonpsychotic patients with the MMPI-2 and Rorschach. *Journal of Personality Assessment*, *90*, 93-101. doi.org/10.1080/00223890701693819

EXNER, J.E. (1969). *The Rorschach Systems*. Grune & Stratton.

EXNER, J.E. (1974). *The Rorschach: A comprehensive system*. John Wiley & Sons.

EXNER, J.E. (1996). Critical bits and the Rorschach response process. *Journal of Personality Assessment*, *67* (3), 464-477. doi.org/10.1207/s15327752jpa6703_3

EXNER, J.E. (2003). The Rorschach: A comprehensive system, *Vol. 1: Basic foundations (4th ed.)*. Wiley.

GOLDFRIED, M.R. (1962). Rorschach developmental level and the MMPI as measures of severity of psychological disturbance. *Journal of Projective Techniques*, *26*, 187-192. doi.org/10.1080/08853126.1962.10381095

HARDER, D.W. & RITZLER, B.A. (1979). A comparison of Rorschach developmental level and 51 form-level systems as indicators of psychosis. *Journal of Projective Techniques*, *43*, 347-354. doi.org/10.1207/s15327752jpa4304_2

KIMHY, D., CORCORAN, C., HARKAVY-FRIEDMAN, J.M., RITZLER, B., JAVITT, D.C. & MALASPINA, D. (2007). Visual form perception: A comparison of individuals at high risk for psychosis, recent onset schizophrenia and chronic schizophrenia. *Schizophrenia Research*, *97*, 25-34. doi.org/10.1016/j.schres.2007.08.022

KIVISALU, T.M., LEWEY, J.H., SHAFFER, T.W. & CANFIELD, M.L. (2016). An investigation of interrater reliability for the Rorschach Performance Assessment System (R-PAS) in a nonpatient U.S. sample. *Journal of Personality Assessment*, *98* (4), 382-390. doi.org/10.1080/00223891.2015.1118380

KLOPFER, B., AINSWORTH, M.D., KLOPFER, W.G. & HOLT, R.R. (1954). *Developments in the Rorschach technique. Vol. 1. Technique and theory*. World Book Co.

LEWEY, J.H., KIVISALU, T.M. & GIROMINI, L. (2018). Coding with R-PAS: Does prior training with the exner comprehensive system impact interrater reliability compared to those examiners with only R-PAS-Based training? *Journal of Personality Assessment*, *101* (4), 393-401. doi.org/10.1080/00223891.2018.1476361

MEYER, G.J., HILSENROTH, M.J., BAXTER, D., EXNER JR, J.E., FOWLER, J.C., PIERS, C.C. & RESNICK, J. (2002). An examination of interrater reliability for scoring the Rorschach comprehensive system in eight data sets. *Journal of Personality Assessment*, *78* (2), 219-274. doi.org/10.1207/S15327752JPA7802_03

MEYER, G.J., SHAFFER, T.W., ERDBERG P. & HORN S.L. (2015). Addressing issues in the development and use of the composite international reference values as Rorschach norms for adults. *Journal of Personality Assessment*, *97* (4), 330-347. doi.org/10.1080/00223891.2014.961603

MEYER, G.J., VIGLIONE, D.J., MIHURA, J.L., ERARD, R.E. & ERDBERG, P. (2011). *Rorschach Performance Assessment System: Administration, coding, interpretation and technical manual*. Rorschach Performance Assessment System.

MIHURA, J.L. & MEYER, G.J. (Eds.). (2018). *Using the Rorschach Performance Assessment System (R-PAS)*. The Guilford Press.

MIHURA, J.L., MEYER, G.J., DUMITRASCU, N. & BOMBEL, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the comprehensive system. *Psychological Bulletin, 139* (3), 548-605. doi.org/10.1037/a0029406

PIANOWSKI, G., MEYER, G.J. & DE VILLEMOR-AMARAL, A.E. (2016). The impact of R-Optimized administration modeling

procedures on Brazilian normative reference values for Rorschach scores. *Journal of Personality Assessment*, *98* (4), 408-418. doi.org/10.1080/00223891.2016.1148701

PIGNOLO, C., GIROMINI, L., ANDO', A., GHIRARDELLO, D., DI GIROLAMO, M., ALES, F. & ZENNARO, A. (2017). An interrater reliability study of Rorschach Performance Assessment System (R-PAS) raw and complexity-adjusted scores. *Journal of Personality Assessment*, *99* (6), 619-625. doi.org/10.1080/00223891.2017.1296844

PIGNOLO, C., VIGLIONE, D.J. & GIROMINI, L. (2021). How reliably can examiners make Form Quality (FQ) judgments in the absence of the Form Quality (FQ) tables? *Rorschachiana*, *42*, 21-34. doi.org/10.1027/1192-5604/a000135

RORSCHACH, H. (1921). *Psychodiagnostik*. Hans Huber.

SHROUT, P.E. & FLEISS, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86* (2), 420-428. doi.org/10.1037//0033-2909.86.2.420

SU, W.-S., VIGLIONE, D.J., GREEN, E.E., TAM, W.-C. C., SU, J.-A. & CHANG, Y.-T. (2015). Cultural and linguistic adaptability of the Rorschach Performance Assessment System as a measure of psychotic characteristics and severity of mental disturbance in Taiwan. *Psychological Assessment, 27* (4), 1273-1285. doi.org/10.1037/pas0000144

VIGLIONE, D.J. (2002). *Rorschach coding solutions: A reference guide for the comprehensive system*. Donald J. Viglione.

VIGLIONE, D.J. (2010). *Rorschach coding solutions: A reference guide for the comprehensive system (2nd ed.)*. www.rorschachcodingsolutions.com

VIGLIONE, D.J., BLUME-MARCOVICI, A.C., MILLER, H.L., GIROMINI, L. & MEYER, G. (2012). An interrater reliability study for the Rorschach Performance Assessment System. *Journal of Personality Assessment*, *94* (6), 607-612. doi.org/10.1080/00223891.2012.684118

VIGLIONE, D.J. & MEYER G.J. (2008). An overview of Rorschach psychometrics for forensic practice. In C.B. Gacono & F.B. Evans with N. Kaser-Boyd (Eds.), *Handbook of forensic Rorschach psychology*. Lawrence Erlbaum Associates.

VIGLIONE, D., MEYER, G., MIHURA, J., ERARD, B., ERDBERG, P. & GIROMINI, L. (2016). *Guidance for coding form quality requiring "Judgment of fit" and an important supplement on FQ coding for former CS user*s. www.r-pas.org.

VIGLIONE, D.J., MEYER, G.J., RESENDE A.C. & PIGNOLO, C. (2017). A survey of challenges experienced by new learners coding the Rorschach. *Journal of Personality Assessment*, *99* (3), 315-323. doi.org/10.1080/00223891.2016.1233559