Psychological Injury and Law (2023) 16:158–176
https://doi.org/10.1007/s12207-023-09472-6

Clarifying Sound and Suspect Use of the Rorschach in Forensic Mental Health Evaluations: A Response to Areh et al. (2022)

Corine de Ruiter[1] · Luciano Giromini[2] · Gregory J. Meyer[3] · Christopher M. King[4] · Benjamin A. Rubin[5]

Received: 13 May 2023 / Accepted: 24 May 2023 / Published online: 9 June 2023

Abstract
Areh et al. (*Psychiatry, Psychology and Law* 29:183–205, 2022) recently commented on what standards should be applied to determine whether a test is appropriate for psycholegal use and concluded that the Rorschach does not meet their proposed standards. Accordingly, they concluded that psychologists should not use it in legal contexts. However, Areh et al.'s (2022) claims are based on a significant misunderstanding of how the Rorschach task works, relative neglect of the last 20 years of Rorschach research, unrealistic psychometric standards for assessing the reliability and validity of a psychological assessment measure, and a single European legal case in which a forensic expert used the Rorschach inappropriately. Our article seeks to clarify and correct some of their errors and misleading assertions. First, we clarify how the Rorschach task works according to more recent and widely accepted conceptualizations. Then, we show that Areh et al.'s (2022) position that Rorschach task data do not meet acceptable validity standards, especially when compared to medical tests, is empirically untenable. Next, we provide a detailed and nuanced account of what the Rorschach has to offer as a performance-based assessment method for forensic evaluators and the legal system, with attention paid to the anecdotal legal case Areh et al. (2022) highlighted. Finally, we provide four reasons why the Rorschach can be a useful tool for forensic mental health assessments when using the Rorschach Performance Assessment System (R-PAS).

Keywords Rorschach · Performance-based assessment · Assessment · R-PAS · Forensic · Legal · Admissibility

The Rorschach Inkblot Task has been the subject of controversy throughout its history (Searls, 2017). The most recent, and probably most vehement, criticisms were voiced in the years between 1995 and 2001 (Garb, 1999; Garb et al., 2001; Wood et al., 2001). The main focus of those critiques was Exner's Comprehensive System (CS; Exner, 2003), which was the coding and interpretive system commonly in use at the time in the United States (US), most European countries, and other parts of the world. It was also the system most focused on gathering psychometric evidence for its normative data, reliability, and validity. These critiques led to debates about the adequacy of CS normative data, reliability, validity, practical utility, and incremental validity (Meyer, 2001; Meyer & Archer, 2001). Partly as a result of these debates, there has been an upsurge in Rorschach research that focused on gaining insight into its underlying response process, improving its psychometric properties, and studying its utility in forensic mental health assessments (FMHAs; e.g., Viglione et al., 2022).

This article draws on this recent research to respond to a critique by Areh et al. (2022) published in *Psychiatry, Psychology and Law*. These authors make several strong claims suggesting that the Rorschach is likely to be or ought to be found inadmissible in legal contexts. Although we agree with some of the points raised by Areh et al. (2022), we argue that many of their claims are based on misconceptions concerning the Rorschach task, outdated science, unrealistic psychometric standards, and a single European legal case (*F v Bevándorlási és Állampolgársági Hivatal,*

✉ Luciano Giromini
luciano.giromini@unito.it

[1] Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands

[2] Department of Psychology, University of Turin, Via Verdi 10, 10123 Turin, TO, Italy

[3] Department of Psychology, University of Toledo, Toledo, USA

[4] Department of Psychology, Montclair State University, Montclair, USA

[5] New York State Office of Mental Health - Rockland Psychiatric Center, Orangeburg, USA

2018) in which a forensic evaluator, in our and Areh et al.'s opinion, used the Rorschach in an inappropriate manner. Because we believe the Areh et al. (2022) article may reflect a view of the Rorschach shared by some lay people, psychologists, forensic practitioners, attorneys, and judges, we offer a detailed and differentiated perspective on what the Rorschach, as a performance-based assessment method, can offer forensic evaluators and the legal system. We also describe its shortcomings and what is still unknown about the test for FMHAs. We start by briefly describing how some of us came to collaborate as authors on this article before listing our points of agreement with Areh et al. (2022), followed by a series of arguments as to why we disagree with a number of their statements.

Most authors of this article did not know each other or publish together until they were asked by the editors of a special issue of the *Journal of Personality Assessment* to join an "adversarial collaborative" team tasked with providing a trustworthy review of the evidence concerning the admissibility of the Rorschach in court (Viglione et al., 2022). We were brought together for that purpose because we encompassed diverse attitudes toward the Rorschach and differing levels of expertise with law and FMHA. The conclusions and views we express in this article draw heavily on the evidence and issues we reviewed and organized for Viglione et al. (2022). However, to avoid redundancy, we do not provide the same evidentiary detail. Rather, we tailor our points to address the issues that more uniquely emerged in Areh et al.'s (2022) critique of the Rorschach.

Briefly, Viglione et al. (2022) reviewed the current psychometric status of the Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011), focusing on normative data, reliability (interrater, interpretive, and retest), and validity (convergent and incremental). They concluded that relative to the CS, R-PAS norms were more accurate and less pathologizing, successfully curtailed variability in the number of responses evaluees give, had *M*s and *SD*s supporting CS research literature generalizing to R-PAS, and used easier to interpret standard scores rather than raw scores. However, additional norms for adults and forensic populations are needed. Interrater reliability among trained evaluators is good overall, though there are a few exceptions; but overall, experienced evaluators tend to draw similar interpretive conclusions about evaluees when reviewing the same Rorschach protocol. The retest literature, in turn, is mostly older and variable, but stability for most scores is likely to be more akin to memory or job performance scores (1- to 2-month *r*s of 0.50 to 0.70) than to intelligence quotient (IQ) subtest scores (mean *r* around 0.80).

Across scales and variables, Viglione et al. (2022) note meta-analyses show good convergent validity, on a par with MMPI(-2) scales. However, Mihura et al.'s (2013, 2015) meta-analyses of 65 specific CS scores found notable variability, with some having support that was good to excellent (46%), modest (15.4%), little to none (20%), or absent because there were no studies targeting their core construct (18.5%). Note that even the most vocal critics of the Rorschach method said Mihura et al.'s meta-analyses "provided an unbiased and trustworthy summary of the literature" (p. 243) and it prompted them to lift their call for a global moratorium on using the Rorschach in clinical and forensic practice (Wood et al., 2015). The Mihura et al. meta-analyses also strongly determined what variables were included in R-PAS, the latest iteration of a research-based coding and interpretive Rorschach system. Individual studies have documented incremental validity for variables in R-PAS over various self-report scores, consistent with the near-independence of these two test methods (i.e., $r = .08$, Mihura et al., 2013) across hundreds of findings.

Against this empirical foundation, Viglione et al. (2022) considered issues related to the use of R-PAS in legal contexts: general acceptance, use by forensic evaluators, and U.S. and select European case law addressing challenges to use of the Rorschach in mental examinations (a term of art for certain US civil litigation) or with respect to admissibility and weight as trial evidence. They concluded that courts will likely rule R-PAS evidence as admissible, assuming the evaluator competently interpreted appropriate variables to evaluate psychological processes that would inform relevant psycholegal questions. They closed their review by identifying effective and ethical uses of R-PAS, as well as inappropriate uses in criminal, civil, juvenile, and family court.

Extent of Our Agreements with Areh et al. (2022)

Areh et al. (2022) rightfully acknowledge the importance of adherence to professional and ethical standards when psychological tests are used in legal contexts. They refer to the *Meta-Code of Ethics* of the European Federation of Psychologists' Associations (EFPA, 2005). They specifically mention Paragraph 3.2.3, which requires psychologists "to practise within, and to be aware of the psychological community's critical development of theories and methods" (p. 186). The EFPA has not promulgated special ethical guidelines for forensic psychologists, similar to the American Psychological Association (APA, 2013). European courts also do not have uniform standards for the admissibility of expert evidence, such as the *Daubert* and related standards used by many jurisdictions in the US (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993; see also Viglione et al., 2022). In their analysis of the *F v Bevándorlási és Állampolgársági Hivatal* case, Areh et al. (2022) consider the use of the Rorschach in light of the *Daubert* standard, which

we agree is a reasonable framework by which to evaluate the appropriateness of the Rorschach method in FMHAs (further down in this document we provide a more detailed discussion of the case and arguments why we think this case does not represent current satisfactory practice for use of the Rorschach in FMHA). For instance, we agree that forensic evaluators should use reliable and valid psychological tests that preferably have also been tested for use in the population to which a forensic evaluee belongs (e.g., Neal et al., 2022).

However, in articulating our support for these arguments by Areh et al. (2022), we place ourselves firmly on the side of forensic evaluators who advocate for empirically defensible use of scores derived from the Rorschach task. In particular, our position is that the most forensically defensible Rorschach system is the Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011), which was developed as a replacement for and correction to Exner's CS. We note, however, that around the world there is wide variability in the extent to which the interpretive systems for the Rorschach are focused on empirical nomothetic evidence to support applied practice versus intuitive and idiographically generated clinical inferences. In this article, unless otherwise indicated, when we refer to "the Rorschach," we are referring to its use in research-based systems founded on nomothetic research (i.e., the CS and R-PAS), and not to intuitively oriented, idiographically focused interpretive systems.

In addition to different formal systems for using the Rorschach, there can be wide variability in the ways that an individual evaluator might use the Rorschach within one of these systems. This includes uses that do not reflect adequate professional or forensic practice standards. Like Areh et al. (2022), we believe that these practices should not be used in psychological evaluations for the legal system, nor should they be considered admissible by legal decision-makers when they occur.

What is "the Rorschach" and How Does it Work?

Areh et al. (2022) consistently refer to the Rorschach as a *projective* test. This outdated terminology is typically used to contrast the task to so-called *objective* tests, though both terms provide seriously misleading inferences about the measures they supposedly represent (for a discussion, see Meyer & Kurtz, 2006). Forensic evaluators considering using the Rorschach as part of an assessment battery need to know how it works and what type of information it adds relative to other psychological tests. Consequently, we outline its key features and its response process to highlight its implications for FMHA after situating the task in the broader context of psychological assessment methods.

Test Methods

Psychological tests use two primary ways to gather information (Bornstein, 2022; Meyer, 2023; Meyer et al., 2018). One is by asking questions; the other is by observing behavior in a particular context. Psychological assessment measures implement these investigative approaches by providing standardized stimuli and guidelines for uniform administration, scoring, norming, and interpretation. This standardization allows researchers to gather evidence of reliability and validity for assessment measure scores, providing a scientific foundation for their use in FMHA (Slobogin, 2007) and distinguishing them from informal questioning or observation, such as from unstructured clinical interviews and observations, and collateral reports and record review.

In standardized question-based information gathering, evaluators either can ask people directly about their experiences, which provides introspective *self-report* information, or they can ask someone who knows the target person to provide their impressions, which provides context- or relationship-dependent *informant-report* information (Bornstein, 2017; Mihura & Meyer, 2018). These methods allow questions across an almost limitless array of concepts or experiences—addressing the past, present, or expected future. They are ideal for assessing the evaluee's beliefs and experiences, but limited by the quality of the reporter's perception and insight, and their potential motives to report information in a biased manner, which is highly salient in FMHA (Heilbrun, 1992; Heilbrun et al., 2009; Rogers & Bender, 2018; Sweet et al., 2021).

Standardized ways of gathering information based on observed behavior can either assess maximum performance or typical performance (Cronbach, 1990; Mihura & Meyer, 2018). *Maximum performance* tasks prompt respondents to express the full extent of their abilities and encompass intelligence and neuropsychological tests. These tasks provide clear criteria about what constitutes success and how to achieve it, a limited number of response options, and testing conditions that foster focused, motivated performance. An example of this category of tasks is the subtests of the Wechsler Adult Intelligence Scale (WAIS–IV; Wechsler, 2008). In contrast, *typical performance* tasks provide minimal guidance for completion, lack clear criteria for successful or desirable performance, provide wide latitude for responding, and offer a context that fosters individualized solutions as opposed to a predetermined goal. For example, in the Strange Situation task (Ainsworth et al., 1978), the focus of observation is on the child's behavior, which is analyzed and coded in a standardized manner. However, children are not given specific instructions on how to behave or respond when the stranger enters the room, when the mother returns, etc. Because of these differences, maximum performance tasks show what a person *can do* when motivated

to perform optimally, whereas typical performance tasks show what a person *chooses to do* when relying on their own resources and preferences (Cronbach, 1990; see also Sackett, 2007).

A strength of both types of performance tasks is that they provide observable, discrete samples of behavior as opposed to the linguistic descriptions provided by question-based measures. As such, they rely on the assessor—not the evaluee—to identify, classify, and interpret relevant behaviors, and they can be ideal for assessing characteristics of which the evaluee is unaware or imperfectly aware. However, the task requirements shape the constructs one can assess, and results depend on the evaluee's authentic engagement with the task during testing. In general, these methods are not helpful for assessing historical events or consciously held attitudes, beliefs, symptoms, or experiences.

Rorschach Basics

The Rorschach is a *typical performance task* that relies on visual-spatial and lexical-conceptual problem solving. It uses a standard set of 10 vertically symmetrical inkblot designs set on white cardstock. Five inkblots are shades of gray, two are shades of gray with prominent bold red areas, and three are fully chromatic with elements ranging from pastels to brightly saturated colors. For administration, an R-PAS assessor asks the evaluee to try to see two or three different things on each card, then hands them each card in a fixed order while asking the question "What might this be?" The evaluee's replies represent their solutions to the task at hand. The assessor records verbatim responses for all 10 cards, generally resulting in 23 or 24 responses in total (Meyer et al., 2011). Next, for each response, the assessor clarifies where objects reside on the card and the inkblot features that contributed to the evaluee's perception. Assessors then code (i.e., classify) each response across multiple dimensions (e.g., use of color, envisioning human activity, coherence of thought processes), and aggregate the codes across all responses to form normed summary scores that contrast the evaluee's scores to the scores indicating what most people see, say, and do when completing the task.

The Nature of the Stimuli

The inkblots are not random designs, as many people assume. Rorschach was an artist who carefully created, pilot tested, and artistically refined each card over time to ensure they would not simply look like inkblots (Searls, 2017). He appeared to have had two intertwined aims when developing them, both based on the suggestive "critical bits" (Exner, 1996) that encompass the prominent inkblot areas and their shapes, color, shading, irregular interior and exterior contours, and symmetrical features. First, within the designed composition of each card, Hermann Rorschach embedded at least one reasonably recognizable object or part of an object, which result in the commonly reported conventional response objects (so-called *Popular* responses). Second, he simultaneously embedded an array of other features that draw one's attention to trigger an alternative perception that contradicts or complicates the more recognizable elements (e.g., a part looks like a person's head, but what would normally be its torso looks more like the head of an animal). Together, these opposing qualities produce evocative but incomplete or imperfect perceptual likenesses that stimulate competition among visual impressions that may form potential responses.

This embedded structure and the competing alternatives combine to create a Zipf or power-law distribution[1] of objects perceived (Meyer et al., 2011). Such a distribution is roughly L-shaped and distinctly non-normal. If one plots frequency on the vertical axis and the rank order of those frequencies on the horizontal axis, the result is a near vertical "arm" on the left for the 250 or so objects that many people regularly see, and a near horizontal "tail" stretching to the right for the thousands of uncommonly reported objects. Indeed, even in very large samples, unique objects seen by a single person account for about 70% of all the distinct objects identified on the Rorschach task (Meyer et al., 2011). Thus, the nature of the task has both clearly embedded structure and remarkably wide latitude for idiographically unique perceptions. This structure provides the mechanism for assessing conventionality in the locations selected for percepts (i.e., the focus of one's attention; Berry & Meyer, 2019) and the quality of the fit of objects to those locations (i.e., perceptual accuracy as coded by *Form Quality*). The idiographic diversity provides personally salient, experience-near imagery that can richly illustrate an evaluee's psychological processing.

The Psychological Operations Engaged

The task of dealing with imprecision in the deliberately contradictory stimuli, as well as uncertainty regarding the adequacy of one's responses, occurs while the evaluee interacts with the inkblots and the evaluator, a relative stranger sitting adjacent who is observing and transcribing the exchange. These features make the task moderately stressful, and more stressful than assessment by self-report methods (Momenian-Schneider et al., 2009; Newmark et al., 1974, 1975).

For the evaluee, solving the problem of what the inkblot might be invokes a series of perceptual and problem-solving

[1] These characterize many natural phenomena, from earthquake magnitudes to city size and website traffic.

operations, which are iterative cycles of prediction and error correction (Meyer & Friston, 2022). These include scanning the stimuli, selecting locations for potential response objects, comparing objects in memory to potential inkblot images, evaluating their inconsistencies or contradictions, formulating and reformulating response options, filtering out options judged less optimal, and articulating a final solution to the assessor (Exner, 2003). The respondent's visual-mnemonic matching of objects in the card to recalled images, conceptual processing of the stimuli, and verbal and nonverbal communication engage all brain regions, encompassing bilateral activity in the frontal, temporal, parietal, occipital, and limbic lobes (Asari et al., 2008, 2010a, b; Giromini et al., 2017).

The available neurophysiological data indicate that completing the task engages both the dorsal and ventral attentional systems (Giromini et al., 2017). The dorsal system is important for directing conceptually guided top-down attentional search processes (i.e., predictions of what it might be) and the ventral system is important for recognizing and reorienting to surprising or unexpected bottom-up input (e.g., misfits with prediction, alternative possibilities). These two attentional systems are negatively correlated with the default mode network (e.g., Zhou et al., 2018), which in humans is implicated in self-referential processing, including the introspective attribution of self-reported characteristics (e.g., Davey et al., 2016). Thus, the cognitive functions engaged while completing the Rorschach task are distinct from those engaged while completing a self-report inventory, which likely contributes to the low correspondence of these methods when assessing conceptually aligned psychological constructs (Mihura et al., 2013).

One can profitably view responding to the Rorschach as a predictive process (grounded in Bayesian probabilities) where the respondent is iteratively refining the fit of conceptual priors (beliefs about what it might be) to noisy, uncertain, and imprecise environmental stimuli (Friston et al., 2012; Meyer & Friston, 2022; Parr & Friston, 2017). Predictive processing accounts of perception posit that brains evolved to predict the hidden features of the environment that generate sensations at the boundary of the organism and its environment. They do so to minimize prediction error and thus survive over time (Clark, 2013, 2016; Friston, 2009; Hohwy, 2016). One can reduce prediction error either by using sensory information to modify the initial predictions about what is in the environment (i.e., by changing one's perceptions) or by taking actions to sample the environment (e.g., shift gaze) and more precisely affirm the prediction (i.e., by gathering better evidence). Model expectations (predictions) propagate down the neural hierarchy to inform sensory receptors of what they should experience, whereas prediction errors (sensory incongruity) propagate from the receptors up the neural hierarchy to register inconsistencies that need resolution, with messages in each direction weighted by their reliability. Importantly, the forward sensory signals register the discrepancy between the predicted sensory expectations and the sensory input encountered, rather than directly registering the external stimuli. The outcome of this iterative calibration process is stabilized perceptions (i.e., beliefs) about the causes of one's sensory stimulation.

The Rorschach task engages this form of active inference. It allows an assessor to see the results of the meaning-making process in action. With each response delivered, the evaluee has iteratively searched (i.e., predicted) the stimuli, encountered ill-fitting incongruities (prediction errors), and reconsidered response possibilities (iteratively refined predictions) to reach their error-corrected, personalized perceptual equilibrium. The evaluee's responses represent their prediction-optimized solutions that address both the context of the assessment occasion and the perceptual ambiguity of what the inkblots might be, with the inkblots serving as analogs to the constant perceptual uncertainties encountered in daily life (Clark, 2016).

In line with this conceptualization of the Rorschach response process, recent eye tracking (Ales et al., 2020) and fMRI (Vitolo et al., 2020) findings indicate that the more complex a response is (in terms of number of response objects, variety of reported contents, etc.), the higher the recorded level of cognitive engagement while the respondent visually scans the inkblot designs. Along similar lines, converging fMRI (Giromini et al., 2019), EEG (Giromini et al., 2010; Pineda et al., 2011), and rTMS (Andò et al., 2015, 2018) data indicate that the more a response involves a "feeling of human movement" (people dancing, waiters setting a table, etc.), the greater the activity in the evaluee's mirror neuron-related areas of the brain. That is, the complex psychological phenomena that occur in the mind of the evaluee while delivering a Rorschach response closely resembles that of a predictive process in which the individual iteratively refines the fit of initial perceptions and expectations (e.g., the evaluee's personal and body-related experiential predictions about what the inkblot might be, in the case of a response involving a human movement; see Kilner et al., 2007) to uncertain and imprecise, visually provocative stimuli.

Coding and Interpretation

A key assumption for coding and interpreting Rorschach responses is that evaluees demonstrate their personal propensities in the aforementioned context of problem solving. Through their responses, they show how they filter and organize information, what they attend to or fail to

recognize, how they make sense of and apply meaning to stimuli and situations, and how they represent people and interactions. They also show how conventionally or idiosyncratically they perceive and how logically or effectively they think and communicate with another person. Furthermore, their behaviors provide examples of how they handle perceptual and conceptual inconsistencies, contradictions, and ambiguity.

R-PAS classifies these manifestations with particular coded features (e.g., instances of misperception; conceptual synthesis; or seeing an object as damaged, harmed, or dysphoric) to identify the psychological operations believed to be active when an individual is generating a response. Normed summary scores thus quantify what occurred in the process of repeatedly attributing meaning to the perceptual stimuli while explaining to another person how one looks at things in the context of multiple competing possibilities.

In turn, the response process foundation for R-PAS interpretation is to infer that the scores summarizing behaviors observed in the microcosm of the task environment will generalize to parallel mental, verbal, perceptual, and interactive behaviors in everyday life (Viglione & Rivera, 2003). Ultimately, the forensic evaluator can explain to an evaluee, referral source, or judge or jury, the reasons for each of their inferences in the form of, "Because they did X this number of times under these task conditions, I think it is likely they will do Y [the everyday analog to X] in their daily life under similar conditions." The everyday life conditions that parallel the task involve being reliant on one's own resources, dealing with ambiguity, and being under at least modest stress, potentially of an interpersonal or evaluative nature. These behaviorally grounded features of performance during the task provide the rationale for distinguishing what the Rorschach provides to a multimethod assessment relative to self-report. Self-report measures reflect how people characteristically describe themselves or wish to be seen by others, which often is neither how they actually behave nor how others see them (e.g., Mihura, 2012). In contrast, based on responses to the Rorschach, R-PAS scores provide inferences regarding likely behaviors, perceptions, and reactions that may emerge in a range of life situations. Thus, the performance-based nature of the Rorschach task can make unique contributions to FMHA (De Ruiter & Kaser-Boyd, 2015).

R-PAS (Meyer et al., 2011) represents a research-based system of Rorschach coding and interpretation that resulted from thorough reviews of prior empirical literature, meta-analyses, and new research. Notably, Areh et al. (2022) only mention R-PAS in passing and largely ignore its research base or the ways in which it improves the psychometric foundation for using the Rorschach in FMHA (see Viglione et al., 2022).

What are Relevant Standards for Reliability (and Validity) of Psychological Tests for FMHA?

Areh et al. (2022) state that they, "propose a set of contemporary standards that psychologists and lawyers can use to determine whether or not a test is sufficiently reliable for use in psycholegal work" and "then use these standards to critically consider the use of the Rorschach in court" (p. 184). The standards they propose are highly desirable for any test user or test developer, as they recommend remarkably strong interrater reliability coefficients that notably surpass commonly cited standards (e.g., Cicchetti, 1994), as well as validity coefficients that are notably larger than the "large" effect size standard in Cohen's (1988, 1992) widely cited guidelines. However, we argue that Areh et al.'s (2022) standards are unrealistic.

Reliability

Areh et al. (2022) recommend interrater reliability coefficients that are no lower than 0.80 and ideally above 0.90. The authors do not state what type of reliability coefficient they have in mind (e.g., percentage agreement, kappa, intraclass correlation), although this makes a huge difference because the possible options can produce very different outcomes. Areh et al. also do not indicate if their guidelines refer to coefficients calculated for individual items or responses, or to scale scores aggregated across items or responses, with the latter producing higher coefficients than the former (see Meyer et al., 2002). Nonetheless, for Rorschach research, the recommended standard is to compute the exact agreement intraclass correlation (*ICC*) on protocol level scores (see Viglione et al., 2022). Given the sources cited by Areh et al. (2022), it is likely that they also intended their guidelines to apply to intraclass correlations.

Meyer (2004) and Meyer et al. (2005) systematically sampled the meta-analytic literature in psychology, psychiatry, and medicine examining interrater reliability. Their review, which encompassed data from meta-analyses of 60 types of measurements at either the item level or scale level, indicated that only six measures out of 72 (8.3%) had average reliability of 0.90 or above, using r or a combination of r and κ or an exact agreement *ICC*. Only one of those was computed at the item level (i.e., measured bladder volume by real-time ultrasound); the other five encompassed summary scores at the scale level. For the five summary scales, three were associated with medical measures (i.e., size of spinal canal and spinal cord on MRI, CT, or X-ray; count of decayed, filled, or missing teeth in

young children; ratings of functional independence using data from joint and separate interviews). The other two were for Rorschach scores (one for a scale of dependency and the other an average across multiple CS scale scores).

Using all available coefficients (i.e., *r,* κ, or *ICC*) and disregarding item and scale considerations, Meyer et al. (2005) summarized reliability by type of measure. On average, they found the following: typical performance personality measure (Rorschach and TAT, including reliability of Rorschach interpretation) = 0.84 from 11 findings; other psychological and psychiatric measure (e.g., semi-structured interviews for levels of depression or diagnoses, job performance ratings, child behavior ratings) = 0.63 from 36 findings; medical measure (e.g., level of drug sedation, stroke classification) = 0.66 from 15 findings; and non-applied judgments (e.g., peer review, grant funding reviews, number of factors to extract based on scree plots) = 0.42 from 8 findings. More recently, Schneider et al. (2020) conducted a meta-analysis of four interrater reliability studies examining all 60 of the primary variables interpreted in R-PAS, finding that the average *ICC* was 0.89 (*SD* = 0.09). These data make two compelling points. First, they suggest that Rorschach coding using the CS or R-PAS is among the more reliable types of applied judgment in psychology, psychiatry, and medicine. Second, collapsing across *r*, κ, and *ICC* and across items and scales, they show that almost all meta-analytic coefficients in the Meyer et al. (2005) literature review (91.7%) are lower than 0.90 and almost 70% are less than 0.80 (49 / 72 * 100 = 68.1%).

Thus, while we agree with Areh et al. (2022) that having *r,* κ, or *ICC* interrater reliability coefficients above 0.80 is desirable, most applied judgments concern complex phenomena that do not lend themselves to that level of interjudge agreement. However, the empirical evidence shows that most Rorschach scoring using the CS or R-PAS typically does meet these standards (see Viglione et al., 2022), contrary to Areh et al.'s (2022) claims.

Validity

Areh et al. (2022) state that the typical validity coefficient for psychological tests is 0.30 to 0.40, and they consider this "poor" validity by asserting that researchers in the field of medicine consider validity coefficients less than 0.70 to indicate "validity problems" (p. 187), citing one individual's personal opinion (Post, 2016). They thus recommend that "psychologists working in the psycholegal context should... ideally use tests that have validity coefficients well above 0.60" (p. 187). Similar to the reliability indices mentioned earlier, Areh et al. (2022) do not indicate to what type of effect size parameter they are referencing for this benchmark (e.g., mean difference, correlation, odds ratio), even though the benchmark would mean very different things depending on the statistic. However, the sources they cited were clearly describing correlations, and an *r* of 0.60 would translate to a standardized mean difference effect size (e.g., Cohen's *d*) value of 1.50.[2]

For perspective, Cohen's (1988, 1992) benchmarks for small, medium, and large effects are quite different than Areh et al.'s (2022). For correlations derived from fully dimensional variables, they are 0.10, 0.30, and 0.50, respectively, for small, medium, and large. For correlations derived from a dimensional variable and a dichotomous variable they are 0.10, 0.24, and 0.37, respectively. For differences between two means in *SD* units (i.e., *d* values), they are 0.20, 0.50, and 0.80, respectively. Cohen considered large effects as generally non-existent in psychological research with two independently measured variables, though they do appear when the same characteristic is measured by two scales that use the same method of assessment (e.g., two scales of depression assessed by self-report, or two IQ scores assessed by maximum performance tasks; see Meyer et al., 2011).

To exemplify the atypical nature of Areh et al.'s (2022) standard for minimal validity of at least $r = .60$ or $d = 1.50$, consider IQ, which has a standard deviation of 15 points on most commercially available measures. Thus, a standardized mean difference between two groups of $d = 1.50$ leads to a raw score mean difference of at least 22.5 IQ points (1.50 * 15 = 22.5).[3] Areh et al.'s (2022) position is that forensic evaluators should not use an IQ test to establish compromised cognitive functioning unless they are using it in a context where research has shown one group or type of person (e.g., with traumatic brain injury [TBI]) has an average IQ score that is at least 22.5 points lower (or higher; i.e., $d \geq |1.50|$) than another group, type of person, or reference standard, such as the average person, with *M* IQ = 100 points. That type of research would produce a validity coefficient of $d = 1.50$ or $r = .60$.

According to research presented in the manual for the WAIS-IV (Wechsler, 2008), this would preclude assessors from using that IQ test to evaluate people with reading disorders (IQ *M* = 89) because their *M* is only 11 raw score points below the mean for people on average, leading to a validity effect size of $d = 11 / 15 = 0.73$ (or $r = .34$). That is less than half the magnitude that Areh et al. recommend on the *d* metric. Using additional data from the test manual, Areh et al. would similarly recommend against forensic evaluators

[2] Consistent with meta-analytic practices, effect sizes can be converted from one metric to the other, such that the relationship (*r*) of a two-group independent variable with a dimensional dependent variable can also be expressed as the difference (*d*) in means on the dependent variable across the two independent variable groups.

[3] The equivalent on the *T*-score metric used by many personality scales, which have *SD* = 10, would be a raw score mean difference between two groups of at least 15 points.

using the WAIS-IV when they want to differentiate normal cognitive functioning from mathematics disorders (IQ $M = 86$, $d = 0.93$, $r = .42$); attention deficit–hyperactivity disorder (ADHD; IQ $M = 97$, $d = 0.20$, $r = .10$); moderate to severe TBI (IQ $M = 84$, $d = 1.07$, $r = .47$); autism spectrum disorder (IQ $M = 80$, $d = 1.33$, $r = .55$); Asperger's disorder (IQ $M = 97$, $d = 0.20$, $r = .10$); mild cognitive impairment (IQ $M = 95$, $d = 0.33$, $r = .16$); or mild Alzheimer's disease (IQ $M = 81$, $d = 1.27$, $r = .54$). Although these groups of people all have genuine cognitive impairments of different types, their average degree of impairment is insufficiently severe to meet Areh et al.'s (2022) standard (i.e., a 22.5-point IQ score *M* difference) for using the WAIS-IV to assess them. Expressed differently, the validity coefficients for these research findings are insufficiently large to consider the WAIS-IV valid for assessing those conditions, per Areh et al.

Surveys of the literature favor Cohen's position on effect sizes much more than Areh et al.'s (2022) position. In fact, these surveys suggest that even Cohen's benchmarks are too high. For instance, $r = .21$ is the average effect size for validation research in social psychology (Richard et al., 2003) and for validation research in the field of communication (Rains et al., 2017). In the field of organizational behavior and human resources, the average validity association is $r = .23$ (Paterson et al., 2016), while the median for measurements in applied psychology (Bosco et al., 2015) or across psychological disciplines (Cafri et al., 2010) is $r = .16$. Perhaps most relevant, Gignac and Szodorai (2016) provide benchmarks for the validity of measures testing for individual differences, which is what evaluators do in FMHA. They found that the median effect size across 708 meta-analytic results was $r = .19$, with just 2.7% of validity effect sizes being $r = .50$ or above and less than 1% being 0.60 or above. Converting effect size metrics, these *M r* values equate to *d* values ranging from 0.32 to 0.47.

Based on findings like these, researchers have argued that psychologists need to recalibrate their expectations for the effect size of research results to realistically appreciate the complexities of human experience and behavior and to discourage expectations that promote finding large effects that will not replicate (Funder & Ozer, 2019, 2020; Götz et al., 2022). Thus, while standards proposed by Areh et al. (2022) sound desirable, they imply a falsely simplified world of cause and effect for human behavior and set forensic evaluators and the consumers of their forensic reports up to foster unrealistic aspirations for their assessment measures.

There also are relevant data speaking to Areh et al.'s (2022) assertion about the validity of medical tests. The APA commissioned a work group to assemble evidence on the efficacy of assessment in practice. As part of its efforts, that group conducted a systematic review of meta-analyses examining test validity coefficients (Meyer et al., 2001). They did this separately for psychological tests and medical tests. In addition, they completed a broad review of effect sizes for various types of phenomena (e.g., associations of gender with assertiveness, arm strength, or height; antihistamine use and symptom reduction) so readers would have a better conceptual map for interpreting the magnitude of test validity coefficients. Although the authors provided many important caveats for why it was hazardous to compute and compare the average effect sizes of various types of tests, they provided these data. Their review encompassed effect sizes from 69 meta-analyses of psychological test validity and 57 meta-analyses of medical test validity.

Meyer et al. (2001) classified tests into five groups and did not find evidence for differences in their average validity. The groups had average *r* values (*SD*, # of effects) as follows: self-report personality tests = 0.24 (0.18, 24); typical performance tests of personality (i.e., Rorschach, picture story tasks, sentence completion) = 0.33 (0.09, 8); maximal performance cognitive or neuropsychological tests = 0.34 (0.17, 26); other psychological tests (e.g., clinician or parent ratings, physical ability for job ratings) = 0.30 (0.08, 7); and medical tests = 0.36 (0.21, 63). Although medical tests had some of the largest validity coefficients (e.g., pulse oximetry readings and arterial oxygen saturation, $r = .84$), they also had some of the lowest coefficients (e.g., routine umbilical artery Doppler ultrasound and reduced perinatal deaths in high-risk women, $r = .03$). However, these data do not support the notion that medical researchers consider validity coefficients less than $r = .70$ as problematic. Physical functioning and medical wellbeing are complicated, multifactorially determined phenomena and medical assessment measures vary in the extent to which they correlate with relevant criterion variables. In conclusion, Areh et al.'s (2022) stance that Rorschach test data do not meet acceptable validity standards, especially compared to medical tests, is empirically untenable.

The *F v Bevándorlási és Állampolgársági Hivatal* Case

Areh et al. (2022) highlight the 2018 case of *F v Bevándorlási és Állampolgársági Hivatal*, decided by the European Court of Justice, as an illustration of their arguments against the Rorschach. The nuances (and vagueness) of the facts and reasoning in this judicial opinion are worth further discussion, which is provided below, and with additional details separately in the Appendix.

Informed Consent and Low Face Validity or Mandated Forensic Evaluations?

According to Areh et al. (2022), the court in *F v Bevándorlási és Állampolgársági Hivatal* commented how because of

"the Rorschach's ability to circumvent examinees' conscious defences" (p. 189), "[it] raises the question as to whether or not the use of the Rorschach leads to violations of the right to avoid self-incrimination and the right to remain silent when questioned, either prior to or during legal proceedings in a court of law" (p. 189).

However, on the issue of consent, the court appeared to be discussing more broadly the contextual pressures of an ordered FMHA, not specifically the use of the Rorschach. The court specifically stated,

> In this regard, it should be noted that a psychologist's expert report, such as that at issue in the main proceedings, is commissioned by the determining authority in the context of the examination of the application for international protection submitted by the person concerned.
> It follows that that report is prepared in a context where the person called upon to undergo projective personality tests is in a situation in which his future is closely linked to the decision that that authority will take on his application for international protection and in which a possible refusal to undergo these tests is liable to constitute an important factor on which the authority will rely for the purpose of determining whether that person has sufficiently substantiated his application.
> Therefore, even if the performance of the psychological tests on which an expert's report, such as that at issue in the main proceedings, is based is formally conditional upon the consent of the person concerned, it must be considered that that consent is not necessarily given freely, being de facto imposed under the pressure of the circumstances in which applicants for international protection find themselves . . . . (*F v Bevándorlási és Állampolgársági Hivatal*, 2018, paras. 51–53)

Thus, it is a stretch to suggest that the court's concerns about the interference with an applicant's right to respect for their private life was related to the Rorschach's purportedly lower face validity relative to self-report personality tests. The more common-sense reading would be that the Court was concerned about the contextual demands of mandated asylum evaluations in general, irrespective of the type of psychological testing used. Furthermore, it is hard to envision how the Rorschach could be used to circumvent protections against self-incrimination in legal proceedings, given ethical guidance for forensic psychology concerning informed consent and mandated evaluations (e.g., *Specialty Guidelines for Forensic Psychology*; APA, 2013). Moreover, the Rorschach is in good company when it comes to low face validity: widely utilized and recommended measures of response style also have this quality by design (e.g., Rogers & Bender, 2018).

Admissibility of What Specifically and for What Purpose? And What Have Other Cases Decided?

Areh et al. (2022) also stated that the court in *F v Bevándorlási és Állampolgársági Hivatal* "further commented that the use of projective tests (i.e. tests that use non-structured, unclear stimuli such as ink blots to induce responses;...) had been vigorously contested during the case" (p. 184). However, this is not so clear from a careful reading of the case. The court specifically said,

> In this context, although interference with an applicant's private life can be justified by the search for information enabling his actual need for international protection to be assessed, it is for the determining authority to assess, under the court's supervision, whether a psychologist's expert report which it intends to commission or wishes to take into account is appropriate and necessary in order to achieve that objective.
> In this respect, it should be noted that the suitability of an expert's report such as that at issue in the main proceedings may be accepted only if it is based on sufficiently reliable methods and principles in the light of the standards recognised by the international scientific community. It should be noted in that regard that, although it is not for the Court to rule on this issue, which is, as an assessment of the facts, a matter within the national court's jurisdiction, the reliability of such an expert's report has been vigorously contested by the French and Netherlands Governments as well as by the Commission. (*F v Bevándorlási és Állampolgársági Hivatal*, 2018, paras. 57–58)

As such, it remains unclear whether the court's concerns, and that of the governments of France and the Netherlands, and the European Commission, were in reference to performance-based personality assessment measures in general, the dubious use of performance-based personality assessment measures to ascertain sexual orientation (which was what it was used for in this specific case), or the dubious use of psychological testing in general to identify sexual orientation.

The nature and extent of the court's concerns with performance-based personality assessment measures in *F v Bevándorlási és Állampolgársági Hivatal* are unclear for the reasons just described. Certainly, the court did not engage in any detailed analysis of the Rorschach, including the current state of the science concerning any particular administration and interpretative system, such as R-PAS. Per our reading, the case simply reflects one or two problematically conceived and conducted asylum evaluations (for reasons involving one psychologist's specific application of the Rorschach, but also more general reasons

beyond the evaluator's use of the Rorschach). And on this point, we agree with Areh et al. (2022) that.

> *F v. Hungary* (2018) is a reminder to psychologists who do psycholegal work of the impact their reports can have on the rights and interests of people and that lawyers and courts will therefore approach their reports and testimony critically. Psychologists should consequently ensure that they use tests which can withstand the scrutiny of both their peers and lawyers. (p. 195)

However, in our opinion, highlighting the *F v Bevándorlási és Állampolgársági Hivatal* case in isolation is not a particularly useful approach for appraising the suitability of the Rorschach in well-conceived forensic evaluations. Indeed, one of us (CK) recently observed that the Rorschach was among the top-10 referenced psychological tests in the history of United States case law, with increasing numbers of case law references to the Rorschach across advancing decades (King & Neal, 2021). In addition, as mentioned earlier, some of us recently conducted a comprehensive review of theoretical, empirical, and legal indicators bearing upon the admissibility of the Rorschach in legal proceedings, which included a detailed review of US case law referencing the Rorschach in relative proximity to admissibility-related language (Viglione et al., 2022). Although ours was not the first review of US case law citing the Rorschach (see Gurley et al., 2014; Meloy, 2008; Meloy et al., 1997; Neal et al., 2019), it is arguably the largest and certainly the most exacting to date, particularly with respect to cases addressing the admissibility of Rorschach evidence.

Based on our and prior US case law reviews, we concluded that the Rorschach did not appear to be especially likely to be legally challenged relative to other psychological assessment tools. Rates of challenge to the Rorschach appeared to be somewhere between self-report personality tests, which are challenged somewhat less often, and risk assessment tools and measures of deviant sexual preferences, which are challenged much more often. Moreover, challenges to the Rorschach were "successful" only about a third of the time. The success of such challenges seemed to relate to several factors, including the nature of the case and evaluee (e.g., a civil plaintiff allegedly subjected to wrongful conduct); the procedural posture of the case at the time of a challenge (i.e., pre-trial, trial, or appeal); and the performance of the involved evaluators (e.g., the range of assessment methods employed and the psycholegal issues addressed). Undoubtedly, we identified some instances that struck us as clearly inappropriate forensic applications of the Rorschach (e.g., to determine whether an evaluee fit the profile of someone likely to have committed a prior criminal act). Such uses are deserving of challenge (and we hope, exclusion) as they are inconsistent with the current empirical basis for the Rorschach, and psychological tests in general.

As part of our review (Viglione et al., 2022), we also conducted a secondary non-exhaustive review of case law citing the Rorschach across several European jurisdictions (Belgium, European Union, France, Germany, the Netherlands, and the UK). Across 16 relevant cases, we identified challenges to the admissibility of the Rorschach in just three, including the *F v Bevándorlási és Állampolgársági Hivatal* case. In the other two cases (Bundesgerichtshof [BGH] [Federal Court of Justice], July 7, 1999; *Lowery v The Queen*, 1974), the courts' decisions appeared to support the admissibility of the Rorschach.

Based on our comparative case law review, as well as theoretical and empirical sources concerning the Rorschach (with special attention to R-PAS), we concluded that the question of whether "the Rorschach" satisfies admissibility standards is overly broad and ultimately misguided. Accordingly, we disagree with Areh et al.'s (2022) overgeneralized conclusion "that the Rorschach does not meet the proposed standards and that psychologists should abstain from using it in legal proceedings even in the absence of a clear judicial prohibition" (p. 183). To us, the more appropriate conclusion is this: an appropriately conceived FMHA incorporating a particular Rorschach system and its variables will likely be considered reliable, valid, and admissible in specific types of cases and toward specific functional legal capacities.

Other Errors and Misleading Information in Areh et al. (2022)

To correct some of the misconceptions and misrepresentations of Areh et al. (2022) regarding the Rorschach, we have attempted in the preceding sections to clarify what the Rorschach task is, what the scientific literature says about the reliability and validity of its variables, what scientific standards should be considered when using it in FMHA, and what the current status of the Rorschach test really is with respect to admissibility in court. In this section, we note that the article by Areh et al. (2022) also contains factual errors and misleading citations that are not evenly distributed in both directions. These inaccuracies consistently result in the Rorschach test being portrayed worse than the true facts suggest. In particular, three major categories of problems deserve mention: (1) inappropriate conclusions based on outdated references; (2) mistaken claims and misconceptions; and (3) citation errors.

Inappropriate Conclusions Based on Outdated References

In the section entitled "Validity of test must be appropriate for the legal question," Areh et al. (2022) state that when "Comprehensive System variables were correlated with externally (e.g., psychiatric diagnosis) and introspectively

(e.g., self-report questionnaires) assessed criteria, the mean validities were 0.27 and 0.08, respectively (Mihura et al., 2013)" (p. 191). Areh et al. (2022) then note that Wood et al. (2015) subsequently responded to Mihura et al.'s (2013) publication to suggest that when adding data from non-peer-reviewed unpublished dissertations to the meta-analyses, they "found no evidence supporting a relationship between Comprehensive System indexes and non-cognitive characteristics such as negative affect and emotionality" (p. 191). The same paragraph is subsequently concluded with a citation to a paper published several years *earlier* (Lilienfeld et al., 2000), in which it was suggested that "the correlations between the Comprehensive System scores and most psychiatric diagnoses were not replicated in later studies" (p. 191). This conclusion is misleading for two reasons. First, both Mihura et al.'s (2013) and Wood et al.'s (2015) meta-analyses were conducted more than 10 years after Lilienfeld et al.'s (2000) study. Second, the data referenced by Lilienfeld et al. (2000) were in fact included in the aforementioned meta-analyses. As such, both Mihura et al. (2013) and Wood et al. (2015) provide a much more accurate and complete picture to appreciate the validity of Rorschach CS variables, compared to Lilienfeld et al.'s (2000) review. In addition, Areh et al. (2022) failed to report that Mihura et al. (2015) published a reply to Wood et al. (2015) in the same journal. Mihura et al. (2015) identified numerous methodological errors, data errors, and omitted studies in the Wood et al. (2015) article, suggesting that one cannot rely on the findings or the conclusions reported therein.

A similar error is found in Areh et al.'s (2022) section entitled "Reliability must be appropriate for the purpose for which test will be used" (p. 193). When describing the empirical literature on interrater reliability, Areh et al. (2022) state, "The interrater reliability of the various Comprehensive System variables is even slightly better, with... Meyer's (1997) meta-analysis of 16 studies finding interrater reliability coefficients ranging from 0.72 to 0.96. Other researchers, however, have found lower coefficients (e.g., W. Perry et al., 1995; Wood et al., 1996) and criticised Meyer's study as flawed" (p. 193). Aside from the fact that Perry et al.'s (1995) study did not actually report any data on interrater reliability (see below), it is evident that Wood et al. (1996) were in fact unable to criticize Meyer's (1997) study because the latter was unpublished in 1996. Wood et al.'s (1996) paper actually commented on a study by Meyer published in 1996 (Meyer, 1996). The article cited by Areh et al. (2022)—i.e., Meyer (1997)—was in fact a reply to Wood et al. (1996), in which Meyer used the statistical approach recommended by Wood et al. and documented how the resulting interrater reliability coefficients still remained highly satisfactory.

Relatedly, it is notable that in their review, Areh et al. (2022) only cited studies from the 1990s, ignoring all subsequent studies on Rorschach interrater reliability published in the third millennium (Kivisalu et al., 2016; Lewey et al., 2019; Meyer et al., 2002; Pignolo et al., 2017; Viglione & Taylor, 2003; Viglione et al., 2012). These studies consistently support the satisfactory interrater reliability of interpretively relevant CS and R-PAS scores.

Along similar lines, the section entitled "Assessment of the Rorschach" in Areh et al. (2022) claims that "Rorschach supporters believe that it circumvents examinees' conscious defenses because they respond to ambiguous stimuli with a minimum of instructions" (p. 189). To substantiate this claim, Areh et al. (2022) cite four papers in that same paragraph: Siipola and Taylor (1952), Weiner et al. (1996), Cerney (1990), Leavitt and Labott (1996). None of these references, however, was published in the third millennium and, in fact, more recent publications conceptualized the test in a very different manner, as we summarized above. Examples of more recent but overlooked references are Exner (2003), Exner and Erdberg (2005), Finn (2012), Meyer and Kurtz (2006), Meyer and Eblin (2012), and Meyer (2017).

Mistaken Claims and Misconceptions

When elaborating on the theoretical basis of the Rorschach, Areh et al.'s (2022) opening sentence states, "the projective hypothesis was the original theoretical assumption behind the Rorschach (1921/1951). At first, it was based on Freud's (1911) theory that people unconsciously assign their characteristics and impulses to others as defence mechanisms" (p. 190). The idea that Hermann Rorschach created his test based on Freud's projective hypothesis is incorrect. The term *projection* is used only once in the entirety of *Psychodiagnostics*, and not in a place describing the nature of the test (Rorschach, 1921). If Rorschach himself believed the test data contained *projected* material, it is reasonable to assume he would have used the term more extensively. According to Schachtel (1966), this misconception likely arose in the American literature on the test and, in fact, the "concept of projection, as originally developed by Freud, plays no important role in any of the 'projective' techniques" (p. 10). Consistent with this position, the idea that Rorschach developed his test based on the projective hypothesis has been refuted several times (e.g., Acklin & Oliveira-Berry, 1996).

Another misconception is evident in Areh et al.'s (2022) statement that the Rorschach "appears to be useful in diagnosing bipolar disorder, schizophrenia and schizotypal personality disorder (Wood et al., 2000). It is, however, less useful in diagnosing post-traumatic stress disorder (PTSD) and other anxiety disorders, major depressive disorder, suicide attempts, dissociative identity disorder, psychopathy, antisocial personality disorder and dependent, narcissistic (see Exner, 1995) or conduct disorders (see Carlson et al., 1997; Hunsley et al., 2015; Wood et al., 2015)" (p. 192).

The idea that the Rorschach should be used "to diagnose" any disorder is problematic, because even though predictive validity is an important psychometric property of any test score, formulating a diagnosis is not the primary goal, and certainly not the sole goal, when using the Rorschach in forensic practice. Recent conceptualizations of the test indeed explicitly note that, like any other psychological assessment measure, the Rorschach should not be used as a stand-alone test to formulate any mental health diagnosis. It would be more appropriate to say that its function is to assist the evaluator to systematically observe and measure what might be referred to as the "personality in action" (Meyer & Eblin, 2012; Meyer et al., 2011), so as to gain valuable information on the evaluee's psychological functioning in terms of information processing, communication, imagery, and self- and interpersonal representations.

Areh et al. (2022) also mistakenly suggest that the use of the Rorschach has dropped down dramatically over the past few years. They state, "surveys of forensic practitioners in North America have... revealed that the Rorschach was the most frequently used unstructured projective test, with up to 36% of research participants reporting that they had used it (see Archer et al., 2006). More recent North American research, however, shows that the frequency of Rorschach use has dropped down to 20% (Viljoen et al., 2010) or even to 3% (Neal & Grisso, 2014)" (p. 194). The problems with this section are twofold. First, Areh et al. (2022) did not provide a comprehensive review of forensic practitioner surveys available in the literature (for such a review, see Viglione et al., 2022). Second, the percentages reported by Archer et al. (2006) are by no means comparable with those by Neal and Grisso (2014). Indeed, Archer et al.'s (2006) percentage values referred to the percentages of respondents who had used a given test, whereas Neal and Grisso (2014) referred to the proportion of respondents using a specific measure in their two most recently conducted forensic evaluations. To provide a yardstick, in Archer et al. (2006), the percentage values for the MMPI, PAI, and Rorschach were 86%, 46%, and 36% respectively; in Neal and Grisso (2014), the corresponding values were 15.2%, 9.6%, and 3.2%. Following Areh et al.'s (2022) approach to interpreting these values, one would thus conclude that the use of the MMPI dropped down from 86 to 15.2% and that of the PAI dropped down from 46 to 9.6%. This interpretation is evidently both mistaken and misleading.

Another factual error in the Areh et al. (2022) article concerns the following claim, "French examinees often see a chameleon in Card VIII, Scandinavian examinees often see Christmas elves in Card II and Japanese examinees often provide a musical instrument-related answer to Card VI. All of these answers are unusual responses according to the Comprehensive System (Weiner, 2014)" (p. 191). First, according to both CS and R-PAS, whether a response is considered to be frequent, common, and ordinary—versus infrequent, unusual, and uncommon—does not solely depend on its content, but also on where it is seen in the inkblot (i.e., its location). For instance, while it would be expected for an evaluee to see a bat in the whole inkblot location of Card V, it would be more surprising if this person provided that same response, i.e., a bat, by focusing on the left portion of the inkblot only. Second, assuming that Areh et al. (2022) meant to suggest that it is common for French evaluees to see a chameleon in a location of Card VIII named "D1," for Scandinavian evaluees to see two Christmas elves in the whole inkblot location of Card II, and for Japanese evaluees to see a musical instrument-related response in the whole inkblot location of Card VI, all of these responses would in fact be classified as frequent, common, and ordinary by both CS (Exner, 2003) and R-PAS (Meyer et al., 2011), and not unusual as claimed by Areh et al. (2022).

Citation Errors

In multiple instances, Areh et al.'s (2022) article cites sources that do not support or have no bearing on the point being made. As citations allow the reader to trace the flow of evidence, these errors are important to document—so we report some of the most problematic ones below. On page 191, Areh et al. (2022) state that "Giromini et al. (2017) found that demographic variables do not influence Rorschach scores, whereas others found significant effects (Delavari et al., 2013; Meyer et al., 2007)." There are two notable errors with this statement. First, Giromini et al. (2017) did not investigate the relationship between Rorschach scores and demographic variables. As its title reveals, *Neural activity during production of Rorschach responses: An fMRI study*, Giromini et al. (2017) investigated what brain areas are differentially activated when the Rorschach is administered, and no Rorschach scores were investigated nor reported in this article. In addition, the aforementioned sentence proposes that Meyer et al. (2007) found significant effects of demographic variables on Rorschach scores, but in reality, the cited paper suggested the opposite, i.e., that the "adult samples from around the world are generally quite similar" (Meyer et al., 2007; p. S201).

On page 193, Areh et al. (2022) state, "The interrater reliability of the various Comprehensive System variables is even slightly better, with... Meyer's (1997) meta-analysis of 16 studies finding interrater reliability coefficients ranging from 0.72 to 0.96. Other researchers, however, have found lower coefficients (e.g. W. Perry et al., 1995; Wood et al., 1996) and criticised Meyer's study as flawed" (p. 193). The citation "Perry et al., 1995" here suggests that the authors reported information on interrater reliability, but Perry et al.'s (1995) article reported κ corrected agreement rates for eight variables that ranged from 0.63 to 0.89, indicating good to excellent reliability. Otherwise, they focused on

test–retest reliability, as its title indicates, *A five-year follow-up on the temporal stability of the Ego Impairment Index*. Additionally, as noted before, referencing "interrater reliability coefficients" in general, without specifying the type of coefficient used (percent agreement, *r*, exact agreement *ICC*, κ, etc.) ignores how the values of these coefficients are interpreted differently.

On page 194, Areh et al. (2022) state, "several researchers have nevertheless criticised the composition of the normative sample (Hibbard, 2003; Viglione & Giromini, 2016)." Hibbard's (2003) article, however, did not criticize CS norms; in fact, it criticized Lilienfeld's criticisms against CS norms. It suffices to provide the title of the article to appreciate Areh et al.'s (2022) misrepresentation: *A Critique of* Lilienfeld et al.'s (2000) *"The Scientific Status of Projective Techniques."*

Lastly, it should be pointed out that on several occasions, Areh et al. (2022) cite books or book chapters, rather than empirical research articles published in peer-reviewed journals, to support their claims. For instance, on page 192, Areh et al. (2022) state, "The validity of the Rorschach overall is not good (Kaplan & Saccuzzo, 2017)." This bold statement, based on a book, is not in line with what is reported in the most extensive meta-analyses on the Rorschach CS variables (Mihura et al., 2013, 2015), which were published in *Psychological Bulletin*, a flagship journal in psychological science (impact factor > 20).

In conclusion, the informed Rorschach scholar readily notices these significant citation errors. These errors misrepresented the empirical research base in ways that supported Areh et al.'s (2022) critical views.

Conclusion

As discussed in this article, Areh et al. (2022) based their conclusion that psychologists should refrain from using the Rorschach in legal proceedings on a significant misunderstanding of how the task works, relative neglect of the past 20 years of Rorschach research, unrealistic psychometric standards for assessing the validity and reliability of a psychological assessment measure, and a single European legal case (*F v Bevándorlási és Állampolgársági Hivatal*, 2018) in which a forensic expert used the Rorschach inappropriately. Still, the article by Areh et al. (2022) has the merit of addressing an important issue in FMHA: not all Rorschach methods are equally reliable (in the legal sense of the word); and we agree with Areh et al. that certain approaches to Rorschach administration, coding, and interpretation are unlikely to meet current psychometric and legal standards for psychological test selection for practice. Accordingly, we strongly recommend that forensic evaluators use R-PAS when using the Rorschach in an applied forensic setting.

We close by offering four reasons why the Rorschach is likely to be a useful tool for FMHA, when R-PAS is used. First, the Rorschach can aid in the assessment of psychological functions, which are relevant in answering psychological questions. Examples include reality testing, coping resources, and thought disorder (for a complete overview, see Table 5 in Viglione et al., 2022).

Second, the Rorschach is of particular value as an alternative method of gathering information relative to self- and informant-report, maximum performance tasks, interview, and case file information. This is critical because multimethod assessment is a best practice for FMHA (APA, 2013). Forensic practitioners are advised to conduct multitrait-multimethod assessments; see, e.g., *Specialty Guideline* 9.02: Use of Multiple Sources of Information (APA, 2013). Heilbrun et al. (2009) explain that the high stakes of a FMHA necessitate using more than one measure to ensure reasonably accurate conclusions. Mihura (2012) notes that there is generally a high degree of statistical correlation between different self-report inventories. The Rorschach, as a performance-based assessment measure with generally minimal overlap with self-report, is a complement to self-report questionnaires. It is not intended to be administered as a stand-alone measure but provides a distinct means of gathering relevant data that is weighed within the context of other assessment data sources (Meyer et al., 2011).

Third, forensic evaluations may involve the assessment of non-mental health constructs such as psychosocial maturity, adaptive personality characteristics, or best interests of the child (Heilbrun et al., 2009). The Rorschach provides broad personality trait information that is relevant beyond the assessment of psychopathology (Meyer et al., 2011).

Fourth and finally, forensic evaluators must always be mindful of the potential distorting impact of evaluee response bias (Heilbrun et al., 2009; Sweet et al., 2021; Viglione et al., 2022). The Rorschach does not have decisive strategies to detect feigned mental illness, though an elevated number of dramatic contents is common (Kiss et al., 2023; Meyer et al., 2011; Sewell & Helle, 2018). In addition, evaluees attempting to simulate positive psychological adjustment are largely unsuccessful due to the ambiguity of Rorschach stimuli and scoring criteria (Benjestorf et al., 2013; Nørbech et al, 2016). The latter finding is important, because positive impression management is common in forensic evaluees (for example, in violence risk assessments or child custody evaluations).

In conclusion, given the results of the review conducted by Viglione et al. (2022) and considering the additional information discussed in this article, we believe that the case of *F v Bevándorlási és Állampolgársági Hivatal* represents something simple yet important: how a poorly conceived FMHA by a psychologist reflects poorly on the profession and psychological assessment. Our analysis of the Areh et al.

(2022) paper uncovered a number of serious errors of omission and commission on the part of the authors. This resulted in a misrepresentation of the current evidence base on the value of the Rorschach, and R-PAS in specific, for use in legal contexts, which we have sought to correct.

Appendix. A More Detailed Description of the European Court of Justice Case Used by Areh et al. (2022)

The *F v Bevándorlási és Állampolgársági Hivatal* (2018) case arose out of an asylum application, in which a Nigerian man sought international protection in Hungary. During his first interview with the Hungarian Office for Immigration and Citizenship, the applicant claimed a well-founded fear of persecution for his homosexual sexual orientation in his country of origin. The man was evaluated by a psychologist, who administered the Rorschach, Draw-a-Person-in-the-Rain test, and Szondi test. In a resultant report, the psychologist opined that it was not possible to confirm the applicant's statements as to his sexual orientation. Although the Office did not find the applicant's statements about his sexual orientation to be contradictory, it nevertheless rejected his asylum application, concluding that he lacked credibility.

The man filed an action challenging the Office's decision with the Hungarian Administrative and Labour Court. He argued that the psychological testing to which he had been subjected seriously prejudiced his protected fundamental rights and could not assess the plausibility of his sexual orientation. The Administrative and Labour Court ordered a new forensic mental health evaluation report that was to utilize methods that were not prejudicial to human dignity, suitably explored the issues presented by the case, and were appropriate for reaching opinions about the applicant's sexual orientation (including his truthfulness about it). Furthermore, although the Administrative and Labour Court did not find that the applicant had specifically shown how the tests administered in the original evaluation had prejudiced his fundamental rights, it nevertheless stayed the proceedings and referred two questions to the European Court of Justice.

The first question was whether a forensic psychologist's expert opinion about a sexual-orientation-based asylum application could be ordered and considered if it was based on "projective personality tests" (a phrase used several times throughout the judicial opinion) but did not involve questions about an applicant's "sexual habits" or a physical evaluation (*F v Bevándorlási és Állampolgársági Hivatal*, 2018, para. 26). If such an expert report could not be considered, the second and related question was whether administrative authorities or reviewing courts were forbidden from considering *any* expert methods to inform their assessments of truthfulness in sexual-orientation-based asylum cases.

The Court of Justice of the European Union began by addressing the second question and held that administrative authorities or reviewing courts were not precluded from ordering a forensic evaluation in such a case. However, the evaluation had to be conducted in a manner consistent with protected fundamental rights (e.g., respect for human dignity and private and family life), and an expert's opinions could neither be the sole basis for a decision nor treated as binding. The Court noted that it was not always necessary to assess the credibility of an asylum applicant's sexual orientation—namely, where an applicant is inaccurately viewed as having a certain sexual orientation by persons allegedly persecuting them in the origin country. Yet, the Court could not rule out the possibility that certain cases may raise a need for a more searching assessment of facts and circumstances to determine an applicant's true need for international protection more accurately. And per the Court, administrative authorities and reviewing courts were generally not restricted with respect to the means to do so. The Court granted that forensic evaluations could be useful in assessing the facts and circumstances of a sexual-orientation-based asylum application (e.g., about the situation of persons with the same sexual orientation in the country of origin), and that such evaluations could be conducted without prejudicing an applicant's fundamental rights.

In turning back to the first question, the Court held that a forensic evaluation that purported to opine about an applicant's sexual orientation based on projective personality tests could not be considered to determine the veracity of an applicant's statements concerning their sexual orientation. The Court reasoned that an applicant for international protection is under contextual pressure to consent to a forensic psychological evaluation, which interferes with the applicant's right to respect for their private life. The Court explained that such an interference must be proportional to what is appropriate and necessary to accomplish legitimate legislative objectives. Moreover, the Court asserted that a forensic evaluation report should only be considered if it is based on sufficiently reliable methods and principles, per standards recognized by the international scientific community. And while such a determination was a factual matter within the jurisdiction of a nation's courts, the Court noted that the governments of France and the Netherlands, and the European Commission, had all contested the reliability of a forensic mental health evaluation of the sort at issue in the case (i.e., a forensic evaluation heavily reliant on projective personality testing to reach opinions about the sexual orientation of an asylum applicant).

On the one hand, the Court acknowledged that a forensic evaluation for the purpose of establishing an applicant's sexual orientation was within the scope of an assessment of the facts or circumstances of an application for international protection. It also acknowledged that there was a legitimate governmental interest in ensuring the competence and appropriate skillfulness of its personnel

(e.g., highly trained psychologists) for assessing sexual-orientation-based asylum applications. On the other hand, the Court reasoned that sexual orientation is an essential element of personal identity and intimate life. It also cited a source of persuasive authority concerning international human rights law and sexual orientation, which provides that persons should not be forced to participate in psychological testing due to their sexual orientation.

On balance, the Court regarded a forensic evaluation, based on projective personality testing, for the purpose of opining about an applicant's sexual orientation, as representing a serious infringement on the right to privacy, and a disproportionate one, considering the potential benefits such an evaluation could offer. It explained that an applicable immigration law provision did not require further confirmation of an applicant's uncorroborated statements about their sexual orientation if the statements were nevertheless consistent and plausible—and this provision made no mention of forensic evaluations. Furthermore, the Court reasoned that even if a forensic evaluation that relied on projective personality tests could contribute to the reliable identification of an applicant's sexual orientation by an initial authority or reviewing court, the nature of the expert's conclusion would still be approximate. These limitations were particularly acute to the Court in this case because the applicant's statements about his sexual orientation had not been contradictory.

The initial forensic psychologist in the case had only administered performance-based "projective" personality assessment measures, and a question posed by the referring nation's court specifically referred to this type of testing. Yet the Court's reasoning with respect to the limited and potentially superfluous nature of forensic evaluations for providing an indication of an applicant's sexual orientation would appear to apply in equal measure to evaluations that incorporate self-report personality tests. Furthermore, the Court seemed to discuss more favorably forensic evaluations that would speak to country-of-origin conditions for members of sexual minority groups, versus forensic evaluations seeking to substantiate an applicant's sexual orientation. Thus, the Court was not clear in its opinion about whether it was sticking closely to the questioned factual situation with which it was presented, taking special issue with performance-based personality assessment measures, or disallowing forensic evaluations claiming to provide evidence on an applicant's sexual orientation in general.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement.

Declarations

Conflict of Interest The third author is a co-owner of the company that sells the Rorschach Performance Assessment System® (R-PAS®) manual and associated products. Luciano Giromini and Corine de Ruiter are members of the R-PAS Research and Development Group for which they receive no monetary compensation.

References

Acklin, M. W., & Oliveira-Berry, J. (1996). Return to the source: Rorschach's psychodiagnostics. *Journal of Personality Assessment, 67*(2), 427–433. https://doi.org/10.1207/s15327752jpa6702_17

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum.

Ales, F., Giromini, L., & Zennaro, A. (2020). Complexity and cognitive engagement in the Rorschach task: An eye-tracking study. *Journal of Personality Assessment, 102*(4), 538–550. https://doi.org/10.1080/00223891.2019.1575227

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *American Psychologist, 68*(1), 7–19. https://doi.org/10.1037/a0029889

Andò, A., Pineda, J. A., Giromini, L., Soghoyan, G., QunYang, B., & M., Maryanovsky, D., & Zennaro, A. (2018). Effects of repetitive transcranial magnetic stimulation (rTMS) on attribution of movement to ambiguous stimuli and EEG mu suppression. *Brain Research, 1680*, 69–76. https://doi.org/10.1016/j.brainres.2017.12.007

Andò, A., Salatino, A., Giromini, L., Ricci, R., Pignolo, C., Cristofanelli, S., Ferro, L., Viglione, D. J., & Zennaro, A. (2015). Embodied simulation and ambiguous stimuli: The role of the mirror neuron system. *Brain Research, 1629*, 135–142. https://doi.org/10.1016/j.brainres.2015.10.025

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*(1), 84–94. https://doi.org/10.1207/s15327752jpa8701_07

Areh, I., Verkampt, F., & Allan, A. (2022). Critical review of the use of the Rorschach in European courts. *Psychiatry, Psychology and Law, 29*(2), 183–205. https://doi.org/10.1080/13218719.2021.1894260

Asari, T., Konishi, S., Jimura, K., Chikazoe, J., Nakamura, N., & Miyashita, Y. (2008). Right temporopolar activation associated with unique perception. *NeuroImage, 41*(1), 145–152. https://doi.org/10.1016/j.neuroimage.2008.01.059

Asari, T., Konishi, S., Jimura, K., Chikazoe, J., Nakamura, N., & Miyashita, Y. (2010a). Amygdalar enlargement associated with unique perception. *Cortex, 46*(1), 94–99. https://doi.org/10.1016/j.cortex.2008.08.001

Asari, T., Konishi, S., Jimura, K., Chikazoe, J., Nakamura, N., & Miyashita, Y. (2010b). Amygdalar modulation of frontotemporal connectivity during the inkblot test. *Psychiatry Research: Neuroimaging, 182*(2), 103–110. https://doi.org/10.1016/j.pscychresns.2010.01.002

Benjestorf, S. T., Viglione, D. J., Lamb, J. D., & Giromini, L. (2013). Suppression of aggressive Rorschach responses among violent offenders and nonoffenders. *Journal of Interpersonal Violence, 28*(15), 2981–3003. https://doi.org/10.1177/0886260513488688

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. https://doi.org/10.1080/00223891.2017.1408016

Bornstein, R. F. (2017). Evidence-based psychological assessment. *Journal of Personality Assessment, 99*(4), 435–445. https://doi.org/10.1080/00223891.2016.1236343

Bornstein, R. F. (2022). *Toward an Integrative Perspective on the Person*. Avance Online Publication. https://doi.org/10.1027/1192-5604/a000160

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*(2), 431–449. https://doi.org/10.1037/a0038047

Bundesgerichtsh of [BGH] [Federal Court of Justice]. (1999, July 7). 1 StR 207/99. https://www.hrr-strafrecht.de/hrr/1/99/1-207-99.php3(Ger)

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*(2), 239–270. https://doi.org/10.1080/00273171003680187

Carlson, C. F., Kula, M. L., & St. Laurent, C. M. (1997). Rorschach revised DEPI and CDI with inpatient major depressives and borderline personality disorder with major depression: Validity issues. *Journal of Clinical Psychology, 53*(1), 51–58. https://doi.org/10.1002/(sici)1097-4679(199701)53:1%3c51::aid-jclp7%3e3.0.co;2-y

Cerney, M. S. (1990). The Rorschach and traumatic loss: Can the presence of traumatic loss be detected from the Rorschach? *Journal of Personality Assessment, 55*(3–4), 781–789. https://doi.org/10.1080/00223891.1990.9674112

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cronbach, L. J. (1990). *Essentials of psychological testing*. Harper & Row.

Daubert v. Merrell Dow Pharmaceuticals Inc. (1993). 509 U.S. 579.

Davey, C. G., Pujol, J., & Harrison, B. J. (2016). Mapping the self in the brain's default mode network. *NeuroImage, 132*, 390–397. https://doi.org/10.1016/j.neuroimage.2016.02.022

De Ruiter, C., & Kaser-Boyd, N. (2015). *Forensic psychological assessment in practice: Case Studies*. Routledge.

Delavari, M., Shairi, M., & Asghari-Moghadam, M. (2013). Role of culture and gender in Rorschach findings in 9 year old Iranian children. *Procedia – Social and Behavioral Sciences, 84*, 1565–1570. https://doi.org/10.1016/j.sbspro.2013.06.789

European Federation of Psychologists' Associations. (2005). *Meta-code of ethics*. https://doi.org/10.1002/9781444306514.app1

Exner, J. E., Jr. (Ed.). (1995). *Issues and methods in Rorschach research*. Lawrence Erlbaum Associates, Inc.

Exner, J. E., Jr., & Erdberg, P. (2005). *The Rorschach: A Comprehensive System. Vol. 2. Advanced Interpretation* (3rd ed.). John Wiley & Sons.

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

Exner, J. E., Jr. (2003). *The Rorschach: A comprehensive system* (4th ed.). Wiley.

F v Bevándorlási és Állampolgársági Hivatal. (2018). Case C-473/16. E.C.R. 36 (Eur. Union).

Finn, S. E. (2012). Implications of recent research in neurobiology for psychological assessment. *Journal of Personality Assessment, 94*(5), 440–449. https://doi.org/10.1080/00223891.2012.700665

Freud, S. (1911). Psycho-analytic notes on an autobiographical account of a case of paranoia (dementia paranoides). In *Standard edition of the complete works of Sigmund Freud* (Vol. 12, J. Strachey, Trans., pp. 9–79). Hogarth.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, Article 151. https://doi.org/10.3389/fpsyg.2012.00151

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Funder, D. C., & Ozer, D. J. (2020). Corrigendum: Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 3*(4), 509. https://doi.org/10.1177/2515245920979282

Garb, H. N. (1999). Call for a moratorium on the use of the Rorschach Inkblot Test in clinical and forensic settings. *Assessment, 6*(4), 313–317. https://doi.org/10.1177/107319119900600402

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment, 13*(4), 433–448. https://doi.org/10.1037/1040-3590.13.4.433

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. Personality and Individual Differences, 102, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Giromini, L., Porcelli, P., Viglione, D. J., Parolin, L., & Pineda, J. A. (2010). The feeling of movement: EEG evidence for mirroring activity during the observations of static, ambiguous stimuli in the Rorschach cards. *Biological Psychology, 85*(2), 233–241. https://doi.org/10.1016/j.biopsycho.2010.07.008

Giromini, L., Viglione, D. J., Pineda, J. A., Porcelli, P., Hubbard, D., Zennaro, A., & Cauda, F. (2019). Human movement responses to the Rorschach and mirroring activity: An fMRI study. *Assessment, 26*(1), 56–69. https://doi.org/10.1177/1073191117731813

Giromini, L., Viglione, D. J., Zennaro, A., & Cauda, F. (2017). Neural activity during production of Rorschach responses: An fMRI study. *Psychiatric Research: Neuroimaging, 262*, 25–31. https://doi.org/10.1016/j.pscychresns.2017.02.001

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science, 17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Gurley, J. R., Sheehan, B. L., Piechowski, L. D., & Gray, J. (2014). The admissibility of the R-PAS in court. *Psychological Injury and Law, 7*(1), 9–17. https://doi.org/10.1007/s12207-014-9182-2

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*(3), 257–272. https://doi.org/10.1007/BF01044769

Heilbrun, K., Grisso, T., & Goldstein, A. (2009). *Foundations of forensic mental health assessment*. Oxford University Press.

Hibbard, S. (2003). A critique of Lilienfeld et al.'s (2000) "The scientific status of projective techniques." *Journal of Personality Assessment, 80*(3), 260–271. https://doi.org/10.1207/S15327752JPA8003_05

Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*(2), 259–285. https://doi.org/10.1111/nous.12062

Hunsley, J., Lee, C. M., Wood, J. M., & Taylor, W. (2015). Controversial and questionable assessment techniques. In S. O. Lilienfeld, S. J. Lynn, & J. M. Lohr (Eds.), *Science and pseudoscience in clinical psychology* (pp. 42–82). Guilford Press.

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues* (9th ed.). Cengage Learning.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2

King, C. M., & Neal, T. M. S. (2021). *The unchecked rise of psychological testing evidence in United States courts [Manuscript submitted for publication]*. Montclair State University.

Kiss, A. B., Mihura, J. L., Meyer, G. J., Pimentel, R. P. F. A., & Kletzka, N. (2023). *Comparing committed forensic inpatients to nonpatients instructed to feign insanity or not using scores from the Rorschach task and self-report*. Avance Online Publication.

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. https://doi.org/10.1080/00223891.2015.1118380

Leavitt, F., & Labott, S. M. (1996). Authenticity of recovered sexual abuse memories: A Rorschach study. *Journal of Traumatic Stress, 9*(3), 483–496. https://doi.org/10.1002/jts.2490090307

Lewey, J. H., Kivisalu, T. M., & Giromini, L. (2019). 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. https://doi.org/10.1080/00223891.2018.1476361

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*(2), 27–66. https://doi.org/10.1111/1529-1006.002

Lowery v The Queen. (1974). AC 85 (PC) (appeal taken from Austl.).

Meloy, J. R. (2008). The authority of the Rorschach: An update. In C. B. Gacono, F. B. Evans, N. Kaser-Boyd, & L. A. Gacono (Eds.), *The handbook of forensic Rorschach assessment* (pp. 79–87). Routledge.

Meloy, J. R., Hansen, T. L., & Weiner, I. B. (1997). Authority of the Rorschach: Legal citations during the past 50 years. *Journal of Personality Assessment, 69*(1), 53–62. https://doi.org/10.1207/s15327752jpa6901_3

Meyer, G. J. (1996). Construct validation of scales derived from the Rorschach method: A review of issues and introduction to the Rorschach Rating Scale. *Journal of Personality Assessment, 67*(3), 598–628. https://doi.org/10.1207/s15327752jpa6703_14

Meyer, G. J. (1997). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment, 9*(4), 480–489. https://doi.org/10.1037/1040-3590.9.4.480

Meyer, G. J. (Ed.). (2001). Special Section II: The utility of the Rorschach for clinical assessment. *Psychological Assessment, 13*(4), 419–502. Retrieved from https://psycnet.apa.org/PsycARTICLES/journal/pas/13/4

Meyer, G. J. (2004). The reliability and validity of the Rorschach and TAT compared to other psychological and medical procedures: An analysis of systematically gathered evidence. In M. Hersen (Ed.-in-Chief) & M. Hilsenroth & D. Segal (Eds.), *Comprehensive handbook of psychological assessment: Vol. 2. Personality assessment* (pp. 315–342). Wiley.

Meyer, G. J. (2017). What Rorschach performance can add to assessing and understanding personality. *International Journal of Personality Psychology, 3*(1), 36–49. Retrieved from https://ijpp.rug.nl/article/download/29881/27195/35551

Meyer, G. J. (2023). Understanding complexity as a construct and as a formally scored variable. *Rorschachiana, Advance Online Publication.* https://doi.org/10.1027/1192-5604/a000166

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*(4), 486–502. https://doi.org/10.1037/1040-3590.13.4.486

Meyer, G. J., & Eblin, J. J. (2012). An overview of the Rorschach Performance Assessment System (R-PAS). *Psychological Injury and Law, 5*(2), 107–121. https://doi.org/10.1007/s12207-012-9130-y

Meyer, G. J., Erdberg, P., & Shaffer, T. W. (2007). Toward international normative reference data for the Comprehensive System. *Journal of Personality Assessment, 89*(Suppl 1), S201–S216. https://doi.org/10.1080/00223890701629342

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*(2), 128–165. https://doi.org/10.1037/0003-066X.56.2.128

Meyer, G. J., & Friston, K. J. (2022). The active Bayesian brain and the Rorschach task. *Rorschachiana, 43*(2), 128–150. https://doi.org/10.1027/1192-5604/a000158

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Jr., 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. https://doi.org/10.1207/S15327752JPA7802_03

Meyer, G. J., Huprich, S. K., Blais, M. A., Bornstein, R. F., Mihura, J. L., Smith, J. D., & Weiner, I. B. (2018). *From screening to integrative multimethod assessment: A framework for conceptualizing and optimally using methods in psychological research and practice* [Unpublished manuscript]. Department of Psychology, University of Toledo.

Meyer, G. J., & Kurtz, J. E. (2006). Advancing personality assessment terminology: Time to retire "objective" and "projective" as personality test descriptors [Editorial]. *Journal of Personality Assessment, 87*(3), 223–225. https://doi.org/10.1207/s15327752jpa8703_01

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*(3), 296–314. https://doi.org/10.1207/s15327752jpa8403_09

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, LLC.

Mihura, J. L. (2012). The necessity of multiple test methods in conducting assessments: The role of the Rorschach and self-report. *Psychological Injury and Law, 5*, 97–106. https://doi.org/10.1007/s12207-012-9132-9

Mihura, J. L., & Meyer, G. J. (Eds.). (2018). Using the Rorschach Performance Assessment System® (R-PAS®). Guilford Press.

Mihura, J. L., Meyer, G. J., Bombel, G., & Dumitrascu, N. (2015). Standards, accuracy, and questions of bias in Rorschach meta-analyses: Reply to Wood, Garb, Nezworski, Lilienfeld, and Duke (2015). *Psychological Bulletin, 141*(1), 250–260. https://doi.org/10.1037/a0038445

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. https://doi.org/10.1037/a0029406

Momenian-Schneider, S. H., Brabender, V. M., & Nath, S. R. (2009). Psychophysiological reactions to the response phase of the Rorschach and 16PF. *Journal of Personality Assessment, 91*(5), 494–496. https://doi.org/10.1080/00223890903088727

Neal, T. M. S., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An international snapshot. *Criminal Justice and Behavior, 41*(12), 1406–1421. https://doi.org/10.1177/0093854814548449

Neal, T. M. S., Sellbom, M., & de Ruiter, C. (2022). Personality assessment in legal contexts: Introduction to the Special Issue. *Journal of Personality Assessment*, *104*(2), 127–136. https://doi.org/10.1080/00223891.2022.2033248

Neal, T. M. S., Slobogin, C., Saks, M. J., Faigman, D. L., & Geisinger, K. F. (2019). Psychological assessments in legal contexts: Are courts keeping "junk science" out of the courtroom? *Psychological Science in the Public Interest, 20*(3), 135–164. https://doi.org/10.1177/1529100619888860

Newmark, C. S., Newmark, L., & Faschingbauer, T. R. (1974). Utility of three abbreviated MMPIs with psychiatric outpatients. *Journal of Nervous and Mental Disease, 159*(6), 438–443. https://doi.org/10.1097/00005053-197412000-00007

Newmark, S. R., Anderson, C. F., Donadio, J. V., & Ellefson, R. D. (1975). Lipoprotein profiles in adult nephrotics. *Mayo Clinic Proceedings, 50*(7), 359–364.

Nørbech, P. C. B., Fodstad, L., Kuisma, I., Lunde, K. B., & Hartmann, E. (2016). Incarcerated violent offenders' ability to avoid revealing their potential for violence on the Rorschach and the MMPI-2. *Journal of Personality Assessment, 98*(4), 419–429. https://doi.org/10.1080/00223891.2015.1129613

Parr, T., & Friston, K. J. (2017). The active construction of the visual world. *Neuropsychologia, 104*, 92–101. https://doi.org/10.1016/j.neuropsychologia.2017.08.003

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership & Organizational Studies, 23*(1), 66–81. https://doi.org/10.1177/1548051815614321

Perry, W., McDougall, A., & Viglione, D. (1995). A five-year follow-up on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment, 64*(1), 112–118. https://doi.org/10.1207/s15327752jpa6401_7

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. https://doi.org/10.1080/00223891.2017.1296844

Pineda, J. A., Giromini, L., Porcelli, P., Parolin, L., & Viglione, D. J. (2011). Mu suppression and human movement responses to the Rorschach test. *NeuroReport, 22*(5), 223–226. https://doi.org/10.1097/WNR.0b013e328344f45c

Post, M. W. (2016). What to do with 'moderate' reliability and validity coefficients? *Archives of Physical Medicine and Rehabilitation, 97*(7), 1051–1052. https://doi.org/10.1016/j.apmr.2016.04.001

Rains, S. A., Brunner, S. R., Akers, C., Pavlich, C. A., & Goktas, S. (2017). Computer-mediated communication (CMC) and social support: Testing the effects of using CMC on support outcomes. *Journal of Social and Personal Relationships, 34*(8), 1186–1205. https://doi.org/10.1177/0265407516670533

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rogers, R., & Bender, S. D. (Eds.). (2018). *Clinical assessment of malingering and deception* (4th ed.). Guilford Press.

Rorschach, H. (1921/1951). *Psychodiagnostics* (5th ed.). Hans Huber.

Sackett, P. R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance, 20*(3), 179–185. https://doi.org/10.1080/08959280701332968

Schachtel, E. G. (1966). *Experiential foundations of Rorschach's test*. Analytic Press.

Schneider, A. M. A., Bandeira, D. R., & Meyer, G. J. (2020). Rorschach Performance Assessment System (R-PAS) interrater reliability in a Brazilian adolescent sample and comparisons with three other studies. *Assessment, 29*(5), 859–871. https://doi.org/10.1177/1073191120973075

Searls, D. (2017). *The inkblots: Hermann Rorschach, his iconic test, and the power of seeing*. Crown Publishers/Random House.

Sewell, K. W., & Helle, A. C. (2018). Dissimulation on projective measures: An updated appraisal of a very old question. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (pp. 301–313). Guilford Press.

Siipola, E., & Taylor, V. (1952). Reactions to ink blots under free and pressure conditions. *Journal of Personality, 21*(1), 22–47. https://doi.org/10.1111/j.1467-6494.1952.tb01857.x

Slobogin, C. (2007). *Proving the unprovable: The role of law, science, and speculation in adjudicating culpability and dangerousness*. Oxford University Press.

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., Suhr, J. A., & Participants, C. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 35*(6), 1053–1106. https://doi.org/10.1080/13854046.2021.1896036

Viglione, D. J., Blume-Marcovici, A. C., Miller, H. L., Giromini, L., & Meyer, G. (2012). An inter-rater reliability study for the Rorschach Performance Assessment System. *Journal of Personality Assessment, 94*(6), 607–612. https://doi.org/10.1080/00223891.2012.684118

Viglione, D. J., de Ruiter, C., King, C. M., Meyer, G. J., Kivisto, A. J., Rubin, B. A., & Hunsley, J. (2022). Legal admissibility of the Rorschach and R-PAS: A review of research, practice, and case law. *Journal of Personality Assessment, 104*(2), 137–161. https://doi.org/10.1080/00223891.2022.2028795

Viglione, D. J., & Giromini, L. (2016). The effects of using the international versus Comprehensive System norms for children, adolescents, and adults. *Journal of Personality Assessment, 98*(4), 391–397. https://doi.org/10.1080/00223891.2015.1136313

Viglione, D. J., & Rivera, B. (2003). Assessing personality and psychopathology with projective methods. In I. B. Weiner (Ed.), *Handbook of psychology* (pp. 600–621). John Wiley & Sons.

Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of Rorschach Comprehensive System coding. *Journal of Clinical Psychology, 59*(1), 111–121. https://doi.org/10.1002/jclp.10121

Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment, 17*(3), 377–395. https://doi.org/10.1177/1073191109359587

Vitolo, E., Giromini, L., Viglione, D. J., Cauda, F., & Zennaro, A. (2021). Complexity and cognitive engagement in the Rorschach task: An fMRI study. *Journal of Personality Assessment, 103*(5), 634–644. https://doi.org/10.1080/00223891.2020.1842429

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). Pearson Assessment.

Weiner, I. B. (2014). *Principles of Rorschach interpretation* (2nd ed.). Routledge.

Weiner, I. B., Exner, J. E., Jr., & Sciara, A. (1996). Is the Rorschach welcome in the courtroom? *Journal of Personality Assessment, 67*(2), 422–424. https://doi.org/10.1207/s15327752jpa6702_15

Wood, J. M., Garb, H. N., Nezworski, M. T., Lilienfeld, S. O., & Duke, M. C. (2015). A second look at the validity of widely used Rorschach indices: Comment on Mihura, Meyer, Dumitrascu, and Bombel (2013). *Psychological Bulletin, 141*(1), 236–249. https://doi.org/10.1037/a0036005

Wood, J. M., Lilienfeld, S. O., Garb, H. N., & Nezworski, M. T. (2000). The Rorschach test in clinical diagnosis: A critical review, with a backward look at Garfield (1947). *Journal of Clinical Psychology, 56*(3), 395–434. https://doi.org/10.1002/(sici)1097-4679(200003)56:3%3c395::aid-jclp15%3e3.0.co;2-o

Wood, J. M., Nezworski, M. T., Garb, H. N., & Lilienfeld, S. O. (2001). The misperception of psychopathology: Problems with norms of the Comprehensive System for the Rorschach. *Clinical Psychology: Science & Practice, 8*(3), 350–373. https://doi.org/10.1093/clipsy.8.3.350

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7*(1), 3–10. https://doi.org/10.1111/j.1467-9280.1996.tb00658.x

Zhou, F., Zhao, Y., Huang, M., Zeng, X., Wang, B., & Gong, H. (2018). Disrupted interhemispheric functional connectivity in chronic insomnia disorder: A resting-state fMRI study. *Neuropsychiatric Disease and Treatment, 14*, 1229–1240. https://doi.org/10.2147/NDT.S162325