Contemplations and discussions on the quality of forensic assessment in sentencing

van der Wolf, Michiel; de Vries Robbé, M.

Published in:
Safeguarding the quality of forensic assessment in sentencing

DOI:
10.4324/9781351266482-2

Chapter 2

Contemplations and discussions on the quality of forensic assessment in sentencing: Puzzling pieces for decision makers

Michiel van der Wolf and Michiel de Vries Robbé

2.1 Introduction: 'state of the art'

A book on safeguards for the quality of forensic assessment in sentencing may suggest that there is something like a state of the art, the use of which is to be safeguarded. The term *state of the art*, in its metaphorical sense, is used for a most recent technique, which is therefore considered the best. As mentioned in Chapter 1, forensic assessment in this book is seen as 'all expertise – either written or oral – provided in establishing psycholegal factors', which in the context of the topic of this book should be relevant to sentencing. In essence, forensic assessment is labour of a diagnostic nature. Diagnosis literally means to discern or to distinguish, and is of course mainly associated with medical conditions. Whilst even in the context of somatic medical science, many a discussion may exist on what the state of the art is in diagnosing a certain pathology,[1] forensic assessment adds at least two layers of complexity to the diagnostic process, represented already in the term 'psycho-legal'.

The first layer is the realm of psychodiagnostics, which covers the discernment of psychopathology – both at the level of functioning and classification – as well as personality traits. Psychiatry, more than any other medical discipline, is prone to philosophical debate. Using the word *discipline* already avoids the debate whether psychiatry is even a science, or may be more suited among the humanities, with all its epistemological consequences.[2] Indeed, psychiatry and clinical psychology do not predominantly study physiological matter, but mind. Already the suggestion that our thoughts, feelings, desires, personalities, and behaviours are manifestations of the brain, as their physiological substrate, would be taking sides in another of such debates. But even those taking a neurobiological view on the discipline won't find it hard to admit that causes of many psychopathologies remain obscure, as the brain is the most complex organ in the human body, and despite all advances of the neurosciences is still largely unknown territory. Without clarity regarding origin or causality, symptomology concerning the mind is based mainly on deviance in functioning, incorporating among other things the risk of societal and normative influences. The misuse of psychiatry in this respect has a history of its own, but is in Western literature especially referred to in relation to the second layer of added complexity: the legal context.[3]

This context confronts the behavioural sciences involved in forensic assessment with a number of challenges, such as differences in language and definition, both between the disciplines but also between jurisdictions. Other differences are for example related to competence concerning the decisions involved, societal and political interests, and stakes added to – and dominant over – the interests of the individual, etcetera. This legal context also provides temporal challenges, as for sentencing the law is generally more interested in the past and the future, than it is in the present.

DOI: 10.4324/9781351266482-2

Therefore, in discussing forensic assessment, the term state of the art does not (only) refer to a most recent technique, but should be interpreted more in a literal sense: the current state of knowledge within the psycholegal disciplines, including the discussions and debates on what its quality is. Quite often in advising legal decision makers, they are left in the dark about many of these discussions, for example, because the advisor in question has already taken a side in a certain debate, or because (sense of) clarity is chosen over transparency. As this chapter aims mainly at illuminating these underlying discussions, it does not aim at providing integral reviews of the literature of the psychometric qualities of certain instruments or methods used in forensic assessment for example, but it will provide the background for understanding such reviews, as there is much more to the quality of the assessment as a whole. In doing that, it hopes to provide decision makers with essential pieces of the puzzle they have to find the best solution for.

In this chapter, first of all, the necessary backdrop to all these discussions will be set through discussing the origins of forensic assessment, types, and measures of quality and possible biases that come with the legal context. Next, discussions on the most common psycholegal concepts relevant for the quality of forensic assessment in sentencing will be described respectively (in its various definitions): mental disorder, criminal responsibility, and dangerousness. As will be explained in the upcoming paragraph, most attention related to the quality of assessment will go out to the last concept.

2.2 Background knowledge: context and quality of assessment

2.2.1 The origins of forensic assessment in (criminal) law[4]

As in the country chapters, the historical traditions in forensic assessment are addressed per country, this paragraph only addresses their common origins. The western world is often said to have a Judeo-Christian tradition, and this is particularly true for its (criminal) laws. The triangle of interrelated concepts that is still to a large extent the 'raison d'être' for forensic assessment in sentencing – mental disorder, criminal responsibility, and dangerousness – is already recognisable in Hebrew law and in (Christian) Church laws, which were actually highly influenced by the morals, myths, and laws of ancient Greece and Rome.[5] On the relation between disorder and responsibility, the Babylonian Talmud (written around 500 AD) mentions:

> *Idiots, lunatics and children below a certain age ought not to be held criminally responsible because they could not distinguish good from evil, right from wrong and were thus blameless in the eyes of God and man.*[6]

As in Hebrew law – similarly in Roman law and many medieval, both the English and Germanic Western European legal traditions – criminal acts were dealt with in a civil law manner, kinsmen of the insane offender were held liable for compensating the victim and were also held responsible for preventing future harm by the offender.[7] It underlines the ancient roots of the presumption of dangerousness, based on the (combined) stigma of offender and mentally disordered. Therefore, from its origins onwards, this triangle has always added both retrospective – criminal responsibility – and prospective complexity – dangerousness – to forensic assessment.

From then on, both developments in (criminal) law and developments in psychodiagnostics – also originating from the ancient Greek ideas of Hippocrates and Galenus which were mainly biological[8] – have shaped forensic assessment, often hand in hand as the

introduction explained that both disciplines are prone to societal and normative influences. For example, the influence of the Church made both area's inherently 'theocratic'. In Medieval times, criminal law became separated from civil law, at first because not all crimes could be compensated, and later shaped by Christian thought under the influence of concepts of sin, personal ethical blame, and guilt. In England, for example, certain crimes which were punishable – even by death – because they could not be wiped out by compensation, could at first not entirely be excused, but through Church influence later could, by absence of intention and voluntariness: 'not out of own free will'.[9] That same influence had another effect on legal insanity through turning herecy into an offence. Some mentally disordered offenders were given harsher punishment than ordinary offenders, but only because they were mistaken for persons possessed by demons, even by doctors.[10] It shows that psychodiagnostics had mostly turned into demonology in those late Middle Ages. The Dutch doctor Johannes Wier is known to be the first to separate the mentally ill from the 'possessed' in the sixteenth century, as a predecessor of French doctor Philippe Pinel who is said to have freed the mentally ill from criminal chains in the dungeons of Bicêtre in the late eighteenth century.[11] The latter event is often being referred to as the birth of forensic psychiatry.

The rise of facilities for psychiatric care went hand-in-hand with the specialisation of the medical discipline. The (lead) psychiatrists of those facilities would also start to be asked for advice by courts. This practice would increase as diagnostics became more refined, and many recognised disorders would no longer be apparent to laypersons, including legal practitioners. The advancements in, also criminological, science would in the meantime also influence criminal law theory, as biological, psychological, and social causes for crime other than rational choice were identified. A modern school no longer propagated proportionate retribution of guilt as grounds for punishment, but dangerousness based on these causes. Obviously, this led to much more demand for advice on these causes in an individual case. The classical criminal law theory, based on responsibility, had led to the origin of forensic assessment – also from the humanitarian point of view of insanity/diminished responsibility as an exclusion criterion for the death penalty.[12] The modern theory however led to its bloom in the beginning of the twentieth century, also because dangerousness became an important concept throughout sentencing.[13] As, especially, the development of the criminal justice system and the consequential tradition of forensic assessment based on these developments differ from this point in time per jurisdiction, more modern historical context will be sketched in the respective country chapters.

2.2.2 Types and measures of quality

The quality of forensic assessment is evidently multifaceted. In the outline of the country chapters in this book, a distinction is made between the quality of the expert and the quality of an individual evaluation. The quality of the justice system itself also determines in part the quality of the assessment, for example, whether there are well-defined psycholegal concepts in place as the outcome of assessments, or whether in criminal procedure requirements regarding assessment exist.[14] From the safeguards described in the following chapters, it can be inferred that quality of forensic assessment can be divided in three major facets or types: contextual quality, procedural quality, and substantive quality.

Under the heading of contextual quality, we refer to the extent to which legal and ethical requirements that are relevant within the jurisdiction have been met in the entire process of the evaluation: from the appointment to possibly testifying in court. These requirements concerning the context of forensic assessment are described in the country chapters, and the extent to which they are followed is amongst other factors dependent on the type of justice

system (explained in more detail in Chapter 10). As in an adversarial system, truth – including conclusions related to psycholegal concepts – is being sought through conflicting opinions; a lot of scrutiny is being directed to the opposing expertise, also on its contextual quality. In an inquisitorial system, where the same outcomes of forensic assessment are often being sought through consensus of opinion, scrutinising the opinion by the court – which is primary in doing the questioning – is less common and the contextual quality is often taken for granted. That is why, for example, it could occur that an omission to inform the defendant about their right to inspection and correction of the report was quite prevalent.[15] In contrast, the requirement of impartiality, which is common in both systems, may be somewhat at odds with adversarialism, considering its proneness to additional biases, as elaborated on in paragraph 2.3.

Procedural quality refers mainly to the adherence to disciplinary standards concerning the entire process of the evaluation: from the collection of data, inferences made on the basis of these data, and reporting and testifying on the conclusions based on these inferences. These disciplinary standards also safeguard the common requirement that an expert remains within the boundaries of his/her expertise. The minimisation of biases and the use of state of the art methods and tools, accepted within the discipline, are generally among such standards. With regards to having the required expertise, an English study on evaluations in juvenile cases, showed a remarkably low percentage of evaluators trained in diagnosing juveniles.[16] When it comes to reporting, again in inquisitorial systems, researchers have mentioned a lack of scrutiny, related to the soundness of argumentation for example, as a Dutch study showed.[17]

Eventually, the substantive quality of the conclusions or outcome of the evaluation is of course key, as advising that a psycholegal criterion, necessary for a certain decision in sentencing, is met or not met impacts legal decision making enormously. Again, especially research from inquisitorial systems shows very high rates of adaptation of conclusions from forensic assessment by decision makers.[18] The substantive quality of forensic assessment is often expressed in similar terms as used for psychodiagnostic tools, such as tests for personality or intelligence, through the psychometric measures of validity and reliability.

Validity relates to the question whether a test actually measures what it claims to measure. Especially for the prospective activity of assessing risk, the assessment may be verified (and researched) by an actual outcome in the future (for example a re-offense). Therefore, the quality of risk assessment is generally expressed through the measure of *predictive* validity. For other concepts, such as mental disorder and especially the retrospective activity of assessing legal insanity/criminal responsibility, no such outcome measure exists. In order to research the validity of conclusions on these matters, one has to resort to proxy-measures (elaborated on in paragraph 3.2), or 'softer' forms of validity, like *construct* validity or *face* validity. Construct validity refers to how well the test relates to underlying theoretical concepts, which given the lack of consensus about underlying theories for many psycholegal concepts is mainly researched when there is a widely accepted theory, such as the RNR principles for the assessment of forensic treatment effectiveness.[19] Face validity refers to how well the test or process (at face value) seems to appear to measure what it claims to measure, for example in the public's eye. Face validity for example plays a role in the discussion about the dichotomous versus dimensional nature of the responsibility doctrine.

The reliability of a test is related to the consistency (of the result) of a measuring test, or in other words: the extent to which a measure or process yields the same results independent from variations in other variables.[20] For example, between different points in time by the same evaluator over repeated measurements (*test-retest* reliability), or between different evaluators using the same test for the same individual (*interrater* reliability). Especially the latter is used in research related to forensic assessment, as the level of agreement between experts

may be related to (the legal concept of) arbitrariness in decision making. More broadly in forensic science, reliability has been distinguished from 'biasability', with the first concept referring to the consistency of expert performance based on relevant information without bias, and the latter referring to decision making that is affected by irrelevant contextual information (which will be elaborated on in the next paragraph). The overall variability is then a function of both reliability and biasability.[21]

Highly relevant for forensic assessment is the finding that the levels of reliability and validity impact each other, especially in individual case decision making. If for example the interrater reliability of a method of risk assessment is low, the predictive validity of a single evaluation is also likely to be impaired.[22] And vice versa, if a concept has a low validity, for example as it is not that clearly defined, it will result in a lower interrater reliability. This latter effect may explain findings on the level of agreement between experts in assessing legal insanity.[23] Both examples will be elaborated on in following paragraphs.

As, of course, it is most important to know how well a test performs in an actual situation of decision making in the 'field', the concepts of *field* validity and field reliability are used to distinguish them from their counterparts based on research in the 'lab'. Studies show that field validity and reliability of tests, or tools, used in forensic assessment tend to be lower than their lab-counterparts, due to methodological issues.[24] It is, therefore, relevant to inquire into the type of research when evaluators report on reliability rates for tools they have used. However, as for specific tools (especially for risk assessment) such rates are generally available, this is much less the case for the more idiographic processes used in forensic assessment, which are less easily, and thus less frequently, researched.[25] Inferences resulting in individualised diagnoses, levels of dangerousness based on an 'offense analysis', or (levels of) criminal responsibility are examples of such processes. Nevertheless, it should be possible for an evaluator to comment at least qualitatively on the reliability of these processes (also based on literature), as is often required for the contextual quality of the report or testimony, even though – again especially in inquisitorial systems – this is often omitted.[26]

'Forensic psychiatry and forensic psychology are referred to as "soft sciences" for which satisfying levels of reliability and validity of findings are suspect'.[27] Nevertheless: 'The judicial system finds psychiatry and psychology, despite their limitations, to be relevant and useful, even indispensable, in a variety of legal issues for which an individual's mental functioning is relevant'.[28] These two quotes – both from the same source, written by Felthous – summarise the scrutinised yet firm position of forensic assessment. They also explain why in (both regular and) forensic psychodiagnostic research often the term 'utility' or 'usefulness' is distinguished from, and preferred, over the term validity.[29]

2.2.3 Legal context: additional biases

In any process of decision making, biases can come into play. Legal decision making, especially within criminal justice, has itself a bad reputation in that respect.[30] And as drawing conclusions based on the collection of information in forensic assessment is a process comparable to decision making, it is no exception. Already within psychodiagnostics outside the legal context, biases, and cognitive distortions have been identified as either: stemming from the structure of the human mental apparatus, expectancy-based, and stemming from learning and experience, or stemming from situational and systemic factors that distort information processing and cause errors in decision making.[31] The legal context of forensic assessment adds 'situational and systemic factors' in a number of ways. Zapf and Dror, for example, mention case-specific factors, such as irrelevant case information, reference materials, and case evidence.[32] In a study into the beliefs about bias among evaluators, these authors (and

colleagues) describe how most evaluators expressed concern over cognitive bias but held an incorrect view on how to mitigate bias (through 'mere willpower'). In addition, they found evidence for a 'bias blind spot', meaning that more evaluators acknowledge bias in their peers' judgments than in their own.[33]

Even though indeed factors influencing the evaluators themselves are many, it is fair to mention that also factors stemming from the defendant, or from the tools evaluators use, add to the variability of outcomes of forensic assessments. Factors stemming from the defendant, which either unconsciously or consciously lead to distorted information, include cognitive distortions, transference, social desirability, and relevant in the legal context simulation, aggravation, or malingering as the stakes can be high.[34] Tools, such as the *DSM* classification system or for structured risk assessment, have also been criticised as being biased, for example racially.[35]

Related to the evaluator, several relevant biases should be mentioned in this context, as mentioned stemming from our nature, nurture, or the context. Bias is already ingrained in human nature, as our cognitive system is such that it processes information in an efficient and effective way. However, the shortcut mechanisms used to do so result in vulnerability to bias and error. Stereotyping is such a mechanism, which in forensic assessment has been demonstrated, for example, for gender, leading to different conclusions for men and women on certain psycholegal factors, depending also on diagnosis.[36] Other well-known mechanisms are primacy – and recency bias – placing too much weight on the first or last information perceived, availability bias – overestimating the probability of an event when other instances of that event or occurrence are easily recalled – and confirmation bias – tunnelling to conclusions that are in accordance with what we believe. In systems with a one-phase trial in which trial and sentencing are combined, confirming the believe that someone is guilty, will impact the evaluation of a defendant who denies the charges.[37]

What we believe may also be the result of nurture, evoking motivational biases or pre-existing attitudes – for example, a firm stance in debates on sentencing related issues such as capital punishment or preventive detention.[38] A very important bias in this context, related to association or affiliation bias, is what is called adversarial allegiance. There is ample evidence that within adversarial justice systems, outcomes of forensic evaluations may be dependent on which party has retained them.[39] Other well-known errors in drawing inferences within forensic assessment include circular reasoning – for example, establishing the mental disorder based on the offense – and base rate expectations. They may be related to case specifics, such as experiences in the interaction with the defendant – like countertransference, or other forms of affective bias.[40] But they could also be due to the evaluator's more general (professional) experiences and knowledge, which may work against him or her if recent insights from literature correcting earlier findings are not processed, or 'rater drift' occurs – the unintentional redefining of criteria, like legal tests which are not consulted for every evaluation because the evaluator considers them internalised.[41]

Decision makers within sentencing have to be aware of the possible impact of bias – either or not due to the legal context – in forensic assessment, as they should also be aware of bias in their own decision making. If suspicious of bias in an evaluation, they could ask for strategies used to mitigate the impact of bias, which have been developed within (forensic) behavioural sciences. Of course, the testing of alternative (or opposite) hypotheses is a well-documented one and the use of more nomothetic scientific knowledge, for example, structured tools. Similarly, training (about bias) and peer review of evaluations will also mitigate susceptibility for bias.[42] But, even with all the de-biasing strategies in the book, error can never be ruled out completely. Not only through the possible unawareness of one's own biases, but also because the state of the art in assessing psycholegal concepts will continue to allow it.

2.3 The assessment of mental disorder

2.3.1 Definition, diagnosis, and classification

'Assessing' a mental disorder, in some definition or another, may be relevant in sentencing, either as a criterion for a certain sanction, disposition, or transfer – sometimes in combination with (a degree of) legal insanity or culpability or need for treatment – or as a factor in determining the height of a sentence. In the latter sense, the concept may be used in two opposite directions: as a mitigating factor related to the concept of culpability or an aggravating factor related to the concept of dangerousness. As, on a group level, there is no evidence that mental disorder is related to incompetence or dangerousness, these consequences within sentencing have often been exposed as stigmatising.[43] However, at the level of specific disorders or through an idiographic approach that relation may well be established in an individual case.[44] The subject of stigma resonates however in the vast body of literature that understands (or denies) mental disorder as being a social construct, stretching from the heydays of antipsychiatry onwards.[45] As not all discussions on the concept of mental disorder are relevant in the context of this book, we focus on a few which may be relevant for the quality of assessment in the forensic context.

Among those who do acknowledge mental disorder as a reality, many different concepts or definitions exist. Already psychology and psychiatry have different ways of looking at the concept, based on their differences in methodology, as more of a deviance from the 'normal' in the bell curve or as an illness in the dichotomy opposing healthy. Traditionally, this can also be explained through the existence of different schools of thought, which all had definitions of mental disorder in line with their theory on behaviour in general. For example, psychodynamic, phenomenological, behavioural, and neurobiological views have all impacted the development of forensic assessment, to a point that nowadays generally a more eclectic biopsychosocial model remains. The fact that there has always been so much discussion on the origins of mental disorder, is one of the underlying factors for the a-theoretical classification system, such as the *DSM*.[46] A possible reason why, especially in forensic assessment, the influences of traditional schools of thought have resonated longer than in general psychodiagnostics may be that, especially for explaining offending behaviour, a mere classification does not provide as much help as an underlying theory.[47]

Another underlying factor for classification of disorders, either or not in a system such as the *DSM*, is traditionally the need for common language. Dating back to the days of Kraepelin, this development based on agreement at the level of description has paved the way for more universal and more nomothetic research.[48] For obvious reasons, however, the endeavour of classification has been criticised as being reductionistic and empiristic,[49] as well as rendering the psychiatric nomenclature with an appearance of validity. Indeed, the boundaries between different 'labels' as well as with 'normality' may be described as 'fuzzy' – exposed for example by the term 'comorbidity' and classifications ending with 'Not otherwise specified' – and the endeavour of diagnosing mental disorder as an 'epistemological uncertainty'.[50] Consequentially, for both clinical and legal decision making, a much more elaborate and individualised description of someone's functioning is needed as diagnosis.

When a mental disorder is itself defined in law, it is generally in a broad sense, rendering it less important what the exact (*DSM*-)classification is. However, it very much depends on the definition and whether there are categories mentioned, or explanatory notes issued on the scope of the definition, whether a certain classification can fall under the criterion.[51] In general, when it is defined in law, it also becomes a legal concept, meaning that it is in the end up to the legal decision maker to establish the concept, often on the advice of a

behavioural expert. Nevertheless, when the definition itself is completely in line with dominant psychiatric terminology, this distinction in competencies becomes more artificial or even problematic.[52] Therefore, in some jurisdictions, to underline this division in competence – and to be independent of the volatile trends in psychiatric lingo – the legal concepts are distinctively not defined in language used within the psychiatric discipline.

As studies into the psychometric quality of diagnoses are generally narrowed down to *(DSM)* categories or labels, and its accessory (semi)structured tools or interviews, the nuances mentioned above might deny the relevance of these findings for forensic assessment. However, not only is classification part of the state of the art of diagnosing psychopathology, it also predetermines the outcome of any following assessment based on the disorder. For example, jurisdiction-specific regulations may disallow certain diagnoses, like substance intoxication or antisocial personality disorder, as the basis for legal insanity. Moreover, also scientific evidence – like the relationship between psychopathy and risk – or practical experience – for example, that legal insanity is mostly based on psychotic disorders – can have such a presorting effect. Therefore, the consequences of misdiagnosis in forensic assessment may be considerable.[53]

2.3.2 Validity, utility, and reliability

As mentioned, in determining the validity of a psychopathological classification, by lack of a certain outcome measure or Delphi oracle, researchers have to resort to proxies. Even though the labels are a-theoretical sometimes such proxies, for example, for psychotic disorders, are biological markers, or the effects of an established (pharmaco)therapy. However, personality disorders, for example, do not have well-established biological markers and do not evidence predictable responses to treatment.[54] There is a vast body of literature on the validity of antisocial personality disorder, often critical because of the overlap with offending behaviour. In recent years, the validity of this classification has been researched as predictive validity for institutional misconduct, with quite opposite results.[55] Of course, validity rates differ per diagnostic categories, but as they are often based on different proxies it is hard to compare such results. When an overview of the clinical utility of the *DSM* classification system is given, the verdict is also based on the intended argument. When, for example, protection is intended against critics 'who use its weaknesses to argue for the complete abolition of psychiatric diagnosis', the clinical utility is portrayed as great.[56] When, for example, other, more modern dimensional methods of diagnostic classification are being propagated, the clinical utility is called overestimated.[57] Forensic utility, as mentioned, may depend somewhat on the type of justice system, the type of decision, and the type of follow-up question. No wonder the DSM has a disclaimer in place that a classification in itself should not have any legal consequences. In general, however, in relation to other psycholegal concepts, mental disorder is one for which legal decision makers tend to really rely on the competence of behavioural experts.

As explained earlier, the limited validity of a concept also impacts the reliability, and the other way around. And researching reliability of diagnostic (categories of) classification has its own methodological obstacles – for example, the unethicality to keep evaluating an individual in person so that audiovisual registration is used – which have been pointed at to explain disappointing initial results of the reliability of DSM-5 classifications.[58] Considering that the forensic context may impact the diagnostic process (see paragraph 2.3), research in that field is most relevant in this respect. In the adversarial realm, however, studies are generally affected by adversarial allegiance and measure 'biasability' rather than reliability.[59] There is one jurisdiction, however, as explained in the American chapter, which provides for

a naturalistic study into the agreement between nonpartisan evaluators. Hawaiian regulations state that the court will appoint three evaluators to a felony case in which forensic mental health assessment is called for. A first-of-a-kind analysis of 240 of such cases on six diagnostic categories showed perfect agreement between the three evaluators in fewer than one of five cases. There was also a difference between the diagnostic categories, with agreement on psychotic disorders being about 72%, agreement on cognitive disorders being highest (90%), and on personality disorders lowest (62%). As next to cognitive disorders, psychotic disorders (72%), substance-related disorders (65%), and mood disorders (65%) are most likely to impact a decision on legal insanity,[60] while for competency also intellectual disorders (95%) are relevant,[61] the authors conclude that 'in terms of field reliability, this means that evaluators reach a consensus on the most pertinent diagnostic categories for pretrial evaluations in fewer than half of all pretrial cases. This low level of agreement is likely to have serious implications for the psycholegal opinions made by the evaluators, and, in turn, the ultimate judicial dispositions made by the court'.[62] In short, evaluators are more likely to disagree than agree on a defendant's total diagnostic picture in pretrial forensic mental health evaluations.

In inquisitorial justices systems, such a naturalistic design is even harder to be found, as it is customary for multiple evaluators to try and reach a consensus before reporting, while initial dissenting opinions are generally not reported. Field reliability has therefore never been researched. A recent vignette-study from the Netherlands in which three actual reports of typical cases, stripped to the level of symptomatology, were presented to 52 evaluators, showed also a very high level of agreement on a case of a schizophrenic suspect of man-slaughter, and much lower levels of agreement on personality disorders in addition with paraphilia or substance abuse, respectively, in cases of a grooming sex-offence and a robbery.[63] These differences in diagnoses had quite an impact in the assessed level of criminal responsibility and the sanctioning advise.

2.4 The assessment of insanity/criminal responsibility

2.4.1 Concept, criteria, and divisibility[64]

In general, as a first step in assessing legal insanity or (diminished) criminal responsibility, some definition of mental disorder is required. So, while all the contemplations of paragraph 3 similarly apply, the next steps add even more complexity. The first additional step is that the mental disorder has had to be present during the time of the offense. Retrospective diagnostics are alien to the regular clinical context, and few tools exist to assist this activity of reconstructing the offender's state of mind during the offense, other than logical inferences, for example, about the chronicity of the disorder diagnosed in the present in combination with information about its onset before the committed act. A second additional step common in provisions of the responsibility doctrine is a specification of the (functional) capacities that the disorder should have impaired at the time of the offence in order to establish legal insanity, often called a 'test'. In provisions in Western nations, it is common to find – either or both – a test of cognition and volition or control. On the other hand, in a few jurisdictions, provisions exist that require a more general causal relationship between the offence and (the product of) the disorder – sometimes called the 'product test'. On the basis of obvious criticisms related to determinism and a demand for restoration, some jurisdictions limit the use of the doctrine, re-label it, or have abolished it altogether. Already in this book all these different models are represented, suggesting that much more so than the concept of mental disorder it is highly culture-specific.[65] It has been argued that the precariousness of the doctrine and its connection to central aspects of criminal law seem to justify that a

national support base is needed.[66] As this suggests a highly normative activity, it is generally accepted that it falls under the competence of the legal decision maker.[67] In several jurisdictions, there are (contextual quality) requirements, which enforce the delineation of the epistemological activity of the evaluator from the normative activity of the legal decision maker, for example, that no advice is given on the ultimate issue.[68]

Apart from the contemplations mentioned earlier, there are more elements surrounding this doctrine that hinder any universal endeavor to grasp the concept empirically: especially its embedding within criminal procedure and (related) discussions on its divisibility. In adversarial jurisdictions, legal insanity is a defense discussed at the trial of fact. Understandably, because of its consequence, it is a dichotomous – all or none – concept. In most inquisitorial jurisdictions of the European continent, the concept is viewed as an excuse, related to the level of culpability and punishability of the offender in relation to the offence. In these jurisdictions, the concept is considered a gradual concept, which could lead to a degree of mitigation of punishment up until a prohibition of punishment in case of a total lack of criminal responsibility. This distinction between adversarial and inquisitorial justice is actually more nuanced, as in adversarial jurisdictions doctrines of diminished responsibility may exist to mitigate the sentence in case of murder,[69] while at the sentencing stage culpability, viewed as a more dimensional concept, may well be mitigated due to mental disorder.[70]

Even though it has been suggested that dichotomous concepts are 'peculiarly foreign' to psychiatry, it is understood that the dichotomy used for the insanity defense is also being preserved to avoid more influence of psychiatrists in legal decision making.[71] The gradual or dimensional approach to responsibility may indeed have more 'face validity', but automatically adopts problems in the reliability of assessment. Indeed, in the Netherlands, there is a lively discussion on how many gradations can scientifically be distinguished.[72]

2.4.2 Utility and reliability

Because of the aforementioned aspects, a universal body of evidence, as there is for risk assessment, will never exist for the assessment of criminal responsibility. Moreover, as no sensible proxy is available for measuring this concept described as 'a legal fiction of a medical fiction',[73] validation will always be impaired. In addition, both legal standards and the psychiatric state of the art are constantly developing. When the need for evidence-based insanity evaluations is described, intended are 'a standard procedural approach, accuracy of diagnosis, and quality monitoring'.[74] With regards to the standardised approach – as the latter two have been discussed – in some countries, tools are in place. For example, guidelines from the American Academy of Psychiatry and the Law (AAPL) in which next to procedures also some substantive guidance is offered, for example, on how to relevantly assess impaired volition.[75] Some psychometric tools exist, like the Rogers Criminal Responsibility Assessment Scale (R-CRAS) from the United States,[76] and the Rating Scale of Criminal Responsibility for mentally disordered offenders (RSCRs) from China.[77] And most recent in Brazil, the 'criminal responsibility scale' has been constructed.[78] Validation of the tool is then being achieved through comparing the outcome of the tool with an expert's opinion, or through construct validity based on major components of existing evaluations.[79] In that sense, the structured method can never be more valid than the expert's opinion, while the question remains how valid that is. Since it is in the end a legal decision, one way of testing the expert's opinion to an external outcome is through the agreement with the legal decision maker. Since in adversarial systems, there are generally multiple, often different, opinions expressed in one case, it is less feasible to research such agreement affected by adversarial allegiance. However, in the Hawaiian system of three court-appointed experts, judges

followed the majority opinion among evaluators in 91% of cases.[80] While in inquisitorial systems, generally one opinion is given – in case of multiple evaluators, based on consensus – research shows a similarly high percentage of following the expert's opinion by the decision maker.[81] In other words, the utility or usefulness of the (impartial) expert's opinion for legal decision making appears to be great, even though there may be some concern about its reliability.

A structured approach is often mentioned as beneficial for the reliability of the assessment,[82] and in doing that enhance the validity of assessments in a single case. However, the mentioned tools are not used in most parts of the world, and not even consistently in the countries they were created for. Reliability rates for such tools are therefore not indicative for field reliability of actual evaluations. For that, similarly designed (or the same) studies as those discussed in paragraph 3.2 are most notable. In 165, again, Hawaiian cases, three evaluators reached unanimous agreement regarding legal sanity in only 55.1% of cases. Agreement was higher when they agreed about diagnosing a psychotic disorder, and lower when a defendant was under the influence of drugs or alcohol at the time of the offense. The authors conclude that 'reliability among practicing forensic evaluators addressing legal sanity may be poorer than the field has tended to assume. Although agreement appears more likely in some cases than others, the frequent disagreements suggest a need for improved training and practice'.[83] That last remark could be tested a year later, when the Hawaiian state adopted more stringent certification standards, of which a rigorous training was part. The overall field reliability increased by 17%.[84] A study from 2015 showed an overall agreement of 63%, labelled as 'fair' or 'moderate', which was however much less than the agreement for competency to stand trial in the same study. This was explained through the retrospective nature of insanity evaluations, which makes it more complex and inferential in comparison to competency assessment. The level of agreement was said to be comparable to complex decision making in (somatic) medicine. 'As task complexity increases, "individuals may use heuristic-based strategies, with associated increases in effort, confusion, error rate, and consequent reduction in performance"'.[85]

When more gradations of criminal responsibility are acknowledged, and there are more (three) potential outcomes, logically agreement would be lower. In a Polish study on field reliability, which is a possibility because in Poland courts may ask for more than one evaluation, however, 57% agreement was reached. When the court asked for a second report knowing the outcome of the first, the conclusion was different in 47% of the cases.[86] This result may suggest that courts are able to identify poor evaluations, or that something exists which may be called 'inquisitorial allegiance': handing the court another conclusion when it is unsatisfied with the first. In the mentioned Dutch vignette study, agreement on the graded concept of criminal responsibility was not really related to agreement on the consequential sanctioning advice. Agreement on this advice was highest in the case of the schizophrenic defendant, even though there was more disagreement on criminal responsibility, with about two out of three of the evaluators drawing the conclusion of non-responsibility and one out of three that of diminished responsibility. In the case of the sex offender and the case of the robber, about four out of five evaluators assessed the defendant to be diminished responsible, and one out of five opted for fully responsible, while there was much more disagreement on the sanctioning advice – ranging from no treatment (in prison) to a severe safety- and treatment-order.[87] Arguably, given the enormous consequences of the dichotomous insanity decision, the sanctioning advice in inquisitorial justice is a better comparison than the graded responsibility assessment, yielding more disappointing results.

2.5 The assessment of risk

2.5.1 Purpose

Risk assessment serves different goals. In a legal sense, it is used predominantly to assess dangerousness. Although, historically, risk assessment was used mainly to predict future re-offending, in recent years, the focus has shifted to the prevention of new offenses through tailored interventions and risk management.[88] Assessment is merely the collection of relevant information that provides insight in the dangerousness of the individual case. Structured assessment (in contrast to unstructured assessment; see paragraph 2.5.4) can be seen as the coat rack to gather and organise this information. Each bit of information is regarded as a piece of the individual's risk puzzle. The task of the assessor is to collect the relevant pieces of information and combine these into a meaningful conclusion regarding the individual's risk. The more reliable the information, the better the quality of the assessment. When relevant information is missing, this should be highlighted in order for decision makers to be able to interpret the findings accordingly and, where necessary, request for additional information.

Risk assessment is often informed by the Risk-Needs-Responsivity model.[89] This model states that (forensic) interventions should: (1) be intensified if risks are more present; (2) focus on those factors most relevant for the individual case – the criminogenic needs; and (3) be offered in a manner that matches the responsivity or learning style of the individual. The complementing theoretical Good Lives Model[90] of rehabilitation states that an individual should work towards positive personal goals. Comprehensive risk assessment for an individual aims to provide insight into each of these aspects in order to be able to draw final conclusions regarding the overall level of risk and inform risk management and intervention decisions. It should be noted that risk assessment is a complex and time-consuming task, which requires extensive training and forensic expertise. As the assessor aims to 'foresee' the future based on the collected information and attempts to formulate a best judgement regarding future behaviour of the assessed individual in the anticipated context, almost by definition risk assessment is an extremely difficult undertaking. In the following paragraphs, several specifically complicating issues are being discussed.

2.5.2 Risk of what type of behaviour is being assessed?

When carrying out or interpreting risk assessment, it should be carefully considered what type of risk is being assessed. Dangerousness regarding what type of undesirable behaviour? Often, the most serious types of offending come to mind when risk is being assessed, such as bodily harm or sexual abuse. However, other types of violence towards others, such as domestic abuse, stalking, fire setting, or verbal threats, are generally also included in the definition of violence risk assessment.[91] While risk assessment measures often differentiate between physical and sexual violence, there are in fact specific assessment measures for a wide range of undesirable behaviours, such as intimate partner violence, stalking, extremist violence, honour-based violence, and so on. Although violent in nature, self-harm and aggression against objects are generally not regarded as 'violence'. Other risk assessment measures consider dangerousness much more broadly and include all types of criminal offending in their definition (e.g. LS/CMI[92]).

Although measures that focus on specific types of violence generally show somewhat more accurate predictive validities,[93] there is no right or wrong in the scope of an assessment measure. However, for both the assessor who formulates conclusions regarding risk and the

decision maker who incorporates the assessment results in his judgement regarding dangerousness, it is of vital importance to clearly define the type of risk that is being evaluated, as the results of an assessment may be altogether different if an alternative definition of risk is employed. In addition, recidivism base-rates vary greatly between different types of offending behaviour. For example, recidivism in sexual violence is relatively rare, compared to recidivism in general criminal behaviour. Knowledge of base-rates for specific offending behaviours in different populations would provide useful background information for decision makers. Surprisingly, this type of information in the form of a base-rate overview is often not easily available. A complicating factor in this regard is the fact that recidivism may go unnoticed and thus official recidivism rates only remain a proxy for actual new offenses that have been committed.

2.5.3 Single versus team-based assessment

As the assessment concerns the collection, weighing, and integrating of relevant information regarding the individual, a risk assessment is as good as the information that is being regarded. Therefore, it is important for the assessor to make use of different sources of information, such as the individual's criminal and psychiatric records, behaviour observations, collateral information from family or friends, and self-reported reflections from interviews with the individual. However, even when multiple sources of information are used and assessors are experienced, they remain susceptible to blind spots, tunnel vision, the (dis)likability of the assessed, dishonest testimonials, or one-sided observations. In order to avoid these, unwanted biases risk assessments are sometimes carried out by multiple people. These team-based assessments help to bring information to the table from different angles, consider this more objectively, and come to a consensus rating regarding the case. Moreover, these discussions often serve as a valuable starting point for risk-management and treatment. Although time consuming and expensive, risk assessment carried out by multiple assessors generally produces more valuable and objective results.[94]

2.5.4 Clinical versus structured risk assessment

In day-to-day life, people carry out personal risk assessment all day long as minimising risk biologically increases the likelihood of survival. Similarly, psychiatrists and psychologists carry out mini-assessments regarding an individual's risk many times a day based on their experience and expertise. These implicit evaluations of risk are considered clinical or unstructured assessment. Although individualised and useful to avoid harmful behaviour in daily interactions, research has shown that unstructured risk assessment has fairly poor predictive validity when it comes to estimating an individual's future violence risk, due to the aforementioned biases that may occur. In the past decades, the science of risk assessment has advanced into structured risk assessment, which provides the assessor with group-level evidence-based guidelines regarding the specific topics to include in an individual assessment and offers clear instructions on how these topics should be evaluated. Validated structured risk assessment instruments have proven to substantially increase the predictive validity of a risk assessment over unstructured clinical judgement.[95] Perhaps somewhat in between these two approaches lies the structured offense analysis, which follows clear guidelines on how to collect information regarding an individual's specific offense and the circumstances that preceded the offense. This offense analysis concerns a structured yet personalised approach.

2.5.5 Actuarial versus structured professional judgement

Structured guidelines for risk assessment exist in various forms. Roughly two main categories of tools can be divided: actuarial measures and those following a structured professional judgement (SPJ) approach. Both kinds of tools include a list of factors that have empirically been shown to be related to an increased likelihood of future offending. The difference between the two approaches concerns the way conclusions are drawn from these empirically based factors.

2.5.5.1 Actuarial approach

In the actuarial approach, the different factors assessed receive a numerical rating (e.g. VRAG[96]). At the end of the assessment, the scores for each factor are tallied-up to come to an overall score on the measure. In more advanced actuarial tools, total scores are then compared to those of reference groups, in order for the assessor to be able to conclude whether the individual falls into a predetermined category of individuals with an increased likelihood of harmful behaviour. This way of actuarially adding up scores and comparison to other similar cases has the advantage that it is straight forward and less susceptible to rater bias and insightful in terms of caseload prioritisation. However, mechanically adding up scores leaves less room for an individualised view, as each concept receives an equal weight in the overall total score. Moreover, comparing to reference groups is only really useful if the individual is sufficiently similar in characteristics to the other individuals in the reference group (e.g. in terms of offending behaviour, psychopathology, gender, age, cultural background, setting, and country), which requires extensive databases of individual ratings, that are often not available in such detail. When interpreting the results from actuarial tools, decision makers should carefully consider whether the reference group that is being applied is indeed sufficiently similar to the assessed individual to warrant this kind of comparison and thus the validity of conclusions drawn from the assessment.

A final concern with this approach is the relative insensitivity to the context for which an assessment is carried out (see paragraph 2.5.8). The latest generation of actuarial tools (e.g. the Static-99R[97] and Stable-2007[98]) offers a more individualised view as the factors assessed are in themselves well-developed mini-judgements regarding specific concepts. However, the reference group issue and context insensitivity remain. Some tools even go as far as to conclude that based on the actuarial rating an individual belongs to a specific subgroup that, based on previous research regarding the applied reference group, has a specific likelihood of reoffending within a specific timeframe (i.e. 30% of this subgroup recidivates with a sexual offense within two years after discharge). This type of conclusion is quite prone to incorrect interpretation by decision makers and should be used with great caution, or better yet be avoided, as it creates an unjustified sense of certainty of the likelihood of future harmful behaviour.

2.5.5.2 Structured professional judgement approach

To overcome the overreliance on evidence gathered from previous research regarding specific groups of individuals, which may not be directly transferable to other individuals, and in an attempt to facilitate more individualised risk assessment, a new method was found in which the approach of assessing structured evidence-based factors is combined with the professional expertise of the assessor: SPJ. Through interpreting, weighing, and integrating

the findings in the structured assessment, the assessor evaluates the individual case and comes to a well-informed final risk judgement regarding the likelihood of re-offending.

There have been many SPJ tools developed for a wide range of different outcomes (e.g. HCR-20^{V399} for violent recidivism, SARA[100] for intimate partner violence, and SAVRY[101] for juvenile offending). Each of these relies on the assumption that a well-trained mental health care professional has the ability to formulate final judgements regarding risk based on carefully evaluating the presence or absence of the factors assessed in the structured checklist. In addition, it is possible for the assessor to add case-specific factors that are not present in the general list of factors. The careful consideration of the meaning and impact of each factor for the specific individual allows for a highly individualised assessment regarding the likelihood of future offending. However, this too has its pitfalls. The possibility for a rater to interpret factors based on his own professional insight or experience brings room for bias in terms of possible overreliance on the presence of specific factors and risk of subjectivity (e.g. a well-behaved assessee is not necessarily low risk).

Given that also in validation studies regarding SPJ measures, the mechanical adding of scores generally predicts future recidivism quite well at group level,[102] the assessor is advised not to stray too far from the overall observed ratings on the factors when arriving at the final conclusion. In order to prevent the actual addition of scores, some tools have moved to descriptive ratings only (e.g. HCR-20^{V3}). Other tools have included the option to highlight critical factors that appear of particular importance to the individual (e.g. START[103] and SAPROF[104]). Regardless, if the assessor does come to a very different conclusion than would be expected from the overall ratings on the factors, it should be carefully explained why this 'clinical override' is justified. It may, for example, be the case that one specific risk factor severely impacts the chance of recidivism (e.g. specific delusions), or that specific protective factors strongly reduce the likelihood of offending (e.g. a physical handicap or support that is in place). The flexibility of the SPJ approach also allows for the evaluator to take into account the influence of context on risk, which is a vital consideration (see paragraph 2.5.8).

Despite the seeming advantages of the SPJ approach, conclusions draw from this approach are unfortunately also prone to incorrect interpretation by decision makers. Many SPJ tools conclude with a final risk judgement regarding future undesirable behaviour (e.g. violence) in terms of 'low-moderate-high'. However, this final conclusion summed up in one word often leaves decision makers puzzled (e.g. how should one interpret 'moderate' risk? See paragraph 2.5.10).

2.5.6 Static factors versus dynamic factors

Many risk assessment tools include static or historical factors. These factors describe the individual's past behaviour or experiences. They are important from a diagnostic viewpoint as figuring out the historical puzzle pieces offers insight into an individual's route to offending behaviour (i.e. risk formulation[105]) and vulnerabilities that should be taken into account in risk management. Historical factors also generally predict quite well, past behaviour provides a fairly good indicator for future behaviour. However, from a psycholegal context, the sole reliance on historical information provides a one-sided view that does not allow for change and offers little optimism for rehabilitation.

Luckily, people can and do change, also those severely impacted by past unfavourable experiences. Therefore, in order to be able to evaluate changes in attitudes and behaviour as well as in contextual factors (e.g. situational and social influences) over time, most risk assessment tools also included dynamic or changeable factors. These factors often provide a more up-to-date picture of the individuals functioning and risks. Dynamic factors can either

consider current or recent functioning, or can concern expected functioning in the near future. Many risk assessment measures compose a combination of historical factors and dynamic factors, to allow for a well-rounded view of the individual that offers room for change over time. Combining the historical findings with dynamic information may either be done through a predefined algorithm in an actuarial way or through the professional insight of the evaluator in an SPJ manner.

2.5.7 Risk factors versus protective factors

Risk assessment measures, even comprehensive ones, have historically been focused predominantly on risk factors. Given the psycholegal context perhaps, this is not surprising as assessors and decision makers are aiming to find out what contributes to dangerousness and investigating deficits seems the most obvious. However, in recent years, clinicians, evaluators, and decision makers have become more aware of the fact that dangerousness may not solely be determined by the presence of risk factors, but also by the absence of strengths or protective factors. In fact, since the early 2000s, understanding has grown regarding the importance of gaining a well-rounded view of the individual as a one-side risk-focused approach may inherently be inaccurate.[106] Scholars are increasingly in agreement that the presence of protective factors is indeed separate from the absence of risk factors and that protective factors should explicitly be evaluated to be able to formulate a clear picture of the individual.[107]

These missing pieces of the risk puzzle have long been ignored or underestimated. One of the first widely used structured risk assessment instruments to incorporate the notion of protective factors, at least to a limited degree, was the SAVRY, an SPJ tool for assessing violence risk for juveniles. Some years later, tools were developed that explicitly incorporate a two-sided view (e.g. START) or even specifically focus on protective factors, in order to complement risk-focus assessment tools (e.g. SAPROF; SAPROF-YV[108]). When interpreting risk assessment results, it should be noted that risk factors and protective factors each provide separate pieces of the risk assessment puzzle, which together provide greater insight into an individual's attitude and behaviour, as well as the supportive elements of their environment. Comprehensive risk assessment that incorporates both risk and protective factors is inherently more accurate and provides more in-depth conclusions regarding dangerousness as well as guidelines for risk management and intervention.[109]

2.5.8 The importance of context

Perhaps the most important protective factor to carefully consider in any risk assessment is context. The protection from situational strengths and limitations is vital to incorporate when evaluating the likelihood of recidivism. For example, an individual who has committed sexual offenses against children in the past who may still have a significant number of risk factors present, nevertheless generally has a 'low' risk of committing new sexual offenses against children while incarcerated or hospitalised, simply because there are no potential victims present. Similarly, an individual with a history of severe intimate partner abuse under the influence of alcohol, might be considered 'low' risk while granted supervised leaves from a forensic hospital, but at the same time be considered 'high' risk for the context of unsupervised leaves to the home environment. These examples highlight the vital importance of considering situation or environmental protection, which may result from legal monitoring, clinical supervision, or social support.

For this exact reason, in many settings, risk assessments are carried out for several contexts at the same time. Especially, dynamic factors that concern an estimation of behaviour in the near future and the final conclusions of a risk assessment regarding future risk are suitable for double rating for multiple contexts simultaneously. For example, if an individual is currently incarcerated, the future-rated factors of the HCR-20^{V3} or the protective factors of the SAPROF can be rated for the in-patient context, but at the same time receive a second set of ratings for the hypothetical context 'what if this individual was released tomorrow'. This comparison within the risk assessment between two different (hypothetical) contexts often provides decision makers with a great deal of insight into the likelihood of recidivism in case certain legal restrictions are dropped or imposed. This way, it can assist in evaluating the necessity of prolonged imposed treatment or probation supervision, or it can provide insight into the expected feasibility of specific interventions or risk management strategies. In forensic clinical practice, sometimes risk assessments are even carried out for three or more different contexts simultaneously, to support decision making regarding the most optimal next step in treatment and supervision. Another area where multiple ratings might be valuable is for pre-trial risk assessments (see paragraph 2.5.9).

2.5.9 Front-end and back-end assessments

Risk assessment is used both at the front end of a forensic trajectory (i.e. pre-court assessments) and at the back end (e.g. assessments preceding leave or discharge from a forensic hospital). The application at the back-end stage is relatively straight forward. First of all, there is more information available at the back end, as the individual has generally been in supervision or treatment for some time and hospital records describe all sorts of observations regarding the individual's behaviour. Secondly, the context for which the assessment is being carried out is generally quite clear and well defined (e.g. unsupervised daytime leaves from the hospital). The better the information and the clearer the context for which the assessment is carried out, the easier and more reliable the assessment. In front-end risk assessments, however, much is often unknown. The information regarding the individual case may be limited, due to incomplete file-information and limited ability for the assessor to speak with the individual and his social network and observe attitudes and behaviours. This may lead to information gaps or one-sided input. An even bigger challenge in pre-court risk assessments concerns the context for which the assessment is carried out. As often, the outcome of the legal decision making is yet unclear to the assessor at the time of the assessment (and sentencing might even be influenced by this assessment), it is complicated for the pre-court assessor to determine the context for which to carry out the assessment. In such a situation, it is often helpful to perform the assessment for different contexts simultaneously (see paragraph 2.5.8), in order to be able to draw conclusions regarding the impact of each of the assessed contexts on the (reduced) likelihood of recidivism. This may help decision makers to oversee the effects of different sentencing decisions and contemplate on the necessity of imposed interventions and/or supervision.

2.5.10 Risk communication and scenarios

Findings from an assessment are often described in a risk assessment report. This report provides an informative narrative for other professionals and decision makers. It is advised to avoid the use of numbers in these assessment reports and instead describe observations and findings in words. Conclusions drawn from assessments in terms of a summarising categorising word (e.g. low/moderate/high) or numerical score (e.g. a risk score of 5) are generally

little informative for decision makers as they highly summarise and simplify information and are susceptible to different interpretations. Also, although low-risk individuals are generally the easiest to identify, the implications of low-risk conclusions may be great (i.e. the reduction of supervision or even release) and thus, for assessors, it is more challenging to draw low-risk conclusions than high-risk conclusions. In turn, decision makers sometimes find it difficult to accurately interpret these low/moderate/high-risk outcomes. Thus, there is a need for more informative and effective risk communication.

The newest advancement in risk assessment in recent years has been to describe the conclusions from the assessment in an informative narrative. This narrative provides a short summary that includes the description of the most likely risk scenarios for the individual. Based on previous routes to violence (or other undesirable outcomes, such as criminal behaviour in general) for the individual and current functioning, as well as the anticipated presence of risk and protective factors for the specific assessment context in the near future, the assessor sets out to contemplate on what could happen by asking himself the question: 'based on all the evidence gathered in this assessment regarding the different puzzle pieces for this individual, what am I mostly afraid of in terms of violent behaviour in the near future?'. General questions that can be posed here are 'what type of harmful behaviour is anticipated?', 'who could become victim?', 'how severe would this violence be?', 'how imminent could this take place?', and 'what factors are most likely to enhance or reduce this risk?'. Describing risk scenarios in this manner offers a great deal of insight into the reasoning of the assessor when contemplating on that one final conclusion 'low/moderate/high'. In fact, for one individual, there may be multiple risk scenarios thinkable at the same time, each with a different type of risk, victim, severity, imminence, and precipitating factors.[110] It would be good for decision makers to carefully consider all of the described risk scenarios for an individual when contemplating on the issue of dangerousness and to realise when an assessment does not include these narrative scenarios that the conclusions drawn from the assessment in numbers or in words 'low/moderate/high' compose a very scarce summary of the real estimation of risk that it aims to describe, which isn't nearly as informative for decision making as the more explicit and nuanced description of risk-scenario narratives.

2.5.11 Change over time

As discussed earlier, risk is not a static concept, but inherently changes over time. It is important for decision makers to take the assessment timeframe of specific measures into account. There are measures that assess imminent risk (e.g. DASA[111]), measures that assess risk in the coming weeks to months (e.g. START, HARM[112]), and measures that provide assessment for the more medium term of the coming six months to a year (e.g. HCR-20^{V3}, SAVRY, and VRS[113]). Thus, different risk assessments also have different 'expiry dates'. Risk far away in time is inevitably more difficult to assess than risk in the near future, as changes in context and individual behaviour may occur. In addition to considering variations in risk between different contexts, it may be useful for the decision maker to take into account changes in risk for an individual over time. This may be informative when aiming to evaluate whether specific interventions result in beneficial risk-reducing effects for the individual and contemplate on possible necessity for alterations in risk management or treatment initiatives.

Measuring change in risk over time, in other words treatment evaluation or routine outcome monitoring, can be accomplished by carrying out repeated assessment with the same measures at different points in time. In an attempt to facilitate this process, some tools have explicitly included a change rating in the assessment procedure (e.g. VRS). By

comparing the results from different assessment timepoints, decreases in risk factors and improvements in protective factors can be monitored. It should be noted, however, that when assessments at different timepoints have been caried out for different contexts, it becomes less straight forward to compare the results between different moments in time, as a new context may also bring forth new (risk-enhancing) challenges and new (risk-reducing) protective circumstances. Nevertheless, comparing assessments over time for a given tool provides the decision maker with valuable insight into the improvements an individual is making over time, resulting in risk reduction over time. This may be helpful when deciding on lifting restrictions or allowing specific leaves or ultimately granting discharge. From a clinical perspective, ideally a large database would be created in which data of multiple timepoint assessments are stored for a great number of individuals. This would then facilitate the comparison of change over time of one individual to that of other individuals on similar developmental pathways and with similar psychopathology and initial risk levels, in order to be able to evaluate whether the assessed individual is still on his anticipated change trajectory in comparison to other similar individuals. However, such 'big data' risk assessment databases are not widely available yet, so for now this largely remains an anticipated opportunity to inform decision making in the future.

2.5.12 Generalisability

It should be noted that most risk assessment measures have been developed in Western European or North American contexts. Often, the initial population an assessment tool was developed on predominantly consisted of Caucasian males. Although culturally informed studies attempt to validate widely used risk assessment measures for a range of cultural and ethnic backgrounds, including immigrants and indigenous people,[114] overall the evidence-base for risk assessment measures still varies widely across different groups. Several risk assessment measures have been translated in many different languages and are being applied in a wide range of cultures and countries (e.g. in Japan[115]). Validation studies in these countries often provide comparable results to those found in Western European or North American samples for people from different backgrounds; however; it cannot be assumed that results are universally generalisable across groups before sound validation studies have been carried out. A specifically difficult group to study are immigrants from different countries, as often immigrant groups present in forensic settings represent a large variety of backgrounds, which cannot be grouped together in research and thus complicates validation.

The same may be true for people with varying psychopathologies. While different studies have focused on people with commonly observed psychopathologies in forensic practice, such as psychotic disorders, personality disorders, or substance abuse, less-abundant disorders often remain understudied (e.g. Autism spectrum disorder). As mentioned earlier, it should be noted that the relationship between psychopathology and dangerousness varies greatly between diagnosis and individuals. For example, the relationship between psychotic disorders and violence is generally limited (i.e. most individuals with psychotic symptoms are not violent); however, for the individual case, this relationship can be quite clear.

Finally, risk assessment measures may not generate the same findings for female offenders as for males.[116] For this reason, specific additional measures have been developed that focus on factors which appear more prevalent for women and are valuable to explicitly take into account when doing risk assessment for a female individual (e.g. FAM[117]). In conclusion, risk assessment measures may in practice be applied to people for whom they have not (yet) been properly validated or study results are less convincing. Assessors and decision makers should

be aware of this and take this limitation into account when drawing conclusions from assessments carried out for individuals from minority groups in forensic settings.

2.5.13 Age

A related topic that might be relevant for the individual case is the question of age cut-offs for risk assessment measures. Traditionally, adult in risk assessment tools have focused on individuals from the age of 18 upwards, while juvenile risk assessment tools focused on younger individuals between the age of 12 and 17 (e.g. SAVRY, SAPROF-YV, and YLS/CMI[118]). Although in many cultures and jurisdictions, at age 18, an individual is legally regarded as an adult, this artificial cut-off remains quite arbitrary. We do not become altogether different individuals overnight on our 18th birthday, with altogether different risk profiles and assessment needs. Moreover, increasingly, studies of the brain highlight the finding that neurologically young adults are still developing until their mid-twenties.[119] In fact, while the age group of young adults (18–23) shows the highest rates of offending and recidivism,[120] surprisingly few studies focus on risk assessment specifically for this age group. In some legal systems, the notion that this group of young adults may be quite diverse in terms of developmental stage and that the adult sentencing system might not be entirely applicable to these young offenders has led to the development of specific 'adolescent law'. In the Netherlands, sentencing has become flexible in the sense that for young offenders between the age of 17 and 23 either juvenile or adult law can be applied, based on the developmental stage of the individual. Similarly, it would make sense if the application of adult or juvenile risk assessment tools would also be informed by evaluating the young individual's developmental age. If a young adult shows predominantly juvenile like behaviour, such as being in school, living at home, having younger friends, and being dependent on parents or caregivers, then the juvenile risk assessments tools are likely the best suited for assessing the individual. However, if the young adult lives independently, goes to work rather than school and relates mostly to older individuals, then the adult instruments are better suited. While research has shown that at group level, juvenile and adult risk assessment tools perform equally well for young adults at group level,[121] at the individual level it is advised to carefully consider which tools seem most applicable. Similarly, for very young juveniles, it could be considered whether child risk assessment measures (e.g. EARL[122] and SAPROF-CV[123]) might be more appropriate to use than juvenile tools. The decision maker should take note of the fit between the developmental age of the assessed individual and the applied risk assessment measure when drawing conclusions based on the findings in an assessment report.

2.5.14 The certainty of uncertainty

To summarise this contemplation of benefits and limitations of risk assessment in the light of legal decision making, perhaps the most important thing to remember when making use of risk assessment is that whatever measure was used and however results have been reported and interpreted by the assessor, it should be assumed that the assessor has attempted to unravel as many puzzle pieces as possible and from that has drawn conclusions to the best of his ability. Since 'assessment' concerns the future, one thing we know for sure is the certainty of uncertainty. Many seemingly high-risk individuals do not go on to recidivate (false positives), while some individuals who are considered low risk do commit new offenses (false negatives). Predictive validity studies aim to analyse correctly versus incorrectly predicted individuals; however, the question of 'what is considered recidivism?' is

also complicated (i.e. what types of offenses are included and within what timeframe after the assessment?). In addition, only a proportion of future crimes lead to convictions. Thus, the balance of correctly identified individuals is a fine one. Societal tolerance for false negatives of serious offenses is limited, while from a legal and ethical perspective, we aim to prevent unnecessary lengthy and costly interventions. From clinical experience, we have learned that gradual community re-integration providing room for learning from mistakes has shown to be the most effective way to prevent future recidivism, which in the end enhances the safety of society as a whole. Risk assessment results should be interpreted in this light by decision makers as well. Generally, personalised interventions and risk-management as offered in a forensic treatment setting are much more effective in terms of recidivism reduction than harsh punishments and lengthy prison sentences. In addition, sometimes slight risks (e.g. granting leaves during prison or hospital stay in order to practice with community re-integration goals) may be acceptable if the anticipated gain is worthwhile (i.e. reduced likelihood of longer-term recidivism) and risks are manageable. In this process, we can only attempt to optimise our assessment of likely future behaviour and from that aim to prevent undesirable outcomes through tailored risk management. Unfortunately, every hint towards certainty in predicting future (criminal) human behaviour is unjustified. Nevertheless, assessors and decision makers should strive for the best possible evaluation of risk and carefully consider the findings from risk assessment when legal decisions are contemplated.

2.6 In sum

In trying to summarise discussions on the current state of knowledge within the psycholegal disciplines and its quality, we realise we may have left the legal decision maker puzzled. But, we feel that transparency about strengths and limitations of forensic assessment eventually enhances the quality of legal decision making based upon it, without mitigating its utility.

In discussing the background of forensic assessment, in what ways the quality of assessment may be judged, and why behavioural assessment in the forensic context is an even more daunting task than in the clinical context, we hope to provide discussions between the disciplines with relevant subject matter to inquire after as well as with appreciation for respective roles and competencies. We have aspired to explain why in general the scientific evidence related to the quality of forensic assessment is hindered by both epistemological and methodological limitations, and why the possibilities for sound, relevant, and universal research on such quality differs enormously per psycholegal concept. Moreover, when the body of knowledge is more vast, for example, regarding risk assessment, it is also because new opportunities present itself to further strengths over limitations, which, however – given also the prospective nature of the endeavor – will never completely be overcome.

In this chapter, we have limited ourselves to three psycholegal concepts relevant for sentencing, as the assessment of other relevant concepts or criteria, for example, related to treatability or the need for treatment, builds on the assessments discussed here. Knowledge on those issues also overlaps with literature from the clinical context or criminology on the effectiveness of interventions, even though such evidence may be less translatable to the forensic context due to the limitations posed by potential legal frameworks.[124]

Indeed, legal decision making on the basis of forensic assessment is a 'puzzling' activity – in more than one meaning – in which some puzzling pieces will always be missing. Nevertheless, we have hoped to provide decision makers with enough guidance on finding pieces of the puzzle to eventually identify the complete picture enough to make a decision with the required certainty, despite remaining uncertainties.

Notes

1 See Meyer, Mihura and Smith, 2005, who performed a meta-analysis of interrater reliability in psychology and medicine to show that clinicians could reliably interpret the Rorschach test.
2 See Cooper, 2008.
3 See for example Halpern, 1980 and Group for the advancement of psychiatry, 1974, regarding legal insanity and competency to stand trial, respectively.
4 Parts of this paragraph are based on Van der Wolf and Van Marle, 2018.
5 McGlen et al., 2015.
6 Cited in Simon and Ahn-Redding, 2006, p. 4.
7 See respectively Walker, 1968 and McGlen et al., 2015.
8 See e.g. Siegel, 1973.
9 Walker, 1968.
10 Robinson, 1996.
11 See for modern influences of Wier and Pinel respectively: Hoorens, 2011 in Dutch, and Weiner, 2010.
12 See Halpern, 1980.
13 Mooij, 1995, in Dutch.
14 See also the Canadian chapter.
15 25% versus 36% respectively, as reported in a questionnaire among Dutch evaluators, Hummelen et al., 2013, in Dutch.
16 As referred to in the English chapter.
17 As for example Van Esch, 2012, in Dutch, found that only a third of the reports in her sample contained an adequate description of the relation between mental disorder and offense – which is the essence of the Dutch concept of criminal responsibility.
18 See for example the country chapters of Germany and the Netherlands.
19 See on a related note; Skeem et al., 2017.
20 Gowensmith et al., 2017a.
21 See Dror, 2016; and Mossman, 2013, specifically for forensic behavioural assessment.
22 Compare Edens and Boccaccini, 2017 and Gowensmith et al., 2017b.
23 See for example Gowensmith, Murrie and Boccaccini, 2013.
24 See Edens and Boccaccini, 2017, in their editorial of a special issue on: Field Reliability and Validity of Forensic Psychological Assessment Instruments and Procedures.
25 See Lamiell, 1998, on how the distinction between the idiographic and nomothetic approach, introduced by the Neo-Kantian philosopher Wilhelm Windelband is used in modern days. In short, the idiographic approach, common in the humanities, is related to the tendency to specify and describes research goals that focus on the individual. The nomothetic approach, common in the natural sciences, is related to the tendency to generalise and fits research goals that focus on generalising individual results to the entire population. Also, in forensic assessment, a combination of these approaches is or should be used to first not omit any relevant generalisable knowledge relevant to the case, while eventually coming to individualised conclusions.
26 See for example the chapter on the Dutch perspective.
27 Felthous, 2012, p. 14.
28 Felthous, 2012, p. 13.
29 See for example Edens and Boccaccini, 2017, and Colins et al., 2017 respectively.
30 See for example Osborne, Davies and Hutchinson, 2017.
31 Bornstein, 2017.
32 Zapf and Dror, 2017.
33 Zapf et al., 2017.
34 Koenraadt and Muller, 2013, in Dutch.
35 See for respective examples Neighbors et al., 2003, and Perrault, Vincent and Guy, 2017.
36 See for an overview of the literature and a conclusion on female retardation Sygel et al., 2015, who found that for people with the diagnosis mental retardation, women found more likely to reoffend.
37 The case in inquisitorial justice systems, see Chapter 10.
38 Zapf and Dror, 2017.
39 See for references the American chapter.
40 Koenraadt and Muller, 2013, mention for example the Horn-effect, the tendency to judge someone (too) negatively and neglect positive traits, and its opposite the Leniency-effect. The Halo-effect and Hawthorne-effect are also relevant in this respect.

41 Zapf and Dror, 2017. As an example of new (and quite opposite) scientific insights they mention the treatability of psychopaths.
42 See for example Bornstein, 2017 and Zapf and Dror, 2017.
43 Most prominently by the French philosopher Foucault, 1978, who has argued that in the nineteenth century, the developing functioning of Western medicine as a public hygiene – often equaling dangerousness with disorder or degeneracy – ensured that safety-measures, especially in continental European jurisdictions, could be used as a 'social defence' against 'non-social' groups in society.
44 See for example Ahonen, 2019.
45 See for example Zsasz, 1960.
46 APA, 2013, there are other systems of classification, like that of the ICD, now up to edition 11.
47 Van Marle and Van der Wolf, 2013, in Dutch.
48 See Ebert and Bär, 2010.
49 Reductionistic, in narrowing the problems of an individual to a certain label. Empiristic, in not incorporating the subjective experience of the person involved. Other common criticism is on the risk of overdiagnosis, possibly in the interest of the pharmaceutical industry.
50 Lane, 2020. As the system is based on cut off scores on longer lists of symptoms, it also allows for very different presentations of mental states under a similar label.
51 See paragraph 2.3 in all the country chapters.
52 See for example the insanity doctrine of Norway for which 'psychosis' is required, which played an important role in the case of the infamous terrorist Breivik, as two teams of experts disagreed on the matter. See Melle, 2013.
53 See also Gowensmith et al., 2017a.
54 See also Rogers et al., 1992.
55 See study in forensic mental health, which found predictive validity (Marin-Avellan et al., 2014), versus a study in prison which did not (Edens et al., 2015), so the type of institution may play a role.
56 See for example Frances, 2016.
57 See for example Maj, 2018.
58 See Chmielewski et al., 2015.
59 For an overview of possible biases especially for the assessment of criminal responsibility, see Meyer and Valença, 2021.
60 See Knoll and Resnick, 2008. See for a Swedish study on the relation between disorder and accountability, Höglund et al., 2009.
61 See Pirelli, Gottdiener and Zapf, 2011.
62 Gowensmith et al., 2017a, p. 697.
63 Van der Wolf, forthcoming.
64 Parts of this paragraph are based on Van der Wolf and Van Marle, 2018.
65 See for example the Dutch and Swedish perspective for deviant doctrines, and Chapter 10 for a comparison.
66 Van der Wolf and Van Marle, 2018.
67 Which is also underlined by prior fault doctrines related to disorders as a consequence of substance use.
68 See for example the American chapter.
69 By changing its qualification to manslaughter. This was derived from the humanitarian approach, originally in Scottish case law, to pardon mentally disordered offenders in capital cases. Walker, 1968.
70 See the chapters of the adversarial countries and Chapter 10.
71 Diamond, 1961.
72 See the Dutch chapter.
73 Compare Halpern, 1980.
74 Knoll and Resnick, 2008. As areas of potential research for evidence-based insanity defense evaluations, they mention studies on threshold criteria for mental disease or defect, malingered insanity (incidence, correlates, and detection methods), and the systematic use of feedback from triers of fact.
75 AAPL, 2002.
76 Rogers and Sewell, 1999.
77 Cai et al., 2014.
78 Meyer et al., 2020.
79 See also Dobbrunz et al., 2020, for a study on criteria used to assess control, related to criminal responsibility, among paraphilic offenders in Germany.
80 Gowensmith, Murrie and Boccaccini, 2013. But when judges disagreed with the majority opinion, they usually did so to find defendants legally sane, rather than insane.

81 See the country chapters of Germany, Sweden and the Netherlands.
82 See for example Guarnera, Murrie and Boccaccini, 2017.
83 Gowensmith, Murrie and Boccaccini, 2013, p. 98.
84 Gowensmith, Sledd and Sessarego, 2014.
85 Acklin, Fuger and Gowensmith, 2015, p. 334. See for an Australian comparison, Large, Nielssen and Elliott, 2009, and for a meta-analysis related to the interrater reliability in competency and insanity cases, Guarnera and Murrie, 2017.
86 Kacperska et al., 2016.
87 Van der Wolf, forthcoming.
88 Hart and Logan, 2011.
89 Andrews and Bonta, 2010.
90 Ward and Brown, 2004.
91 Douglas et al., 2013.
92 Andrews, Bonta and Wormith, 2004.
93 Singh et al., 2013.
94 De Vogel, Van den Broek and De Vries Robbé, 2014.
95 Douglas et al., 2013.
96 Quinsey et al., 1998.
97 Hanson and Thornton, 1999.
98 Fernandez et al., 2012.
99 Douglas et al., 2013.
100 Kropp and Hart, 2000.
101 Borum, Bartel and Forth, 2002.
102 Douglas and Otto, 2021.
103 Webster et al., 2009.
104 De Vogel et al., 2012.
105 See Douglas et al., 2013.
106 Rogers, 2000.
107 De Ruiter and Nicholls, 2011.
108 De Vries Robbé et al., 2015.
109 De Vries Robbé and Willis, 2017.
110 Douglas et al., 2013.
111 Ogloff and Daffern, 2006.
112 Chaimowitz and Mamak, 2011.
113 Wong and Gordon, 2003.
114 Shepherd et al., 2014.
115 Kashiwagi et al., 2018.
116 De Vogel and Nicholls, 2016.
117 De Vogel et al., 2012.
118 Hoge and Andrews, 2006.
119 Diamond, 2002; Steinberg and Icenogle, 2019.
120 Piquero, Farrington and Blumstein, 2007.
121 Kleeven et al., 2020.
122 Augimeri et al., 2001.
123 De Vries Robbé et al., 2021.
124 See also Weisburd, Farrington and Gill, 2016.

References

AAPL (2002). AAPL practice guideline for forensic psychiatric evaluation of defendants raising the insanity defense. American Academy of Psychiatry and the Law. *Journal of the American Academy of Psychiatry and the Law*, 30(2), pp. s3–s40.

Acklin, M.W., Fuger, K., and Gowensmith, W. (2015). Examiner agreement and judicial consensus in forensic mental health evaluations. *Journal of Forensic Psychology Practice*, 15, pp. 318–343.

Ahonen, L. (2019). *Violence and Mental Illness: An Overview*. New York: Springer.

Andrews, D.A., and Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology Public Policy and Law*, 16(1), pp. 39–55.

Andrews, D.A., Bonta, J., and Wormith, S.J. (2004). *The Level of Service/Case Management Inventory (LS/CMI)*. Toronto, Canada: Multi-Health Systems.

APA (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington: APA.

Augimeri, L.K., Koegl, C.J., Webster, C.D., and Levene, K.S. (2001). *Early Assessment Risk List for Boys: EARL-20B, Version 2*. Toronto, Canada: Earlscourt Child and Family Centre.

Bornstein, R.F. (2017). Evidence-based psychological assessment. *Journal of Personality Assessment*, 99, pp. 435–445.

Borum, R., Bartel, P., and Forth, A. (2002). *SAVRY: Manual for the Structured Assessment of Violence Risk in Youth*. Tampa, FL: University of South Florida, Florida Mental Health Institute.

Cai, W., Zhang, Q., Huang, F., Guan, W., Tang, T., and Liu, C. (2014). The reliability and validity of the rating scale of criminal responsibility for mentally disordered offenders. *Forensic Science International*, 2014(236), pp. 146–150.

Chaimowitz, G.A., and Mamak, M. (2011). *Companion Guide to the Aggressive Incidents Scale and the Hamilton Anatomy of Risk Management (HARM)*. Hamilton, Canada: St. Joseph's Healthcare Hamilton.

Chmielewski, M., Clark, L.A., Bagby, R.M., and Watson, D. (2015). Method matters: Understanding diagnostic reliability in DSM-IV and DSM-5. *Journal of Abnormal Psychology*, 124(3), pp. 764–769.

Colins, O.F., Fanti, K.A., Andershed, H., Mulder, E., Salekin, R.T., Blokland, A., and Vermeiren, R.J.M. (2017). Psychometric properties and prognostic usefulness of the youth psychopathic traits inventory (YPI) as a component of a clinical protocol for detained youth: A multiethnic examination. *Psychological Assessment*, 29(6), pp. 740–753.

Cooper, R. (2008). *Psychiatry and Philosophy of Science*. Montreal: McGill-Queen's University Press.

Diamond, B. (1961). Criminal responsibility of the mentally ill.*Stanford Law Review*, 14, pp. 59–86.

Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. Stuss and R. Knight (Eds.), *Principles of Frontal Lobe Function*. New York, NY: Oxford university press, pp. 466–503.

De Ruiter, C., and Nicholls, T.L. (2011). Protective factors in forensic mental health: A new frontier. *International Journal of Forensic Mental Health*, 10, pp. 160–170.

De Vogel, V., De Ruiter, C., Bouman, Y., and De Vries Robbé, M. (2012). *SAPROF. Guidelines for the Assessment of Protective Factors for Violence risk* (2nd ed.). Utrecht, The Netherlands: Van der Hoeven Stichting.

De Vogel, V., De Vries Robbé, M., Van Kalmthout, W., and Place, C. (2012). *Female Additional Manual (FAM). Additional Guidelines to the HCR-20 for Assessing Risk for Violence in Women*. Utrecht, The Netherlands: Van der Hoeven Stichting.

De Vogel, V., Van den Broek, E., and De Vries Robbé, M. (2014). The use of the HCR-20^{V3} in Dutch forensic psychiatric practice. *International Journal of Forensic Mental Health*, 13, pp. 109–121.

De Vogel, V., and Nicholls, T.L. (2016). Gender matters: An introduction to the special issues on women and girls. *International Journal of Forensic Mental Health*, 15(1), pp. 1–25.

De Vries Robbé, M., Geers, M.C.K., Stapel, M., Hilterman, E.L.B., and De Vogel, V. (2015). *SAPROF - Youth Version. Structured Assessment of Protective Factors for Violence Rrisk – Youth Version. Guidelines for the Assessment of Protective Factors for Violence Risk in Juveniles*. Utrecht, The Netherlands: Van der Hoeven Kliniek.

De Vries Robbé, M., Smaragdi, A., Hilterman, E., Walsh, M., and Augimeri, L. (2021, in press). *The Structured Assessment of Protective Factors for violence risk – Child Version (SAPROF-CV)*.

De Vries Robbé, M., and Willis, G. (2017). Assessment of protective factors in clinical practice. *Aggression and Violent Behavior*, 32, pp. 55–63.

Dobbrunz S., Daubmann A., Müller, J.L., and Briken, P. (2020). Predictive validity of operationalized criteria for the assessment of criminal responsibility of sexual offenders with paraphilic disorders: A randomized control trial with mental health and legal professionals. *Frontiers in Psychology*, 11, p. 613081.

Douglas, K.S., Hart, S.D., Webster, C.D., and Belfrage, H. (2013). *HCR-20^{V3}Assessing Risk for Violence. User Guide*. Vancouver, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Douglas, K.S., and Otto, R.K. (Eds.) (2021). *Handbook of Violence Risk Sssessment* (2nd ed.). New York, NY: Routledge.

Dror, I.E. (2016). A hierarchy of expert performance. *Journal of Applied Research in Memory and Cognition*, 5(2), pp. 121–127.

Ebert, E., and Bär, K.-J. (2010). Emil Kraepelin: A pioneer of scientific understanding of psychiatry and psychopharmacology. *Indian Journal of Psychiatry*, 52(2), pp. 191–192.

Edens, J.F., and Boccaccini, M.T. (2017). Taking forensic mental health assessment "Out of the Lab" and into "the Real World": Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment*, 29(6), pp. 599–610.

Edens, J.F., Kelley, S.E., Lilienfeld, S.O., Skeem, J.L., and Douglas, K.S. (2015). DSM-5 antisocial personality disorder: Predictive validity in a prison sample. *Law and Human Behavior*, 39(2), pp. 123–129.

Felthous, A. (2012). The diagnosis of psychopathology: Improving reliability and validity in forensic examinations. In T. I. Oei, and M. S. Groenhuijsen (Eds.), *Progression in Forensic Psychiatry*. Deventer: Kluwer, pp. 13–48.

Fernandez, Y., Harris, A.J.R., Hanson, R.K., and Sparks, J. (2012). *STABLE-2007 Coding Manual: Revised 2012*. [unpublished scoring manual]. Ottawa, Canada: Public Safety Canada.

Foucault, M. (1978). About the concept of the 'Dangerous Individual' in 19th-century legal psychiatry. *International Journal of Law and Psychiatry,* 1, pp. 1–18.

Frances, A. (2016). A report card on the utility of psychiatric diagnosis. *World Psychiatry*, 15(1), pp. 32–33.

Gowensmith, W.N., Murrie, D.C., and Boccaccini, M.T. (2013). How reliable are forenisc evaluations of legal insanity? *Law and Human Behavior*, 37(2), pp. 98–106.

Gowensmith, W.N., Sledd, M., and Sessarego, S. (2014). The impact of stringent certification standards on forensic evaluator reliability. *Paper presented at the 122nd annual meeting of the American Psychological Association*, Washington, DC.

Gowensmith, W.N., Sessarego, S.N., McKee, M.K., Horkott, S., MacLean, N., and McCallum, K.E. (2017a). Diagnostic field reliability in forensic mental health evaluations. *Psychological Assessment*, 29(6), pp. 692–700.

Gowensmith, W.N., Murrie, D.C., Boccaccini, M.T., and McNichols, B.J. (2017b). Field reliability influences field validity: Risk assessments of individuals found not guilty by reason of insanity. *Psychological Assessment*, 29(6), pp. 786–794.

Group for the advancement of psychiatry (1974). *Misuse of Psychiatry in the Criminal Courts: Competency to Stand Ttrial*. New York: Group for the advancement of psychiatry.

Guarnera, L.A., and Murrie, D.C. (2017). Field reliability of competency and sanity opinions: A systematic review and meta-analysis. *Psychological Assessment*, 29(6), pp. 795–818.

Guarnera, L.A., Murrie, D.C., and Boccaccini, M.T. (2017). Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations. *Translational Issues in Psychological Science*, *3*(2), pp. 143–152.

Halpern, A.L. (1980). The fiction of legal insanity and the misuse of psychiatry. *Journal of Legal Medicine*, 1(4), pp. 18–74.

Hanson, R.K. and Thorton, D. (1999). *Static-99: Improving Actuarial Risk Assessments for Sex Offenders*. Ottawa, Canada: Department of the Solicitor General of Canada.

Hart, S.D., and Logan, C. (2011). Formulation of violence risk using evidence-based assessments: The structured professional judgment approach. In: P. Sturmey and M. McMurran (eds.), *Forensic Case Formulation*. West Sussex, UK: John Wiley and Sons Ltd., pp. 83–106.

Hoge, R.D., and Andrews, D.A. (2006). *Youth Level of Service / Case Management Inventory (YLS/CMI): User's Manual*. Toronto, Canada: Multi-Health Systems.

Höglund, P., Levander, S., Anckarsäter, H., and Radovic, S. (2009). Accountability and psychiatric disorders: How do forensic psychiatric professionals think? *International Journal of Law and Psychiatry*, 32(6), pp. 355–361.

Hoorens, V. (2011). *Een ketterse arts voor de* heksen. Jan Wier (1515-1588). Amsterdam: Bert Bakker.

Hummelen, J.W., Van Esch, C.M., Schipaanboord, A.E., and Van der Veer, T.S. (2013). Het inzage- en correctierecht bij gedragsdeskundig onderzoek in strafzaken. *Expertise en Recht*, 2, pp. 44–49.

Kacperska, I., Heitzman, J., Bak, T., Le´sko, A.W., and Opio, M. (2016). Reliability of repeated forensic evaluations of legal sanity. *International Journal of Law and Psychiatry*, 44, pp. 24–29.

Kashiwagi, H., Kikuchi, A., Koyama, M., Saito, D., and Hirabayashi, N. (2018). Strength-based assessment for future violence risk: A retrospective validation study of the Structured Assessment of Protective Factors

for violence risk (SAPROF) Japanese version in forensic psychiatric inpatients. *Annals of general psychiatry*, 17(1), pp. 1–8.

Kleeven, A.T.H., De Vries Robbé, M., Mulder, E.A., and Popma, A. (2020). Risk assessment in juvenile and young adult offenders: Predictive validity of the SAVRY and SAPROF-YV. *Assessment*, 29, pp. 1–17.

Knoll, J.L. IV, and Resnick, P.J. (2008). Insanity defense evaluations: Toward a model for evidence-based practice. *Brief Treatment and Crisis Intervention*, 8(1), pp. 92–110.

Koenraadt, F. and Muller, E. (2013). Het psychologisch onderzoek en de daarop gebaseerde rapportage pro justitia. In H.J.C. Van Marle, P.A.M. Mevis and M.J.F. Van der Wolf (Eds.), *Gedragskundige rapportage in het strafrecht*, *Tweede herziene druk*. Deventer: Kluwer, pp. 269–346.

Kropp, P.R., and Hart, S.D. (2000). The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders. *Law and Human Behavior*, 24(1), pp. 101–118.

Lamiell, J.T. (1998). 'Nomothetic' and 'Idiographic': Contrasting Windelband's understanding with contemporary usage. *Theory & Psychology*, 8(1), pp. 23–38.

Lane, R. (2020). Expanding boundaries in psychiatry: Uuncertainty in the context of diagnosis-seeking and negotiation. *Sociology of Health and Fitness*, 42(s1), pp. 69–83.

Large, M., Nielssen, O., and Elliott, G. (2009). Reliability of psychiatric evidence in serious criminal matters: Fitness to stand trial and the defence of mental illness. *The Australian and New Zealand Journal of Psychiatry*, 43(5), pp. 446–452.

Maj, M. (2018). Why the clinical utility of diagnostic categories in psychiatry is intrinsically limited and how we can use new approaches to complement them. *World Psychiatry*, 17(2), pp. 121–122.

Marin-Avellan, L.E., McGauley, G.A., Campbell, C.D., and Fonagy, P. (2014). The validity and clinical utility of structural diagnoses of antisocial personality disorder with forensic patients. *Journal of Personality Disorders*, 28(4), pp. 500–517.

McGlen, M., Brown, J., Hughes, N.S., and Crichton, J. (2015). *The Classical Origins of the Insanity Defence (Homicide and Mental Disorder Lecture Series, Book 2)*. Edinburgh: Bahookie Publishers.

Melle, I. (2013). The Breivik case and what psychiatrists can learn from it. *World Psychiatry*, 2013, pp. 16–21.

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

Meyer, L.F., Leal, C.C.S., Almeida Souza Omena, A. de, Mecler, K., and Valença, A.M. (2020). Criminal responsibility scale: Development and validation of a psychometric tool structured in clinical vignettes for Criminal Responsibility Assessments in Brazil. *Frontiers in Psychiatry*, 11, p. 579243.

Meyer, L.F., and Valença, A.M. (2021). Factors related to bias in forensic psychiatric assessments in criminal matters: A systematic review. *International Journal of Law and Psychiatry*, 75, p. 101681.

Mooij, A.W.M. (1995). TBS en Rapportage pro Justitia. Een historische beschouwing. *Ontmoetingen. Voordrachtenreeks van het Lutje Psychiatrisch-Juridisch Gezelschap*, pp. 21–31.

Mossman, D.M. (2013). When forensic examiners disagree: Bias, or just inaccuracy? *Psychology, Public Policy, and Law*, 19(1), pp. 40–55.

Neighbors, H.W., Trierweiler, S.J., Ford, B.C., and Muroff, J.R. (2003). Racial differences in DSM diagnosis using a semi-structured instrument: The importance of clinical judgment in the diagnosis of African Americans. *Journal of Health and Social Behavior*, 44(3), pp. 237–256.

Ogloff, J.R., and Daffern, M. (2006). The dynamic appraisal of situational aggression: An instrument to assess risk for imminent aggression in psychiatric inpatients. *Behavioral Sciences & the Law*, 24(6), pp. 799–813.

Osborne, D., Davies, P.G., and Hutchinson, S. (2017). Stereotypicality biases and the criminal justice system. In C.G. Sibley and F. K. Barlow (Eds.), *The Cambridge Handbook of the Psychology of Prejudice*. Cambridge: Cambridge University Press, pp. 542–558.

Perrault, R.T., Vincent, G.M., and Guy, L.S. (2017). Are risk assessments racially biased?: Field study of the SAVRY and YLS/CMI in probation. *Psychological Assessment*, 29(6), pp. 664–678.

Piquero, A.R., Farrington, D.P., and Blumstein, A. (2007). *Key Issues in Criminal Career Research: New Analyses of the Cambridge Study in Delinquent Development*. Cambridge, UK: Cambridge University Press.

Pirelli, G., Gottdiener, W.H., and Zapf, P. A. (2011). A meta-analytic review of competency to stand trial research. *Psychology, Public Policy, and Law*, 17(1), pp. 1–53.

Quinsey, V.L., Harris, G.T., Rice, M.E., and Cormier, C.A. (1998). *Violent Offenders: Appraising and Managing Risk*. Washington, DC: American Psychological Association.

Robinson, D. (1996). *Wild Beasts & Idle Humours. The Insanity Defense from Antiquity to the Present*. Cambridge, Massachusetts: Harvard University Press.

Rogers, R. (2000). The uncritical acceptance of risk assessment in forensic practice. *Law and Human Behavior*, 24, pp. 595–605.

Rogers, R., Dion, K.L., and Lynett, E. (1992). Diagnostic validity of antisocial personality disorder: A prototypical analysis. *Law and Human Behavior*, 16(6), pp. 677–689.

Rogers, R., and Sewell, K.W. (1999). The R-CRAS and insanity evaluations: A reexamination of construct validity. *Behavioral Sciences & the Law*, 17(2), pp. 181–194.

Shepherd, S.M., Luebbers, S., Ferguson, M., Ogloff, J.R.P., and Dolan, M. (2014). The utility of the SAVRY across ethnicity in Australian young offenders. *Psychology Public Policy and Law*, 20(1), pp. 31–45.

Siegel, R.E. (1973). Galen on psychology, psychopathology, and function and diseases of the nervous system: An analysis of his doctrines, observations and experiments. In *Galen's System of Physiology and Medicine*. Basel: Karger.

Simon, R.J., and Ahn-Redding, H. (2006). *The Insanity Defense, the World Over*. Lanham: Lexington Books.

Singh, J.P., Yang, S., Bjorkly, S., Boccacini, M.T., Borum, R., Buchanan, A., et al. (2013). *Reporting standards for Risk Assessment Predictive Validity Studies: The Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Statement*. Tampa, FL: University of South Florida.

Skeem, J.L., Kennealy, P.J., Tatar II, J.R., Hernandez, I.R., and Keith, F.A. (2017). How well do juvenile risk assessments measure factors to target in treatment? Examining construct validity. *Psychological Assessment*, 29(6), pp. 679–691.

Steinberg, L., and Icenogle, G. (2019). Using developmental science to distinguish adolescents and adults under the law. *Annual Review of Developmental Psychology*, 1, pp. 21–40.

Sygel, K., Sturup, J., Fors, U., Edberg, H., Gavazzeni, J., Howner, K., Persson, M., and Kristiansson, M. (2015). The effect of gender on the outcome of forensic psychiatric assessment in Sweden: A case vignette study. *Criminal Behavior and Mental Health*, 27(2), pp. 124–135.

Van Esch, C.M. (2012). *Gedragsdeskundigen in strafzaken* (diss. Leiden). Assen: Koninklijke Van Gorcum.

Van Marle, H.J.C., and Van der Wolf, M.J.F. (2013). Forensisch psychiatrische ziekteleer: een inleiding, een handleiding, een handreiking. In H.J.C. Van Marle, P.A.M. Mevis and M.J.F. Van der Wolf (Eds.), *Gedragskundige rapportage in het strafrecht, Tweede herziene druk*. Deventer: Kluwer, pp. 77–130.

Van der Wolf, M.J.F. (forthcoming). The level of agreement between experts in forensic assessment in the Netherlands.

Van der Wolf, M.J.F., and Van Marle, H.J.C. (2018). Legal approaches to criminal responsibility of mentally disordered offenders in Europe. In K. Goethals (Ed.), *Forensic Psychiatry and Psychology in Europe. A Cross-Border Study Guide*. Basel: Springer International Publishing, pp. 31–44.

Walker, N. (1968). *Crime and Insanity in England. Volume One: The Historical Perspective*. Edinburgh: University Press.

Ward, T., and Brown, M. (2004). The good lives model and conceptual issues in offender rehabilitation. *Psychology, Crime and Law*, 10, pp. 243–257.

Webster, C.D., Martin, M.L., Brink, J., Nicholls, T.L. and Desmarais, S.L. (2009). *Short-Term Assessment of Risk and Treatability (START) (Version 1.1)*. Coquitlam, Canada: British Columbia Mental Health and Addiction Services.

Weiner, D.B. (2010). Philippe Pinel in the 21st Century: The myth and the message. In E.R. Wallace, and J. Gach (Eds.), *History of Psychiatry and Medical Psychology*. New York: Springer pp. 305–312.

Weisburd, D., Farrington, D.P., and Gill, C. (Eds) (2016). *What Works in Crime Prevention and Rehabilitation: Lessons from Systematic Reviews*. New York, NY: Springer.

Wong, S., and Gordon, A. (2003). *Violence Risk Scale (VRS)*. Saskatoon, Canada: Regional Psychiatric Centre.

Zapf, P.A., and Dror, I.E. (2017). Understanding and mitigating bias in forensic evaluation: Lessons from forensic science. *International Journal of Forensic Mental Health*, 16(3), pp. 227–238.

Zapf, P.A., Kukucka, J., Kassin, S.M., and Dror, I.E. (2017). Cognitive Bias in forensic mental health assessment: Evaluator beliefs about its nature and scope. *Psychology, Public Policy, and Law*, 24(1), pp. 1–10.

Zsasz, T. (1960). *The Myth of Mental Illness*. New York: HarperCollins.