EDITORIALS

An agitation of contrary opinions
Peter Tyrer

Dangerous and severe personality disorder and in need of treatment
Paul E. Mullen
Br J Psychiatry 2007 190: s3-s7.

Dangerous and severe personality disorder: antecedents and origins
Anthony Maden
Br J Psychiatry 2007 190: s8-s11.

Persistent violent offending: what do we know?
Sheilagh Hodgins

SPECIAL ARTICLES

Understanding the structure of the Psychopathy Checklist – Revised: An exploration of methodological confusion
David J. Cooke, Christine Michie, and Jennifer Skeem

PAPERS

Value of standard personality assessments in informing clinical decision – making in a medium secure unit
Conor Duggan, Lauren Mason, Penny Banerjee, and John Milton

Altered memory and affective instability in prisoners assessed for dangerous and severe personality disorder
Tim Kirkpatrick, Eileen Joyce, John Milton, Conor Duggan, Peter Tyrer, and Robert. D. Rogers

Developmental trajectories associated with juvenile sexually abusive behaviour and emerging severe personality disorder in childhood: 3 – year study
Eileen Vizard, Nicole Hickey, and Eamon McCrory

Aetiology of the relationship between callous–unemotional traits and conduct problems in childhood
Essi Viding, Paul J. Frick, and Robert Plomin
Critical developments in the assessment of personality disorder
Peter Tyrer, Natalie Coombs, Fatema Ibrahimi, Anand Mathilakath, Priya Bajaj, Maja Ranger, Bharti Rao, and Raana Din

Precision of actuarial risk assessment instruments: Evaluating the `margins of error' of group v. individual predictions of violence
Stephen D. Hart, Christine Michie, and David J. Cooke

Assessment and treatment of violence–prone forensic clients: an integrated approach
Stephen C. P. Wong, Audrey Gordon, and Deqiang Gu

Collecting service use data for economic evaluation in DSPD populations: Development of the Secure Facilities Service Use Schedule
Barbara Barrett and Sarah Byford
Br J Psychiatry 2007 190: s75-s78.
An agitation of contrary opinions

PETER TYRER

Summary Those people who are dangerous often have personality disorders. Should these individuals be dealt with by criminal justice or mental health services? England (note not Scotland) has taken the mental health route with the Dangerous and Severe Personality Disorder Programme. Is this bold move wise or foolish? To answer this question we have both evidence and opinion — neither is conclusive.

Declaration of Interest P.T. is Editor of the British Journal of Psychiatry and has received a grant for the evaluation of aspects of the DSPD Programme. The decision to publish this supplement, and the peer review of its content, was made independently by other members of the Editorial Board. The views expressed here, and in the rest of the supplement, are those of the authors and not necessarily those of the Home Office or Department of Health (nor do they represent Government policy). The publication of this supplement was supported by the DSPD Programme but that played no part in the selection and review of the papers included.

The initials DSPD (dangerous and severe personality disorder) have only been with us for 7 years, and although the term is specific to services in England, it is becoming better known through international communication (Moran, 2001; Maden & Tyrer, 2003; Fountoulakis & Kaprinis, 2006). For those who have not yet ventured beyond the initials to the context, this supplement is an essential read. This is not a specious attempt to get you to read something you would otherwise ignore but an attempt to expose a range of important issues that have attracted a good deal of controversy, and no doubt will continue to do so, to fuller scrutiny. Reading this will help you to decide, or at least consider your previous opinions, about important questions in everyday practice, such as ‘Is the care of those with severe personality disorder a responsibility for mental health services? Is dangerousness predictable?’ and ‘How do we protect society from those who pose a severe risk of violence?’

As a profession, psychiatry has not been slow to express its opinions about DSPD. The Home Office and Department of Health stress that it is not a diagnostic term but a description of a new programme to provide both specialised care for a small but important group and to protect the public from risk. Early on negative views outweighed positive views by a factor of 3 to 1 (Haddock et al, 2001), and among the concerns the ethical ones of detaining people for long periods were very prominent (Moran, 2001). Indeed, some have gone right back to the Hippocratic oath and questioned whether doctors might be abusing their position if they ‘give patients, or recommend to them, an investigation or treatment which you know is not in their best interests’ (Haddock et al, 2001).

Others are more sanguine, believing that the concerns about doctors being granted unlimited powers over people’s freedom is never likely to happen, or as Maden (2002) put it in the context of worries expressed by the philosophy of Michel Foucault, ‘the experience of working in the National Health Service soon reduces one’s ability to empathise with a man who worries about doctors taking over the world.’

There is genuine concern about the ethics of opting out of care for a needy group of individuals if we fail to address these problems, ‘if we as clinicians refuse to treat people who are clearly unwell and distressed, we would be failing in our duty of care and pushing these vulnerable patients into the criminal justice system’ (Mahapatra, 2001). On an even more positive note Mullen (2007) comments in this supplement that the DSPD Programme ‘now represents a genuine attempt to address the psychological and interpersonal difficulties of recidivist violent offenders in a manner which it is hoped will decrease the damage these people do to others and to themselves.’ However, you will note from the punctuation of his title that joining the three distinct entities remains a hypothesis, not a proof.

The group we are talking about is small (hence those in less than 5000 of the relevant population). This translates to 350 people and we know that there are perhaps 7–10 times as many in prison. This remains a tiny proportion of the 2.2 million people in the UK who have personality disorders according to a nationwide survey (Coid et al, 2006), but one which is highly significant in political and social terms. There may be arguments over the best way of managing this small group but the need for management cannot be ignored, either in terms of public protection or in meeting their mental health needs. What is done in this area will no doubt have a significant impact on the broader development of personality disorder services.

In Scotland the introduction of legislation to cater for this group has been relatively smooth: the adoption of the MacLean Committee’s recommendations of ‘an order for lifelong restriction for offenders likely to pose a continuing and serious risk to the public’ has avoided the complexities created by the term ‘personality disorder’ (Darjee & Crichton, 2002). In England, Chiswick’s prediction (2001) that ‘the Government has created a personality disorder monster that the public wish to see slain and we can expect Members of Parliament dutifully to approve the legislation’ has not yet taken a form that critics feared. However, the legal landscape in England and Wales has been transformed (Criminal Justice Act 2003) with the introduction of April 2005 of public protection sentences whereby indeterminate detention can be imposed on the basis of convictions for serious sexual or violent offences, with release, after the tariff period has been served, being dependent on the parole board being satisfied that the risk has reduced.

In Holland, as Maden (2007) reports in these pages, they have been quietly dealing...
with this problem in a different way for many years, and in Canada, that haven of good order and violence assessment capital of the world, they have pioneered assessment and treatment of this group ahead of all others, and we are pleased to report one of the core programmes in these pages (Wong et al, 2007).

The management and reduction of risk is central to the success of any Programme to prevent violence, and this has always been at the core of the DSPD Programme. The problem here is that we are not yet very good at prediction and are a long way from the film Minority Report in which the exact nature and timing of violent offences was identified by ‘precogs’ with advance knowledge. Our precogs are primitive and do not go much beyond basic demographic details and legal history (Buchanan & Leese, 2006), and as a consequence we are likely to detain people for longer than might be necessary to address the needs of public protection (Buchanan & Leese, 2001). The hope is that we are getting better at prediction, but it will take a long time to be sure. The best ways of measuring risk and personality disorder and the interpretation of the data are discussed at length in this supplement.

We are very pleased to be able to have all sides of this controversy exposed for the readers of the British Journal of Psychiatry to mull over at their leisure.

Samuel Johnson, in his first dictionary of the English language, separated the word ‘dispute’ from ‘controversy’ as the former was ‘commonly oral’ whereas controversy, nicely defined as ‘an agitation of contrary opinions’, was ‘in writing’. Whatever your interest in the subject of personality disorder or dangerousness we hope that you will allow yourself to be diverted into this controversy and be stimulated, informed and possibly edified.

REFERENCES


Dangerous and severe personality disorder
and in need of treatment

PAUL E. MULLEN

Summary  The Dangerous and Severe Personality Disorder Programme was born out of a populist law and order reaction, developed on false premises, but is now evolving into an exciting initiative for providing effective services to a group of offenders with mental illness who psychiatry, and the justice services, have so long ignored. Enthusiasm, flexibility and an evidence-based approach may yet lead to real progress towards the improved management of disturbed high-risk offenders, improving the psychological and social functioning of the offenders as well as delivering a safer community.

Declaration of interest  None.

The Dangerous and Severe Personality Disorder (DSPD) Programme represents an enormous commitment of money and resources by the British Government to what is essentially a mental health initiative. The programme was born out of an ill-conceived attempt to hide the imposition of preventive detention and indefinite sentences behind the veneer of respectability provided by a mental health context (Eastman, 1999; Mullen, 1999). However, just as good intentions can pave the road to hell so cynical calculation can occasionally lead, if not to heaven, at least to progress for the good.

The DSPD Programme has come a long way from its murky beginnings. It now represents a genuine attempt to address the psychological and interpersonal difficulties of recidivist violent offenders in a manner which it is hoped will decrease the damage these people do to others and to themselves. This is a therapeutic experiment on a massive scale. A lot is riding on the outcomes, not just the credibility of those who advised and advocated for these programmes, but the credibility of mental health itself. One eminent academic at the York conference made it quite clear that she considered that if the target was decreasing violent crime public money would have been far more effectively spent on improving services for those with the schizophrenic syndrome at high risk of offending. She is right, but that is now irrelevant. The DSPD juggernaut is on the move, and if it fails I doubt if there will be any further money for initiatives directed at offenders with mental illness for some considerable time. Conversely if it succeeds this may well improve the public and political stock of mental health considerably. The DSPD initiative is being watched by forensic mental health professionals, and by governments, around the world. Will it become the model or a warning?

How the DSPD units will be judged, at least by government, is obvious: a decrease in violent recidivism among those who are managed by these services. How large a decrease? For the money they are spending it will have to be substantial, but whether this amounts to expectations of 20, 50, 80 or even 100% is not yet clear (see Barrett & Byford, 2007, this issue). How will these services reach such targets? By correctly identifying individuals with DSPD and making them harmless people, with or without personality disorders or keeping them contained for so long that time and the depredations of age render them pacific.

The York conference from which the papers in this special issue were drawn was to mark the official beginning of the DSPD initiative and so can hardly be expected to offer outcome data. If the enthusiasm and commitment of the clinicians at the conference from the various DSPD units were an indicator of future success, then that success is assured. But what do the papers presented here and the international experience augur for the future?

THE DSPD MODEL

The models which drive the DSPD initiative are still developing but are grounded in the belief that specific pathologies of personality initiate and sustain criminal behaviours. The programmes assume that those with these psychopathologies can be reliably identified, that unless something changes these individuals will commit further serious crimes and that by treating their personality disorders the risks they present will be substantially decreased. The key construct is psychopathy.

The psychopathy at issue is a putative criminogenic mental disorder. There is discussion as to the best means for identifying such psychopathy and factors that might modify its manifestations (Cleckley, 1964; Hare, 1991, 1999). The advocates for the psychopathy concept present it as a modern scientific discovery whose existence and relevance is vouchedsafe by nosological, epidemiological, neuropsychological, neurological and genetic research (Rhee & Waldman, 2002; Blair, 2003; see also Hodgins, 2007, and Viding et al, 2007, this issue). The unfortunate histories of previous attempts at medicalising crime (moral insanity, degeneration and conceptualisations which shared the name psychopathy) are presumed to be irrelevant (Lewis, 1974; Pick, 1989). Perhaps the only link between the failed systems of the past and today’s ‘psychopathic sciences’ is the enthusiasm which greets each and every biological explanation of crime. The attractions of explaining criminal propensities in terms of individual pathology are multiple. If some disorder drives the behaviour then curing or suppressing the disorder stops the criminality. If crime is a malfunction at an individual level it reduces the need to seek explanations in social, cultural and familial influences, a search that might suggest that lawlessness is bred by inequality, disadvantage and oppression. The fact that biomedical explanations of crime are perennially appealing to the elite does not make them wrong. Genetic factors and the environment probably do contribute to some crime but the attributable risk or even the potential mechanisms of genetic influences are not so clear. The much heralded announcement of genetically mediated links between childhood disadvantage and adult criminality await more substantial attempts at refutation (Caspi et al, 2002). The occasional reports of brain abnormalities underlying murder and mayhem do not usually have the methodological sophistication to have justified their publication let alone the bother of refutation. However, substantial or silly, the continuing appearance of
ON BEING DANGEROUS

The DSPD Programme does depend on being able to correctly identify those who will, unless something is done, go on to commit serious crimes. In broad terms, predicting recidivism is easy enough. Reoffending is associated with being young, male, unpartnered, poorly educated with few work skills and the amount and versatility of prior criminality, having substance misuse problems, antisocial attitudes and a criminal peer group. The importance of psychopathy is that it may provide the missing ingredient which separates out those destined both to greater criminal persistence and to more violent predation. These are the ones who will not mature, not settle into abiding intimate relationships, not stop seeking sensation and novelty, not exchange empathy for callous indifference and above all not abandon their exploitative and predatory ways for more modest satisfactions. The psychopathy ‘thermometer’ which measures how much of this dangerous ingredient is present is the Psychopathy Checklist–Revised (PCL–R; Hare, 1991). This was not designed as a risk assessment tool but as a personality test. However, the PCL–R provides the core and legitimation of many approaches to risk assessment, including the Historical/ Clinical/Risk Management 20-item scale (HCR–20; Webster et al., 1997) and the Violent Offenders: Appraising and Managing Risk (VRAG; Quinsey et al., 2005). The VRAG is important because it was central to the early development of the DSPD initiative. It was even rumoured that specific VRAG scores were being considered for legally identifying individuals with a DSPD. If an individual’s score on the VRAG is not a reliable guide to the extent of their dangerousness, and if the PCL–R does not identify that risk, there may be a problem.

The paper by Hart and colleagues (2007, this issue) on the accuracy of predictions of violence based on actuarial assessments is of immense importance. The authors succeed where others have repeatedly failed (Mullen, 2000; Szmukler, 2001) in providing a clear, convincing and elegant demonstration of why you cannot simply move from group to individual risk evaluations. By the expedient of calculating the 95% confidence limits of group and individual estimates obtained from the VRAG and the Static–99 they expose a fatal flaw in the reasoning which underlies risk assessment. Their conclusions deserve to be in lights ‘At the individual level, [the margins of error] were so high as to render risk estimates virtually meaningless’. All actuarial risk instruments can properly be used for is to identify the group into which the individual is placed by their scores. Even here Hart et al identify a problem. In the VRAG, for example, the claimed nine ‘bins’, as they are so unfortunately termed are not significantly different from each other with at best only three groupings being statistically distinct. It is important to realise that this devastating critique will apply equally to other actuarial risk assessment instruments that are used to predict any particular individual’s likelihood of being violent in the future. This includes both the MacArthur Classification of Violence Risk (Monahan et al., 2003) and the so-called structured professional judgment instruments such as the HCR–20 (Webster et al., 1997). The caveat also applies to the PCL–R itself when you attempt to use it for prediction of risk. There is a word for assuming the characteristics you attribute to a group of people applies to each and every member, and that word is prejudice.

Does this mean that the whole literature on risk factors for criminal and violent behaviour and the multiple risk assessment instruments can now be consigned to the waste bin of history? Fortunately, or unfortunately according to your perspective, it does not. If you have a young man with a schizophrenic syndrome who misuses cannabis and alcohol, who is symptomatic, uncooperative with treatment, denying of illness, interpersonally callous and living a disorganised life in a high crime neighbourhood, then he is at risk of acting in an antisocial and violent manner. The risk will be reduced by moderating or removing the substance misuse, by improving symptom control, by stable accommodation in a low crime neighbourhood, by structuring his day with meaningful activity, and working on his attitudes towards others (Mullen, 2006). Whether this particular patient would ever have actually committed a crime is a moot point. A reasonable expectation is that if your services manage effectively all such patients with these risk factors then the total level of criminal behaviour committed by the patients as a group will fall. In the DSPD context this implies that if you have a large enough group of individuals sharing established risk factors for recidivist offending, and if you reduce the prevalence of the risks factors in the group, then you will reduce the future offending by the group as a whole. The risk assessment instruments become guides to whether intervention is justified, and towards which risk factors to address your efforts.

The margins of error in every actual, or conceivable, risk assessment instrument are so wide at the individual level that their use in sentencing, or any form of detention, is unethical. Perhaps legislators and judges will now return to that time-honoured principle of English law that you are sentenced for the acts you have committed not the acts you might commit. Health practitioners are, however, habituated to acting on the basis of probability estimates for groups applied to specific individuals. The patient has a right-sided lower abdominal pain moving centrally, combined with anorexia, guarding and rebound tenderness. The probability based on case series of an inflamed appendix may be 50% but the surgeon will almost certainly advise 100% that the patient be taken to theatre. This situation does not raise ethical or practical problems because: (a) the group-based probability estimate is being used exclusively for the benefit of the patient; (b) the patient is in a position to reject the advice based on that estimate.

As soon as we move away from using group-based probability estimates for the individual’s benefit and toward compulsion we are in ethical and practical difficulties. At the York conference there was a troubling assumption that the primary or even sole purpose of the DSPD Programme was community safety. This would only begin to be defensible if each individual inducted into the programme could reliably be ascertained as at high risk of future serious violence. It is now hopefully clear this cannot be accomplished. This need not affect the DSPD initiative. What is not sufficiently emphasised is the high price these people with DSPD pay for their own limitations and deviance. This group have a high mortality rate comparable to that of people with schizophrenia or bipolar illness (Ruschera et al., 1998). As to
morbidity, the fact that they are in prison on long sentences is evidence enough. Without help many will continue to lay waste their own lives as well as the lives of those around them. Offering treatment is legitimate. Offering inducements to bring them into treatment is legitimate, given the potential advantages to the wider community. Respecting their refusal is essential, as is avoiding adding any punitive measures in response to such a refusal. If the inducements are sufficient then there will be enough recruits. Those inducements will in part depend on successful participation in the programmes and become part of the therapeutic culture of reward.

ON HAVING DSPD

The paper of Tyrer et al (2007, this issue) challenges the assumptions which underpin the dominant categorisations of personality disorder in DSM–IV (American Psychiatric Association, 1994) and ICD–10 (World Health Organization, 1992) as well as many of those which sustain the psychopathy concept. Further they question one of the fundamental tenets of personality theory, that traits, and disorders, are stable over time. Edward Gibbon (1969 edn) appears to have been right when he wrote that there is nothing so unlike the man of today than that man yesterday. Tyrer et al argue that even the notion of qualifying a description of personality disorder with the term serious is problematic. Their conclusions do, however, hold out scintilla of hope. The way forward is suggested to be a focus not on the rickety notions of disorder but on personality function conceptualised as both dimensional and inherently variable over time.

What then of the key construct of the DSPD Programme, namely psychopathy? The paper by Cooke et al (2007, this issue) is steeped in the arcane language and statistics of the PCL–R industry. The paper is, however, well worth the effort of reading for the jewels caught up in the complex skeins of data and argument. Cooke et al suggest that psychopathy as a construct has become conflated with its putative measure the PCL–R which ‘forecloses on the possibility of examining the mapping of the theoretical construct (psychopathy) onto the empirical observations (PCL–R)’. The situation is, in fact, even more dire because thorough training is mandatory to employ the PCL–R. This training focuses on reliability, which amounts to ensuring raters see what they are supposed to see, and see it in the same manner. Which would be fine if there were some external validation of both construct and instrument, but there is no such gold standard (see also Duggan et al, 2007, this issue). The whole psychopathy/PCL–R enterprise teeters on the edge of falling into the logical fallacy of the ‘self-sealing argument’ (Fogelin & Sinnott-Armstrong, 1991; Hacking, 1995).

Cooke and colleagues emphasise that in the DSPD Programme individuals are detained because of the assumption of a functional link between their personality disorder and the risk they pose. Thus if the measure and the construct of psychopathy are not distinct and the personality measure incorporates the behaviour it is supposed to be accounting for we have a double circularity. Wootton (1959) many years ago remarked on the potential circularity inherent in the concept of psychopathy. To paraphrase ‘Why does he keep committing crimes? Because he is a psychopath. How do you know he’s a psychopath? Because he keeps committing crimes’.

Cooke and colleagues have proposed an approach which might both rescue the concept of psychopathy and generate a more independent and productive approach to its recognition and quantification. In the advantages they claim for their approach they omitted what seems to me the most important. The clear separation of the underlying personality vulnerabilities in psychopathy, and potentially their hierarchical arrangements, should allow the development of more focused and effective therapeutic interventions. In the end the conceptual framework may totter and fall, the instruments may be found wanting, but what really matters is will the DSPD service both aid a distressed and disturbed group of offenders and deliver a safer community?

ON BEING TREATED

The York conference had a large number of presentations about treatment issues. They were characterised by offering theoretical structures around which treatment programmes could be developed. Most were sensible, plausible, promising and without adequate empirical evaluation. The paper by Wong et al (2007, this issue) is the only representative in this issue of what was a major thrust of the conference. Stephen Wong has extensive clinical experience and advocates well structured and pragmatic approaches to assessment and management. He has been influential in the development of the DSPD therapeutic programmes, particularly in prison-based units. Dr Wong has also advised our service in Australia on programme development. Clinicians warm to his non-sense practical approaches and to the framework used to organise the evaluation and interventions. Clinicians tend to feel his approaches will be effective. Our clinicians are now comfortable with the language of responsivity principles, of the theoretical models of change and of acronyms such as VRP (Violence Reduction Programme) and VRS (Violence Risk Scale). This generates confidence and optimism, both of which are invaluable, and a belief that such approaches work, but where is the empirical evidence for such efficacy.

Wong and colleagues promise that results of outcome evaluation support the stated objectives of the programme and in the discussion write of outcomes which are encouraging and substantial. The sketchy presentation of the methods and statistical approaches make evaluating such claims problematic. However, if we put such quibbles aside and take the claims at face value then encouraging might be a fair description of the reported results. You would not be rushing to introduce a new pharmacological agent on the basis of such outcome data, but neither would you be abandoning further trials.

In her presentation Sheilagh Hodgins used the evocative image of the psychopath as the fearless child grown up (see Hodgins, 2007 and Vizard et al, 2007, this issue): the defiant child who will not back down in response to intimidation and/or punishment but will walk through fire to reach their desired goal. Such children do not necessarily turn into villains, clearly some become heroes, some captains of industry and, hopefully, some psychiatrists. These are reward-driven people for whom anxiety, if it arises, is a hurdle to be ignored or jumped. They are not frightened of social exclusion and accept intimacy only on their terms. They crave battle and learn early that the bruises and breaks inflicted by physical and emotional blows heal quickly enough and are of no real account. What is of account is the joy of inflicting damage on your adversary. Influencing such people depends on offering them rewards they value and avoiding becoming embroiled in.
their fights. Conversely, antisocial people who are anxious, suspicious and resentful can be controlled by intimidation and trapped into conformity by giving them something they fear losing. In clinical practice you rarely if ever see the psychopath totally untrammelled by guilt or anxiety, nor the anxious antisocial without elements of sensation-seeking and tendency for interpersonal exploitation. There may be enough truth in the stereotypes, however, to direct therapeutic strategies.

If the DSPD units continue to select those with marked psychopathic traits then the therapeutic context had better be reward driven. Threatening indefinite detention for failure to comply will be far less effective than offering early release for cooperation. Removing privileges will provoke defiance, offering valued rewards may motivate and produce progress. All contracts will be broken, but some freely given promises will be kept. Battles are inevitable, all that can be hoped is they are for the right causes. Recidivist offenders are in a rut where damaging reactions and rigid attitudes are repeatedly exhibited at inopportune moments. Greater flexibility and acquiring a larger repertoire of coping mechanisms and responses in the dedicated pursuit of personal goals is one route to decreased antisocial behaviour. In my experience it is easier to add to the behavioural and attitudinal repertoire of the individual with personality disorder than to inhibit or remove ingrained approaches. The changeability and plasticity of personality highlighted by Tyrer and colleagues is the basis for therapeutic hope. In the end, however, it is behaviour that matters not personality. The good news is that behaviour is easier to change than what we call personality. The even better news, if Aristotle was right, is that virtue is a habit to be acquired like any other skill by practice and repetition (Hutchinson, 1986; Aristotle, 1991 edn). Those who learn to behave well.

**REFERENCES**


Paul E. Mullen, DSc, Monash University, Clayton and Victorian Institute of Forensic Mental Health, Thomas Embling Hospital, Locked Bag 10, Fairfield, Victoria 3078, Australia. Tel: +61 3 9495 9136; fax: +61 3 9495 9195; email: paulmullen@forensicare.vic.gov.au


Dangerous and severe personality disorder: antecedents and origins

ANTHONY MADEN

Summary The origins of the Dangerous and Severe Personality Disorder (DSPD) Programme can be traced to developments in structured assessment and services for the cognitive – behavioural treatment of sexual and violent offenders in other countries. A comparison with these other services highlights the strengths and weaknesses of DSPD. The decision to use a medical model raises ethical and financial questions that may jeopardise the Programme’s future.

Declaration of interest None.

Politicians and civil servants invented dangerous and severe personality disorder (DSPD) in 1999. The announcement of a new service coincided with the conviction of a notorious offender, Michael Stone, but it is a mistake to attribute too much significance to this piece of political theatre. The true motivation was not a single case but longstanding frustration within government at the refusal of psychiatrists to address the problem of high-risk offenders with personality disorder. The profession was seen as cynically hiding behind the ‘treatability’ clause in the Mental Health Act 1983 in order to avoid responsibility for dangerous and difficult patients.

The DSPD diagnosis and service appeared suddenly but not from nowhere. Too much has been made of the claim that all is new and we know nothing, the attraction of the tabula rasa being that any old scribble can pass for wisdom. Certainty may be hard to come by but an educated estimate of the strengths and weaknesses of the DSPD Programme emerges readily from a study of its antecedents which include: (a) decreasing tolerance for the risks associated with crime and violence; (b) standardised risk assessment and the Psychopathy Checklist; (c) cognitive–behavioural programmes for sexual and violent offenders; (d) the Dutch TBS system; (e) the Violence Reduction Programme in Saskatoon, Canada.

DECREASING TOLERANCE FOR RISKS ASSOCIATED WITH CRIME AND VIOLENCE

Blaming the government is the UK’s fastest growing indoor sport, but decreasing tolerance of violence risk is a global phenomenon that extends far beyond mental health. We assumed that the violent, like the poor, would always be with us, until California and New York City, along with other parts of the USA, broke with this consensus in the 1990s and achieved significant reductions in crime rates. The costs were high as the prison population soared, but voters saw it as a bargain and other countries have adopted similar policies. Most changes in the UK have been to criminal sentencing. Although mental health has not been singled out for special treatment it could not expect to remain untouched by such sweeping social changes. The pressure on mental health services comes from new expectations rather than new law, and the only legal change is likely to be loss of the ‘treatability’ clause in the Mental Health Act.

STANDARDISED RISK ASSESSMENT AND THE PSYCHOPATHY CHECKLIST

Without the Psychopathy Checklist – Revised (PCL–R; Hare, 1991) there could have been no DSPD service. The diagnosis of psychopathy has a long history but had fallen into disuse for want of a reliable diagnostic method. Much of the problem lay with the DSM diagnosis of antisocial personality disorder which was hopelessly confounded with criminality (American Psychiatric Association, 1994). Hare resurrected the concept of psychopathy by developing a simple yet reliable and valid method for diagnosis.

Although the DSPD Programme embraced the concept of psychopathy it did not heed the advice of most practitioners about the implications for treatment. High psychopathy scores suggest poor response to all aspects of treatment. The evidence base, scanty though it is, favours treatment of people with a range of scores rather than an attempt to treat large numbers of people with high psychopathy scores together. The decision to go against this evidence is probably the greatest strategic risk taken by the programme.

The development of the PCL–R coincided with massive growth in the risk assessment industry. It began as a welcome improvement on unstructured clinical assessment, but was damaged by overenthusiastic claims that actuarial methods should replace inferior clinical approaches. The debate has settled to a consensus view that structured clinical judgment is the best way of assessing violence risk in an individual. Structured methods should inform the clinical judgement but cannot replace it (Monahan, 1992).

The DSPD service is ahead of most UK forensic services in its use of standardised measures of personality disorder and risk. One of the strongest aspects of the service is its use of dynamic measures to assess the progress of treatment, which is less contentious than the use of risk measures in sentencing. On the negative side, planners and civil servants expect too much from actuarial measures and worry when a man with ‘the right scores’ is not considered suitable for admission, or vice versa. Further education is needed and Mullen (2007) has made a good start.

Fortunately when it comes to detention in hospital, science and the law are on the same side. Human rights legislation ensures that doctors have the final say in deciding on ‘unsound mind’ – which provides reassurance against the spectre of politicians using mental health services as a means of social control, but raises a more difficult series of questions for the profession. How do doctors define treatment needs in personality disorder? Which treatments do doctors believe are suitable for which types of patient? When and how do doctors judge that coercive treatment is likely to be effective? The Mental Health Act 1983 allowed doctors to opt out of involvement with patients with psychopathy, so many
doctors are considering these issues for the first time. There is no consensus within the profession. Some experienced psychiatrists recommend treatment of patients with DSPD whereas other experienced psychiatrists advise tribunals that these same patients are untreatable. Standardisation still has some way to go in this field.

COGNITIVE-BEHAVIOURAL PROGRAMMES FOR SEXUAL AND VIOLENT OFFENDERS

The criminal justice system of the 1980s laboured under the pessimistic belief that nothing could be done about the tendency of offenders to carry on reoffending until they grew out of it or died. This approach was not helpful for prisoners or prison staff. In the 1990s the pendulum swung the other way. The search for ‘What Works’ was a corrective to the nihilism that had gone before – and the answer to that question appeared to be cognitive-behavioural programmes.

Offending behaviour programmes have swept through the criminal justice systems of several countries, including the UK (see McGuire, 1995). The prison system is ahead of the National Health Service, where forensic services emphasised medical treatment of the underlying disorder and were slower to take up cognitive-behavioural therapy (CBT). Some hospital practitioners even argued that the prevention of reoffending should not be a primary goal of forensic mental health services, although this argument appears to have been resolved in favour of the common sense view that even the best mental health treatment serves little purpose if it is rapidly followed by reoffending and a return to custody. Most forensic services are now adopting cognitive-behavioural programmes similar to those found in many prisons.

If the criterion is number of customers then cognitively-behavioural/offending behaviour programmes are a smash hit that will run and run. However, the scientific evidence falls short of a rave review and there is no adequate evidence from randomised controlled studies. The following examples refer to the treatment of sex offenders but there is no reason to hope for better results in violent offenders. Moreover, as violence has always been more socially acceptable than sexual offending, treatment of violent offenders will probably be more difficult.

Hanson et al (2002) reviewed 43 studies that included at least a matched, untreated group, yielding a total of 9454 sexual offenders (5078 treated and 4376 untreated). Meta-analysis showed a significant reduction in recidivism rates in the treated group: treated v. untreated recidivism rates were 12.3 v. 16.8% for sexual offending and 27.9 v. 39.2% for all offending. If the analysis includes only treatments meeting current standards for offending behaviour programmes the results are better: 9.9 v. 17.4% for sexual recidivism and 32 v. 51% for all offending.

The existence of an apparently effective treatment for the management of sexual offenders has led to optimism, but we need to keep it within proper bounds. Marshall & McGuire (2003) emphasise that we do not know ‘with which types of offenders’ treatment is most likely to be effective (p. 654). We can assume that outcome will be best in those whose offending was confined to a domestic setting, in the absence of personality disorder or an otherwise deviant lifestyle. Treatment is most likely to succeed in those who have most to lose in terms of employment, self-image and social standing. It is likely to be most difficult in predatory offenders, and in those with ‘stranger’ victims and high psychopathy scores – in other words, those for whom DSPD services were designed.

Brooks-Gordon & Bilby (2006) highlight the ‘enormous political and institutional pressure to prove that treatment works’ but note the methodological limitations of the evidence to date as well as the rather poor results revealed by the Cochrane meta-analysis (Kenworthy et al, 2004). Cognitive-behavioural therapy may be the best hope for success in DSPD treatment but further results are awaited.

Enthusiastic advocates of ‘What Works’ should also bear in mind that although CBT has a growing reputation as the wonder treatment for all psychological problems more evidence of its efficacy is required. A good example of this need for caution would be its use in the treatment of schizophrenia. A review of 20 randomised controlled trials indicated modest effect sizes, with the strongest evidence available for chronic patients (Tarrier & Wykes, 2004). It may be too early to say, but nobody should be surprised if similar conclusions emerge in relation to CBT in the population with DSPD.

If this all sounds very negative, it is important to remember the positive aspects of CBT. Its structured approach and comprehensibility make it suitable for a difficult client group. Its explicit procedures and aims facilitate evaluation. At present it is probably still the best available treatment but it has much to prove.

THE DUTCH TBS SYSTEM

An important stimulus for the development of the DSPD Programme in the UK was the feeling that the Dutch, just across the North Sea, were doing things so much better. The TBS system has been managing high-risk violent and sexual offenders in institutions and in the community since 1928 and nobody could disagree with the suggestion that we have a lot to learn from them. The big question is precisely what we should learn. The danger is that we use their approach as a giant Rorschach ink blot in which we see only what suits us but ignore the obvious and inconvenient.

Under TBS legislation, offenders convicted of a serious sexual or violent offence and judged to present a high risk of reoffending can be sentenced by the criminal court to a TBS order. They serve a prison sentence appropriate to the offence and are transferred to a TBS facility for treatment at the end of that sentence. They remain within the TBS system indefinitely (subject to regular review by a tribunal), first in a secure institution and later as conditionally discharged, supervised patients in the community. Dutch courts rarely give a sentence of life imprisonment and the TBS order is in many ways a substitute.

Treatment within the TBS system is eclectic, with a strong emphasis on therapeutic community principles and on work: patients are expected to spend about half the week in paid employment. In recent years the service has embraced the PCL–R and other standardised measures, as well as CBT offending behaviour programmes. Antilibidinal medication is commonly prescribed to sex offenders and accounts for a considerable proportion of medical input, as most other treatments are delivered by psychologists or specially trained (non-medical) therapists. There is a well-developed community service and most patients progress rapidly to supervised leave from the in-patient units.

The civil servants who created the DSPD Programme were obviously impressed by the TBS system so it is remarkable how little of it they chose to adopt. No new legislation was introduced so the
criminal courts play no role in sending offenders to DSPD units. Paid work remains deeply unfashionable within most English psychiatric hospitals, and in the notoriously fashion-resistant high-secure estate its availability is severely restricted by security considerations. The DSPD Programme developed as a high-security in-patient service with no clear pathway to the community.

The most important difference may be that the DSPD Programme lacks the legitimacy that the Dutch courts give to TBS. English and Welsh life-sentenced prisoners are in a similar position to Dutch TBS patients, but prisoners with determinate sentences feel aggrieved to be transferred to hospital during their sentence then detained beyond its end. The grievance has some justification. The (historical) information to support detention on the grounds of risk was available at the time of sentencing, so it is reasonable to ask why indeterminate detention is appropriate now if it was not considered appropriate by the sentencing judge.

A further problem arises from the Mental Health Act 1983 which requires that treatment be likely to ‘ameliorate or prevent deterioration’ if a patient is to be detained on the grounds of psychopathic disorder. This unfortunate wording has led many patients to seek discharge on the grounds that they are untreatable, a quest in which they are often supported by their legal representatives and some doctors who testify to tribunals. In practice, tribunals do not discharge many high-risk patients, but much time and money is wasted along the way. It may be many years before the realisation dawns that the tribunal route is not the best way out of hospital, and tremendous damage can be done to the service if those years are spent by the patient in a determined effort to prove intractability. By contrast, the TBS system delivers a clear message that the only way out is by demonstrating reduced risk, hence sustained non-adherence to treatment is rare. The viability of DSPD services may depend on a change in UK law that brings similar clarity.

Another major difference is that TBS units are not dominated by doctors and nurses. Conventional medical attitudes are inappropriate for people without a mental illness. English high-security hospitals have tended to infantilise patients and remove all responsibility from them. Such an approach is unhelpful to any patient, but it is disastrous in a personality disorder service whose primary aim is to encourage patients to accept responsibility for their own behaviour.

The TBS units can also teach us about long-stay patients. About 20% of their patients are judged unlikely ever to be discharged after an initial treatment period of about 6 years. The assumption is that they will be detained indefinitely in units that give priority to quality of life. As the DSPD Programme has higher threshold criteria than the TBS system, it is likely that a much higher proportion of DSPD patients will never be considered suitable for discharge. Unless plans are made for them they will become a source of dissatisfaction and instability within the service.

THE VIOLENCE REDUCTION PROGRAMME IN SASKATOON, CANADA

The Violence Reduction Programme operates within a regional psychiatric centre in Saskatoon, a specialised unit of the Canadian correctional System, and it has been contrasted with the DSPD Programme (Maden et al, 2004). It is important as a model of a service based on cognitive-behavioural principles and using standardised measures of change.

There are two major differences from the DSPD Programme. First, the unit is part of the prison system, so all patients are serving prisoners who have volunteered for the service and can be sent back to ordinary prisons if they are violent within the unit or if treatment is not progressing. Second, the Canadian correctional system includes a ‘Supermax’ prison that effectively provides back-up for dealing with the most disruptive or violent prisoners who are not suitable for the regional psychiatric centre. Such a facility exists within the UK prison system and can provide support for prison DSPD units, but there is no equivalent resource within the high-security hospitals. The techniques required to deal with violent and disruptive prisoners inevitably involve restraint and are quite different from the techniques used within the regional psychiatric centre in Saskatoon, where any violent act results in transfer out. The Canadian system manages these types of offender separately because they have different needs. The DSPD Programme may run into problems because it has committed itself to managing them together in the hospital units. The skills, techniques and building needed for the two groups are very different. Hence the DSPD Programme is effectively committed to providing two parallel services when one would provide more than enough of a challenge.

THE FUTURE

Missing from the list above are those services that had little influence on the development of the DSPD Programme and deserved consideration. Chief among these is the discretionary lifer system in the UK, consisting of offenders who were given a life sentence not because of their offence alone but because of their perceived risk. Most had psychopathy, and the system has a good record of safe management and rehabilitation through the criminal justice system, with little medical input. There was scope to increase that input without radical change to the system. Anecdotal evidence suggests that many men within this system with personality disorders spent several years fighting the system before settling in to an active programme of rehabilitation. The challenge for DSPD services is to show that they can do better with similar men, and they face a major handicap from the stigma attached to the label.

Comparison with other services shows that the architects of the DSPD service took some major strategic risks when they decided to go down a different road. They took those risks in a high-stakes game, and the most important risk may turn out to be financial. The cost of DSPD treatment in high-security hospitals is over £200 000 per patient per year, a sum that can buy a lot of risk management in other settings. People who work in DSPD programmes already feel their practice is subject to close examination, but it may be nothing compared with the scrutiny heading their way from the accountants.

REFERENCES


Persistent violent offending: what do we know?

SHEILAGH HODGINS

Summary  A great deal is known about men who display a stable pattern of antisocial behaviour since childhood. However, more research is needed to identify subtypes within this population so as to further understanding of the causal processes that initiate and maintain violent behaviours and to identify interventions that specifically target the deficits presented by each subtype. Evidence-based practice means not only using treatments proven to be effective but also basing conceptualisations of disorders on scientific evidence.

Declaration of interest  None.

We know that many more men than women become persistent violent offenders and that a very small group, less than 5% of the male population, commit between 50% and 70% of all the violent crimes (Moffitt, 1993; Hodgins, 1994). These men fulfil criteria for diagnoses of conduct disorder before 10 years of age, antisocial personality disorder and psychopathy (according to the Psychopathy Checklist–Revised (PCL–R); Hare, 1991) in adulthood, and have been labelled as life course persistent offenders. This population, however, is heterogeneous. Identifying the distinct subgroups is essential to unravel the complex and dynamic interactions of biological, psychological and social factors that initiate and maintain the violent behaviour over the life span and to establish interventions that address the specific deficits characterising each subgroup.

SUBGROUPS OF PERSISTENT VIOLENT OFFENDERS

Within the population of persistent violent offenders with a stable pattern of antisocial behaviour since childhood, evidence suggests that there are at least three subgroups with distinct aetiologies and response to treatments. This population first divides in half based on the presence or absence of comorbid anxiety disorders (Sareen et al., 2004). This evidence from adults is consistent with developmental studies of young children showing that among those with early-onset conduct problems, one subgroup displays high levels of internalising problems and another does not (Gilliom & Shaw, 2004). The non-anxious half of this population contains a subgroup who present the syndrome of psychopathy as diagnosed by the PCL–R, which requires full-blown expression of the traits of arrogant and deceitful interpersonal behaviour and deficient affective experience (Cooke & Michie, 2001) and a much larger group who do not meet the diagnostic cut-off for psychopathy, but who obtain higher scores on the traits than men in the general population. By contrast, the highly anxious males with antisocial personality disorder or conduct disorder present a distinctively different set of traits.

During the past 25 years, evidence has accumulated that men within this population with psychopathy diagnosed according to PCL–R differ markedly from the others. They commit more violent offences, engage in both instrumental and reactive aggression and more quickly reoffend when released. They display a profound emotional dysfunction characterised by impaired recognition of fear and sadness in faces and voices, low levels of stress reactivity indexed by heart rate, skin conductance and cortisol, and a cognitive style dominated by reward and a failure to recognise punishment. Performance on IQ and standard neuropsychological tests is within normal limits (Patrick, 2006).

Among children with early-onset conduct disorder, there is a subgroup with stable, callous–unemotional traits that are thought to represent the antecedents of the psychopathic traits. There are as yet no studies that have followed such children into adulthood, but the available evidence suggests that boys who present both conduct disorder and callous–unemotional traits share many features with adult psychopathy. Compared with boys with conduct disorder alone, boys with both conduct disorder and such traits present more severe conduct problems, more aggressive behaviour and violent crimes at an earlier age. They demonstrate a preference for risky behaviours that are novel and exciting to them, less reactivity to threatening and emotionally distressing stimuli and a muted cortisol stress response. They show less ability to recognise sadness in faces and vocalisations and are more impaired on tests of moral reasoning and empathy. In addition, these boys have difficulty changing behaviours that are initially rewarded and subsequently punished and emphasise the positive consequences of aggressive behaviour but fail to appreciate the negative consequences. Not surprisingly given these characteristics, the boys with both conduct disorder and callous–unemotional traits are less responsive to parenting practices and most particularly to punishment. In addition to the emotional deficits, children with callous–unemotional traits also display cognitive abnormalities similar to those observed among adult men with psychopathy, and have higher IQ scores than other boys with conduct disorder (Blair, 2003; Frick & Marsee, 2006).

Boys with conduct problems who do not present callous–unemotional traits display aggressive behaviour that is emotionally charged in response to provocations that may be real or result from their tendency to perceive hostility in others, even in neutral faces (Dadds et al., 2006). These children report emotional distress, are more reactive to distress and especially to negative emotional stimuli. They are, however, responsive to good parenting practices and benefit when their parents complete parent training programmes (Hawes & Dadds, 2005). These boys present high levels of impulsivity, a tendency towards anger and like children with anxiety disorders have lower than average verbal abilities (Frick & Marsee, 2006).

The place of attention-deficit hyperactivity disorder (ADHD) in persistent violent offending and in the characterisation of the proposed subgroups remains unclear (Washbusch, 2002). This is often comorbid with conduct disorder and the combination
is predictive of criminality in adulthood. Among boys with conduct disorder and ADHD, it is callous–unemotional traits and not ADHD symptoms that are associated with aggression and delinquency. Further, boys with such traits without ADHD have been shown to have the highest rates of aggressive behaviour and delinquency (Frick & Marsee, 2006). Recent evidence demonstrates that both the genotype and the phenotype of ADHD are heterogeneous, and that IQ mediates, at least in part, psychosocial functioning, including violent offending, in adulthood (Mill et al., 2006). Boys with conduct disorder and ADHD obtain lower IQ scores and are more impulsive than boys with only conduct disorder. Inattention precedes the onset of conduct disorder, allowing for the possibility that early intervention could interrupt the development of conduct disorder (Simonoff, 2000; Waschbusch, 2002).

Available evidence suggests that the aetiology in the three subgroups of males with early-onset stable antisocial behaviour differs. A meta-analysis of twin and family studies reported a modest genetic contribution (0.41 heritability) to the development of early-onset and stable antisocial behaviour, including substance misuse (Rhee & Waldman, 2002), but depending on the definition of the phenotype other studies report much higher heritability coefficients (Young et al., 2000). Recent evidence suggests that callous–unemotional traits are also heritable. For example, a recent study of young twins in the UK has shown that the combination of conduct problems and callous–unemotional traits is much more heritable (heritability coefficient 0.81) than conduct problems alone (heritability coefficient 0.30) (Viding et al., 2003). This is consistent with studies of older participants showing high heritability for psychopathic traits (Blonigen et al., 2003; Taylor et al., 2003; Larsson et al., 2006). These results indicate that from conception onwards individuals who display early-onset stable antisocial behaviour differ from the rest of the population and that the subgroup who will develop callous–unemotional traits differ from the others. Genes modify the individual’s interaction with their environment, hence each of the subtypes would be reacting to and selecting environments differently from conception onwards. Interestingly, this would mean that the subgroups would react differently during the prenatal period when individual thresholds for stress reactivity of the hypothalamic–pituitary–adrenal axis are set (Susman, 2006). This could be one of the reasons why one sub-group displays high anxiety levels and another presents abnormally low levels.

As is hypothesised for other complex disorders, each of the genes involved in antisocial and aggressive behaviour may interact with a specific environmental factor to determine outcome. For example, a functional polymorphism in the promoter of the monoamine oxidase A gene had previously been associated with persistent aggressive behaviour in animals and in one human pedigree. In a birth cohort, neither the low nor high activity allele was associated with violent behaviour in adulthood. The males in the cohort who carried the low activity allele and who experienced physical abuse during childhood were three times more likely than the men with the same allele who had not experienced abuse to present conduct disorder and ten times more likely to commit violent criminal offences. Physical abuse in the absence of the gene did not increase the risk of conduct disorder or violent crime (Caspi et al., 2002). Two studies have replicated this finding in White males (Foley et al., 2004; Widom & Bzustowicz, 2006). Thus, individuals who differed in genetic profile reacted differently to their environments, and even to a severe event such as physical abuse.

**IMPLICATIONS OF THE FAILURE TO IDENTIFY SUBTYPES**

The failure to distinguish subtypes within the population of males who show an early-onset and stable pattern of antisocial and aggressive behaviour blocks progress in research aimed at furthering understanding of persistently violent offenders and in efforts to identify effective treatments. Consider for example, studies using single-photon emission computed tomography (SPECT), magnetic resonance imaging (MRI), and functional MRI to investigate brain structure and functioning. These studies have focused largely on men who met criteria for psychopathy. Participants have been characterised using different cut-off scores and different scales, and little attention has been paid to the composition of the comparison group. Results have been inconsistent and difficult to interpret. One reason might be that the comparison group included offenders with and without anxiety disorders. A recent fMRI study comparing boys with and without conduct disorder illustrates the problem. Contrary to the hypothesis, no differences in activity in the amygdala were observed when viewing negative pictures. This absence of a difference, however, was due to high anxiety scores among some of the boys with conduct disorder. Post hoc analyses showed that anxiety scores, as would be expected, were related to amygdala activation and aggressive behaviour scores were related to amygdala inactivity (Sterzer et al., 2005). These results clearly suggest that a more accurate description of the subtypes is needed to unravel the aetiology of early-onset stable antisocial behaviour.

The failure to characterise subtypes within this population may also lead to difficulty in interpreting the evaluations of treatment programmes. It has been known for a number of years that parent training programmes reduce conduct problems among young children (Farmer et al., 2002). A recent study showed that children with conduct disorder but without callous–unemotional traits benefited most, whereas those with such traits only learned when reward was used. Time-out, although an effective intervention for the children without callous–unemotional traits, failed to reduce inappropriate behaviours displayed by the children with such traits (Dadds et al., 2005). Insensitivity to punishment is a key feature of psychopathy and of children with conduct disorder and callous–unemotional traits, but not of the other subgroups (Dadds & Salmon, 2003). Taking account of the presence or absence of such characteristics when developing treatment programmes will increase their specificity and thereby effectiveness.

Although several offender rehabilitation programmes have been shown to be as effective as most accepted medical treatments, there is little evidence about the characteristics of offenders who benefit and those that do not (Welsh et al., 2002). Richer and more complete characterisations of the subtypes of persistent violent offenders would lead to the development of treatments that directly target deficits. For example, one of the characteristics of boys with conduct disorder and callous–unemotional traits and offenders with psychopathy is their altered perception of reward and punishment. Both in neuropsychological tests and in real-life situations, they focus on rewards and ignore punishments. Consequently, they persistently miss
the signal – the punishment – that a behaviour is inappropriate. As children, this may be one of the key mechanisms that promotes their antisocial behaviour and that limits their access to the usual socialising experiences such as sports and other community activities, and eventually even to school. The problem persists into adulthood and is present, for example, later in life when they are incarcerated and enrolled in an offender rehabilitation programme. Many of these programmes include a module designed to develop problem-solving skills. Problem-solving is divided into four steps. Step one involves identifying the problem, step two generating as many responses as possible, step three assessing the likely positive and negative consequences of each possible response, and the final step selecting the best response. At step three, men with psychopathy and boys with conduct disorder and callous-emotional traits would focus on rewards and be less able than others to identify possible negative outcomes. To help them learn to solve problems would first require teaching them to identify negative outcomes. For the adults, this may be particularly difficult as they have had so many years without this skill.

REFERENCES

Value of standard personality assessments in informing clinical decision-making in a medium secure unit

CONOR DUGGAN, LAUREN MASON, PENNY BANERJEE and JOHN MILTON

Background Assessing those with personality disorder for treatment in secure settings is known to be unsatisfactory.

Aim To examine the utility of a standardised assessment of offenders with personality disorder referred for treatment in secure care in a naturalistic study.

Method A consecutive series of 89 men were assessed with a battery of four recommended instruments measuring personality and risk. Decisions on whether or not to admit were based on a multidisciplinary discussion informed by these assessments.

Results Of the 89 comprehensively assessed referrals, 60 (67%) were offered admission. High scores on the Psychopathy Checklist—Revised (especially on Factor I) was the only measure that was associated with rejection. Of 44 patients discharged, 29 (66%) failed to complete treatment; none of the pre-admission assessments distinguished ‘completers’ from ‘non-completers’. Although skills were acquired on the unit, follow-up of 24 men in the community showed that this had only a marginal effect on re-offending rate (58%).

Conclusions Current recommended assessment methods appear unsatisfactory in identifying those who either (a) complete treatment or (b) benefit from treatment. Our results throw doubt on their value.

Declaration of interest None.

The detention of those with personality disorder for treatment in secure institutions has been criticised in the past because the assessments were considered too subjective with insufficient use of standardised measures (Reed, 1994; Fallon et al 1999). This has been supported empirically by Collins (1991), Berry et al (1999) and Milton (2000).

This lack of a standardised process led Reed (1994) and the Dangerous and Severe Personality (DSPD) Programme (Home Office & Department of Health, 2001) to recommend that personality disorder be assessed with multiple standardised measures. Despite these recommendations, there have been few empirical reports of their implementation. Milton et al (2007), in a preliminary study, described the results of this process in a medium secure setting. We expand on this sample and evaluate the process both with regard to treatment adherence and short- to medium-term outcome following discharge.

METHOD

The data for this report come from a dedicated facility that provides interventions for mentally disordered offenders with a personality disorder, the Personality Disorder Unit at Arnold Lodge in Leicester. This provides brief interventions to men with a personality disorder who are transferred primarily from prison, but also from the National Health Service with the intention of reducing their rate of re-offending and improving their social functioning. Those admitted had to have at least one personality disorder according to DSM–IV (American Psychiatric Association, 1994) or ICD–10 (World Health Organization, 1992) and be willing to have treatment. The main exclusion criteria were IQ < 80, history of psychosis, a score of 25 or more on the Psychopathy Checklist–Revised (PCL–R; Hare, 1991) and a history of sadistic offending. These criteria were adopted in the hope of ensuring that those selected for the unit had a reasonable chance of responding to the treatment offered.

Assessment process

Referrals were made by medical practitioners in forensic or prison settings. Each referral was initially screened by a consultant psychiatrist to determine if there was sufficient evidence to proceed to a multidisciplinary assessment. Referrals deemed as inappropriate were screened out at that stage and the remainder had an independent assessment from the psychiatrist, psychologist and nurse attached to the unit. The psychiatrist, in addition to obtaining the usual psychiatric history, administered the Schedule for Affective Disorders and Schizophrenia–Lifetime Version (SADS–L; Endicott & Spitzer, 1979) or Structural Clinical Interview for DSM–IV Disorder (SCID–I; First et al, 1997) to determine the presence of important mental state (Axis I) disorders and the interview version of the International Personality Disorders (IPDE; Loranger et al, 1994) to document personality (Axis II) psychopathology. The psychologist assessed intellectual ability with the Wechsler Adult Intelligence Scale (3rd edn, revised) (WAIS–R; Wechsler et al, 1998) and psychopathy with the PCL–R (Hare, 1991). The nurses assessed risk with the Histrionic/Clinical/Risk Management 20-item scale (HCR–20; Webster et al, 1997), and motivation and manageability using prison records and interviewing staff. Each professional made an independent recommendation as to whether or not the individual ought to be admitted and the final decision was made by the full multidisciplinary team.

Treatment

The general principle underpinning the treatment approach was that these individuals offended because they lacked the necessary skills to do otherwise. Hence, any augmentation of their skills repertoire (e.g. improved problem-solving, social skills or better anger management) ought to lead to a reduction in their rate of re-offending. Treatment was based primarily on principles of cognitive–behavioural therapy informed by a therapeutic community approach modified for a secure setting. Specific provision was made to tackle criminogenic needs. Treatments were...
generally provided in a group which was mostly nurse led. Individual work also took place if this was indicated. A battery of psychometric tests was completed at each care programme approach or case conference meeting (i.e. 3 months after admission and 6 months thereafter).

The programme accepted individuals for a maximum of 2 years, after which they were discharged either back to prison, hospital or the community. All participants volunteered for treatment and could choose to leave the unit and return to the host institution at any time during their admission. There was an expectation that they would engage actively with the treatment regime. The unit had a zero tolerance for physical violence, the use of illicit drugs or alcohol. All these rules were actively enforced so that any departures resulted in individuals being returned to their host institution (usually prison). Those discharged prematurely were usually offered a second opportunity to be readmitted after a period of reflection.

Follow-up
As outcome is one of the validating criteria of any classification process (Robins & Guze, 1970), we selected adherence to treatment and a reduction of re-offending after discharge as the main criteria to evaluate the assessment process. Thus, we assumed that those who were deemed to be suitable for the unit, following this complex battery of assessments, would be likely (a) to complete treatment and (b) to benefit from it. Hence, we examined the completion rate of treatment among those admitted and followed-up all those discharged from the service (whether they had completed or not) annually for five years after discharge. The follow-up usually comprised a face-to-face interview that documented the mental health status over the previous 12 months, any criminal activity and completion of relevant psychometric tests. Permission for the follow-up study and for the use of routine clinical data for research purposes was obtained from the North Nottinghamshire Ethics Committee.

RESULTS
Comparison of those who were and were not accepted
There were 122 men referred to the service between its opening on 1 February 1999 and 30 September 2005 (the time of completion of this report). Of these, 33 met one or more of the exclusion criteria, 13 (38%) either had personality disorder or a history of psychosis or both, 11 (34%) were deemed to lack motivation and 5 (14%) had a low IQ. The remaining 89 were seen by the multidisciplinary team. These 89 men had a mean age of 27.9 years and suffered from multiple disadvantages. Their educational attainment was poor (only 18 (20%) had obtained at least one GCSE), they had high rates of sexual or physical abuse (65, 73%) often while in the care of social services) and were usually violent at their index offence (73, 82%). There was a high frequency of Axis 1 disorders with major depression (42, 47%), drug misuse (37, 42%) and alcohol dependency (26, 29%) being the most common diagnoses. On the IPDE (interview version), antisocial personality disorder was diagnosed in 57 (64%), borderline in 43 (48%), paranoid in 23 (26%) and avoidant personality disorder in 19 (21%). Several different personality disorders were frequently diagnosed in a single individual, so that 37 (42%) satisfied the criteria for diffuse or complex personality disorder (Tyrer & Johnson, 1996). On the PCL–R, the mean total score was 19.7 (range 6–32), with scores of 6.7 (range 1–17) and 11.5 (range 2–15) on Factors 1 and 2 respectively. Most of those assessed were at high risk of future violence, with a mean HCR–20 score of 27.1 (s.d.=5.5).

Of the 89 men assessed by the multidisciplinary team, 60 were offered admission (67%) and 29 were declined admission (33%). Table 1 shows the characteristics of those who were and were not accepted for admission. Those with a high total PCL–R score (and especially a high score on the PCL–R Factor 1) and a high HCR–20 score were unlikely to be offered admission.

Adherence to treatment
Adherence to treatment was poor. Of the 44 patients who had been discharged from the unit by 30 September 2005, 29 (66%) left treatment prematurely, with only 15 (34%) of those discharged completing the treatment originally recommended by the multidisciplinary team. The reasons for this failure comprised: (a) 4 (9%) who were subsequently found to be inappropriate for treatment despite being deemed suitable at the initial assessment; (b) 3 (7%) who chose to leave despite clinical advice to the contrary; (c) 11 (25%) who disengaged from the programme; (d) 6 (14%) who showed violent behaviour; and (e) 5 (11%) who indulged in illicit drug taking. None of the original assessment variables was useful in identifying those who were likely to complete treatment. While on the unit, those in the treatment programmes showed a positive response as measured by self-report.

Follow-up
Only 1 of the 44 patients discharged refused to be followed-up after discharge and by September 2005 (5 years after the first discharge), 37 continued to be part of the follow-up process. By that stage, 6 (14%) had either withdrawn their consent for any further follow-up, could not be traced or had died (1 from a heroin overdose).

The rate of re-offending (primary outcome) at 5 years showed that of the 24 that

Table 1 Characteristics of those who were and were not accepted for admission.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Men offered admission</th>
<th>Men not offered admission</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n=60)</td>
<td>(n=29)</td>
<td></td>
</tr>
<tr>
<td>PCL–R: mean (range; s.d.)</td>
<td>18.7 (6–31; 5.9)</td>
<td>21.9 (9–32; 6.3)</td>
<td>0.03</td>
</tr>
<tr>
<td>Factor 1</td>
<td>5.7 (1–17; 2.9)</td>
<td>9.1 (1–16; 3.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Factor 2</td>
<td>11.2 (2–20; 3.8)</td>
<td>11.9 (5–25; 4.0)</td>
<td>0.7</td>
</tr>
<tr>
<td>HCR–20: mean (range; s.d.)</td>
<td>26.4 (7–38; 5.6)</td>
<td>29 (18–36; 5.1)</td>
<td>0.07</td>
</tr>
<tr>
<td>WAIS–R: mean (range; s.d.)</td>
<td>88 (62–131; 13.3)</td>
<td>87.6 (69–126; 14.5)</td>
<td>0.54</td>
</tr>
<tr>
<td>Physical/sexual abuse, %</td>
<td>76.7</td>
<td>65.5</td>
<td>0.27</td>
</tr>
<tr>
<td>Schooling (at least GCSE), %</td>
<td>15</td>
<td>13.7</td>
<td>0.33</td>
</tr>
</tbody>
</table>

were alive and had lived at some time in the community (and hence were at risk of re-offending), 10 (42%) had not reoffended, 12 (50%) had reoffended to a similar or lesser degree compared with their index offence, and 2 (8%) had reoffended more seriously. Although there was roughly a similar num-
ber among those who did or did not reoffend who completed the programme, both men who reoffended more seriously were non-completers. Despite the relatively high rate of reoffending, there was evi-
dence that at least to 1 year of followup (when we had the largest sample) the gains on social problem-solving on the self-report Social Problem-Solving Inventory (SPSI; D’Zurilla & Nezu, 1990) persisted.

DISCUSSION

We present some preliminary data on an inpatient service set up to treat men with a personality disorder and a history of off-
fending. Although the results from our as-
essment will not surprise those who work with this group (i.e. that they have multiple disadvantages and a broad range of psycho-
pathology), to our knowledge this is the first systematic description of a complete cohort referred for inpatient treatment using measures with scientific acceptability. Although there are major limitations, our main findings were that: (a) few of these measures helped to discriminate between those who were or were not admitted; (b) of those who were admitted, there was a considerable attrition (66%); and (c) although patients appeared to gain skills during the inpatient admission, some of which persisted during the followup, at least to 1 year, their impact on the rate of reoffending was modest.

Limitations

This is a prospective naturalistic descriptive study and hence is subject to the limitations of such a design. First, the continued collection of new data might result in these preli-
nary results changing in the future, particularly as the small size of this initial data-set makes it sensitive to changes in whatever data are subsequently entered. For instance, if a few more men were con-
icted of reoffending (or a few less), this would result in a major change in the pro-
portions in the reoffending outcome cate-
gory. Second, we are describing a clinical service that is likely to take account of regular feedback on its effectiveness and change its practice accordingly. There was evi-
dence that this indeed occurred: when the results of the early followup demon-
strated that reoffending was a common outcome, a specific programme was intro-
duced that focused on criminogenic needs to reduce such high rates of reoffending. However, a proper evaluation of the impact of this programme will take some further time.

Third, the absence of a control group makes interpretation of the data difficult. For instance, is our rate of reoffending (58%) in those with a personality disorder on discharge over a variable followup per-
iod either high or low? We simply do not know as comparable data from other studies are not available. However, we do know that 58.2% of prisoners released in 2001 (the period during which this study took place) were reconvicted of a standard list offence within 2 years (Home Office 2002), a rate that is equivalent to our find-
ings. Fourth, although the assessment provided guidance as to who should or should not be admitted, such guidance was applied loosely, with a degree of clini-
cal override, and not as would have been the case in a clinical trial.

Pre-admission assessment

Examination of the data in relation to ad-
mision shows that few of the measures separated the groups, apart from a high score on the PLC-R (especially a high score on Factor 1), which led to the individual being rejected. (This was not a surprise as a high score on the PLC-R was one of our exclusion criteria.) However, these comparisons of mean scores tell us very little about the decisionmaking process, as these measures were combined to provide a composite score which was used to decide whether or not to admit. Thus, it is impos-
ible to tell from the data in Table 1 as to whether, and in what way, the multidisci-
plinary team was informed by the assess-
ment process (with the exception of the PCL-R). This is because all of the data were fed into a complex decisionmaking process and, as this was not hierarchically designed, it is not possible to tell whether a low score on the WAIS-R was more important than a high score on the PCL-R in leading to someone being rejected. Hence, our failure to find differences in the individual assess-
ment measures is neither an argument for abandoning or retaining a systematic assessment using these measures.

Post-admission

Two main points emerged when the patients were on the unit. The first was that there was an improvement in patients’ skills acquisition and the second that there was considerable attrition. One criticism of the first finding was that skills acquisition was measured by selfreport, and hence was likely to be inaccurate as patients would wish to portray themselves in the best poss-
ible light. We believe that this is unlikely for two reasons. First, all patients were volun-
teers and so were not being detained or re-
leased on the basis of their improvement (or the reverse). Hence, there was little incen-
tive to portray themselves positively. (A possible exception was those serving life sentences or on restriction orders where a favourable report would be of benefit. However, this group comprised less than 10% of the sample.) Second, there was an observed improvement in patients’ behav-
ior when they were on the unit and this paralled the improvement in their self-
report. In addition, the selfreported im-
provement on the social problem measure was enhanced further a year after discharge when those completing the form could not have achieved any further advantage by a positive response.

However, the rate of attrition from the service paints a less favourable picture: with twothirds of the sample dropping out from treatment prematurely. Dropouts from treatment are infrequently reported, with the exception of clinical trials where the dropout from psychological treatments in those with personality disorder ranges between 30 and 70% (Garfield, 1986). As those who enter trials are often a selfselected sample in the community that are enthusiastic and thus more likely to engage, our dropout rates are surprisingly concordant with those from trial data. Moreover, as the most robust predictors of an increased dropout in those with per-
sonality disorder are low educational at-
tainment (Berrigan & Garfield, 1981), young age and hostility (Smith et al., 1995), all of which were common in our sample of young antisocial offenders, our finding of a 50–66% noncompletion rate (depending on the definition) is perhaps better than might be expected. None the less, it is disappointing. What is most trou-
bling is that the failure to select those who are likely to complete treatment is, not only an inappropriate use of an expensive facil-
ity, but may be damaging as there are data.
to suggest that non-completion of treatment predicts a less favourable course compared with no treatment (McMurran & Theodosis, 2007).

Finally, the follow-up results were also disappointing as the re-offending rate was similar to that of those released directly from prison to the community (Home Office, 2002). This high rate of re-offending occurred despite our data showing – at least for the social problem-solving measure – that gains in treatment persisted during the first year of follow-up.

There are a number of reasons to explain this discrepancy. First, re-offending may either be weakly related or unrelated to the acquisition of skills. Many mental health programmes that treat offenders have been criticised in not focusing sufficiently on the core criminological issues that need to be addressed if future violence is to be reduced (Maden et al., 2004). Hence, one of the reasons for our lack of success might have been the initial injudicious selection of the treatment programme, and this has changed since the unit was formed. Second, the reason that the programme was unsuccessful might have been, not that it was inappropriately selected, rather that it was inappropriately applied. There are many reports indicating that programmes fail because of poor implementation rather than deficiencies in the programmes themselves (Hollin, 1995). Third, patient selection might have failed to identify those who would be likely to benefit from what was on offer. We recognise that there needs to be a more systematic assessment of the individual’s motivation and if treatment resistance is strong little progress is likely to be made with programmes designed to change personality (Tyrer et al., 2003). Deciding which is the best explanation for these negative findings requires more careful investigation than was possible in this pilot project.

Implications

Clearly, one needs to be cautious in drawing too many definite conclusions from such a small sample. In addition, one might ask about their relevance to the DSPD Programme as this clearly has a very different remit. None the less, we believe that the findings of this study are relevant to the DSPD Programme for the following reasons. First, the DSPD Programme currently employs similar assessments to those used in this study as entry criteria to its service (i.e. scores on the PCL–R and IPDE), albeit in a different direction. That is, that the selection process employed by the Personality Disorder Unit was predicated on criteria (i.e. being volunteers, having a low PCL–R score, absence of sadistic offending, etc.) so that those admitted ought to have had a favourable outcome – certainly compared with those admitted to the DSPD services. Hence the evidence that this treatment programme had little effect on the re-offending rates after release – even in this group chosen to optimise outcome – ought to cause some pause for thought.

Second, our finding that a relatively comprehensive examination had only a very modest impact – if any impact at all – on either adherence to treatment or its outcome is noteworthy. This must call into question the assessment process (that is being replicated within the DSPD Programme), as those selected for admission showed neither the expected adherence to treatment nor a significant benefit in primary outcome. These data provide additional ammunition for those who believe that our current conceptions of personality disorder that underpin the DSM–IV categorical system are fundamentally flawed (Malik & Beutler, 2002).

Third, although one could argue that this selection procedure and treatment implementation bear little resemblance to the DSPD Programme, many patients transferred from hospital DSPD services will pass through ‘step-down’ services such as our unit. Hence, our very modest reduction in re-offending after release, if replicated elsewhere, ought to be a cause for concern in the light of the high cost of these services.

Future directions

There is a general principle that offenders with active psychosis ought to be transferred to a mental health setting and treated, rather than remain in prison. The position for offenders with a personality disorder is less clear. Should they remain in custodial settings or be transferred to mental health settings for treatment? Those who argue for the former make the following points: (a) that those referred to mental health settings are so similar to other incarcerated inmates that whatever treatment is offered in the former (at considerably increased cost) ought to be offered to all prison inmates; (b) that programmes addressing criminogenic factors within prisons may be as (or even more), effective in reducing re-offending than those provided in mental health settings. Hence, if a reduction in re-offending is the primary outcome, is it not more sensible to treat all people with personality disorders in custodial settings? There is also a concern that the medicalisation of personality disorder may be used as an excuse for antisocial behaviour and thereby encourage irresponsible individuals to take even less responsibility for their behaviour than they might otherwise do.

Although all of these arguments are plausible, what is clearly lacking at present are empirical studies. Although the DSPD Programme has many detractors, it has at least caused mental health professionals, service providers and other policy makers to place the connection between personality disorder and offending centre stage and thereby wrestle with several different questions that all of these groups have hitherto managed to avoid.

REFERENCES

American Psychiatric Association (1994) Diagnostic and Statistical Manual of Mental Disorders (4th edn) (DSM–IV), APA.
CONOR DUGGAN, FRCPsych, Section of Forensic Mental Health, University of Nottingham, LAUREN MASON, BSc, Arnold Lodge, Nottinghamshire Healthcare NHS Trust, PENNY BANERJEE, MRCPsych, Arnold Lodge, Nottinghamshire Healthcare NHS Trust, JOHN MILTON, MRCPsych, Rampton Hospital, Nottinghamshire Healthcare NHS Trust

Correspondence: Professor Conor Duggan, Arnold Lodge RSU, Cordelia Close, Leicester LE5 0LE, UK. Email: conor.duggan@nottshc.nhs.uk


Altered memory and affective instability in prisoners assessed for dangerous and severe personality disorder

TIM KIRKPATRICK, EILEEN JOYCE, JOHN MILTON, CONOR DUGGAN, PETER TYRER and ROBERT. D. ROGERS

Background Previous studies of borderline personality disorder report neuropsychological impairments in several domains, including memory. No studies have compared memory functioning in high-risk prisoners with borderline personality disorder with similar prisoners with other personality disorders.

Aims To explore mnemonic impairments in prisoners undergoing personality assessment as part of the dangerous and severe personality disorder initiative or detained in a medium secure facility.

Method We investigated memory function in 18 prisoners with borderline personality disorder and 18 prisoners with other personality disorders.

Results Prisoners with borderline personality disorder exhibited a pattern of multi-modal impairments in the immediate and delayed recall of verbal and visual information, with some association with affective instability. These deficits were not associated with the severity of personality disturbance.

Conclusions These data suggest that memory deficits have some specificity in relation to the constituent traits of borderline personality disorder and indicate that neuropsychological assessment may be a source of useful adjunctive information for distinguishing between the cognitive and psychological difficulties of individual prisoners.

Declaration of interest None. Funding detailed in Acknowledgements.

Individuals undergoing assessment for dangerous and severe personality disorder (DSPD) are likely to be a heterogeneous group, with divergent psychological and treatment requirements. Clinical assessment should, therefore, make provision for distinguishing between individuals with different personality characteristics and cognitive difficulties. Individuals with borderline personality disorder constitute an important subgroup within the DSPD service. In our studies of 31 prisoners from the initial sample assessed for DSPD at HMP Whitemoor, 17 had a definite diagnosis of borderline personality disorder according to DSM-IV criteria (American Psychiatric Association, 1994). Several sources of evidence attest to a link between borderline personality disorder and serious antisocial behaviour. Borderline personality disorder is highly comorbid with antisocial personality disorder (Becker et al., 2000), and may involve common psychological mechanisms such as affective instability and impulsiveness (Paris, 1997). Second, it is over-represented in surveys of forensic psychiatric services (Coid et al., 1999) and, third, its most prominent characteristics – unstable, intense relationships and affective instability – are exaggerated in individuals with histories of extreme violence (Raine, 1993).

Other evidence suggests that borderline personality disorder is mediated by disturbances within neural (Driessen et al., 2000; Donegan et al., 2003) and neurochemical systems (Hollander et al., 1994; Soloff et al., 2000) that support cognitive and emotional functions, although the specificity of these phenomena remains uncertain. Neuropsychological investigations suggest that borderline personality disorder is associated with memory disturbances, including difficulties in the encoding and retrieval of complex multi-modal information (O’Leary et al., 1991; Burgess 1992; Judd & Ruff, 1993; Kurtz & Morey, 1999). However, to date positive results have been demonstrated only in comparisons of patients with borderline personality disorder and non-clinical healthy controls, again raising the question as to whether such impairments tell us anything about the disorder itself or are merely indicative of the non-specific consequences of psychological distress. Similarly, there have been no studies of memory function in prisoners with borderline as compared with other personality disorders.

In this study, we used traditional pencil-and-paper neuropsychological instruments to investigate visual and verbal memory function in prisoners undergoing assessment for DSPD. All prisoners were screened for neurological illness, learning disability, current mood disorder, and current or previous psychoses. We sought to compare the performance of prisoners with DSM-IV borderline personality disorder with that of prisoners with other DSM-IV personality disorders.

METHOD

Participants Thirty-six male prisoners participated, 18 fulfilled DSM-IV criteria for borderline personality disorder (borderline group) and 18 had other personality disorders (non-borderline group). Thirty-one prisoners were voluntarily undergoing assessment at a maximum security prison (Whitemoor) as part of the UK government’s DSPD policy initiative. Five prisoners were undergoing assessment at a medium secure unit (Arnold Lodge, RSU) as a part of an assessment for prisoners with clear personality difficulties. All provided written, informed consent.

Clinical assessment Participants underwent a full psychiatric assessment of DSM-IV Axis I and Axis II disorders. Axis I disorders were assessed using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I; First et al., 1996). One participant from the borderline group and 4 from the non-borderline group were screened for Axis I disorders using the Schedule of Affective Disorders and Schizophrenia (Endicott & Spitzer, 1979). Exclusion criteria were as follows: previous or current neurological illness or injury; previous or current psychotic illness (including schizoaffective disorder and bipolar I or II disorder); current unipolar depression; any evidence of learning disability. Of the 18
participants with borderline personality disorder, I received a comorbid diagnosis of gender identity dysphoria, I had an adjustment disorder, I had a history of attention-deficit hyperactivity disorder (ADHD) and I had alcohol dependency. Of the non-borderline participants, I was diagnosed with ADHD, and I with polydrug misuse and generalised anxiety disorder. Significant previous alcohol or substance misuse was reported by 3 men with borderline and 1 with other personality disorders. The number of participants who had been dependent upon alcohol, stimulants (i.e. amphetamine) or opiates was 4 and 3 respectively.

**Personality disorders**

Personality disorder was assessed with the International Personality Disorder Examination (IPDE; Loranger et al, 1994), which is a semi-structured interview that provides diagnostic information about DSM–IV personality disorders and indicates both ‘definite’ and ‘probable’ diagnoses within the personality disorders specified in Axis II. The participants with borderline personality disorder had more definite diagnoses of personality disorder than the non-borderline participants (2.06 ± 0.29 vs. 1.33 ± 0.27; F (1, 34) = 3.40, P = 0.074). The distribution of ‘definite’ and ‘probable’ diagnoses is shown in Table 1.

Each participant was also rated for severity of personality disturbance according to the procedure described by Tyrer & Johnson (1996) in which the number of diagnoses within and between DSM–IV clusters is transformed into four levels of severity: no personality disorder, personality difficulty, simple personality disorder and diffuse personality disorder (Psychiatric Assessment Schedule (PAS)). The PAS has been used to specify the presence, cluster type and severity of personality disorders in a variety of clinical settings (Cuesta et al, 2001; Seivewright et al, 2002). It has been shown to have adequate interrater and test–retest reliability (Tyrer & Alexander, 1979; Tyrer et al, 1983; Hill et al, 2000) and to predict treatment outcome (Tyrer & Seivewright, 1988). The participants with borderline personality disorder were rated as having a significantly greater breadth of personality disturbance than the non-borderline participants (2.56 ± 0.12 vs. 2.00 ± 0.20; F (1, 34) = 5.743, P < 0.05).

Specifically, in the non-borderline group, 1 participant was classified as having no personality disorder but had a total score on the Psychopathy Checklist–Revised (PCL–R; Hare, 1991; Hart et al, 1992) of 32.6, indicating a very high degree of psychopathy, 3 were classified as having a personality disturbance, 9 as having a simple personality disorder and 5 as having a diffuse personality disturbance. Eight participants in the borderline group were classified by the PAS as having a simple personality disorder and 10 as having a diffuse personality disorder.

Participants were also assessed using the PCL–R which is a 20-item checklist that measures the personality traits and behaviours characteristic of psychopathy. All items are scored on the basis of a file review and semi-structured interview. Notwithstanding recent controversy surrounding the psychometric structure of the PCL–R (Cooke & Michie, 2001; Cooke et al, 2004; Neumann et al, 2005), total scores on the PCL–R in this study were taken to represent the composite of just 2 factors: Factor 1 relates to interpersonal and emotional deficits characteristic of psychopathy whereas Factor 2 relates to a history of criminality and a propensity towards an antisocial lifestyle. Generally, participants PCL–R scores were rated in the high range but were matched between the borderline and non-borderline groups (26.34 ± 1.23 vs. 24.96 ± 2.06; F < 1.00). There were 11 participants from the non-borderline group and 10 from the borderline group with a PCL–R of 27 or above. The two groups were comparable for Factor 1 (9.62 ± 0.70 vs. 10.06 ± 1.22; F < 1.00) and Factor 2 scores (12.91 ± 0.69 vs. 11.84 ± 0.81; F < 1.10).

**Medication**

Twelve participants in the non-borderline group were unmedicated; 2 were prescribed atypical antipsychotics; 2 selective serotonin reuptake inhibitors (SSRI) antidepressants, 1 a tricyclic antidepressant, 3 mood stabilisers, 1 beta-blockers, 1 hypnotics and 1 a stimulant. Six participants from the borderline group were not being prescribed any medication; 1 was being prescribed a standard antipsychotic, 6 atypical antipsychotics, 2 SSRI antidepressants, 8 tricyclic antidepressants, 4 mood stabilisers, 3 beta-blockers and 2 hypnotics. Significantly more participants with borderline personality disorder were receiving antidepressants (χ² = 5.9, d.f. = 1, P < 0.05) and there was a tendency for a greater proportion to be prescribed antipsychotics (χ² = 3.704, d.f. = 1, P = 0.054). However, the two groups were well-matched in terms of mood stabilisers (χ² = 0.177, d.f. = 1, P = 0.674), beta-blockers (χ² = 1.125, d.f. = 1, P = 0.289), hypnotics (χ² = 0.364, d.f. = 1, P = 0.546), and stimulants (χ² = 1.029, d.f. = 1, P = 0.31).

| Table 1 Number of concurrent DSM–IV personality disorders diagnosed in 18 prisoners with borderline personality disorder and 18 with other personality disorders. |
|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| Diagnosis                                       | Definite diagnosis | Probable diagnosis | Definite diagnosis | Probable diagnosis |
| Cluster A                                        |                    |                    |                    |                    |
| Paranoid                                        | 6                   | 3                   | 4                   | 2                   |
| Schizoid                                        | 0                   | 3                   | 0                   | 2                   |
| Schizotypal                                     | 1                   | 2                   | 1                   | 0                   |
| Cluster B                                        |                    |                    |                    |                    |
| Antisocial                                      | 14                  | 2                   | 11                  | 4                   |
| Borderline                                      | 18                  | 0                   | 0                   | 1                   |
| Histrionic                                      | 4                   | 0                   | 2                   | 1                   |
| Narcissistic                                     | 5                   | 2                   | 3                   | 0                   |
| Cluster C                                        |                    |                    |                    |                    |
| Avoidant                                        | 6                   | 1                   | 2                   | 2                   |
| Dependent                                       | 0                   | 2                   | 0                   | 0                   |
| Obsessive–Compulsive                            | 1                   | 4                   | 2                   | 0                   |
Psychometric assessments and self-report measures of mood and impulsivity

All participants were assessed for general cognitive ability using the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999). They also completed the Welsh Anxiety Scale (WAS; Welsh, 1956), which is a measure of anxiety/negative affect and is derived from the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943). High scores on the WAS are thought to relate to a dysphoric and dysphoric nature in which anxiety is a prominent feature. Depressive symptoms were assessed using the short form of the Beck’s Depression Inventory (BDI; Beck et al., 1961). Impulsivity was assessed using the Barratt Impulsiveness Scale (Version 11) (BIS–11; Patton et al., 1995). The BIS–11 has been used to distinguish between violent and non-violent parolees (Cherek et al., 1997), and has been reported to be a reliable measure of impulsivity in inmate populations (Patton et al., 1995).

Neuropsychological instruments

Verbal memory
Logical Memory and Word Lists sub-tests of the Wechsler Memory Scale–III (Wechsler, 1998) were administered to assess retention of verbal material under immediate and delayed conditions.

Visual memory/Rey–Osterrieth Complex Figure Test
The Complex Figure Test (CFT; Osterrieth, 1944) was used to assess visual memory. Participants were scored according to the strict marking scheme described by Bennett-Levy (1984) on the quality of their copy, immediate recall (3 min after copy) and delayed recall (25–30 min after copy). Additional measures of the participants’ recall of features with good continuation and symmetry were also derived. Previous results have found that recall deficits in the non-verbal tasks, such as the CFT, have been associated with right-sided temporal damage (Lezak, 1983).

Statistical analysis
Age, WASI scores, WAS, BDI, BIS–11 and WMS–III subtest scores were analysed using one-way ANOVA with the between-participant factor of group (borderline and non-borderline). Scaled WMS–III scores were derived from the raw scores adjusted for age according to published norms (Wechsler, 1998). Scores on CFT were analysed by repeated-measures ANOVA with the between-participant factor of group and the within-participant factor of trial (copy, immediate and delayed recall score).

To investigate which criteria had the largest impact on neuropsychological test performance, we performed backward regression analyses, including both variables expected to influence memory performance – age (except where the dependent measure was scaled), full WASI score and current mood (BDI score) – and other personality-related variables such as the breadth of personality disorder (PAS score), Factor 1 score of the PCL–R and IPDE ratings of impulsiveness and affective instability (scored as absent, present at a sub-criteria level and present at full criteria level). There were no significant differences between the ages, full WASI scores and current mood ratings for participants rated with the IPDE for different levels of trait impulsivity (all F (2, 35), s < 1.8) or for different levels of trait affective instability (F (2, 35) s < 1.90). Correlations between ratings of impulsiveness and affective instability and other regressors included in the models were, in the main, modest (−0.10 to −0.19 and 0.01 to 0.28). However, affective instability was significantly correlated with the Factor 2 score of the PCL–R, so the latter was dropped. We report only regressors retained in the final model for the dependent measures.

RESULTS

The participants from the non-borderline group were marginally younger than those from the borderline group (F (1, 34) = 14.357, P < 0.001) and depressive/angry symptoms on the BDI compared with participants without borderline personality disorder (F (1, 34) = 6.987, P < 0.05) (Table 2). Different aspects of impulsiveness were also increased, as reflected in higher scores on the attentional (F (1, 34) = 31.83, P < 0.01) and planning impulsiveness scores of the BIS–11 (F (1, 34) = 14.357, P < 0.01)

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Demographic and psychometric characteristics of 18 prisoners diagnosed with borderline personality disorder and 18 diagnosed with other personality disorders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borderline participants</td>
<td>Non-borderline participants</td>
</tr>
<tr>
<td>Age, years: mean (s.d.)</td>
<td>34.44 (1.70)</td>
</tr>
<tr>
<td>Full-scale IQ (WASI) score: mean (s.d.)</td>
<td>93.28 (3.06)</td>
</tr>
<tr>
<td>Verbal IQ (WASI)</td>
<td>90.67 (3.25)</td>
</tr>
<tr>
<td>Performance IQ (WASI)</td>
<td>97.17 (2.76)</td>
</tr>
<tr>
<td>Beck’s Depression inventory score: mean (s.d.)</td>
<td>15.11 (2.26)*</td>
</tr>
<tr>
<td>WASI Anxiety Scale score: mean (s.d.)</td>
<td>15.39 (1.35)**</td>
</tr>
<tr>
<td>BIS-11 total score: mean (s.d.)</td>
<td>77.44 (3.48)**</td>
</tr>
<tr>
<td>Motor impulsiveness</td>
<td>27.28 (1.46)**</td>
</tr>
<tr>
<td>Attentional impulsiveness</td>
<td>18.89 (1.04)**</td>
</tr>
<tr>
<td>Planning impulsiveness</td>
<td>31.83 (1.45)**</td>
</tr>
<tr>
<td>PAS score: mean (s.d.)</td>
<td>2.56 (0.12)**</td>
</tr>
<tr>
<td>PCL–R total score: mean (s.d.)</td>
<td>26.43 (1.21)</td>
</tr>
<tr>
<td>Factor 1</td>
<td>9.62 (0.70)</td>
</tr>
<tr>
<td>Factor 2</td>
<td>12.91 (0.69)</td>
</tr>
</tbody>
</table>

WASI, Wechsler Abbreviated Scale of Intelligence; BIS–11, Barratt Impulsiveness Scale (version II); PAS, Psychiatric Assessment Schedule; PCL–R, Psychopathy Checklist–Revised.

*P < 0.05; **P < 0.01; ***P < 0.001.
Table 3: Performance on verbal memory (WMS–II; Wechsler, 1998) tests by 18 prisoners diagnosed with borderline personality disorder and 18 prisoners diagnosed with other personality disorders.

<table>
<thead>
<tr>
<th></th>
<th>Mean (s.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Borderline participants</td>
</tr>
<tr>
<td>Logical memory (WMS–III)</td>
<td></td>
</tr>
<tr>
<td>Immediate recall</td>
<td></td>
</tr>
<tr>
<td>Immediate recall total score</td>
<td>8.28 (0.91)*</td>
</tr>
<tr>
<td>Immediate first recall total score</td>
<td>8.06 (0.78)**</td>
</tr>
<tr>
<td>Immediate recall learning slope</td>
<td>10.06 (0.54)</td>
</tr>
<tr>
<td>Immediate thematic score**</td>
<td>7.72 (0.68)**</td>
</tr>
<tr>
<td>Delayed recall</td>
<td></td>
</tr>
<tr>
<td>Delayed recall total score</td>
<td>8.78 (0.75)</td>
</tr>
<tr>
<td>Delayed recall percentage retention</td>
<td>8.89 (0.82)</td>
</tr>
<tr>
<td>Delayed thematic score*</td>
<td>7.72 (0.68)*</td>
</tr>
<tr>
<td>Recognition score</td>
<td>23.89 (0.86)*</td>
</tr>
<tr>
<td>Word Lists (WMS–III)</td>
<td></td>
</tr>
<tr>
<td>Immediate recall</td>
<td></td>
</tr>
<tr>
<td>List A first recall score</td>
<td>8.33 (0.86)</td>
</tr>
<tr>
<td>List A recall total score</td>
<td>8.72 (0.78)</td>
</tr>
<tr>
<td>Learning slope</td>
<td>10.39 (0.85)</td>
</tr>
<tr>
<td>Interference from List A (List B)</td>
<td>11.72 (0.64)*</td>
</tr>
<tr>
<td>Delayed recall</td>
<td></td>
</tr>
<tr>
<td>Recall total score</td>
<td>11.61 (0.67)</td>
</tr>
<tr>
<td>Recognition score</td>
<td>10.78 (0.48)</td>
</tr>
<tr>
<td>Percentage retention</td>
<td>11.72 (0.59)*</td>
</tr>
</tbody>
</table>

The total impulsiveness score of the BIS–11 was also increased (F (1, 34)=12.668, P < 0.01).

Verbal memory/logical memory

Participants with borderline personality disorder recalled significantly fewer story elements when recalling the stories immediately compared with participants without borderline personality disorder (F (1, 33)=4.54, P < 0.05) (Table 3). This was particularly evident when recalling the stories for the very first time (F (1, 33)=10.21, P < 0.005). They also performed significantly worse when recalling thematic elements of the test, both immediately (F (1, 33)=9.65, P < 0.005) and after a delay of 25–35 min (F (1, 33)=6.03, P < 0.05). There were no significant differences between groups in the improvement in immediate recall when hearing the stories for the second time compared with the first time (F < 1.00), in the total number of story elements recalled after 25–30 min (F (1, 33)=1.57), or the proportion of information retained in delayed recall (F < 1.00). However, the participants with borderline personality disorder showed significantly poorer recognition of story elements (F (1, 32)=6.21, P < 0.05).

Regression analysis indicated that both clinical ratings of breadth of personality disorder (PAS score) and affective instability (IPDE criterion score) were significant or near-significant predictors of poor immediate story recall (R^2=0.28; standardised β=−0.32, t=−2.03, P < 0.05 and standardised β=−0.37, t=−2.37, P < 0.05, respectively). Affective instability alone was a significant predictor of poor immediate recall after hearing the stories for the first time (R^2=0.29; standardised β=−0.54, t=−3.53, P < 0.001; see Fig. 1), immediate thematic recall (R^2=0.41; standardised β=−0.64, t=−4.56, P < 0.001; see Fig. 1) and recognition of story items (R^2=0.20; standardised β=−0.44, t=−2.77, P < 0.01). Cognitive ability (WASI full IQ) predicted better delayed thematic recall scores (R^2=0.44; standardised β=−0.49, t=−3.51, P < 0.005) whereas higher Factor 1 scores from the PCL–R and affective instability both predicted worse delayed thematic scores (standardised β=0.32, t=−2.21, P < 0.05 and standardised β=0.35, t=−3.51, P < 0.05 respectively; see Fig. 1).

Verbal memory/word lists

Participants with borderline personality disorder tended to show poorer scores on the verbal list learning test compared with participants without borderline personality disorder. Specifically, they tended to recall fewer words after the first learning trial (F (1, 34)=3.29, P<0.08) and after four learning trials (F (1, 34)=3.10, P<0.09). However, there was no difference in terms of the improvement in recall over the course of the four learning trials (F<1). In addition, the participants with borderline personality disorder showed greater levels of interference from the first list (List A) while recalling the second list (List B) (F (1, 35)=4.29, P<0.05; see Table 3). They showed only modest and non-significant reductions in recall scores after a delay of 25–30 min but showed significantly poorer retention at delayed recall compared with the non-borderline group (F (1, 34)=4.28, P<0.05).

Cognitive ability and current mood (BDI score) showed significant but opposing relationships with the number of words recalled after the first learning trial (R^2=0.46; standardised β=0.45, t=3.37, P<0.005; standardised β=−0.42, t=−3.05, P<0.01, respectively). Current mood alone predicted poorer list recall after the fourth learning trial (R^2=0.14; standardised β=−0.37, t=−2.27, P<0.05).

Visual memory/Rey–Osterrieth Complex Figure Test

Participants with borderline personality disorder obtained significantly lower scores on all aspects of performance compared with participants without borderline personality disorder (F (1,33)=8.11, P<0.01; see Fig. 2). Planned comparisons indicated significantly lower scores on the copy component of the CFT (F (1,33)=6.86, P<0.05). Specifically, the copies of the two groups showed comparable evidence of good continuation (12.00±0.84 v. 11.22±0.86; t<1) but those with borderline personality disorder tended to show poorer reproduction of the symmetrical
components of the CFT (9.61 ± 0.87 v. 12.06 ± 0.82; F (1,35) = 4.29, P < 0.05). In addition, they produced lower scores on the immediate recall (F (1,33) = 5.93, P < 0.05) and delayed recall measures (F (1,33) = 5.52, P < 0.05).

Cognitive ability and affective instability were significant predictors of good and poor copy scores (R² = 0.52; standardised β = 0.40, t = 3.19, P < 0.005; standardised β = −0.47, t = −3.59, P < 0.005; see Fig. 1). Breadth of personality also tended to show some association with poorer copy scores (standardised β = −0.22, t = −1.73, P < 0.10). However, affective instability was the single best predictor of poor immediate recall score of the CFT (R² = 0.13; standardised β = −0.36, t = −2.18, P < 0.05; Fig. 1).

**DISCUSSION**

These results indicate that a sample of prisoners being assessed for DSPD who were given a diagnosis of DSM-IV borderline personality disorder exhibited significant impairments in the recall of thematic and complex information from verbal and visuospatial memory. These impairments cannot be attributed to a greater incidence of previous or current psychotic disorder, or current mood disorders, since these were exclusion criteria for our study. Neither are they attributable to a lack of motivation since the pattern of test performance did not show the kind of generalised impairment consistent with a motivational deficit. Therefore, since we have identified significant memory impairments in comparison with age and IQ-matched controls with other forms of personality disturbance, these data confirm and extend the findings of O’Leary (1991) and other investigators (Burgess, 1992; Kurtz & Morey, 1999), by demonstrating that memory impairments have some specificity for borderline personality disorder.

We acknowledge that our two groups of prisoners were inevitably different in ways that may have influenced cognitive function. First, the participants with borderline personality disorder scored significantly higher on the BDI; therefore, their greater ratings of depressive symptomology may account for some of the deficits observed here. However, although BDI scores predicted poorer recall of items from the word lists test on the first and fourth presentations across the entire sample, there was no indication of state mood effects in the case of the logical memory or the Rey–Osterrieth CFT tests. Rather, these impairments seemed most closely tied to the presence or otherwise of a diagnosis of borderline personality disorder and the presence of trait affective instability scored as part of the IPDE assessment. Second, we also note that participants with borderline personality disorder were receiving more antipsychotic and antidepressant medication. Previous studies have shown that such medications tend to have mixed effects on cognitive measures (Markovitz & Wagner, 1995), depending upon participants’ baseline performance, their dose and the treated psychopathology. However, in the present study, the differences in drug regimens between the two groups were relatively modest and unlikely to account for the pattern of our observations.

Finally, participants with borderline personality disorder had a higher number of ‘definite’ and ‘probable’ comorbid personality disorders so that the increased multiplicity of personality disorders associated with borderline personality disorder (Becker et al, 2000) might also account for the observed memory impairments. However, our analyses indicated that breadth of personality disorder – as measured by the PAS (Tyrer & Johnson, 1996) – accounted for only a very modest amount of the variance of memory performance, being a significant predictor only of the immediate recall scores of the logical memory test. We did not find that individuals with a diffuse pattern of personality disorder scored consistently worse than those with a simple personality disorder. Finally, it is unlikely that gross differences in clinical ratings of psychopathy can explain our results as these were comparable between our two groups. Instead, the data indicate that poor recall was more closely associated with the presence of borderline personality disorder and its diagnostic features.
Verbal and visuospatial mnemonic function

Prisoners with borderline personality disorder demonstrated significant reductions in the total recall, the first recall and the immediate and delayed thematic recall measures of the WMS–III Logical Memory test, and on the copy, immediate and delayed recall of Rey–Osterrieth CFT. Performance of the WMS–III Word Lists test was more equivocal, showing statistically unreliable impairments in immediate and delayed recall but a significant reduction in the delayed recall of words successfully reproduced at immediate recall. This pattern is broadly similar to those reported previously in community samples, and is consistent with the proposal that the memory impairment in borderline personality disorder is most strongly expressed in the recall of complex material (O’Leary et al., 1991). Extensive evidence indicates that declarative memory is supported by functionally dissociable sub-systems supported by overlapping neural substrates (Schacter, 1996). The present data suggest that borderline personality disorder in individuals with significant histories of antisocial behaviour is associated with relatively generalised memory impairments, perhaps reflecting dysfunction in some of these sub-systems.

These data do not tell us too much about whether memory problems relate primarily to the encoding of information, the consolidation or the recall of already sampled information. On the one hand, deficits were most marked in the immediate story recall of the Logical Memory and Rey–Osterrieth CFT tests and there was little sign that these differences were increased at delayed recall (the single exception being the poorer retention of previously remembered items of word lists). This indicates that the prisoners with borderline personality disorder did not show greater forgetting of material that has been successfully recalled earlier, at least over relatively brief intervals. Similarly, there were no significant between-group differences in the improvement seen in performance over the successive learning trials of the WMS–III Word Lists test. However, the participants with borderline personality disorder did show evidence of greater susceptibility to interference while recalling word lists, indicating that the retrieval of verbal information might also be compromised (see Della Rocchetta & Milner, 1993).

Further information about the nature of the mnemonic deficits is provided by the observation that participants with borderline personality disorder were most clearly impaired on the thematic elements of the Logical Memory test. These relate to the ability to remember the gist of the stories, which is more general than the specific and literal information required by other components of the test. This finding may have real consequences for assessment and treatment in centres such as those in the DSPD initiative. In these and other forensic settings, treatment intervention might involve exposure to sets of concepts, organised around one single theme, relevant to a particular treatment target (e.g. stop-and-think in cognitive skills training). The present results suggest that some prisoners may find it difficult to assimilate even such general ideas when delivered in a verbal format. Treatment development might involve assessing memory for both the themes and the details of treatment programmes over the short and longer-term to optimise delivery in different individuals.

In addition, it was notable that the participants with borderline personality disorder were impaired in the copy component of the Rey–Osterrieth CFT. Previous results involving patients with borderline personality disorder have been equivocal, with two studies reporting that community patients were impaired relative to non-clinical controls (Judd & Ruff, 1993; Dinn et al., 2004) but three reporting no differences (O’Leary et al., 1991; Driessen et al., 2000; Sprock et al., 2000). Given that the participants with borderline personality disorder in the present study also reproduced significantly fewer symmetrical elements of the CFT, further investigations in antisocial populations should at least explore the possibility that memory impairments are accompanied by more basic deficits in visuospatial construction, as well as using tailored tasks to probe the cognitive mechanisms that underlie problems in mnemonic function related to borderline personality disorder.

Our analyses sought to provide a preliminary test of the extent to which state or trait factors were most closely associated with memory dysfunction. We found only limited effects of current mood or trait impulsiveness but more consistent evidence for the involvement of affective instability in poor memory performance. Thus, these data suggest an association between failing regulation of emotion and generalised multi-modal memory deficits in prisoners with a diagnosis of DSM–IV borderline personality disorder. This is consistent with the findings of a complementary study – conducted in an overlapping sample of prisoners assessed for DSPD – demonstrating that affective instability was associated with reduced attention towards emotional cues while making risky choices (Kirkpatrick et al., 2007). This, and the present study, utilised clinical ratings of affective instability – scored across a relatively restricted range – but future studies should explore more recently developed and multi-faceted structured clinical measures of emotional regulation (Koenigsberg et al., 2002) to examine its influence upon altered cognition in borderline personality disorder.

Finally, the selection of neuropsychological instruments provides only poor information about dysfunction within underlying neural systems. However, the pattern of memory impairments observed is broadly consistent with the proposal that borderline personality disorder involves dysfunction of bilateral frontotemporal systems. The Logical Memory test and Rey–Osterrieth CFT are sensitive to anterior left-sided anterior temporal lobe damage (Treynery et al., 1996; Griffith et al., 2004) and right-sided hippocampal and association cortex damage (Bobbot et al., 1998) respectively. Recent research has also suggested that the amygdala plays a significant role in memory for themes or gist, particularly for emotionally salient information (Adolphs et al., 2001). This is consistent with the finding that higher Factor 1 scores, indicative of more marked interpersonal and affective deficits in psychopathy, were associated with impaired thematic memory on the Logical Memory test. Intriguingly, Donegan et al. (2003) have demonstrated altered neuronal activity within the amygdala in response to emotional facial expressions in individuals with borderline personality disorder. Therefore, the observed memory impairment might be mediated by dysfunction within association cortices that subserve learning, memory and visuospatial capabilities, and limbic circuits that route emotional information into other forms of cognitive processing.

ACKNOWLEDGEMENTS

We thank Dr Val Hayes, Psychology Department, and the prison officers on Red-Spur, HMIP Whitehouse, as well as the staff at Arnold Lodge RSU.
study was supported by a 2-year grant from the Home Office, UK.

REFERENCES


 Correspondence: Dr Robert D. Rogers, Department of Psychiatry, Warneford Hospital, Oxford, OX3 7JX, UK. Email: robert.rogers@psychohelp.oxford.ac.uk
Developmental trajectories associated with juvenile sexually abusive behaviour and emerging severe personality disorder in childhood: 3-year study

EILEEN VIZARD, NICOLE HICKEY and EAMON MCCORRY

Background Little is known about the developmental trajectories of juveniles presenting with sexually abusive behaviour or emerging severe personality disorder traits.

Aims To investigate whether ‘age at onset’ of sexually abusive behaviour and whether emerging severe personality disorder traits are associated with specific developmental profiles.

Method A retrospective file review of 280 juveniles presenting with sexually abusive behaviour was conducted and follow-up Offenders Index data were analysed.

Results Juveniles with early onset (<11 years) of sexually abusive behaviour had higher levels of psychosocial adversity and early childhood antisocial behaviour compared with those with late onset. Emerging severe personality disorder traits were associated with higher levels of psychosocial adversity, antisocial behaviour, convictions and predatory sexually abusive behaviour.

Conclusions Preliminary evidence supports the existence of distinct developmental trajectories within this population and points to a key role for traits of emerging severe personality disorder.

Declaration of interest None. Funding detailed in Acknowledgements.

A substantial minority of sexual offences against adults and children are perpetrated by young people (Home Office, 2003; Vizard, 2006) who tend to be subsequently convicted of non-sexual rather than sexual offences (Sipe et al., 1998). Juvenile sexually abusive behaviour may therefore represent a marker in a subgroup of children for later antisocial behaviour. In general delinquency research the concepts of age at onset (often, synonymously, ‘age of onset’ (Moffitt, 1993)), and ‘emerging personality disorder traits’ in childhood (Frick et al., 1994; Vizard et al., 2004) have informed the identification of those children most at risk of embarking on chronic antisocial behaviour (Moffitt, 1993; Broidy et al., 2003). It is not known if either concept can identify clinically relevant subgroups of juveniles with sexually abusive behaviour. Gretton et al. (2001) found that juvenile sex offenders with psychopathy traits were at increased risk for violent, non-sexual recidivism but not sexual recidivism. There has, however, been no examination of whether emerging personality disorder traits influence the type of sexually abusive behaviour exhibited. The aims of the current study were to investigate these concepts in a sample of children and adolescents presenting with sexually abusive behaviour.

Data collection

Psychosocial and behavioural data were gathered from the services' files which included reports from multiple informants across a range of domains. The file data were also used to score the Psychopathy Checklist–Youth Version (PCL–YV, Forth et al., 2003), a 20-item rating scale for assessing psychopathy traits in 12–18-year-olds. Clinical items are usually scored following a combined interview and file review. However, for research purposes file review alone is acceptable as long as information is sourced from multiple informants and domains. Adequate internal consistency (Cronbach’s alpha=0.94) has been reported for this measure in community samples (Forth et al., 2003).

Data on convictions were obtained from the Offenders Index and covered the period up to December 2003. This is a database administered by the Home Office, containing details of convictions for standard list offences recorded in England and Wales. Prevalence rates for ‘lifetime’ convictions were calculated based on convictions that occurred between the individual’s tenth birthday (the age of criminal responsibility in the UK) and December 2003. ‘Time at risk’ excluded any time spent in custody. Two mutually exclusive categories of offences (sex and violence) were examined, as well as the composite ‘any offences’.

Data analysis

There were three strands to the data analysis. Individuals were categorised as either early onset (n=93) or late onset (n=120), depending on whether their sexually abusive behaviour began before or after their 11th birthday. In 67 individuals the age at onset was not known and these were excluded from the analysis. Comparisons were made of psychosocial, sexual and non-sexual antisocial behaviour and conviction characteristics. It was hypothesised that the early-onset group would have higher rates of difficult temperament and maltreatment, show more indiscriminate sexually abusive behaviour, but during adolescence have similar antisocial behaviour and conviction profiles as the late onset group.

Individuals were categorised as presenting with (n=54) or without (n=149) emerging severe personality disorder traits. These traits were operationalised as scores above the sample mean for conduct disorder symptoms and on the PCL–YV. A
total of 77 juveniles were excluded from the analysis either because they were too young (i.e. were under 12 years) or because of insufficient information. Comparisons were made of psychosocial, sexual and non-sexual antisocial behaviour and conviction characteristics. It was hypothesised that the group with emerging severe personality disorder would have higher rates of difficult temperament, engage in more predatory disorder would have higher rates of difficult characteristics. Given the small sample sizes these comparisons were necessarily exploratory in nature.

Between-group comparisons were conducted using χ² or t-tests as appropriate.

RESULTS

Age at onset trajectories for sexually abusive behaviour

Psychosocial risk factors

As shown in Table 1, the early-onset group experienced significantly higher rates of psychosocial adversity than the late-onset group. As predicted, they had higher rates of difficult temperament and, apart from exposure to domestic violence, also had higher rates of maltreatment. In addition, the early-onset group were more likely to have been exposed to poorer parenting models, inappropriate sexualisation (childhood sexual abuse and/or inadequate family sexual boundaries) and were more likely to display behavioural problems. The only variable on which the late-onset group had a significantly higher rate than the early-onset group was substance misuse.

Sexually abusive behaviour

To ensure developmental comparability for the early- and late-onset groups, only

---

### Table 1 Psychosocial characteristics of juveniles with sexually abusive behaviour according to age at onset and emerging severe personality disorder traits

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Age at onset</th>
<th>ESPD traits</th>
<th>Age at onset x ESPD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early, % (n=93)</td>
<td>Late, % (n=120)</td>
<td>With, % (n=54)</td>
</tr>
<tr>
<td>Parental and family factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parental criminality</td>
<td>33</td>
<td>29</td>
<td>35</td>
</tr>
<tr>
<td>Parental childhood abuse</td>
<td>44</td>
<td>33</td>
<td>54</td>
</tr>
<tr>
<td>Parental mental health problems</td>
<td>50</td>
<td>33*</td>
<td>69</td>
</tr>
<tr>
<td>Parental time in care</td>
<td>27</td>
<td>15*</td>
<td>24</td>
</tr>
<tr>
<td>Inconsistent parenting</td>
<td>77</td>
<td>53***</td>
<td>76</td>
</tr>
<tr>
<td>Lack of parental supervision</td>
<td>65</td>
<td>30***</td>
<td>46</td>
</tr>
<tr>
<td>Inadequate family sexual boundaries</td>
<td>59</td>
<td>25**</td>
<td>39</td>
</tr>
<tr>
<td>Marital separation/divorce</td>
<td>77</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>Attachment related factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early difficult temperament</td>
<td>38</td>
<td>22*</td>
<td>46</td>
</tr>
<tr>
<td>Removal to local authority care</td>
<td>83</td>
<td>73</td>
<td>93</td>
</tr>
<tr>
<td>6+ changes in home placement</td>
<td>50</td>
<td>30***</td>
<td>62</td>
</tr>
<tr>
<td>Insecure attachment</td>
<td>68</td>
<td>33**</td>
<td>72</td>
</tr>
<tr>
<td>Child factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peri-natal problems</td>
<td>32</td>
<td>21</td>
<td>35</td>
</tr>
<tr>
<td>Hyperactive/impulsive behaviour</td>
<td>75</td>
<td>61*</td>
<td>87</td>
</tr>
<tr>
<td>Disruptive behaviour primary school</td>
<td>61</td>
<td>36*</td>
<td>63</td>
</tr>
<tr>
<td>Excluded from school</td>
<td>51</td>
<td>40</td>
<td>67</td>
</tr>
<tr>
<td>Any sexual cruelty to animals</td>
<td>15</td>
<td>3**</td>
<td>20</td>
</tr>
<tr>
<td>Any physical cruelty to animals</td>
<td>30</td>
<td>12**</td>
<td>35</td>
</tr>
<tr>
<td>Sexual and physical cruelty to animals</td>
<td>10</td>
<td>0**</td>
<td>15</td>
</tr>
<tr>
<td>Any substance misuse</td>
<td>15</td>
<td>27*</td>
<td>32</td>
</tr>
<tr>
<td>Learning disability (IQ ≤70)</td>
<td>31</td>
<td>21</td>
<td>35</td>
</tr>
<tr>
<td>Trauma factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Childhood sexual abuse</td>
<td>83</td>
<td>58**</td>
<td>69</td>
</tr>
<tr>
<td>Physical abuse</td>
<td>77</td>
<td>55**</td>
<td>72</td>
</tr>
<tr>
<td>Emotional abuse</td>
<td>84</td>
<td>63**</td>
<td>83</td>
</tr>
<tr>
<td>Physical neglect</td>
<td>72</td>
<td>41**</td>
<td>54</td>
</tr>
<tr>
<td>Exposure to domestic violence</td>
<td>51</td>
<td>44</td>
<td>52</td>
</tr>
</tbody>
</table>

ESPD, emerging severe personality disorder.

*P < 0.05, **P < 0.01.
Table 2  Sexually abusive behaviour according to age at onset and emerging severe personality disorder traits

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Age at onset(^1)</th>
<th>ESPD traits(^1)</th>
<th>Age at onset × ESPD(^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early, % (n=93)</td>
<td>Late, % (n=120)</td>
<td>With, % (n=54)</td>
</tr>
<tr>
<td>Female</td>
<td>80</td>
<td>82</td>
<td>93</td>
</tr>
<tr>
<td>Male</td>
<td>71</td>
<td>48**</td>
<td>67</td>
</tr>
<tr>
<td>Male and female</td>
<td>55</td>
<td>33**</td>
<td>61</td>
</tr>
<tr>
<td>Child and adult</td>
<td>31</td>
<td>19*</td>
<td>44</td>
</tr>
<tr>
<td>Female only</td>
<td>25</td>
<td>49**</td>
<td>32</td>
</tr>
<tr>
<td>Male only</td>
<td>16</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>Strangers</td>
<td>8</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Adult women</td>
<td>4</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>≥ 5 years younger</td>
<td>63</td>
<td>72</td>
<td>56</td>
</tr>
<tr>
<td>≥ 5 years younger only</td>
<td>12</td>
<td>29**</td>
<td>6</td>
</tr>
<tr>
<td>Abuse characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any penetration (anal or vaginal)</td>
<td>46</td>
<td>63*</td>
<td>61</td>
</tr>
<tr>
<td>Predatory(^2)</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Excessive force(^2)</td>
<td>17</td>
<td>7*</td>
<td>17</td>
</tr>
<tr>
<td>Verbal coercion</td>
<td>25</td>
<td>45**</td>
<td>52</td>
</tr>
<tr>
<td>Physical coercion</td>
<td>5</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Prior grooming</td>
<td>25</td>
<td>35</td>
<td>56</td>
</tr>
<tr>
<td>With co-abusers</td>
<td>7</td>
<td>14</td>
<td>13</td>
</tr>
</tbody>
</table>

ESPD, emerging severe personality disorder.
1. Comparisons are only in relation to abuse committed during adolescence.
2. Comparisons are in relation to abuse committed at any time during childhood and adolescence.
3. Data on abuse and excessive force not available for adolescent period.
*P < 0.05, **P < 0.01.

Table 3  Developmental patterns of non-sexual antisocial behaviour according to age at onset of sexually abusive behaviour

<table>
<thead>
<tr>
<th></th>
<th>0–3 years</th>
<th>4–6 years</th>
<th>7–10 years</th>
<th>11–17 years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early, % (n=93)</td>
<td>Late, % (n=120)</td>
<td>Early, % (n=93)</td>
<td>Late, % (n=120)</td>
</tr>
<tr>
<td>Difficult temperament(^1)</td>
<td>26</td>
<td>13*</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Oppositional</td>
<td>8</td>
<td>3</td>
<td>19</td>
<td>8*</td>
</tr>
<tr>
<td>Insecure attachment</td>
<td>13</td>
<td>3**</td>
<td>36</td>
<td>4**</td>
</tr>
<tr>
<td>Physically aggressive</td>
<td>24</td>
<td>11*</td>
<td>41</td>
<td>18**</td>
</tr>
<tr>
<td>Physical cruelty to animals(^2)</td>
<td>–</td>
<td>–</td>
<td>5</td>
<td>0*</td>
</tr>
<tr>
<td>Fire-setting(^2)</td>
<td>–</td>
<td>–</td>
<td>8</td>
<td>2*</td>
</tr>
<tr>
<td>Stealing(^2)</td>
<td>–</td>
<td>–</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Excluded from school(^2)</td>
<td>–</td>
<td>–</td>
<td>11</td>
<td>2**</td>
</tr>
<tr>
<td>Impulsivity(^2)</td>
<td>–</td>
<td>–</td>
<td>19</td>
<td>9*</td>
</tr>
<tr>
<td>Reckless behaviour(^2)</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Socially isolated(^2)</td>
<td>–</td>
<td>–</td>
<td>9</td>
<td>5</td>
</tr>
</tbody>
</table>

1. Only collected in the 0–3 years developmental periods.
2. Only collected in the developmental periods covering 4–17 years.
*P < 0.05, **P < 0.01.

sexually abusive behaviour displayed during adolescence (11–17 years) was investigated here. As shown in Table 2, during adolescence the early-onset group was more likely to abuse multiple types of victims, (males, females, children and adults). The late-onset group had higher rates of only abusing female victims, only abusing much younger children (i.e. those at least 5 years
younger) and using verbal coercion. The late-onset group also appears to have higher rates of penetrating their victims.

**Developmental patterns of non-sexual antisocial behaviour**

The results of the developmental stage analysis in relation to non-sexual antisocial behaviours are presented in Table 3. During infancy, early and middle childhood (0–10 years) the early-onset group tended to have higher rates of antisocial behaviour than the late-onset group, but as predicted, by adolescence both groups had generally similar rates. However, reckless behaviour and insecure attachment were significantly higher in the early-onset group at every stage of development. It is notable that physical aggression in the early-onset group begins early and increases consistently across every developmental stage. By contrast the level of physical aggression in the late-onset group presents at relatively modest levels but doubles between middle childhood and adolescence.

**Conviction profile**

Owing to differences in time spent ‘at risk’ between the groups, 58 juveniles in the early-onset group were compared with 50 in the late-onset group. Both of these subgroups spent a mean of 8 years at risk (s.d.=1.5). During the ‘lifetime’ period their conviction profiles for ‘any offending’ did not differ significantly (early onset 45 v. late onset 56%). There were no significant differences between the groups for sexual convictions (early onset 9 v. late onset 18%) or violent convictions (early onset 22 v. late onset 34%).

**Emerging severe personality disorder traits in juveniles with sexually abusive behaviour**

**Psychosocial risk factors**

As shown in Table 1, the group with emerging severe personality disorder traits had higher rates of difficult temperament, were more likely to have had parents with abuse histories and mental health problems, to have been exposed to inconsistent parenting, removed to local authority care, and be insecurely attached. They were also more likely to display disruptive behaviour in school, hyperactivity and cruelty towards animals. Notably, both groups were equally likely to have experienced maltreatment.

**Sexually abusive behaviour**

As predicted, the group with emerging severe personality disorder traits were more likely to abuse multiple types of victims (male and female victims, child and adult victims), and were more likely to engage in predatory sexual behaviour (Table 2). They also had higher rates of abusing strangers, using verbal coercion and grooming behaviours.

**Developmental patterns of non-sexual antisocial behaviour**

The results of the developmental stage analysis are presented in Table 4. Within each developmental period the group with emerging severe personality disorder traits was significantly more likely to display antisocial behaviour. Even in infancy a considerable proportion displayed difficult temperaments (33%) and physical aggression (30%). By middle childhood (7–10 years) over three-quarters were displaying physical aggression, whereas in adolescence over a quarter were physically cruel to animals.

**Conviction profile**

Those with and without emerging severe personality disorder spent comparable periods ‘at risk’ during the lifetime, i.e. a mean of 10.3 years (s.d.=3.5). During that time those with such traits were significantly more likely to be convicted of any offence (63% of n=196, \(\chi^2=0.231, P=0.001\)). Although this group had higher rates of sexual convictions (20%) than the
group without such traits (17%), this difference was not statistically different. However those with emerging severe personality disorder traits were significantly more likely to be convicted of violent offences (44 v. 19% of n=196, $\chi^2=0.259$, $P=0.001$).

**Age at onset trajectories and emerging severe personality disorder**

**Psychosocial risk factors**

A total of 32 (64%) of those with emerging severe personality disorder traits were on the early-onset trajectory, while only 18 (36%) were on the late-onset trajectory. Those young people with such traits on the early-onset trajectory were more likely to have experienced lack of parental supervision, inappropriate family sexual boundaries, sexual victimisation, physical neglect, multiple changes in home placement and to display cruelty to animals than those with those traits but on the late-onset trajectory. Those with such traits on the late-onset trajectory were only significantly more likely to misuse substances.

**Sexually abusive behaviour**

Juveniles with emerging severe personality disorder traits on the early-onset trajectory were significantly more likely to have abused both male and female victims whereas juveniles with such traits on the late-onset trajectory tended to target specific victim groups with more force. For example, they had higher rates of only abusing females, abusing strangers, raping adult women and using verbal or physical coercion.

**Emerging severe personality disorder traits**

The current study used both behavioural and personality criteria to define emerging severe personality disorder. It was predicted that juveniles with such traits would show a characteristic set of developmental risk factors, as well as serious antisocial behaviours. The results supported these predictions. Those with such traits, although no more likely to experience maltreatment, were more likely to have had parents with mental health problems who had also been abused. These factors may have served to compromise their ability to provide good parenting and a secure attachment base.

Higher levels of poorer attachment, and impulsive and disruptive behaviour may contribute to feelings of social isolation from peers in adolescence, partly motivating subsequent aggressive and sexually abusive behaviour. It is striking that these young people present with difficulties, even before school, that persist across development. The sexually abusive behaviour of this group tended to be more predatory, more likely to entail excessive force and verbal coercion, and be targeted at a range of victims.

The significantly higher levels of convictions, particularly violent convictions, suggests that the construct of emerging severe personality disorder may prove useful in helping to identify those young people most at risk of later serious offending.

**Prevention**

The established cost benefits of preventing childhood conduct disorder (Scott *et al.*, 2001) and investing in multi-systemic treatment for juvenile sexual offenders (Borduin & Schaeffer, 2001) indicate the economic viability of primary prevention. Therefore, there is a clear case for investment in prevention and treatment resources that enable local services to identify and intervene early with vulnerable children.

**Early identification**

Early assessment and intervention is highly recommended if there is evidence of an early onset of sexually abusive behaviour or emerging severe personality disorder traits. The findings presented here indicate that an assessment of emerging severe personality disorder traits may need to become part of the assessment portfolio of child and adolescent mental health services (CAMHS) which will require appropriate measures to be developed and resources to provide preventative input for children identified as high risk (Royal College of Psychiatrists, 1999: pp. 34–46). Furthermore, the early identification of children under 10 years of age displaying sexually abusive behaviour may help prevent a trajectory of development leading to contact.
with the criminal justice system, with significant benefits for the child and society. However, any risk assessment should be coupled with a needs assessment (Kroll et al, 1999) which includes planning for the input of appropriate resources. Concerns about the labelling of children at an early age should be set firmly against the lifelong preventative benefits to the child and society of the early identification of need.

Service provision

Juvenile sexually abusive behaviour and emerging severe personality disorder traits in children are complex, multi-dimensional problems requiring coordinated responses from a range of community-based agencies including local authorities and CAMHS. At present, few local services are willing to accept such cases, and many are not convinced that the needs of these children fall within their remit.

There is a pressing need for more active involvement of CAMHS in the assessment and treatment of these children (Royal College of Psychiatrists, 1999: pp. 34–46). It has been noted that particular attention should be given to the delivery of services to children whose complex needs span different specialties, such as juvenile sexual offenders and people with learning disabilities, since care planning may require close cooperation between a wide range of services (Royal College of Psychiatrists, 1999: pp. 34–36). These children require a range of services, including non-residential services for the majority and specialist residential services for the few with a more disturbed presentation. Strategic thinking is needed to clarify how these will be achieved.

Research strategies

Further research is required to improve early identification of sexually abusive behaviour and emerging severe personality disorder traits, to develop appropriate interventions, and to determine long-term outcome. The sample in this study may not be representative of all children with sexually abusive behaviour and there was a limited follow-up for antisocial and sexual conviction. From a practical perspective research is required to develop and evaluate developmentally sensitive measurement tools for assessing such traits in young people. Retrospective studies of adults with severe personality disorders, and not just antisocial personality disorder, would prove helpful in identifying predisposing childhood and adolescent developmental characteristics. Such retrospective studies would build on an existing evidence base that has already explored developmental models of antisocial personality and has identified childhood-onset conduct disorder as a factor that increases the relative risk of developing adult antisocial personality disorder (Loeb et al, 2003). However, the limitations of both longitudinal (Robins, 1966) and adult retrospective studies (Zoccolillo et al, 1992) have been discussed and the dearth of prospective, longitudinal studies of relevant antecedents to antisocial personality disorder has been noted (Loeb et al, 2003). The results of the present study strongly support the case for prospective, longitudinal research. Such prospective studies with high-risk children and adolescents would improve understanding of factors that allow some children to move off a severe personality disorder trajectory.

Policy development

Given the complexity and diversity of need within populations of young people with sexually abusive behaviour and emerging severe personality disorder, such as those reported here, it is essential that government policy addresses service provision within community services and residential care services. A dedicated interdepartmental government committee may be required to coordinate and monitor progress, and to facilitate inter-agency liaison.

ACKNOWLEDGEMENTS

This study was supported by a research grant from the Home Office DSDP Programme.

REFERENCES


Aetiology of the relationship between callous–unemotional traits and conduct problems in childhood

ESSI VIDING, PAUL J. FRICK and ROBERT PLOMIN

Background A callous and unemotional disposition is an indicator of early-onset antisocial behaviour.

Aims To investigate the extent to which genetic influences contribute to the overlap between callous–unemotional traits and conduct problems in a large population sample of 7-year-old twins.

Method Teachers provided ratings of callous–unemotional traits and conduct problems for 3434 twin pairs from the Twins Early Development Study. Model-fitting analyses were performed across the continuum of scores and at the extremes.

Results The phenotypic relationship was primarily genetically mediated, both across the continuum and at the extremes and was substantial.

Conclusions At 7 years of age, genetic influences on callous–unemotional traits overlap substantially with genetic influences on conduct problems. This combination should guide selection criteria in future molecular genetic studies.

Declaration of interest None. Funding detailed in Acknowledgements.

One delineator of heterogeneity within children with early-onset antisocial behaviour is a callous and unemotional disposition (Frick & Morris, 2004; Lynam & Gudonis, 2005). This designates a subgroup of children/youths with a more-severe, aggressive and stable pattern of antisocial behaviour and a specific neurocognitive profile indicative of defects in affect processing (Lynam & Gudonis, 2005; Blair, 2006). These are all markers that could be considered precursors of adult psychopathy and as such warrant careful study. We recently conducted the first twin study of callous–unemotional traits and conduct problems in childhood. High levels of callous traits were found to be under strong genetic influence (Viding et al, 2005). This finding was consistent with behavioural genetic studies of psychopathic personality in youth and adults (Bloningen et al, 2003; Taylor et al, 2003; Larsson et al, 2006). Furthermore, when twins with conduct problems were divided according to the presence of callous traits, a strong genetic influence on conduct problems was found.

These results provide strong support for the use of callous–unemotional traits to designate children with early-onset conduct problems who may have distinct causal processes leading to their antisocial behaviour. The present study expanded on these findings by examining the extent of genetic and environmental influences on the relationship between these two important dimensions in 7-year-old twins. Extremes in combination could be highly heritable simply because individual differences across the continuum are highly heritable, even if they are genetically uncorrelated. If common genes are important mediators of the relationship, molecular genetic analyses should focus on finding the common genes that mediate the risk.

Two twin studies to date have addressed the extent of overlap in the genetic influences on callous–unemotional traits and antisocial behaviour/lifestyle (Taylor et al, 2003; Larsson et al, 2006). In both studies the genetic influences on the two domains showed substantial overlap, although independent genetic influences were also observed. Both studies were conducted on youths and young adults only, some of whom may have had a childhood onset to their antisocial behaviour. In addition, neither study focused on extreme of the distributions. Given the risk associated with early-onset antisocial behaviour, we focused on the relationship with callous–unemotional traits in childhood and analysed data from extreme groups in addition to the entire continuum of scores.

METHOD

Participants Participants were drawn from the Twins Early Development Study (TEDS), a longitudinal study of twin pairs ascertained from population records of twin births in England and Wales between 1994 and 1996 (Trouton et al, 2002). The sample consisted of 3434 twin pairs, born between January 1994 and August 1996, who had teacher ratings for callous–unemotional traits and conduct problems. Any twin pairs where either twin had parental reports of medical or neurological conditions were not included (Dale et al, 1998), leaving a sample of 3232 twin pairs for analysis.

For the bivariate DeFries–Fulker extremites analysis (Defries & Fulker, 1985, 1988), same-gender twin pairs with at least one proband with callous–unemotional traits were included in the trait–conduct problems analysis (selecting on trait and measuring co-twins’ conduct problems); pairs with at least one proband with conduct problems were included in the conduct problems–trait analysis (selecting on conduct problems and measuring co-twins’ callous–unemotional traits). Probands were selected above the 90th percentile, a cutoff designated as ‘abnormal’ according to the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). The trait probands scored 1.31 or more standard deviations above the mean on the trait scale (612 probands, 459 twin pairs). The conduct problem probands scored 1.28 or more standard deviations above the mean on the conduct problems scale (444 probands, 364 twin pairs). This selection procedure guaranteed that the probands would score beyond the ‘average range’ (i.e. not within 1 s.d.), yet yielded enough probands to perform the twin analyses.
Zygosity was ascertained by parental ratings with an error rate of 5%, as validated by DNA typing of 8–10 microsatellite polymorphisms (Price et al., 2000). Unclear cases were resolved through genotyping a multiplex of 12 highly polymorphic markers (Freeman et al., 2003). Despite attrition, the TEDS sample that provided data at 7 years of age is closely matched to UK population in terms of ethnicity and maternal education (Harlaar et al., 2005).

Testing procedures

Informed, written consent was obtained from all families who agreed to take part in the study. The families were informed that the TEDS encompasses assessment of cognitive ability, behavioural problems and pro-social behaviours and that all of the data would be anonymised and published in a way that did not identify an individual child. Teachers were approached only if there was family consent for teacher involvement. The consent procedure was approved by the Institute of Psychiatry and Maudsley Ethics Committee.

Measures

Teachers provided ratings of callous-unemotional traits and conduct problems. The response rate of teachers was high: 88% of those approached responded by completing the TEDS assessment. There are several reasons for relying on teacher report. First, teachers are familiar with a broad range of children and have expertise regarding normative child development. Second, twin analyses indicate that teacher ratings show less rater bias than typically found in parent ratings (Nadder et al., 2001). Third, and most importantly for the purposes of this study, there is evidence that teacher ratings of callous-unemotional traits lead to a more valid differentiation of subgroups of children with conduct problems in preadolescent samples (Barry et al., 2000). Consistent with these theoretical reasons for relying on teacher report, parent ratings of callous-unemotional traits and conduct problems showed much poorer levels of internal consistency (α=0.45 and α=0.58 respectively) than teacher ratings (α=0.74 and α=0.71 respectively).

The TEDS 7-year assessment of callous-unemotional traits included three items (‘Does not show feelings or emotions’, ‘Feels bad or guilty if he/she does something wrong’ (reverse scored), ‘Is concerned about how well he/she does at school’ (reverse scored)) from the callous–unemotional traits scales of the Antisocial Process Screening Device (APS; Frick & Hare, 2001) and four selected items from the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) (e.g. ‘Considerate of other people’s feelings’ (reverse scored)). None of the items overlapped with any of the conduct problem items (see Viding et al. (2005) for the complete list of items on both scales).

We used the SDQ 5-item scale to assess conduct problems (e.g. ‘Often fights with other children or bullies them’, ‘Often has temper tantrums or hot tempers’). The SDQ is a widely used screening instrument in the UK and its reliability and validity have been demonstrated on a large, national sample (Goodman, 2001). Three of the conduct problem items reflected tendency for aggression or bad temper, whereas the remaining two assessed lying and stealing. The callous-unemotional traits and conduct problem scales correlated 0.50 in this sample.

Genetic analyses

ACE model fitting

We fitted a correlated factors model directly to the individual observations by full-information maximum-likelihood function estimation (Enders & Bondalos, 2001) in the program Mx (Neale et al., 2003).

In addition to yielding maximum-likelihood parameter estimates for the effects of latent additive genetic (A), shared environmental (C), and non-shared environmental (E) influences on callous-unemotional traits and conduct problems, the correlated factors model also provides estimates of the genetic correlation (r_g), shared environmental correlation (r_e), and non-shared environmental correlation (r_n) between a pair of measures (see data supplement 1 to the online version of this paper). The genetic correlation indicates the extent to which genetic effects on one measure overlap with genetic effects on another measure.

It is also possible to estimate the extent to which genetic factors contribute to the observed phenotypic correlation between the measures (bivariate heritability). Shared and non-shared environmental mediation of the phenotypic correlation can also be estimated (Neale et al., 2003).

Because mean effects of age and gender can spuriously inflate twin resemblance, all analyses used age- and gender-adjusted residual scores from multivariate linear regression modelling (McGue & Bouchard, 1984). Gender-related influences on individual differences can none the less be investigated (see data supplement 2 to the online version of this paper).

The relationship of extremes of callous-unemotional traits and conduct problems can be assessed with an extension of the DeFries–Fulker extremes analysis (DeFries & Fulker, 1985, 1988). This addresses the genetic and environmental causes of the mean difference on a quantitative trait score between probands and the rest of the population. Univariate analysis yields a statistic called group differences heritability (h^2_g), which is the proportion of the phenotypic difference between the probands as a group and the population that can be attributed to genetic factors. The bivariate extension of the group analysis addresses the etiology of co-occurrence of two traits for the extremes of dimensions (DeFries et al., 1991). Rather than selecting probands as extreme on X and comparing the quantitative scores of their monozygotic and dizygotic co-twins on X as in univariate group analysis, bivariate analysis selects probands on X and compares the quantitative scores of their co-twins on Y. The extent to which the cross-twin regression to the population mean is greater for dizygotic co-twins than monozygotic co-twins indicates the extent to which proband deficits in X are a result of genetic factors that also influence the co-twins’ quantitative scores on Y (group cross-familiality). An important point to note is that bivariate extremes analysis is not bi-directional. The group genetic correlation can be derived from group heritability estimates (Knopik et al., 1997). The DeFries-Fulker regression analysis is performed on same gender twin pairs and thus a test of gender differences is not incorporated (see data supplement 3 to online version of this paper).

RESULTS

Descriptive statistics

Descriptive statistics for the standardised conduct problems and callous-unemotional

*As some twins shared a teacher, whereas others were in different classrooms, we repeated the analyses using same and different teacher rated pairs. This did not affect the results and we therefore report data from the whole sample to increase the power of the analyses.
traits scores are summarised in Table 1. On both measures, all zygosity and gender groups showed similar mean scores (dizygotic opposite-gender twins showed slightly lower mean scores), but mono- and dizygotic female pairs and dizygotic opposite-gender pairs showed less variance than male mono- and dizygotic pairs, particularly on conduct problems. Although we observed some significant mean differences between our zygosity groups, these are not of a sizeable magnitude and the statistical significance probably reflects our sample size.

The phenotypic correlation between callous–unemotional traits and conduct problems scales was moderate ($r = 0.50$ (0.53 for boys, 0.46 for girls)) in this sample. One twin from each pair was randomly selected for the analyses. When we replicated this correlation with the previously unselected twin, the results were very similar ($r = 0.47$ (0.48 for boys, 0.46 for girls)).

**Genetic analyses**

Although variances and covariances are used in model-fitting analyses of twin data, correlations are useful for comparing resemblances between twins as a function of genetic relatedness. Twin correlations for callous–unemotional traits and conduct problems ratings are shown by gender and zygosity in Table 2. Monozygotic within-trait correlations were consistently greater than the corresponding dizygotic correlations for callous–unemotional traits and for conduct problems, suggesting substantial genetic influence on both. For both, dizygotic opposite-gender correlations were only slightly lower than correlations for dizygotic males and females, suggesting no important qualitative genetic differences between genders. However, quantitative gender differences are suggested by the pattern of correlations for dizygotic males and females, pointing to higher heritability and lower shared environment for males.

Cross-twin, cross-trait correlations for callous–unemotional traits and conduct problems were 0.41 and 0.38, for monozygotic males and females respectively, which

---

### Table 1  Age and gender-regressed z-scores for callous–unemotional traits and conduct problems according to gender and zygosity

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monozygotic ($n = 534$)</td>
<td>Dizygotic ($n = 508$)</td>
<td></td>
</tr>
<tr>
<td>Callous–unemotional traits</td>
<td>0.05 (1.07)</td>
<td>0.06 (1.06)</td>
<td></td>
</tr>
<tr>
<td>Conduct problems</td>
<td>0.00 (1.14)</td>
<td>0.04 (1.25)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monozygotic ($n = 612$)</td>
<td>Dizygotic ($n = 562$)</td>
<td></td>
</tr>
<tr>
<td>Dizygotic opposite gender</td>
<td>-0.01 (0.81)</td>
<td>-0.01 (0.81)</td>
<td></td>
</tr>
</tbody>
</table>

1. One twin from each pair was randomly selected for the analysis. Main effect for zygosity group was found for callous–unemotional traits ($F (4, 3157) = 4.32, P < 0.01$ (two-tailed)), reflecting the mean difference between monozygotic males and dizygotic opposite gender and monozygotic females and dizygotic opposite gender groups (both comparisons significant after correcting for multiple comparisons at $P < 0.025$ and $P < 0.01$ respectively). Marginal main effect for zygosity was found for conduct problems ($F (4, 3157) = 2.25, P = 0.06$ (two-tailed)), reflecting the difference between dizygotic males and dizygotic opposite gender groups. However, this did not survive correction for multiple comparisons.

### Table 2  Within trait (intraclass) and cross-trait twin correlations between callous–unemotional traits and conduct problems according to gender and zygosity

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monozygotic ($n = 534$)</td>
<td>Dizygotic ($n = 508$)</td>
<td></td>
</tr>
<tr>
<td>Callous–unemotional traits</td>
<td>0.72</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>Conduct problems</td>
<td>0.69</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>Cross-trait</td>
<td>0.41</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monozygotic ($n = 612$)</td>
<td>Dizygotic ($n = 562$)</td>
<td></td>
</tr>
<tr>
<td>Dizygotic opposite gender</td>
<td>0.44</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>

1. $n = 756$–1539 twin pairs/cell, based on pairwise deletion. For each pair of traits, the average of two reciprocal cross-correlations is presented. All correlations significant at $P < 0.01$. 

### Table 3  Model fit indices

<table>
<thead>
<tr>
<th>Model</th>
<th>$-2\text{ LL}$</th>
<th>d.f.</th>
<th>Number of parameters</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
<th>$\text{AIC} - \chi^2$</th>
<th>$\Delta\chi^2$ (d.f.)</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully saturated</td>
<td>39828.57</td>
<td>16658</td>
<td>70</td>
<td>55.53</td>
<td>46</td>
<td>3</td>
<td>-36.88</td>
<td>3.59 (2)</td>
<td>0.17</td>
</tr>
<tr>
<td>ACE general gender-limitation model $r_e$ free</td>
<td>39884.10</td>
<td>16704</td>
<td>24</td>
<td>59.12</td>
<td>48</td>
<td>0.13</td>
<td>-36.88</td>
<td>3.59 (2)</td>
<td>0.17</td>
</tr>
<tr>
<td>ACE common effects gender-limitation model $r_e$ fixed</td>
<td>39887.69</td>
<td>16706</td>
<td>22</td>
<td>59.12</td>
<td>48</td>
<td>0.13</td>
<td>-36.88</td>
<td>3.59 (2)</td>
<td>0.17</td>
</tr>
<tr>
<td>ACE no effects $^1$</td>
<td>40181.14</td>
<td>16715</td>
<td>13</td>
<td>352.56</td>
<td>57</td>
<td>$&lt;0.001$</td>
<td>238.56</td>
<td>293.44 (5)</td>
<td>$&lt;0.001$</td>
</tr>
</tbody>
</table>

1. $r_e$ free, genetic correlation between dizygotic males and females is allowed to depart from 0.50 (this model allows qualitative and quantitative gender differences); $r_e$ fixed, genetic correlation between dizygotic males and females is fixed to 0.50 (this model allows quantitative, but not qualitative gender differences). 

1. This model does not allow gender differences. 
ACE model-fitting analyses

Model fitting statistics comparing the gender-limited bivariate correlated factors model with a fully saturated model, as well as comparing nested submodels are presented in Table 3, with parameter estimates of the best-fitting model in Table 4. (Additional results are available from E.V. upon request). The best-fitting model (with the least number of parameters but no decrease in the model fit as compared with a model with more parameters) indicated that, for both callous–unemotional traits and conduct problems, there were quantitative but not qualitative gender differences. That is, the same genetic influences were important for males and females but in different degrees. The bivariate statistics, however, appeared remarkably similar for both genders.

Tables 4 and 5 show the total variance accounted for by genetic and environmental influences, in boys and girls. As expected from the pattern of cross-twin, within-trait correlations, both callous–unemotional traits and conduct problems were significantly heritable but somewhat more heritable in boys than girls ($h^2 = 0.67$ and $h^2 = 0.61$ for boys, and 0.48 and 0.57 for girls, for callous–unemotional traits and conduct problems respectively). Shared environmental influences were not statistically significantly different from zero for boys ($c^2 = 0.04$ for callous–unemotional traits and $c^2 = 0.06$ for conduct problems). For girls, there was modest, significant shared environmental influence for callous–unemotional traits ($c^2 = 0.08$). Non-shared environmental influences accounted for most of the environmental variance ($e^2 = 0.29$ and $e^2 = 0.34$ for boys, and 0.32 and 0.35 for girls, for callous–unemotional traits and conduct problems respectively).

Table 4 also summarises the extent of overlap between genetic and environmental influences. The genetic correlation ($r_g$) is significant as indicated by the confidence intervals and the estimates of 0.57 (boys) and 0.65 (girls) suggesting substantial overlap between genetic influences contributing to individual differences in both boys and girls. The shared environmental correlation ($r_e$) is not significant for either gender. Finally, non-shared environmental influences show significant overlap across callous–unemotional and conduct problems, in slightly greater magnitude for boys ($r_e = 0.40$), than for girls ($r_e = 0.19$). The $r_e$ estimate could also reflect measurement error common to both domains.

Finally, Table 4 summarises the extent to which genetic and environmental influences mediate the phenotypic relationship. The bivariate heritability estimates (biv $h^2$) of 0.71 (boys) and 0.77 (girls) indicate that the phenotypic relationship between the two traits is primarily mediated genetically for both genders. In other words, co-occurrence of callous–unemotional traits and conduct problems is mainly mediated by genetic influences. Non-shared environmental influences (and common error) make a modest contribution to the phenotypic relationship (biv $e^2 = 0.25$ (boys) and 0.14 (girls), although the contribution of shared environmental influences is negligible (biv $c^2 = 0.04$ (boys) and 0.09 (girls)).

---

Table 4  Standardised parameter estimates from the full ACE correlated factor model for boys

<table>
<thead>
<tr>
<th>Parameter estimates</th>
<th>Callous–unemotional traits</th>
<th>Conduct problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total variance resulting from</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additive genetic factors (A)</td>
<td>0.67 (0.58–0.72)</td>
<td>0.61 (0.50–0.69)</td>
</tr>
<tr>
<td>Shared environmental factors (C)</td>
<td>0.04 (0.00–0.11)</td>
<td>0.06 (0.00–0.17)</td>
</tr>
<tr>
<td>Non-shared environmental factors (E)</td>
<td>0.29 (0.26–0.33)</td>
<td>0.34 (0.31–0.38)</td>
</tr>
<tr>
<td>Correlations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic ($r_g$)</td>
<td>0.57 (0.48–0.63)</td>
<td></td>
</tr>
<tr>
<td>Shared environmental ($r_e$)</td>
<td>0.56 (0.00–1.0)</td>
<td></td>
</tr>
<tr>
<td>Non-shared environmental ($r_n$)</td>
<td>0.40 (0.34–0.46)</td>
<td></td>
</tr>
<tr>
<td>Phenotypic relationship mediated by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bivariate heritability (biv $h^2$)</td>
<td>0.71 (0.54–0.80)</td>
<td></td>
</tr>
<tr>
<td>Bivariate shared environmental factors (biv $c^2$)</td>
<td>0.04 (0.00–0.20)</td>
<td></td>
</tr>
<tr>
<td>Bivariate non-shared environmental factors (biv $e^2$)</td>
<td>0.25 (0.20–0.30)</td>
<td></td>
</tr>
</tbody>
</table>

1. As the shared environmental estimates for callous–unemotional traits and conduct problems did not significantly differ from 0.00 for boys, it was possible to drop the C path for boys without significant decrease in model fit. The same held for conduct problems for girls, as well as for the $r_e$ and biv $c^2$ estimates. In this reduced model, most of the $C$ variance ends up in the A term (results available from E.V.). Despite the acceptability of this nested model in model-fitting analyses comparing nested submodels, there were only slightly less than the within-individual correlation of 0.50 (Table 2). The dizygotic cross-trait correlations were only 0.22, 0.23, and 0.17 for males, females and opposite-gender twins respectively. This suggests substantial genetic mediation of the phenotypic correlation. The similar cross-trait correlations for dizygotic twins indicate neither qualitative nor quantitative gender differences.

---

Table 5  Standardised parameter estimates from the full ACE correlated factor model for girls

<table>
<thead>
<tr>
<th>Parameter estimates</th>
<th>Callous–unemotional traits</th>
<th>Conduct problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total variance resulting from</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additive genetic factors (A)</td>
<td>0.48 (0.37–0.60)</td>
<td>0.57 (0.45–0.68)</td>
</tr>
<tr>
<td>Shared environmental factors (C)</td>
<td>0.20 (0.08–0.29)</td>
<td>0.08 (0.00–0.19)</td>
</tr>
<tr>
<td>Non-shared environmental factors (E)</td>
<td>0.32 (0.29–0.35)</td>
<td>0.35 (0.32–0.38)</td>
</tr>
<tr>
<td>Correlations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic ($r_g$)</td>
<td>0.65 (0.52–0.78)</td>
<td></td>
</tr>
<tr>
<td>Shared environmental ($r_e$)</td>
<td>0.33 (0.00–0.95)</td>
<td></td>
</tr>
<tr>
<td>Non-shared environmental ($r_n$)</td>
<td>0.19 (0.12–0.25)</td>
<td></td>
</tr>
<tr>
<td>Phenotypic relationship mediated by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bivariate heritability (biv $h^2$)</td>
<td>0.77 (0.58–0.97)</td>
<td></td>
</tr>
<tr>
<td>Bivariate shared environmental factors (biv $c^2$)</td>
<td>0.09 (0.00–0.26)</td>
<td></td>
</tr>
<tr>
<td>Bivariate non-shared environmental factors (biv $e^2$)</td>
<td>0.14 (0.09–0.190)</td>
<td></td>
</tr>
</tbody>
</table>
DeFries–Fulker extremes analyses

Application of bivariate DeFries–Fulker group analysis selecting on callous–unemotional traits and measuring co-twin conduct problems yielded a bivariate group differences heritability estimate of 76% (95% CI 0.39–1.13). In other words, 76% of the mean difference between the extreme group with regard to callous–unemotional traits and the population on the conduct problems scale can be attributed to genetic factors. The bivariate group shared environment estimate was 4% (95% CI −0.37 to −0.45). The remainder of the mean difference was a result of non-shared environmental factors. The converse analyses — selecting on conduct problems and measuring co-twin callous–unemotional traits — yielded a similar bivariate group differences heritability estimate of 82% (95% CI 0.49–1.14), and bivariate group shared environment estimate of 2% (95% CI −0.31 to 0.35). The extremes genetic correlation estimate is 1, indicating complete commonality of genetic influences at the extremes. The confidence interval for this bivariate DeFries–Fulker extremes estimate of a group genetic correlation has not yet been worked out (Knopik et al, 1997) but is likely to be large, and this finding should thus be treated as instructive rather than definitive.

DISCUSSION

As noted previously, children with callous–unemotional traits seem to constitute an important subgroup of children with early-onset conduct problems (Frick & Morris, 2004). Previously, we demonstrated that antisocial behaviour is highly heritable in the group with such traits but not in children with conduct problems only (Viding et al, 2005). The present study attempted to expand on these findings by examining the extent of genetic and environmental influences on the relationship between these two important dimensions in 7-year-old twins.

Our present findings demonstrated, most importantly, that there is substantial genetic overlap between callous–unemotional and conduct problems in both boys and girls. Common genetic influences operate to bring about both of these problems, assessed as a dimension in the entire sample and even more so at the high extremes. These common genetic influences also appear to be largely responsible for the phenotypic relationship. Our study was unique in that its large sample size enabled us to study genetic and environmental influences at the extremes of the distribution, as well as across the entire continuum. We replicated findings from studies of adults and youths which show substantial heritability of individual differences in callous–unemotional traits (Blinenberg et al, 2003; Taylor et al, 2003; Larsson et al, 2006) and of genetic mediation of the phenotypic relationship with antisocial behaviour (Taylor et al, 2003; Larsson et al, 2006).

Unlike in an earlier study (Larsson et al, 2006), there was a gender difference in the magnitude of genetic and shared environmental effects on individual differences in callous–unemotional traits in childhood and this warrants further investigation. One target for future research is to identify specific shared environmental influences that may affect the level of such traits in girls and whether these influences relate to low or high levels (e.g. these could be influences encouraging prosocial behaviour in girls). However, and most importantly, callous–unemotional traits and conduct problems were associated at the phenotypic level in both boys and girls and the mediation of the relationship was strongly driven by common genes for both.

The shared genetic influences suggest that molecular genetic studies should concentrate on polymorphisms associated with callous–unemotional traits and conduct problems.

Shared environmental influences could not be reliably detected as an aetiological factor mediating the relationship between callous–unemotional traits and conduct problems either across the continuum or at the extremes. This does not mean that environmental influences present in the family are not important. However, these influences appear to operate in a child- and trait-specific manner. As an example, parental treatment may differ for twins and this differential treatment may cause differences in levels of callous–unemotional traits and conduct problems considered separately. A recent study demonstrated that elevated maternal negative emotionality was an environmental variable that influenced the extent of differences in conduct problems in genetically identical monozygotic twins (Caspì et al, 2004). Finally, it is likely that the latent addictive genetic influence (\( A' \) parameter) also includes effects of gene–environment correlation. For example, children with a particular genotype may evoke a certain reaction from their environment or may actively seek out certain kinds of activities, all of which would reinforce the measured trait.

In line with earlier findings (Taylor et al, 2003; Larsson et al, 2006), not all genetic influences on the individual differences in callous–unemotional traits and conduct problems were overlapping in our study. The non-overlapping genetic variance has been proposed to imply some independence in the underlying biological substrates (Taylor et al, 2003). However, both previous studies and our own individual differences analysis addressed the entire continuum of scores. Our analysis of extreme groups suggests that genetic overlap may be complete at the extremes, although we acknowledge that such estimates entail substantial confidence intervals. None the less, we would not rule out the possibility that unique genetic influences may be important.

Some general limitations of the study should be mentioned. Our scale for assessing callous–unemotional traits was not a standard instrument. However, teacher ratings on this scale showed good internal consistency and distinguished an aetiological distinctly group of children with early-onset antisocial behaviour in our earlier study (Viding et al, 2005). Relying on a single source of measurement could be considered a limitation. As the parent ratings of such traits did not show good internal consistency, it seemed dubious to base conclusions on these (Viding et al, 2005). Collection of data at a single age is a limitation, which precludes commenting on the aetiology of the stability of the association or whether the genetic links are of different magnitude in childhood than later in development. We are currently following up the twins at 9 years of age and will thus be able to add a longitudinal aspect in the future.

Within the context of these limitations, the present findings have several important implications. The finding of genetic overlap for callous–unemotional traits and conduct problems suggests that although distinct brain anatomical substrates or cognitive operations may be associated with these dimensions, genetic influences for the two are largely overlapping. Developing a better understanding of genes–brain–cognition–behaviour pathways will enable us to tailor individualised prevention and treatment strategies for children who show the combination of callous–unemotional traits and
conduct problems. This genetically vulnerable subgroup with persistent antisocial behaviour requires early intervention. Given the negligible influence of shared environment for the antisocial behaviour in such children (Viding et al., 2005), prevention and treatment programmes may benefit from identifying and targeting child-specific environmental risk factors, such as differential parental treatment or developing programmes that capitalise on the specific cognitive and affective style of the child. For example, programmes that intervene early to promote the development of empathy and the internalisation of values or that use motivational strategies that capitalise on reward-oriented response style and appeal to self-interest may be particularly important for this group of children (Frick, 2001).

ACKNOWLEDGEMENTS

We thank the twins, their parents and teachers, Dr Frühling Rijsdijk for invaluable support with the statistical analyses, and the TEDS team, particularly Patricia Busfield, Jane McKay and Andrew McMillan. The TEDS receives support from the British Medical Research Council (G0424799). Research into development of antisocial behaviour also receives support from the UK Department of Health and Home Office (MRD 12-37). EV was supported by an ESRC postdoctoral fellowship during the analysis phase.

REFERENCES


Understanding the structure of the Psychopathy Checklist – Revised
An exploration of methodological confusion

DAVID J. COOKE, CHRISTINE MICHEL and JENNIFER SKEEM

Summary  Psychopathy is the key construct in the Dangerous and Severe Personality Disorder (DSPD) Programme. The Psychopathy Checklist – Revised is used as a primary means of selection for the programme. The Checklist confounds two distinct constructs – personality disorder and criminal behaviour. This confound is important both practically and theoretically. For example, under the criteria for DSPD it is necessary to demonstrate that personality disorder has a functional link with future risk of criminal behaviour. The confound has been exacerbated recently by claims that criminal behaviour is a core feature of psychopathic disorder. This contention is based on inappropriate analytical methods. In this paper we examine the source of this confound and illustrate how inappropriate methods can mislead.

Declaration of interest  None.
Funding detailed in Acknowledgements. Psychopathy is the key construct in the Dangerous and Severe Personality Disorder (DSPD) Programme, yet considerable debate surrounds the core features of this construct (e.g., Cooke, et al, 2004; Skeem et al, 2003; Hare & Neumann, 2005). Within DSPD assessment, a measure of psychopathy, the Psychopathy Checklist – Revised (PCL–R; Hare, 2003), is central for selection. A fundamental criticism of the PCL–R as a measure of psychopathy is that it confounds two distinct constructs – personality disorder and criminal behaviour (Lilienfeld, 1994). Failure to disaggregate the measurement of these two constructs renders it impossible to argue persuasively that psychopathic personality disorder produces criminal behaviour (Blackburn, 1988; Cooke et al, 2004). The demonstration of such a link is necessary to derive an individual under Article 5 of the European Convention on Human Rights. Definitions of psychopathy that entail criminal behaviour have long been recognised as essentially tautological (Ellard, 1988). In this paper we argue that measures and constructs of psychopathy are confused and that the introduction of inappropriate statistical methods has led to criminal behaviour being placed at the centre of the definition of the disorder. The tautology is thereby perpetuated.

MEASURES AND CONSTRUCTS
There are significant dangers when measures and constructs are confused; this is particularly the case under operationalism, where the measure is conflated with the construct (Campbell, 1960). Scores on the PCL–R are now being confused with the construct of psychopathy (see Skeem & Cooke, 2007). This forecloses on the possibility of examining the mapping of the theoretical construct (psychopathy) onto the empirical observation (PCL–R scores). Factor analysis is a tool that can inform our understanding, given its explicit recognition that all measures are fallible indicators of constructs; manifest variables (measures) are the product both of latent variables (constructs) and error. Hence, factor analytical approaches assume that latent variables produce the thoughts, feelings and modes of behaviour that are measured or recorded by item scores plus error (Edwards & Bagozzi, 2000). Factor analysis can partition the variance associated with each item into two parts: common variance, or variance associated with latent variables, and unique variance, or variance specific to that item and random error. Factor analysis thus explicates the multivariate relationships among the latent variables (constructs) that together influence the item ratings (empirical observations).

The structure of the PCL–R and its antecedents has been the subject of some debate. The original PCL–R manual (Hare, 1991) lacked clarity about the structure of the test (Cooke & Michie, 2001). For a number of years a two-factor model dominated the literature (Harpur et al, 1988). Unfortunately, the support for this model was over-reliant on congruence coefficients; these provide inadequate tests of the similarity of factor solutions across samples. Cooke & Michie (2001), using item response theory, confirmatory factor analysis and cluster analytical methods, argued that 13 of the 20 PCL–R items are conceptually distinct and psychometrically non-redundant indicators of psychopathy. Since they were found to be relatively poor indicators of psychopathy, items that tapped antisocial behaviour largely were excluded. Cooke & Michie (2001) developed a well-fitting hierarchical structure in which the superordinate trait, psychopathy, overarched three highly correlated symptom factors: arrogant and deceitful interpersonal style, deficient affective experience and impulsive and irresponsible behavioural style (see Fig. 1). The first factor was specified by glibness/superficial charm, grandiose sense of self-worth, pathological lying, and conning/manipulative, the second factor by lack of remorse or guilt, shallow affect, callous/lack of empathy and failure to accept responsibility for own actions, and the third factor by need for stimulation/proneness to boredom, irresponsibility, impulsivity, parasitic lifestyle and lack of realistic, long-term goals. This model, despite being described by some as ‘controversial’ (Salekin et al, 2006), is conceptually coherent (Skeem & Cooke, 2007) and consistent with clinical tradition (Cooke & Michie, 2001). Moreover, it has been replicated in a number of independent samples and by independent researchers using both the PCL (Jackson et al, 2002; Skeem et al, 2003) and other measures of psychopathic traits (Andershed et al, 2002). The three-factor model also has been shown to relate to external
correlates in a theoretically coherent manner (Hall et al., 2004).

There can be little doubt that the three-factor model has stimulated a number of researchers to reconsider the structure of the PCL–R measure and its implications for our understanding of the construct of psychopathic personality disorder. This must be regarded as positive: the definition and validity of constructs must be revisited as knowledge advances (Smith et al., 2003).

CRIMINAL BEHAVIOUR IN THE DIAGNOSIS OF PSYCHOPATHIC PERSONALITY DISORDER

Hare (2003) and his colleagues (Hare & Neumann, 2005; Neumann et al., 2007) have argued against the three-factor model and proposed a number of four-factor models. Essentially, within these models the three factors of Cooke & Michie (2001) are supplemented with a fourth ‘factor’ comprising five items related to criminal behaviour, i.e. poor behavioural controls, early behavior problems, juvenile delinquency, revocation of conditional release and criminal versatility. Previously, Hare and his colleagues had argued that psychopathy and criminality are distinct but related constructs (Hart et al., 1995; emphasis in original) and that psychopathy should not be confused with antisocial and criminal behaviour (Hare, 1999).

More recently, Hare & Neumann (2005) have argued that PCL–R items that capture antisocial tendencies, including criminality, are indicators of important psychopathic traits, asserting that the ‘real core of psychopathy has yet to be uncovered’ (p. 62). They observe that the exclusion of antisocial behaviour in the three-factor model decreases the utility of the PCL–R for predicting violence and aggression (see Skeem et al., 2003). Furthermore, they assert that ‘current findings suggest that the four-factor model has incremental validity over the three-factor in predicting important external correlates of psychopathy’ (Neumann et al., 2007: p. 22). This logic is confused. Adding variables, for example, gender, age or a history of substance misuse, would also improve prediction. However, such an improvement would not imply that these characteristics are core to psychopathic personality disorder. A measure’s validity in representing the construct of psychopathic personality disorder. A measure’s validity in representing the construct of psychopathic personality disorder. A measure’s validity in representing the construct of psychopathic personality disorder. A measure’s validity in representing the construct of psychopathic personality disorder. A measure’s validity in representing the construct of psychopathic personality disorder.

To inform the debate, we consider the appropriateness of various analytical strategies and demonstrate their impact on the model selected to describe the PCL–R. In the interests of transparency, we append as data supplements to the online version of this paper the code for all models tested (data supplement 1) together with the correlation matrix for the dataset we used (data supplement 2). This will allow other researchers to replicate – or reject – our conclusions. Our goal is to address three of the difficulties that confront the field in this debate about the structure of psychopathy. First, never are the competing models considered are often imprecise and thus it is hard for independent researchers to parameterise these models accurately. Third, contentious analytical approaches such as parcelling are adopted. We begin by describing the competing models, then consider key issues of method and conclude
by presenting analyses to illustrate these issues of method.

THE COMPETING PCL–R MODELS

The two-factor model

The two-factor model proposed by Harpur et al (1988) suggests that the interpersonal and affective items of the PCL–R coalesce to form a factor described as ‘the selfish and remorseless use of others’ (Hare, 1991, p. 76) and the items relating to behavioural instability, lack of planfulness and criminal behaviour coalesce to form a factor described as ‘the chronically unstable and antisocial lifestyle; social deviance’ (Hare, 1991: p. 76). The use of modern techniques of confirmatory factor analysis has demonstrated that this model is untenable (Cooke & Michie, 2001; Cooke et al, 2005a, b). Hare (2003) amended the two factors by adding an extra item, criminal versatility to the second factor (in the original two-factor model, this was one of three items included in PCL–R total scores, but omitted from factor scores). Below we refer to the original and amended 2-factor models.

The three-factor model

The three-factor model is illustrated in Fig. 1. There are perhaps four points of emphasis regarding this three-factor model. First, the structure is hierarchical, with a superordinate construct ‘psychopathy’ that is sufficiently unidimensional to be regarded as a coherent psychopathological construct or syndrome (Cooke & Michie, 2001; Cooke et al, 2005a, b). This hierarchical structure reflects a common model of personality and personality disorder in which traits of different levels of generality, from general to more specific, are structured in a hierarchical manner (McCrae & Costa, 1995). Second, the three factors can be regarded as having reliable general variance as a consequence of the influence of the broad psychopathy construct shared with the other factors. In addition, however, there is reliable specific variance unique to each particular factor. The value of refining the broad construct into specific factors has advantages in that the specificity between aspects of the disorder and external variables may be clearer (Raine et al, 2000; Soderstrom et al, 2002; Dolan & Anderson, 2003; Hall et al, 2004). Thus, this hierarchical model highlights ‘differential relations between the psychopathy factors and a variety of important criteria’ (Neumann et al, 2007: p. 24), but requires that the factors investigated are actually components of the general disorder of psychopathy.

Third, although some variants of the original three-factor model exclude testlets for the sake of parsimony (Skeem et al, 2003; Odgers, 2003; see Fig. 2), below the level of specific factors, and between the items, are testlets. Testlets occur when items are more highly associated than can explained by their relationship with the underlying latent trait; thus, a pair of items that form a testlet can be construed as being somewhere between one and two items (Chen & Thissen, 1997). Indeed, the use of item response theory demonstrated that all PCL–R items other than poor behavioural controls form testlets (Cooke & Michie, 2001). Testlets do not merely capture shared error variance, instead testlets are conceptually meaningful. Testlets combine specific indicators to form higher-order facets within the hierarchy of personality features.

Fourth, the model entails only 13 of the 20 PCL–R items. The 7 excluded items primarily reflect antisocial behaviour rather than core traits; it is possible to achieve maximum scores on some of these items without any evidence that the behaviour is trait-like, i.e. persistent and pervasive. These items failed to coalesce into a coherent syndrome and form a clear structure with the three factors, and generally demonstrated poor discrimination (Cooke & Michie, 1997, 2001).

The four-factor models

The current debate is frequently described as a choice between a three- and a four-factor model. This is misleading as there are two three-factor models and many four-factor models (Hare, 2003; Hare & Neumann, 2005). Frequently, authors fail to distinguish between these models and this creates conceptual confusion. We can identify at least ten four-factor – or equivalent – models in the literature. We describe these as: (a) a four-factor hierarchical model; (b) a two-factor, four-facet hierarchical model; (c) a four-factor correlated model. Since each of these models can be ‘parcelled’, we also describe a four-factor parcelled model.

A four-factor hierarchical model

Hare (2003) implied that four factors (i.e. the three factors from the Cooke & Michie (2001) model together with a criminality ‘factor’ specified by five items that tap criminal behaviours) are in a hierarchical relationship with the superordinate psychopathy factor (Fig. 3). Although Hare (2003) argued that the model ‘envelopes’ the three-factor model, to date no results have been provided to support this position.

A two-factor, four-facet hierarchical model

Hare (2003) asserted that ‘a superior test structure’ (p. 85) is one in which two superordinate factors (i.e. minimal modifications of the original two factors) are interposed between the four factors and the superordinate psychopathy construct (Fig. 4). Although Hare (2003) asserts that this is an improvement over ‘a model that goes directly from factors to the overall superordinate factor’ (p. 85), no results are provided to support this contention. Support is particularly important because the model specified by Hare (2003: Fig. 7.1) has several equivalents (models that yield the same covariances or correlations but have different paths among the latent variables). Indeed, the two-factor, four-facet hierarchical model has six equivalent models. For example, a correlation between factor 3 and factor 4 – or indeed, any other pair of factors – is mathematically equivalent to the one that Hare (2003) selected. When a model has mathematically equivalent versions, the models cannot be distinguished statistically. No model can be shown to be statistically superior. Instead, model selection must be based on sound theory: it is an established principle that researchers must justify their preference for a particular model over mathematically identical ones (Kline, 1998; Martens, 2005).
A four-factor correlated model

A number of researchers (e.g., Hare, 2003; Hill et al, 2004; Hare & Neumann, 2005) have presented correlated factor models in which the three factors from the three-factor model, together with the fourth criminality ‘factor’, are all inter-correlated (Fig. 5). Hence, each factor (e.g. factor 1) is correlated with all of the other factors (e.g. factors 2, 3 and 4). Neumann et al (2007) contend that correlated factor models are superior to the hierarchical models previously offered.

Parcellled variant of the four-factor correlated model

All of these four factor models can be subjected to parcelling, a procedure where items are summed to form composites prior to factor analysis (Hare, 2003). Parcelling creates even more conceptually and mathematically distinct models. Although these are described as four-factor or four-facet models, parcelling essentially yields one-factor models (i.e. one latent factor specified by four manifest composites). For illustrative purposes, we present the parcellled two-factor four-facet hierarchical model (Fig. 6).

ISSUES OF METHOD IN TESTING MODEL FIT

The debate raises methodological issues about how best to model the structure of a test such as the PCL–R.

Correlated v. hierarchical models

The work of Hill et al (2004) highlights the emerging difficulty in distinguishing between hierarchical and non-hierarchical models (Hare & Neumann, 2005; Neumann et al, 2007). A key feature of a hierarchical model is the demonstration that the higher order construct of interest is sufficiently unidimensional to be regarded as a coherent psychopathological syndrome. For two- or three-factor models, all correlated models are inherently hierarchical in that they are mathematically equivalent to models with a superordinate factor overarching subordinate factors. For models with four or more factors, this is no longer the case. In terms of statistical modelling, a PCL–R three-factor correlated model has three correlations among the factors, and the hierarchical model also has three loadings on the superordinate psychopathy factor. In contrast, the PCL–R four-factor correlated model has six correlations among the factors, whereas the hierarchical model has four loadings on the superordinate psychopathy factor. The hierarchical model is more parsimonious, more constraints being placed on the model, and thus it is a more demanding model to fit.

Nevertheless, proponents of the four-factor model strongly favour nonhierarchical models: ‘we recommend using first-order models with correlated factors in future research’ (Neumann et al, 2007). The assumption is that ‘the strong correlations between the factors . . . reveal that they are indicators for a second-order latent variable’ (Neumann et al, 2007). This assumption that correlated and hierarchical models are the same is misleading. It is necessary to explicitly compare a four-factor hierarchical model with a four-factor correlated model.

This issue has fundamental conceptual importance. The three-factor hierarchical model implies that psychopathic personality disorder (the superordinate factor) is underpinned by distinct constellations of interpersonal, affective and lifestyle traits (the first-order factors): the expression of these trait constellations is caused both by the overarching disorder and specific variance associated with the factor. The four-factor correlated model does not imply the presence of an overarching disorder that produces particular symptoms. Instead, these symptoms could be merely a hodgepodge of domains that co-occur. Essentially, the correlated model implies a compound trait composed of distinct constructs without a common cause (Smith et al, 2003). For example, measures of psychopathy are often associated with indices of alcohol and drug misuse and/or addiction.
There are no reasonable arguments for neglecting this distinction. Proponents of the four-factor model argue that if the four factors have differential associations with external correlates, then it would be unwise to employ a superordinate model to seek out such differential associations. This argument conflates the scoring and application of a measure (PCL–R total scores or PCL–R factor scores) with the understanding of a construct (psychopathy, with specific trait constellations). Hierarchical models represent specific factors that underpin superordinate constructs. Unlike correlated models, hierarchical models require that the specific factors included in the model be part of a coherent construct. Thus, hierarchical models have great potential for understanding both psychopathic personality disorder and its specific components.

The use of testlets v. correlated errors

Local dependency occurs when there is consistency among item responses that cannot be explained by individual differences on the latent trait being measured. Testlets are groups of items that show local dependence (a testlet formed of two items may be viewed as somewhere between one and two items). Although local dependence can emerge for a variety of reasons, with a rating scale the most common reason is the overlap of item definitions. PCL–R definitions are often overlapping. For example, ‘lack of remorse or guilt’ and ‘failure to accept responsibility for own actions’ both require consideration of whether the individual externalises blame. In creating the screening version of the PCL–R (the PCL–SV), Hare and colleagues recognised this issue and grouped conceptually related PCL–R items to produce distinct PCL–SV items (Hart et al, 1995; Cooke et al, 1998).

If there is consistent evidence that testlets exist within a scale this indicates local dependence. Local dependence is an undesirable property of a scale for three reasons. First, local dependence complicates the structure underpinning the data and can incorrectly challenge the assumption that a unidimensional trait underpins the test. This is crucial if data are to be subjected to parcelling. Second, local dependence leads to an overestimation of the true amount of information provided by the test. That is, the test appears to be more accurate than it actually is because
information is double-counted. Third, the ratings do not allow clinicians to adequately distinguish between conceptually distinct symptoms. Although testlets can be confused with correlated errors, the two concepts are conceptually and mathematically distinct. Conceptually, unlike correlated errors, testlets explicitly describe the measurement model, specifying theoretically meaningful sub-facets within the hierarchical description of the disorder. For example, 'pathological lying' and 'conning/manipulative' combine to describe a deceptive interpersonal style. This was implicitly acknowledged when these two PCL–R items were combined to create the one PCL–SV item called ‘deceitful’ (Hart et al, 1995). Correlated errors are more opaque – they do not provide this additional level of description. Mathematically, testlets are more elegant than correlated errors. A model with a three-item testlet is more parsimonious than a model with three items with correlated errors that load on the same factor: the former requires two parameters, the latter three parameters.

We are criticised for including testlets in our three-factor model (Neumann et al, 2007). In our view, any attempt to provide an accurate model of the structure of the PCL–R should consider testlets, even if only to reject the need for their inclusion in any model.

The use of parcelling

In structural equation modelling a parcel is an aggregate-level indicator derived by combining individual items (e.g. adding individual PCL–R items to derive a new manifest variable; Little et al, 2002). This is a controversial technique (Bandalos, 2002; Little et al, 2002). Proponents of the technique argue that parcelling has two broad advantages. First, combining
items results in composite variables with better psychometric properties than item variables (e.g. greater reliability, a higher ratio of common-to-unique factor variance, smaller and more equal intervals between scale points and distributions that are less likely to violate distributional assumptions; Little et al., 2002; Kim & Hagtvet, 2003). Second, parcelling results in models with better fit indices. This is because they reduce sources of sampling error, require fewer parameters and are less likely to be affected by correlated residuals or dual loadings. Broadly, the number of variances and covariances that the model must account for is reduced (Bandalos, 2002; Little et al., 2002; Kim & Hagtvet, 2003; Martens, 2005).

Opponents of parcelling point out five problems. First, parcelling can obscure the multidimensional nature of the items. Bandalos (2002) noted that when the assumption of unidimensionality is not met (an assumption that is rarely tested) ‘the use of parcels can obscure rather than clarify the factor structure of the data’ (p. 80). This is clearly a problem when there is evidence of local dependency, as there is with the PCL–R items (Cooke & Michie, 2001). Second, the improvement in fit is more apparent than real; models that do not fit at an item level can be made to appear to fit with parcelling. Bandalos & Finney (2001) noted that parcelling improves model fit, irrespective of whether the model is specified correctly or not; this also reduces our ability to detect mis-specified models. Kim & Hagtvet (2003) demonstrated empirically that when parcellled and item models were compared, the parcellled models yielded better fit statistics. Unlike the item models, the parcellled models pointed to the acceptance of mis-specified models. Third, Bandalos (2002) reported that parcelling can bias estimates of structural parameters (e.g. path coefficients; ‘factor loadings’). Fourth, comparisons of factor structure across groups for parcellled variables vary considerably from those based on individual items (Bagozzi & Edwards, 1998). This will affect our understanding of important issues, including cross-cultural variation in psychopathy and variation across gender, age and race. Fifth, even if the use of parcelling is defensible statistically, from a clinical perspective it can result in the loss of important information about the condition being considered (Little et al., 2002). When one sums across several items and then examines the relation between that sum (e.g. parcellled interpersonal facet) and an external variable (e.g. dominant behaviour), it is impossible to know that one aspect of the sum (e.g. grandiosity/charm) strongly predicts the external variable, whereas the other (e.g. deception) is unrelated. Data aggregation results in a loss of information.

When is it legitimate to use parcelling to analyse PCL–R data? Justification of the approach depends on (a) the purpose of the analysis and (b) the analytical strategy adopted before parcelling is undertaken. Little et al. (2002) noted that ‘careful delineation of the goals of the study is clearly the paramount issue’ (p. 6). If the purpose of the analysis is to explicate the interrelations among items on a test for construct validation purposes, then parcelling is inappropriate (Rogers & Schmitt, 2004). However, if the goal is to examine the interrelations among well-established measures of latent traits then parcelling may be appropriate. In the latter case it is assumed that the structure of the latent traits is well established (i.e. is not the subject of debate) and the primary interest is in putative causal relations among the latent traits rather than the measurement model (Little et al., 2002; Rogers & Schmitt, 2004; Martens, 2005). Rogers & Schmitt (2004) warned, however, that even when a measure has been validated at the item level in terms of a measurement model, parcelling of these items can still result in ‘undesirable or unpredictable effects on estimates and fit when testing the theoretical model’ (p. 380).

If parcelling is to be attempted, an essential prerequisite is an analysis of the dimensionality of the parcel (Bandalos & Finney, 2001; Bandalos, 2002; Little et al., 2002; Rogers & Schmitt, 2004). Unfortunately, this is more honoured in the breach than the observance. Kim & Hagtvet
(2003) demonstrated that the unidimensionality of a parcel must be established before it is entered into a more complex model. In the absence of unidimensionality the structural relations among latent traits cannot be interpreted (Little et al., 2002). Kim & Hagerty (2003) propose a method that explicitly models the single item and parcel indicators simultaneously, allowing a comprehensive evaluation of the unidimensionality of the parcel. Interestingly, this approach is formally equivalent to the testlet approach adopted by Cooke & Michie (2001). The three-factor approach with testlets provides greater understanding of the PCL–R items than a four-factor parcelled approach.

In summary, our understanding of the structure of the PCL–R, and to some degree our understanding of psychopathy, may be adversely influenced by the application of inappropriate methods for specifying the basic framework of the measurement model (correlated rather than hierarchical factors) and specific components (parcels rather than items and/or testlets).

**DEMONSTRATION OF METHOD ISSUES**

We now report a series of analyses of PCL–R data on a large sample of male offenders to illustrate the impact of using correlated $\nu$ hierarchical models, of parceling variables prior to model fitting and of modeling local dependency with testlets. Overall these analyses demonstrate that appropriate methods are necessary for appropriate conclusions.

**Participants**

The sample comprised a total of 1212 adult male offenders. The largest subsample comprised 608 adult male offenders from seven prisons in Her Majesty’s Prison Service (HMPS) in England and Wales, selected to be representative of the HMPS population. Additional subsamples included a representative sample of 246 offenders from the Scottish Prison Service (Cooke & Michie, 1999), a stratified random sample of 253 offenders from Scotland’s largest prison (Michie & Cooke, 2006) and a sample of 105 incarcerated Scottish offenders who volunteered to participate in a study of early childhood experiences (Marshall & Cooke, 1998). Only complete cases were used ($n=827$) to ensure that the same data were used irrespective of the model being tested.

**The procedure**

The PCL–R ratings were made according to instructions in the test manual (Hare, 1991). All PCL–R evaluations were conducted by trained raters.

**Confirmatory factor analysis**

Confirmatory factor analysis permits quantification of a particular factor structure’s fit within a particular sample. We assessed quality of fit using multiple indices, as each index has limitations (Kline, 1998; MacCallum & Austin, 2000). Different aspects of fit were evaluated, including absolute fit ($\chi^2$), fit adjusted for model parsimony (non-normed fit index, NNI) and fit relative to a null model (comparative fit index, CFI), and root mean square error of approximation (RMSEA). Following convention, the criterion for adequate

---

4. Another problem in the literature is the proliferation of underpowered studies. Confirmatory factor analysis requires moderate-to-large samples (Kline, 1998). Many of the attempts to explore the structure of the PCL measures have been seriously underpowered in terms of sample size, with samples at, or even well below, 150 individuals (e.g. Jackson et al., 2002; Hill et al., 2004; Salekin et al., 2006; Vitacco et al., 2006). Kline (1998) provides guidance on the issue and indicates that 20 cases per free parameter is desirable, $10:1$ is just acceptable and the statistical stability with a $5:1$ must be regarded as suspect. The three-factor hierarchical model with testlets has 36 free parameters, suggesting a minimum sample size of between 360 and 720. The four-factor hierarchical model has 40 free parameters (minimum $n=400–800$); the four-factor correlated model has 42 free parameters (minimum $n=420–840$) and the two-factor, four-facet hierarchical model has 41 free parameters (minimum $n=410–820$). Underpowered studies will mislead (Floyd & Widaman, 1995). In addition to the problem of lack of stability is the problem of Heywood cases. Small samples are prone to improper solutions in which estimated correlations are greater than 1 or estimated error variances are less than 0. Solutions may also fail to converge.
fit was defined as CFI and NNFI $\geq 0.90$ and RMSEA $\leq 0.08$ (Byrne, 1994). Following Kim & Hagvet (2003) we classified RMSEA values into four categories: close fit ($0.00-0.05$), fair fit ($0.06-0.08$), mediocre fit ($0.08-0.10$), and poor fit ($>0.10$).

Confirmatory factor analysis was performed using EQS for Windows (Bentler & Wu, 1995). Participants with missing data were deleted listwise for these analyses. Maximum likelihood estimation with robust-fit statistics and standard errors was used. The correlations were polychoric. Recommendations in the electronic help manual for the EQS 6 software suggested that this estimation approach is the best EQS approach for data of this type. This differs from the approach used in Cooke & Michie (2001). We also ran the analyses using the MPlus program with robust-weighted least-squares estimation and the same pattern of results was obtained.

### Table 1  
EQS Categorical Variables

<table>
<thead>
<tr>
<th>Model</th>
<th>Sartorra–Bentler $\chi^2$</th>
<th>d.f.</th>
<th>AIC</th>
<th>NFI</th>
<th>NNFI</th>
<th>CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Models with testlets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hierarchical three-factor with testlets</td>
<td>1497</td>
<td>170</td>
<td>1157</td>
<td>0.69</td>
<td>0.68</td>
<td>0.71</td>
<td>0.10</td>
</tr>
<tr>
<td>One-factor</td>
<td>743</td>
<td>118</td>
<td>507</td>
<td>0.82</td>
<td>0.82</td>
<td>0.84</td>
<td>0.08</td>
</tr>
<tr>
<td>Two-factor traditional</td>
<td>948</td>
<td>134</td>
<td>680</td>
<td>0.79</td>
<td>0.79</td>
<td>0.82</td>
<td>0.09</td>
</tr>
<tr>
<td>Hierarchical three-factor</td>
<td>277</td>
<td>62</td>
<td>153</td>
<td>0.91</td>
<td>0.91</td>
<td>0.93</td>
<td>0.06</td>
</tr>
<tr>
<td>Hierarchical four-factor</td>
<td>669</td>
<td>131</td>
<td>407</td>
<td>0.85</td>
<td>0.86</td>
<td>0.88</td>
<td>0.07</td>
</tr>
<tr>
<td>Hierarchical two-factor, four-facet</td>
<td>629</td>
<td>130</td>
<td>369</td>
<td>0.86</td>
<td>0.87</td>
<td>0.89</td>
<td>0.07</td>
</tr>
<tr>
<td>Correlated four-factor</td>
<td>623</td>
<td>129</td>
<td>365</td>
<td>0.86</td>
<td>0.87</td>
<td>0.89</td>
<td>0.07</td>
</tr>
<tr>
<td>Two-factor, four-facet parcelled</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Two-factor, four-facet 'wrong' factor</td>
<td>981</td>
<td>130</td>
<td>721</td>
<td>0.79</td>
<td>0.77</td>
<td>0.81</td>
<td>0.09</td>
</tr>
<tr>
<td>Two-factor, four-facet parcelled 'wrong' factor</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

AIC, Akaike information criterion; NFI, normed fit index; NNFI, non-normed fit index; CFI, comparative fit index; RMSEA, root mean square error of approximation.

5. The level of fit achieved on the development sample using this method is excellent. S-B $\chi^2=817$, d.f. = 56, AIC = 55, NFI = 0.98, NNFI = 0.98, CFI = 0.99, RMSEA = 0.04.

6. Some commentators have advocated the use of MPlus; the rationale for their preference is unclear. The same pattern of results was achieved using MPlus. The level of fit achieved with the three-factor model with testlets was good: $\chi^2=181$, d.f. = 40, CFI = 0.95, RMSEA = 0.06; for the three-factor model without testlets it was fair: $\chi^2=261$, d.f. = 40, CFI = 0.92, RMSEA = 0.08; and for the four-factor hierarchical model the fit was poor: $\chi^2=692$, d.f. = 73, CFI = 0.82, RMSEA = 0.10. Results for all models are available from the authors.

### Comparison of models

We started our analysis by estimating a one-factor model with all 20 items loading on a single latent trait. Fit statistics (Table 1) indicate that this is not a plausible model. We then tested the traditional two-factor model, which contains 8 items that load on the factor described as ‘the selfish, callous and remorseless use of others’ and 9 items that load on a factor termed ‘the chronically unstable and antisocial lifestyle; social deviance’ factor (Harpur et al., 1988). Fit statistics (Table 1) indicate that this too is not a plausible model. We then tested the amended two-factor model (Hare, 2003), i.e. we added the item ‘criminal versatility’ to the second factor. Fit statistics (Table 1) again indicate that this is not a plausible model.

We then fitted the full three-factor hierarchical model with testlets (Cooke & Michie, 2001). This model achieves a close fit with a CFI of 0.96 and an RMSEA of 0.05 (Table 1).

### Examining the fit of the fourth criminality ‘factor’

We then tested the three unparcelled four-factor models described in the literature; the four-factor hierarchical model, the two-factor, four-facet hierarchical model and the four-factor correlated model. None of these models achieve conventionally acceptable levels of fit (Table 1). Their level of fit is poorer than the level of fit achieved by even the degraded three-factor model.

### Exploring the impact of parcelling

Parcelling, or adding individual items together prior to model fitting, was used to achieve the fit indices presented for the four-factor models in the PCL–R manual (Hare, 2003: Figs 7.1, 7.3, 7.4). A prerequisite to parceling is to demonstrate unidimensionality for the items being parcelled (Bandolos, 2002; Kim & Hagvet, 2003; Rogers & Schmitt, 2004). The presence of testlets in PCL–R data (Cooke & Michie, 2001) means that this assumption is not met; that is, multiple latent constructs are tapped by items that are parcelled within individual factors of the four-factor models.
To examine the effects of parcelling on our dataset, we first estimated the two-factor, four-facet hierarchical model. Fit statistics (Table 1) indicate that this does not provide an adequate fit. We then parcellled the items by adding them together within their respective factor. We tested the fit of this model and fit statistics (Table 1) reveal a very good fit, with a non-significant $x^2$ value, and a CFI of 1.0.

We next tested the potential of parcelling to mislead. To do so, we compared the fit of an incorrect model that did, and did not, involve parcelling. The incorrect model involved swapping two item pairs within the two-factor, four-facet model: ‘pathological lying’ with ‘poor behavioural controls’ and ‘irresponsibility’ with ‘failure to accept responsibility for own actions’. Therefore, in this incorrect model, 4 of the 18 items loaded on the wrong factors. We then parcelled these 4 items into the same wrong factors.

Not surprisingly, the fit statistics for the un parcelled model indicate that swapping items substantially degraded the model’s fit (Table 1). Indeed, the fit statistics suggest that this is an incorrect model. In contrast, the fit statistics for the parcelled model indicate an extremely good fit with a non-significant $x^2$ value, a CFI of 1.00 and a RMSEA of 0.00. We concluded that parcelling is an inappropriate technique when the intent is to understand the interrelations among PCL–R items.

**DISCUSSION**

These analyses yield three broad conclusions. First, of the un parcelled models, the original three-factor model with testlets achieves the best fit. Second, in keeping with the methodological literature, parcelling can achieve excellent fit even when items are placed on the wrong factor: it is thus an inappropriate analytical approach for understanding the structure of the PCL–R. Nonsensical models of PCL–R psychopathy can masquerade as ‘excellent fits’ when items are parcelled. Third, the analyses provide no evidence that models that add a criminal behaviour factor to the three-factor model achieve satisfactory fit.

**Is criminal behaviour central to the construct of psychopathy?**

Proponents of the four-factor model(s) embrace the three Cooke & Michie factors within their models. The point at dispute is the role of additional items that essentially enumerate antisocial acts. None of the non-parcelled PCL–R models that add these antisocial behaviour items achieve acceptable fit. This indicates that these models do not provide adequate measurement models for the PCL–R. This is true whether the models involve hierarchical factors or (the less demanding) correlated factors. In our view there is no compelling empirical evidence to support the conclusion that antisocial behaviour is a central feature of psychopathy.

In addition, there are good conceptual (logical, theoretical) reasons for considering antisocial behaviour to be causally downstream from psychopathic personality disorder (Cooke et al., 2004; Skeem & Cooke, 2007). First, classical clinical descriptions of psychopathy do not give a central role to antisocial behaviour (Schneider, 1950; Karpman, 1961; Arieti, 1963; Cleckley, 1988). As Blackburn (2005) noted ‘criminal behavior was not intrinsic to Cleckley’s concept’ (p. 279). Indeed, Cleckley (1988), referring to the propensity to be antisocial in general, and seriously criminal in particular, indicated that ‘such tendencies should be regarded as the exception rather than as the rule, perhaps, as a pathologic trait independent, to a considerable degree, of the other manifestations which we regard as fundamental’ (p. 262). The critical feature for Cleckley was not the occurrence of criminal behaviour in itself but rather the occurrence of criminal behaviour for which the motivation is obscure. Simple counts of criminal acts cannot address this subtlety.

Second, it is plausible that the characteristic traits of psychopathy have a functional link with antisocial behaviour. Individuals who are grandiose frequently have a strong sense of entitlement that permits them to steal from, rape and exploit others. Those who lack empathy and anxiety may fail to inhibit violent thoughts and urges. Impulsivity increases the likelihood that these individuals will carry out criminal acts without considering the consequences (Cooke et al., 2004; Skeem & Cooke, 2007).

Third, specific socially deviant acts are qualitatively different from the pervasive and persistent personality traits that underpin the three factors within the PCL–R (Blackburn, 1988). As Lilienfeld (1994) noted it is important not to conflate ‘basic tendencies’ (traits) and ‘characteristic adaptations’ (specific acts).

Fourth, antisocial behaviour has been linked to a number of mental disorders (e.g. psychotic disorders, learning disability, substance misuse and other personality disorders). It is thus a non-specific indicator (Blackburn, 1988; Skeem & Mulvey, 2001). Theories of crime have implicated a multitude of factors in relation to antisocial behaviour (Gottfredson & Hirschi, 1990).

We would argue that there are thus strong theoretical and empirical reasons for excluding measures of criminal and antisocial acts from attempts to measure the construct of psychopathy; not least because it represents significant construct drift (Blackburn, 2005).

**The importance of understanding the structure of a measure**

Some commentators have argued that the three-factor model differs very little from the two-factor model and it is of little importance whether two-, three-, or four-factor models are used. For example, Jones, et al (2007) expect the three- and four-factor models to ‘perform alike’ because the ‘models are quite similar’. They opin
that the decision to use either model will hinge on personal preference or ‘researchers’ underlying conceptualisation of psychopathy’. We disagree. Given that the set of symptoms being modelled is the same, the content of any derived structure will inevitably be similar. This does not mean that the underlying structures are the same. Obtaining greater understanding of the structural properties of a disorder can yield many advantages (Watson et al., 1994).

First, it can serve as a starting point for the identification of fundamental psychological structures or processes. Structural research on IQ tests revealed distinct verbal and spatial factors; subsequent neuropsychological research indicated that these factors measured separate neural sub-systems (Watson et al., 1994).

Second, understanding structure can inform theories of causation: are the distinct facets products of some common underlying tendency towards psychopathy or are they not true facets but merely a number of distinct constructs without a common cause?

Third, explication of the structure can improve investigations of construct validity. If items are not grouped into unidimensional constructs, their relation to other variables in the nomological net may not be readily apparent. Associations with cognate variables based on a broad measure of a construct effectively average the associations underpinned by distinct factors; it is not clear which part of the measure drives the association, with the average association frequency being weaker than that of the strongest factor.

Fourth, an appreciation of structure can improve scales; and can provide direction on where new variables should be added to improve construct representation or removed to reduce construct-irrelevant variance (Lilienfeld, 1994; Floyd & Widaman, 1995; Little et al., 2002). Construct under-representation occurs when a measure fails to capture key aspects of the latent construct: it has been argued elsewhere, for example, that the PCL–R fails to adequately assess problems of self, attachment and interpersonal style which are central to the construct of psychopathy (Cooke et al., 2006). Construct-irrelevant variance occurs when the measure captures aspects of latent constructs other than the target latent construct. It is our contention (Cooke et al., 2004; Skeem & Cooke, 2007) that the inclusion of counts of criminal and other antisocial behaviour in the PCL–R represents construct-irrelevant variance.

Conclusion

The validation of a construct is never complete. Validation is important for reasons of theory and for reasons of practice. The field is in danger of falling into the trap of operationalism: conflating a fallible measure of psychopathy (PCL–R) with the construct of psychopathy. Psychopathy and criminal behaviour are distinct constructs. If we are to understand their relationships and, critically, whether they have a functional relationship, it is essential that these constructs are measured separately. This is particularly critical within the context of the function between their personality disorder and the risk that they pose. Recently, we have endeavoured to develop a more comprehensive model of the construct of psychopathy. Using clinical informants and a trait-descriptive algorithmic approach we have identified – after numerous iterations – a list of 33 symptoms that are grouped rationally into six domains of functioning (interpersonal attachment; behavioural; cognitive; interpersonal dominance; emotional; and self). This model is currently being subjected to empirical evaluation.

This study is not without limitations. First, we were not able to test the various models on the data from the PCL–R manual (Hare, 2003). Second, we were unable to demonstrate the chief problem inherent in correlated (rather than hierarchical) PCL–R model; that any correlate, whether essential to psychopathy or not, will fit. In future research, we will determine whether adding a non-psychopathic factor (e.g. addiction) to core PCL–R factors yields adequate fit indices in correlated factor models and (appropriately) poor fit indices in hierarchical models. Third, the results focus only on adult males prisoners; the generalisability of the results to other groups, including female offenders, remains unclear (Forouzan & Cooke, 2005).

ACKNOWLEDGEMENTS

D.C. received support from the Research and Development Directorate of the Greater Glasgow Primary Care NHS Trust. We thank Brian Rae for his continued support and Stephen Hart for his comments and suggestions for particular analytical strategies.

REFERENCES


Cooke, D. J., Michie, C., Hart, S. D. et al. (2005b) Searching for the pan-cultural core of psychopathic personality disorder: Continental Europe and North
factor analytic findings. Psychological Assessment, in press.


Critical developments in the assessment of personality disorder

PETER TYRER, NATALIE COOMBS, FATEMA IBRAHIMI, ANAND MATHILAKATH, PRIYA BAJAJ, MAJA RANGER, BHARTI RAO and RAANA DIN

Background  The assessment of personality disorder is currently inaccurate, largely unreliable, frequently wrong and in need of improvement.

Aims  To describe the errors inherent in the current systems and to indicate recent ways of improving personality assessment.

Method  Historical review, description of recent developments, including temporal stability, and of studies using document-derived assessment.

Results  Studies of interrater agreement and accuracy in diagnosis in complex patients with independently established personality status using document-derived assessment (PAS—DOC) with a four personality cluster classification, showed very good agreement between raters for the flamboyant cluster B group of personalities, generally good agreement for the anxious/dependent cluster C group and inhibited (obsessional) cluster D group, but only fair agreement for the withdrawn cluster A group. Overall diagnostic accuracy was 71%.

Conclusions  Personality function or diathesis, a fluctuating state, is a better description than personality disorder. The best form of assessment is one that uses longitudinal repeated measures using a four-dimensional system.

Declaration of interest  P.T. is Editor of the British Journal of Psychiatry, but had no part in the assessment of this manuscript for publication.

The assessment of personality and its range of abnormality, a range that is much greater than that implied by the conventional label of ‘personality disorder’, is one of the critical elements of a psychiatric examination. However, it is frequently omitted in clinical assessments, and even in research studies it is rarely assessed formally, even now, at a time when personality disorder is highly topical and its diagnosis possibly a reason for compulsory admission and treatment. When personality is assessed it is often done in a cursory and brief manner, and again this extends to research studies. Thus, for example, a review of all the 152 original papers published in the British Journal of Psychiatry in 2005, revealed 13 (8.6%) in which personality assessment was at least part of the focus of the paper, in 5 of which (3.3%) it was the main subject, and 14 other papers (9.2%) in which general psychopathology was assessed but personality status was omitted. One might have expected that most of the papers addressing personality status would have used a formal assessment instrument. However, only 3 of the papers did so. These were: (a) a careful review (Cooke et al, 2005) of cross-national variations with the Psychopathy Checklist – Revised (Hare, 1991); (b) a study of risk factors for repeated self-harm (Sokero et al, 2005), which used a structured interview for personality disorders (SCID–II; Spitzer et al, 1987); and (c) an examination of personality comorbidity (Khan et al, 2005), which assessed personality by self-ratings using Cloninger’s Tri-Dimensional Questionnaire (Cloninger et al, 1991) and Eysenck’s Personality Questionnaire (Eysenck & Eysenck, 1975). The other two papers devoted to personality described new methods of assessment (Bradley et al, 2005; Thompson-Brenner & Westen, 2005), which reflects the low level of belief in existing ones. Seven of the studies merely used standard ICD (editions 8–10; World Health Organization, 1992) or DSM–III–R/DSM–IV (American Psychiatric Association, 1994) diagnoses of personality disorder from case records. Perhaps the most interesting revelation came from what is probably the first structural neuroimaging study of lying (Yang et al, 2005). One might have thought that this was a subject for which standard personality assessment would have been both natural and essential. However, the authors felt it necessary to construct a portmanteau instrument derived from the PCL–R, DSM–IV and an extra criterion for malinger. It is hardly surprising that the findings of the study (increased pre-frontal white matter in liars) has attracted a great deal of attention when the authors are unable to find an existing rating instrument that can even make a passable attempt at discriminating liars from non-liars.

Nevertheless, there have been advances in the assessment of personality disorder and currently a great deal is expected of it in terms of accuracy and precision, particularly in forensic psychiatry. Indeed, a great deal was expected of it in the past, particularly in military psychiatry during the USA in the Second World War, but there it had a poor record of success and had to be abandoned (Wessely, 2005).

Assessment is linked closely to classification and the two subjects need to be discussed in tandem before examining ways of improving current assessment strategies, particularly in the context of new forensic initiatives.

BRIEF HISTORY OF CLASSIFICATION AND ASSESSMENT

Classification of personality has a long history. Hippocrates hypothesised that all illness was a result of imbalance in the four humours of yellow bile, black bile, phlegm and blood, and Galen extended this further to personality by describing personality types linked to excess of each of these: choleric (yellow bile), melancholic (black bile), phlegmatic (phlegm) and sanguine (blood). Although other attempts were made to formalise groupings of abnormal personality, they really did not attract any following until Schneider (1923) formulated his famous list of psychopathic personalities that he conceptualised as distinct from other mental illnesses. He regarded the term ‘psychopathic’ literally (i.e. as a pathology of mind) rather than
as a synonym for ‘antisocial’ as was commonly used by English-speaking writers. Schneider’s ten categories of psychopathic personality were: hyper-thymic, depressive, insecure (sensitive and anankastic sub-categories), fanatical, attention-seeking, labile, explosive, affectionless, weak-willed and asthenic. Many of these have persisted in one form or another since 1923 and Standaage (1979) found that the asthenic, explosive, depressive and affectionless were the most reliably rated. The current categories of dependent, impulsive (ICD only), depressive (extended DSM only) and schizoid are very similar to Schneider’s descriptions of these four personalities.

When DSM–III was formulated (American Psychiatric Association, 1980) two critical decisions were made. The first was to give personality disorders a separate axis (Axis II) in the classification. The official reason for this was a pragmatic rather than a scientific reason. There was concern, probably justified in view of subsequent developments, that the diagnosis of personality disorder would be forgotten when it competed with other disorders.

The unofficial reason was that the psychotherapists advising the task force were very unhappy with much of DSM–III and were offered a separate axis as a *quid pro quo* for accepting the main Axis I descriptions. ICD–10 (World Health Organization, 1992) retained personality disorder on Axis I and introduced Axis II for disability and function, so in this respect, and this only, did it differ fundamentally from DSM. Which is right remains open to much debate, and, after reviewing the arguments Kendell (2002) wrote:

> ‘It is impossible to conclude with confidence that personality disorders are, or are not, mental illnesses; there are ambiguities in the definitions and basic information about personality disorders is lacking’.

The second decision was to use clearly defined operational criteria to define the behavioural elements of personality disorder according to the 11 chosen categories in the classification. This was understandable in view of the success of this approach in depression and schizophrenia, but was a mistake with personality disorder. The main reason for the failure of the classification was that the definitions of personality disorder used heterogeneous descriptions, and when all their operational criteria were assessed carefully their distribution was quite unlike that of DSM ( Livesley et al., 1994). The alternative of a dimensional classification, most commonly based on traits rather than behaviour, existed before the introduction of DSM–III and has been revised and reformulated many times since (Persly & Walton, 1973; Tyrer & Alexander, 1979; Clark et al., 1996; Mulder & Joyce, 1997; Widiger & Simonsen, 2005), but only now is beginning to have a realistic possibility of being adopted by the world community.

The dimensional system contemplates personality as a continuum, with normal variation at one extreme and what is currently called personality disorder at the other. The best fit is based on four dimensions which are not unlike the original classification system of Hippocrates and Galen (Table 1), particularly when one realises that in the past ‘sanguine’ or ‘full of blood’ was synonymous with confidence and stubborn determination, and ‘phlegmatic’ was equivalent to dull and cold indifference. There continues to be some debate over whether the normal/abnormal personality continuum is best served by three, four or five dimensions (Widiger & Simonsen, 2005), but a very strong case can be made for sticking to four to maintain historical continuity as well as general accuracy (Table 1).

In examining the assessment of personality disorder it is therefore necessary to examine both dimensional and categorical approaches even though at present both world classifications in psychiatry adopt the categorical model of disorder. However, even if DSM–V and ICD–11 persisted with the present unsatisfactory system, an alternative one would have to be used to link with studies of normal personality and its variation. As Widiger and Simonsen (2005, p. 126) stated:

> ‘It is impossible to conclude with confidence that personality disorders are, or are not, mental illnesses; there are ambiguities in the definitions and basic information about personality disorders is lacking’.

The first problem arising in the assessment of personality disorder is the level of agreement between different systems of diagnosis. Others include the stability (or, more accurately, the instability) of current assessment methods in personality disorder,

### Table 1

<table>
<thead>
<tr>
<th>Description</th>
<th>Personality dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hippocrates/Galen</td>
<td>Choleric</td>
</tr>
<tr>
<td>Tyrer &amp; Alexander (1979)</td>
<td>Sociopathic</td>
</tr>
<tr>
<td>Eysenck (1987)</td>
<td>Extraversion</td>
</tr>
<tr>
<td>Livesley (1990)</td>
<td>Dissocial</td>
</tr>
<tr>
<td>Costa &amp; Macrae (1990)</td>
<td>Disagreeable</td>
</tr>
<tr>
<td>Mulder &amp; Joyce (1997)</td>
<td>Antisocial</td>
</tr>
<tr>
<td>Current cluster model (DSM–IV)</td>
<td>Cluster B</td>
</tr>
<tr>
<td>Proposed cluster model</td>
<td>Cluster B (Dissocial)</td>
</tr>
</tbody>
</table>

1. Cloninger et al (1991) proposed a tridimensional structure to personality in 1987, and expanded this subsequently to four and then seven dimensions. The four-dimensional model (novelty-seeking, reward dependence, harm avoidance and perseverance) has some similarities to the above dimensions but is not a good fit.
2. A fifth factor, openness, is identified in this scale but not included here.
the problem of defining severity, particularly relevant in forensic psychiatry, and the source of information for assessing personality status.

**ASSESSMENT OF PERSONALITY DISORDER BY CATEGORIES AND DIMENSIONS**

The first basic requirement of an assessment is that it should be accurate. Accuracy includes elements of both reliability and validity. The latter is often more difficult to determine, as it requires a true measure of that which is being measured, and this genuine ‘gold standard’ is very hard to find in personality research (Cicchetti & Tyrer, 1988). However, reliability, the extent of agreement between assessors (inter-rater or test–retest reliability) is an essential first step. Zimmerman (1994) and Clark & Harrison (2001) have carried out an extensive review of published studies and their results are similar. Personality is assessed by a combination of self-report questionnaires, check-lists and interviews, of which the structured interview is currently considered the most robust.

The best possible level of interrater reliability should therefore come from a structured interview in which assessments are carried out jointly (i.e. the same material is assessed by the two assessors). The disappointing level of agreement shown in such settings is illustrated in Table 2; only one study (carried out with the interview schedule’s creator; Zanarini et al, 1987) reached the kappa agreement of 0.75 or above necessary to confirm excellent agreement (Cicchetti & Sparrow, 1981) for clinical purposes.

However, the level of agreement for the presence or absence of personality disorder is more satisfactory (Table 2), and this tends to be a uniform finding across a range of studies (Bronsich & Mombour, 1994; Zimmerman, 1994, Clark & Harrison, 2001). The mean kappa values for the categorical diagnoses (Table 2) hide tremendous variation as agreement for individual diagnostic categories varies from 0.25 to 0.9. By contrast, when similar assessments are made using the dimensional system the level of agreement tends to show agreement that is consistently 0.1–0.2 correlation points higher than categorical diagnoses (Loranger et al, 1991; Vittengl et al, 1999). This even applies to individual traits. Thus, for example, in a cross-national reliability study of the Personality Assessment Schedule (PAS; Tyrer et al, 1984) the individual levels of agreement across the separate ratings of 24 traits with both informant and participant interviews (i.e. 48 assessments) ranged from 0.52 to 0.94, with a mean agreement of 0.82 (informant assessment) and 0.75 (participant assessment) (Cicchetti & Tyrer, 1988: p. 71).

If these levels of agreement for categorical diagnosis are the best that can be achieved in ideal research settings with generally cooperative patients using instruments that take between 90 and 360 min to complete, it bodes ill for their reliability in general clinical practice. The major reason for the poor agreement is clear, if the operational criteria for individual diagnoses overlap then their identification will lead to the diagnosis of several personality disorders, even when they may be assessing the same single clear construct. The presence of multiple personality disorders is euphemistically called comorbidity, implying the presence of several independent disorders. However, when a diagnostic system fails and splits a common condition into several, the outcome is still called comorbidity when the correct term is consanguinity (Tyrer, 1996). An attempt to redress the confusion created by multiple personality disorders (a term that also cannot be used as it has been appropriated by dissociative disorders in the international classifications) is the cluster model. This has been used in the DSM classification for many years (Reich & Thompson, 1987) and has the advantages of reducing the overlap a little, bringing the odd, eccentric, withdrawn group (paranoid, schizoid and schizotypal; cluster A), the flamboyant, erratic and dissocial group (antisocial, histrionic, borderline and narcissistic; cluster B) and the anxious fearful group (dependent,
avoidant and obsessive-compulsive; cluster C), into more natural groupings. However, to fit in well with the four-factor dimensional model (Table 1) there should be a fourth cluster (cluster D) devoted to the obsessional group alone.

**CATEGORIES AND CLUSTERS OF PERSONALITY DISORDER**

Although DSM experts give only the faintest of praise to the cluster model (Widiger, 2005) and it has not been endorsed by ICD–10, it is becoming increasingly used (Evans et al., 1999; Bowden-Jones et al., 2004; Simeon et al., 2004; Bradley et al., 2005; Moran et al., 2006) in both clinical and research contexts because it simplifies what otherwise becomes a morass of comorbidity. To use this with ICD it is necessary to exclude schizotypal from cluster A, narcissistic from cluster B (but adding impulsive) and renaming antisocial as dissocial and obsessive-compulsive as anankastic. The advantages of the cluster system follow mainly from its links to basic personality structure (Table 1) but also can be helpful in improving reliability, even though this can only be a qualified improvement as the basic disorders remain unaltered. This is illustrated by a recent comparison of the reliability of a short assessment of personality (Quick Personality Assessment Schedule (PAS-Q; Tyrer, 2000a) with a longer structured version based on ICD–10 (PAS-I; Tyrer, 2000b) in 72 patients in an assertive outreach team. All had one or more prominent mental state diagnoses, as well as many personality disorders (Ranger et al., 2004), and approval for assessments of personality were agreed by the patients and by St Mary’s Hospital Ethical Committee. Both assessments were carried out by M.R. using a clinical informant interview. Informants had all known the patients closely for at least 2 years and to reduce carry over of information assessments were separated in time by a mean period of 9 months. The results showed the expected high variation in the reliability of individual diagnoses (kappa=0.26–0.70) (another reason for avoiding use of these in clinical practice) but somewhat greater agreement (kappa= 0.4–0.78) for the three clusters (Table 3). In general the cluster D diagnoses tend to be rated more reliably than cluster C as there is less overlap between their clinical features and those of other mental illness. This overlap is one of the main sources of difficulty when attempting to improve the accuracy of diagnosis (Tyrer et al., 1983; Hassiotis et al., 1997). With the separation of cluster D (inhibited or obsessional group) from cluster C the level of agreement is improved. For those involved in forensic assessment, the higher level of reliability for dissocial personality disorder is encouraging; the same level of superior agreement has been found in a forensic sample (Tyrer et al., 2005a).

**Table 3** Agreement between two personality interviews (Quick Personality Assessment Schedule (PAS–Q) and a longer version based on ICD–10 (PAS–I)) separated by personality category, cluster and severity in 72 patients with severe mental illness tested a mean of 9 months apart.

<table>
<thead>
<tr>
<th>Personality group or category using ICD–10 criteria</th>
<th>Agreement level, kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paranoid</td>
<td>0.45</td>
</tr>
<tr>
<td>Schizoid</td>
<td>0.59</td>
</tr>
<tr>
<td>Dissocial</td>
<td>0.70</td>
</tr>
<tr>
<td>Impulsive</td>
<td>0.45</td>
</tr>
<tr>
<td>Borderline</td>
<td>0.50</td>
</tr>
<tr>
<td>Histrionic</td>
<td>0.28</td>
</tr>
<tr>
<td>Anankastic</td>
<td>0.28</td>
</tr>
<tr>
<td>Anxious</td>
<td>0.26</td>
</tr>
<tr>
<td>Dependent</td>
<td>0.33</td>
</tr>
<tr>
<td>Cluster A</td>
<td>0.53</td>
</tr>
<tr>
<td>Cluster B</td>
<td>0.78</td>
</tr>
<tr>
<td>Cluster C</td>
<td>0.40</td>
</tr>
</tbody>
</table>

1. For severity of personality disorder (0–4) the intraclass correlation coefficient was 0.66

**INSTABILITY OF PERSONALITY ASSESSMENT**

One of the main defining features of personality disorders in both ICD and DSM classifications is that they are ‘pervasive’ and ‘ingrained’. It now looks as though this definition is also wrong, as we now have abundant evidence that personality status, at least that assessed by our current instruments, is unstable (Paris, 2002, 2003; Seivewright et al., 2002; Shea et al., 2002; Shea & Yen, 2003). Whereas in the past this lack of stability was regarded as a ‘contaminating’ effect of mental state or a poor assessing instrument, the evidence now that it seems to be universal has prompted a change in view. A consistent finding from all studies is that both in the short and longer term those patients who present for treatment with their personality disorders show a steady improvement (Table 4). This is generally greater for those with borderline personality disorder than others, but in the Collaborative Longitudinal Personality Disorder Study similar improvement was found in all four personality disorders (borderline, schizotypal, avoidant and obsessive-compulsive) after 2 years, with the highest rate of remission being 61% in schizotypal personality disorder (regarded as belonging to the schizophrenias in ICD–10) and the lowest 30% in avoidant personality disorder (Shea et al., 2002; Grilo et al., 2004). However, in personal studies using a self-rated instrument for dependant personality (Tyrer et al., 2004) dependent personality features show greater stability (Seivewright, 2003). In the longer term we have very clear accumulating evidence that borderline personality disorder in a treatment setting has a good outcome, but still have to be aware that suicide, the worst of outcomes, can occur at any stage, often late in the course of illness when the worst pathology seems to be over (Paris & Zweig-Frank, 2001).

The high level of instability of personality pathology, only a little less than that of major depressive disorder and more so than anxiety (Shea & Yen, 2003), has led to doubts that current instruments, working as they do with a failed classification system, do really indicate that personality is quite so unstable (Widiger, 2005), and there is also evidence of greater stability of social dysfunction in longer-term studies (Nur et al., 2004; Seivewright et al., 2004; Skodol et al., 2005a; Tyrer et al., 2005b). However, the genie is out of the bottle. We can no longer plod forwards developing new instruments that we hope will take us to the Holy Grail of temporal stability (Tyrer, 2005a) and refuse to accept that spontaneous change in personality features can take place independent of any treatment effects.

However, in acknowledging the improvement in clinical samples being treated for disorder we must also note that these populations are relatively uncommon in epidemiological terms. Most individuals with personality disorder (3 out of 4) in contact with services are treatment resistant (Type R) rather than treatment seeking (Type S; Tyrer et al., 2003), and in the normal population this proportion is even higher (C. Kirby, personal communication,
Recent studies demonstrating change in personality status in both clinical and population samples

<table>
<thead>
<tr>
<th>Authors</th>
<th>Population</th>
<th>Duration of study</th>
<th>Personality change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shea et al (2002); Skodol et al (2005a,b) (Collaborative Longitudinal Personality Disorder Study)</td>
<td>Clinical treatment-seeking population with borderline, schizotypal, avoidant and obsessive–compulsive personality disorders (n = 573)</td>
<td>6 months, 1 year and annually thereafter</td>
<td>10% of borderline patients remitted in first 6 months, 50% of all personality disorders within 2 years</td>
</tr>
<tr>
<td>Bernstein et al (1993); Cohen et al (2005) (Children in the Community Study)</td>
<td>Random community sample of children (n = 733)</td>
<td>5 assessments at (mean age of 14, 16, 22 and 33 years</td>
<td>General decline in personality pathology from 12 to 27 years (but 1 in 5 get worse; little change in cluster A and C personalities, more improvement in cluster B)</td>
</tr>
<tr>
<td>Zanarini et al (2003, 2006)</td>
<td>Treatment-seeking in-patients (n = 362)</td>
<td>2–12 years at 2-year intervals</td>
<td>35% in remission at 2 years, 49% at 4 years, 69% at 6 years</td>
</tr>
<tr>
<td>Links et al (1998)</td>
<td>Treatment-seeking in-patients (n = 57)</td>
<td>7 years</td>
<td>53% in remission at time of follow-up</td>
</tr>
<tr>
<td>Seiaviewright et al (2002)</td>
<td>Treatment-seeking patients with anxiety and depressive disorders (n = 201)</td>
<td>12 years</td>
<td>Significant improvement in patients with cluster B personality disorders, significant increase in cluster A personality disorders</td>
</tr>
<tr>
<td>Paris &amp; Zweig-Frank (2001)</td>
<td>Borderline hospital-treated patients (n = 64)</td>
<td>27 years</td>
<td>Only 5 (8%) still met the criteria for borderline personality disorder</td>
</tr>
</tbody>
</table>

1. Only those studies which had a formal assessment of personality status at baseline and follow-up are included.

2007). Those with borderline (mainly) and avoidant personality disorders (less prominently) (Emmelkamp et al, 2006) are the ones involved in most of the recent studies, although other approaches, particularly nido-therapy, which changes the environment, not the patient (Tyrer, 2002; Tyrer & Bajaj, 2005), may be suitable for the Type R majority. The findings that one in five children with abnormally personalities get worse in the Children in the Community Study (Cohen et al, 2005) and that older people who have had anxiety and depressive disorders in the past have a higher rate of cluster A personalities than when young (Schwartz et al, 2002) is a reminder that personality pathology can go in different directions. There is also evidence from epidemiological studies that cluster A pathology persists into older age (Reich et al, 1988).

### MEASUREMENT OF SEVERITY OF PERSONALITY DISTURBANCE

Epidemiological studies suggest that between 5% and 13% of the population has at least one personality disorder (Casey & Tyrer, 1986; de Girolamo & Reich, 1993; Torgersen et al, 2001; Coid et al, 2006a), so it is clear that it is a common condition. It is also equally apparent that some form of severity assessment is necessary to decide on priorities for management. This has become increasingly necessary when expensive provision is being made for small groups, such as those in the Dangerous and Severe Personality Disorder (DSPD) Programme in England (Home Office & Department of Health, 1999). The concept of dangerousness is often invoked when deciding on the severity of personality disorder, but this is mistaken. Dangerousness is not a function of personality disorder, as it can be present with many other mental disorders, or indeed, in the absence of disorder.

Unfortunately there is no measure of severity of personality disorder in the ICD or DSM classifications. This absence has caused significant concern, as it is highly relevant to the planning and provision of services. What is clear from empirical research studies is that those with more severe personality disorder do not have stronger manifestations of one single disorder as often postulated (Tyrer & Johnson, 1996), but instead their personality disturbance extends, ripple-like, across all domains of personality, so that in the most severe disorders there is virtually no satisfactory personality function in any area (Oldham et al, 1992; Dolan et al, 1995; Tyrer & Johnson, 1996). By using this measure of severity, and by giving special attention to those with marked antisocial personality features, thereby giving a separate level of ‘severe personality disorder’, it is possible to use the cluster system to get a measure of severity and a reasonable level of agreement (Table 3). This assessment is also relevant in assessing those with the most severe personality disorders, as there is some evidence of a different response in this group in high secure settings (Tyrer et al, 2006).

### PAS–DOC STUDY OF DOCUMENT-DERIVED PERSONALITY ASSESSMENT

Who provides the information for personality assessment is often overlooked. It is commonly assumed that the patient is the best source of information but, following the Robert Burns dictum, ‘O what gift would the lordie gie us, to see ourselves as others see us’, a close informant may be a much more accurate judge. Although there is no clear way of deciding whether an informant’s ratings are more accurate than those of the patient (Zimmerman, 1994), the additional information derived from interviewing an informant can be extremely valuable (Zimmerman et al, 1986), particularly if the informant is closely related and is female (Brothwell et al, 1992).

However, the value of written records describing the patient’s attitudes and habitual behaviour has only been appreciated fully by one group, those who make psychopathy with the Psychopathy Checklist (Hare, 1991). Although the record of interrater reliability and predictive reliability of instruments assessing personality disorder is disappointingly poor, the PCL–R, and its briefer fellow traveller, the screening
version (PCL–SV; Hart et al., 1995) go against the trend. These instruments attach great importance to written records without which the full PCL–R cannot be completed. The success of the PCL–SV in being the best single predictor of violence following the discharge of a psychiatric patient from hospital (Monahan et al., 2001) is unlikely to result from just the presence of superior psychometric properties; the bonus of the additional information derived from records is almost certainly critical, and helps over other methodologies (Moran et al., 2003). This is also important when the data show that half of all people with antisocial personality disorder show no significant violence (Coid et al., 2006b).

We have developed a document-derived version of the PAS (PAS–DOC) (Tyrer, 2005b) that has the same underlying structure as the parent instrument (Tyrer & Alexander, 1979) but has been adapted for written records, including those about children and adolescents. This latter process has been helped by the modification of the original PAS for use in adolescents where it has been of value (Rangel, 2005) and helps over other methodologies.

Reliability study
A single typed summary (2 pages) from the case notes of 20 patients involved in the earlier study with patients in an assertive outreach team (Table 3) was selected at random by an independent administrator and given to two assessors (N.C. and F.I.) who scored them independently using the PAS–DOC, which gives personality status after completion of a computer algorithm. The dimensional ratings of the four main clusters (A=withdrawn cluster, B=flamboyant cluster, C=dependent cluster, D=inhibited (obsessional) cluster) were rated for agreement using the intraclass correlation coefficient and also tested for rater bias (Cicchetti et al., 1976). The results are shown in Table 5. The best agreement was reached for the obsessional cluster (RI = 0.83), with the cluster B group (RI = 0.74) close behind. The scores for the withdrawn cluster, A, showed the lowest level of agreement (RI = 0.41).

Validity study
The 20 patients examined in the reliability study were all very well known to the clinical team and a consensus agreement of personality status had been agreed and recorded. A team of five raters, who had received some prior training only in the original PAS (this included N.C. and F.I. after they had completed the reliability study and before their data were analysed), each made an independent assessment of one volume of case notes (which contained none of the research information on personality status) using the PAS–DOC. In assessing the validity of the raters’ assessments it was assumed that a satisfactory assessment would make a correct

<table>
<thead>
<tr>
<th>Personality cluster</th>
<th>Intraclass correlation coefficient</th>
<th>Clinical significance (^\dagger)</th>
<th>Rater bias (^\ddagger)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.41</td>
<td>Fair</td>
<td>2.0</td>
</tr>
<tr>
<td>B</td>
<td>0.74</td>
<td>Good</td>
<td>5.4*</td>
</tr>
<tr>
<td>C</td>
<td>0.67</td>
<td>Good</td>
<td>2.6</td>
</tr>
<tr>
<td>D</td>
<td>0.83</td>
<td>Excellent</td>
<td>3.8</td>
</tr>
</tbody>
</table>

2. F ratio with 1 and 19 degrees of freedom.

Table 6 Comparison of the accuracy for both personality type and disorder of the Document-Derived Version of the Personality Assessment Schedule (PAS–DOC) using masked assessment of one volume of case notes for 20 patients whose personality status had been determined independently by consensus meetings of a clinical team.

<table>
<thead>
<tr>
<th>Patient number</th>
<th>Consensus personality status (gold standard)</th>
<th>Diagnostic accuracy (^\ddagger)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disorder present Cluster</td>
<td>Correct Incorrect</td>
</tr>
<tr>
<td>1</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>N</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>Y</td>
<td>2</td>
</tr>
</tbody>
</table>

1. With four and five raters using PAS–DOC for both type of personality disturbance and presence of personality disorder.
2. Overall diagnostic accuracy (for all patients) = 67/94 (71%); diagnostic accuracy for primary cluster A patients (n = 2) = 5/9 (56%); diagnostic accuracy for primary cluster B patients (n = 9) = 37/42 (88%); diagnostic accuracy for primary cluster C patients (n = 4) = 12/19 (63%); diagnostic accuracy for primary cluster D patients (n = 3) = 7/15 (47%); diagnostic accuracy for patients with no personality disorder (n = 2) = 6/9 (67%). There was also considerable variation between the overall diagnostic accuracy of the five raters, being 87%, 84%, 75%, 60% and 55%.
decision as to whether personality disorder was present and, if so, in which of the four clusters it would be placed, or, in the case of more complex personality disorders, which ones. Diagnostic accuracy was only regarded as positive if both type and presence or absence of personality disorder were correct.

The results showed that overall diagnostic accuracy was 71%, cluster B personalities were the most accurately identified (88%) and, in contradistinction to the reliability study, cluster D (obsessional/inhibited group) were the least well detected (47%). There was also considerable variation in accuracy between the raters (Table 6). In the context of the results it should be emphasised that all 20 patients had complex pathology (schizophrenia or schizoaffective disorder (11), bipolar disorder (5), recurrent self-harm (1), psychotic depression (1), multiple phobias (1) and obsessive-compulsive disorder (1), with 9 also having a history of drug misuse.

**IMPLICATIONS FOR FUTURE ASSESSMENT OF PERSONALITY**

There are two main conclusions arising from this review of studies and recent experimental work. The first is that personality and its disorder can no longer be regarded as a clear and stable entity that will yield eventually to the right form of assessment. What can be assessed accurately at a point in time is personality function, not disorder. Just as mental state can be dependent on environmental influences, so can personality status, and this can be made use of in therapy (Tyzer, 2002; Tyzer & Bajaj, 2005). The notion of personality function, first expressed clearly by Bronisch & Klerman (1991), has been confirmed by recent studies showing that personality functions in different ways at different ages and in response to different needs. At the same time we must also recognise that there are some underlying characteristics, best described as traits, which do show some tendency to stability, but it must be acknowledged that this is not an absolute tendency and cannot be allowed to form the only prediction of the future. At the same time it should not be ignored, as although personality assessment is still defective, it is still a strong predictor of outcome when present with other mental disorders (Newton-Howes et al, 2006).

The second conclusion is that a revision of the current classification of personality disorder is overdue. Any changes must take account of the abundant evidence that normal and abnormal personalities merge into each other and it is not appropriate to have one classification for normal variation and another for pathological variation. It is suggested here that four dimensions cover the range of normal and abnormal pathology and that this is the best separation available.

In future, for better assessment we need to have improved global assessments of personality status that can be applied across all age-groups. At present, many investigators, particularly in assessments of children and adolescents, are compelled to pick one aspect of personality functioning at the neglect of others and this may lead to different results between investigators. Thus the study by Viding et al (2007, this issue) describing the significance of calrous-unemotional traits in the onset of conduct disorder, would be helped greatly by having a much greater breadth of personality assessed, not least because the presence of some more adaptive traits may alter the progression of the maladaptive ones. Similarly, the follow-up of the Aberdeen Children’s cohort has had to rely on the Rutter Scale (Rutter, 1967) for recording personality pathology in the flamboyant cluster (Wiles et al, 2005), something that was unlikely to have been anticipated by its originator. In other childhood studies, such as those in which internalising and externalising features are examined (Fergusson et al, 2006), grouping these features by personality status might help to explain much of subsequent pathology (Mervielde et al, 2005; Westen et al, 2005). At the very least this hypothesis should be tested.

With greater awareness of the variability of personality function over time it is also necessary to take more notice of written and other independent evidence about personality status at successive points in time. At present, reliability remains hamstrung by the deficiencies of the current classification, so all attempts to meld and merge diagnoses are bound to fail to some extent because the building blocks are faulty. However, the results with the PAS-DOC suggest that personality pathology in the flamboyant and antisocial group can, as with the PCL-R, be rated both reliably and accurately, but this is more difficult for those aspects of pathology that do not ‘hit the headlines’ as it were, and are confined to more private settings where documentation is poor.

These problems need to be resolved. The work described here suggests that they are being addressed, and this is essential if clinicians are to feel confident about diagnosing clinical problems comprehensively, planning care and predicting outcome for the disorders they commonly treat.

**ACKNOWLEDGEMENTS**

We thank Sandra O’Sullivan for help in organising the selection of notes and papers, Drs Andrew McDonald & Katarina Miloseka for coordinating the reliability studies, and the patients and members of the Paddington Outreach Rehabilitation Team (PORT) and Community Outreach Rehabilitation Team (COR), Central North West London Trust for their good-natured cooperation. This work was partly funded by a grant from the National Programme on Forensic Mental Health.

**REFERENCES**

American Psychiatric Association (1994) Diagnostic and Statistical Manual of Mental Disorders (4th edn) (DSM IV), APA.

American Psychiatric Association (1980) Diagnostic and Statistical Manual of Mental Disorders (3rd edn) (DSM III), APA.


Precision of actuarial risk assessment instruments

Evaluating the ‘margins of error’ of group v. individual predictions of violence

STEPHEN D. HART, CHRISTINE MICHEL and DAVID J. COOKE

Background Actuarial risk assessment instruments (ARAs) estimate the probability that individuals will engage in future violence.

Aims To evaluate the ‘margins of error’ at the group and individual level for risk estimates made using ARAs.

Method An established statistical method was used to construct 95% CI for group and individual risk estimates made using two popular ARAs.

Results The 95% CI were large for risk estimates at the group level; at the individual level, they were so high as to render risk estimates virtually meaningless.

Conclusions The ARAs cannot be used to estimate an individual’s risk for future violence with any reasonable degree of certainty and should be used with great caution or not at all. In theory, reasonably precise group estimates could be made using ARAs if developers used very large construction samples and if the tests included few score categories with extreme risk estimates.

Declaration of interest None. Funding detailed in Acknowledgements.

Many years ago the physicist, Niels Bohr, observed dryly, ‘Predicting is very difficult, especially about the future.’ What is true in the field of physics appears to be true in the field of forensic mental health. Predicting whether or not individual people will engage in violence is one of the most practically and ethically troublesome of all clinical responsibilities (Grisso & Applebaum, 1992; Szmukler, 2001). Research indicates that predictions of violence made using unaided (i.e. informal, impressionistic or intuitive) judgement are seriously limited with respect to both inter-clinician agreement and accuracy. This has motivated the development of a number of psychological tests commonly referred to as actuarial risk assessment instruments (ARAs).

The ARAs conceptualise violence risk solely in terms of probability of future violence, ignoring other facets of risk, such as the possible nature, severity, imminence, duration or frequency of future violence (Hart, 2001, 2003). They use fixed and explicit algorithms, developed on the basis of data from known groups of recidivist and non-recidivist violent offenders and patients, to estimate the specific probability or absolute likelihood that a person will engage in violence in the future. The ARAs increasingly are being used to determine whether a person should be incapacitated to prevent future violence. For example, in England and Wales ARAs may play a central role in evaluations by psychiatrists and psychologists to determine whether a person should be committed indefinetly as a dangerous person with severe personality disorder, as well as whether these people, once committed, are now ready for release into the community (Maden & Tyrer, 2003; Tyrer, 2004). In the United States, they are used in sex offender civil commitment and even capital sentencing evaluations (Janus, 2000; Hart, 2003).

The ARAs differ from most psychological tests. Rather than being descriptive or diagnostic in nature, they are predictive or prognostic, designed solely to forecast the future. Findings of ARAI tests typically are interpreted using inductive logic, which can be expressed in the form of a syllogism, as follows.

Major premise In the samples used to construct Test X, 52% of people with scores in Category Y were known to have committed violence during the follow-up period.

Minor premise Jones has a score on Test X that falls in Category Y.

Conclusion Therefore, the risk that Jones will commit future violence is similar to the risk of people in Category Y.

Findings of ARAI tests could also be interpreted using deductive logic, but few people appear to make the strong or rigid assumptions required for such an interpretation – namely, that all people belong to one of several naturally occurring discrete classes or categories, each class having a different probability of future violence, and that ARAs determine the class to which a person belongs.

Given the high stakes of violence risk assessment, including evaluations of severe and dangerous personality disorders, forensic mental health professionals have an ethical responsibility to familiarise themselves with the limitations of ARAs (Heilbrun, 1992). Perhaps the most critical limitation is the ‘margin of error’ in risk estimates made using test scores. Staying with the example above, the findings of Test X for Jones indicate that he falls in a category for which the estimated risk of violence was 52%. This sounds ominous. But how precise or credible is this prediction? How much faith or confidence should we have in the test findings?

There are two major types of error relevant in the case of violence predictions made using ARAs. The first is group error. The construction samples for Test X were just that – samples drawn from a larger population. The findings from the samples are used to draw inferences about the population parameter (i.e. the true rate of violence for the entire population of people who have scores in Category Y). We need to know the margin of error – typically expressed as a 95% CI – for the estimated violence risk associated with Category Y in the original construction samples.
The second type of error is individual error. Moving the focus of analysis from groups to individuals changes the way in which risk is conceptualised. According to ARAIs, violence risk is defined as the probability of violence. When considering groups, probability is defined in frequentist terms as the proportion of people who will commit violence (i.e. the relative frequency of events in a reference class; see Hajek & Hall, 2002), and the margin of error is uncertainty regarding the proportion of people who will commit violence. However, these definitions do not make sense for individuals, who either will or will not commit violence. (For a discussion of this ‘problem of the single case’ see Hajek & Hall, 2002.) When considering individuals the margin of error is uncertainty regarding whether a given person will commit violence. According to this view, the margin of error or uncertainty for an individual prediction is not the same as – and indeed, logically, must be considerably greater than – that for groups. Suppose a public opinion survey of 500 eligible voters found that 54% expressed their intent to cast ballots for candidate Smith in an upcoming election. This information allows one to forecast with reasonable confidence that candidate Smith will be elected by another group – namely, the general electorate. However, this same information does not allow one to predict the behaviour of a randomly selected voter with great confidence. Even though, in the absence of other relevant information, the most rational prediction is that every single voter will cast a ballot for candidate Smith, these individual predictions frequently will be wrong. So, to return to the ARAI example above, we need to know the margin of error for predictions made using Test X that a given person, such as Jones, will commit violence.

It is simply impossible to make rational, reasonable and legally defensible decisions based on the results of tests or statistical models without understanding the errors inherent in those results for both groups and individuals (with respect to forensic mental health, see Heilbrun, 1992; with respect to medicine more generally, see Henderson & Keiding, 2005). However, surprisingly, these issues are rarely discussed in journal articles about ARAs or in ARAI test manuals (but for noteworthy exceptions, see Monahan et al, 2005, Mossman, 2006). In this paper, we re-analyse data from the development samples of the most commonly used ARAIs to calculate the margins of error for group and individual estimates of violence risk.

**METHOD**

**Measures**
We estimated the precision of violence predictions for two ARAIs, both constructed using a criterion groups design in which multivariate statistics were used to select and weight test items to maximise the discrimination between known groups of recidivists and non-recidivists. The tests were selected because they are frequently used, researched and discussed in Europe and North America.

**Violence Risk Appraisal Guide**
The Violence Risk Appraisal Guide (VRAG; Quitsey et al, 1998) is a 12-item test designed to assess risk for general violence over periods of 7–10 years. It was developed in a sample of patients released from a maximum-security forensic psychiatric hospital in Ontario, Canada. We evaluated the precision of violence predictions for violent recidivism over a 10-year follow-up period, following Quitsey et al (1998: Table A-1). The number of people and the corresponding proportion of recidivists for each of the nine score categories are presented in Table 1.

**Static-99**
The Static-99 (Hanson & Thornton, 1999) is a 10-item test designed to assess risk for violence and sexual violence over periods of 5–15 years. It was developed from re-analyses of data from four diverse samples of offenders and forensic psychiatric patients released from institutions in Canada and the UK. We evaluated the precision of risk estimates for sexually violent recidivism over a 15-year follow-up period, following Hanson & Thornton (1999: Table 5). The number of people and the corresponding proportion of recidivists for each of the nine score categories are presented in Table 2.

**Statistical analyses**
If one assumes that for a given ARAI score category group estimates of violence risk are binomial proportions, then it is possible to calculate the 95% CI using a method first outlined by Wilson (1927). This method is relatively simple, carries a relatively low assumption burden and can be used without access to raw data. A recent review by Agresti & Coull (1998) (see also Brown et al, 2001) indicates that it is superior to some alternatives, such as the exact and Wald methods, because it not strongly influenced by extreme values with respect to sample size or the proportion of recidivists, and because it does not yield impossible values (e.g., negative lower limits).

According to Wilson’s method, the upper limit (UL) and lower limit (LL) of the confidence interval are:

\[
UL = \frac{\hat{\theta} + \frac{z_{1-\alpha/2}}{n} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}}{1 + \frac{z_{1-\alpha/2}^2}{n}}
\]

and

\[
LL = \frac{\hat{\theta} - \frac{z_{1-\alpha/2}}{n} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}}{1 + \frac{z_{1-\alpha/2}^2}{n}}
\]

where \(n\) is the number of people in a given ARAI score category, \(\hat{\theta}\) is the proportion of recidivists in the score category and, for the purpose of constructing a 95% CI, \(z_{1-\alpha/2} = 1.96\).

We applied Wilson’s method to the VRAG and Static-99. Based on published reports describing the construction of the tests, for each score category of the VRAG and Static-99 we calculated the precision of group estimates of violence risk with \(n\) equal to the number of people in the category and \(\hat{\theta}\) equal to the proportion of recidivists in the category. This is the standard and accepted application of Wilson’s method. For group estimates of violence risk, the 95% CI is interpretable as follows: ‘Given a group of \(n\) people with ARAI scores in this particular category, we can state with 95% certainty that the proportion of recidivists will fall between the upper limit and lower limit.’

There are various ways to calculate the precision of individual estimates of violence risk. Perhaps the best methods come from logistic regression and event history analysis. With these methods, it is possible to model at the group level the occurrence of violence over a fixed time period (logistic regression analysis) or as a function of time (event history analysis), then to derive individual regression or survival scores and their respective margins of error. Unfortunately, the VRAG and Static-99 were not constructed using logistic regression or event history analysis, so it is impossible to evaluate the tests using these methods. Indeed, it appears to be impossible to...
calculate directly the precision of individual estimates of violence risk for any of the existing ARAs using any standard statistical method, and so the only alternative is to use ad hoc procedures. The ad hoc procedure we selected was to apply Wilson’s method to each score category of the VRAG and Static-99 with \( n \) equal to 1 and \( \theta \) equal to the proportion of recidivists in the score category. For individual estimates of violence risk, we interpret the 95% CI as follows: ‘Given an individual with an ARAI score in this particular category, we can state with 95% certainty that the probability he will recidivate lies between the upper and lower limit.’ We piloted this application of Wilson’s method in several prediction data-sets of our own and it yielded findings very similar to those obtained using logistic regression or event history analysis.

To illustrate our use of Wilson’s method for determining group and individual margins of error, let us take an example. Suppose that Dealer, from an ordinary deck of cards, deals one to Player. If the card is a diamond, Player wins; but if the card is one of the other three suits, Player loses. After each deal, Dealer replaces the card and shuffles the deck. If Dealer and Player play 10 000 times, Player should be expected to win 75% of the time. Because the sample is so large, the margin of error for this group estimate is very small, with a 95% CI of 74–76% according to Wilson’s method. Put simply, Player can be 95% certain that he will win between 74 and 76% of the time. However, as the number of plays decreases, the margin of error gets larger. If Dealer and Player play 1 000 times, Player will only expect to win 75% of the time, but the 95% CI increases to 72–78%; if they play only 100 times, the 95% CI increases to 66–82%. Finally, suppose we want to estimate the individual margin of error. For a single deal, the estimated probability of a win is still 75% but the 95% CI is 12–99%. The simplest interpretation of this result is that Player cannot be highly confident that he will win – or lose – on a given deal.

**RESULTS**

**Precision of group estimates**

The 95% CI for group estimates for the score categories of the VRAG and Static-99 are shown in Tables 1 and 2 and Figs 1a and 2b. Looking first at the VRAG, the 95% CIs for score categories ranged from 13 to 30 percentage points in width, with a mean of about 20 percentage points. For the Static-99, the 95% CIs for score categories ranged from 8 to 19 percentage points in width, with a mean of about 13 percentage points. The somewhat smaller 95% CI for the Static-99 highlights the benefit of large sample sizes: increasing the number of people in a score category yields more precise group estimates.

Overlap among 95% CIs indicates that the group estimates for score categories did not differ significantly. Looking at the VRAG, the 95% CIs overlapped considerably and adjacent score categories almost always overlapped. This is most apparent in Fig. 1a. Categories 1–4 had overlapping 95% CIs. The 95% CIs for categories 5–7 overlapped with each other, but not with those of categories 1–4. The 95% CI for category 8 did not overlap with those of categories 1–6, but did overlap with that of category 7. The 95% CI for category 9 did not overlap with those of categories 1–7, but did overlap with that of category 8. These findings suggest that the VRAG score categories yield three reasonably distinct group estimates of risk: low (categories 1–4), moderate (categories 5–7) and high (categories 8–9).

Looking next at the Static-99, and in particular Fig. 2a, categories 0, 1, 2 and 3 had overlapping 95% CIs; categories 4, 5 and 6+ had 95% CIs that overlapped with each other but not with those of categories 0–3. Thus, the Static-99 yielded only two distinct group estimates of risk: low (categories 0–3) and high (categories 4–6+).

The greater number of distinct group estimates of risk on the VRAG highlights the importance of identifying extremely high risk or low risk groups: Even if score categories contain many people, unless the proportions of recidivists in various score categories differ substantially, their confidence intervals will overlap.

---

**Table 1 Estimates of risk for groups and individuals with the Violence Risk Appraisal Guide**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of people</th>
<th>Proportion of recidivists</th>
<th>95% CI Group</th>
<th>95% CI Individual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>0.00</td>
<td>0.00–0.26</td>
<td>0.00–0.79</td>
</tr>
<tr>
<td>2</td>
<td>71</td>
<td>0.08</td>
<td>0.04–0.17</td>
<td>0.00–0.82</td>
</tr>
<tr>
<td>3</td>
<td>101</td>
<td>0.12</td>
<td>0.07–0.20</td>
<td>0.00–0.84</td>
</tr>
<tr>
<td>4</td>
<td>111</td>
<td>0.17</td>
<td>0.11–0.35</td>
<td>0.01–0.86</td>
</tr>
<tr>
<td>5</td>
<td>116</td>
<td>0.35</td>
<td>0.27–0.44</td>
<td>0.03–0.91</td>
</tr>
<tr>
<td>6</td>
<td>96</td>
<td>0.44</td>
<td>0.34–0.54</td>
<td>0.04–0.93</td>
</tr>
<tr>
<td>7</td>
<td>74</td>
<td>0.55</td>
<td>0.44–0.66</td>
<td>0.07–0.96</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
<td>0.76</td>
<td>0.58–0.88</td>
<td>0.12–0.99</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1.00</td>
<td>0.70–1.00</td>
<td>0.21–1.00</td>
</tr>
</tbody>
</table>

**Table 2 Estimates of risk for groups and individuals with the Static-99**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of people</th>
<th>Proportion of recidivists</th>
<th>95% CI Group</th>
<th>95% CI Individual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>107</td>
<td>0.13</td>
<td>0.08–0.21</td>
<td>0.00–0.84</td>
</tr>
<tr>
<td>1</td>
<td>150</td>
<td>0.07</td>
<td>0.04–0.12</td>
<td>0.00–0.82</td>
</tr>
<tr>
<td>2</td>
<td>204</td>
<td>0.16</td>
<td>0.12–0.22</td>
<td>0.01–0.85</td>
</tr>
<tr>
<td>3</td>
<td>206</td>
<td>0.19</td>
<td>0.14–0.25</td>
<td>0.01–0.86</td>
</tr>
<tr>
<td>4</td>
<td>190</td>
<td>0.36</td>
<td>0.30–0.43</td>
<td>0.03–0.91</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>0.40</td>
<td>0.31–0.50</td>
<td>0.04–0.92</td>
</tr>
<tr>
<td>6+</td>
<td>129</td>
<td>0.52</td>
<td>0.43–0.60</td>
<td>0.06–0.95</td>
</tr>
</tbody>
</table>
Our findings likely come as no surprise to many people. The difficulties of predicting the outcomes for groups versus individuals – whether in the context of games of chance or of violence risk assessments – are intuitively obvious. Take, for example, the following quotation from Sir Arthur Conan Doyle’s novel, The Sign of the Four:

‘[W]hile the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant.’

Limitations

The method we used to estimate margins of error did not translate into distinct risk estimates made using the tests were sub-what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant.

With respect to estimating the precision of group predictions, Wilson’s method assumes that people with scores in the same ARAI score category are homogeneous. However, ARAIs of 10 or 12 items almost certainly exclude potentially important information about risk, such as information about dynamic factors (e.g., Hart, 1998, 2001) – and this is acknowledged by most authors (see Quinsey et al, 1998; Hanson & Thornton, 1999; Monahan et al, 2005). Also, Wilson’s method assumes that people are classified into ARAI score categories with perfect reliability. However, what little information is available in the published literature concerning the inter-clinician agreement for ARAI scores suggests that they are not perfect. If either of these assumptions is violated, then Wilson’s method is overly conservative, and the tests’ margins of error for groups are either larger than reported here or may even be inacalculable.

With respect to estimating the precision of individual predictions, we were forced to use Wilson’s method in an ad hoc manner. We recognise that some readers may object to this approach but our pilot testing suggested that Wilson’s method yields findings very similar to those obtained using more sophisticated methods for estimating the error of individual predictions based on raw data, such as logistic regression or event history analysis, which also suggest that individual prediction errors are extremely large (e.g., Henderson & Keiding, 2005). The only apparent alternatives to this ad hoc approach are: (a) to acknowledge that it is impossible to estimate the margin of error for individual predictions made using existing ARAIs and (b) to construct and evaluate new ARAIs using procedures that permit the direct estimation of the margin of error for individual predictions.

Also with respect to estimating the precision of individual predictions, some readers may object to our application of Wilson’s method because they interpret individual risk estimates as a person’s propensity for future violence, not as a prediction of future violence. The problem is that this sort of ‘propensity’ bears no direct conceptual or statistical relation to an individual’s actual behaviour (which, of course, has not yet occurred), making the entire concept a sort of metaphysical abstraction that is divorced from empirical reality (for clear and concise critiques of propensity approaches to probability, see Hajek & Hall, 2002 and Hajek, 2003). Thus anyone who relies on a propensity view of probability must also accept that it is impossible to use propensities to make specific predictions about the future violent behaviour of individuals with any reasonable degree of certainty.

Implications for forensic mental health evaluations

The potential implications of these findings for the practice of forensic mental health are profound. At best, they suggest that professionals should be extremely cautious when using ARAIs to estimate or draw
inferences about an individual’s risk for violence. This means, as Henderson & Keiding (2005) have recommended, ‘avoiding use of a single quantity to characterise a probability distribution, whether a point or categorical prediction, prognostic index, relative risk, or probability of surviving a given time’ (p. 705). At worst, they suggest that professionals should avoid using ARAs altogether, as the predictive accuracy of these tests may be too low to support their use when making high-stakes decisions about individuals. Low predictive accuracy not only makes reliance on ARAs ethically problematic, it also means that they may not meet legal standards for the admissibility of expert or scientific evidence. (For outlines of such criteria in the UK, see Mackay et al, 1998 and Zeedyk & Raitt, 1998; for a discussion of criteria in the USA, see Faigman, 1995 and Melton et al, 1997.) Admissibility is also a problem if one concludes that margin of error for individual predictions is inescapable.

Another counter-argument presented to us is that ARAs can be used appropriately as long as professional judgement or discretion is used to modify or override test-based decisions in the presence of relevant rare, case-specific or dynamic risk factors. According to Meehl (1998), ‘This sounds amicable, tolerant, and even-handed, but it’s actually stupid.’ The problem here is that it does not make sense to ‘fudge’ the results of a statistically derived estimate on the basis of personal preference; in addition, there is simply no empirical evidence that this improves the accuracy of predictions.

Finally, some professionals argue that it is appropriate to use ARAs to make relative risk estimates concerning individuals (e.g. ‘Jones has a higher risk for violence than does Smith’). However, our findings indicate that the margin of error in group findings is substantial, leading to overlap among ARA score categories. This means that it is perhaps difficult to state with a high degree of certainty that one individual’s risk for future violence is higher than that of other individuals.

Test users should be very careful when using ARAs to make sure that consumers of test findings (other mental health professionals, patients, courts, etc.) understand that it is, at least at present, impossible to make accurate predictions about individuals using these tests; this may help to minimise their potentially prejudicial impact on decision-making. Also, it may be wise to limit or avoid the use of ARAs in situations where the cost of potential decision errors is high. An appropriate use of ARAs may be for making administrative decisions regarding the frequency or intensity of risk management strategies recommended for a given individual (e.g. number of office visits, priority for admission into treatment groups). In such low-stakes circumstances, it may be reasonable to overlook numerous prediction errors at the individual level and focus on aggregate benefits at the group level.

**Implications for the development and evaluation of ARAs**

Our findings also have implications for the development of ARAs. First, they highlight the importance of large sample sizes. It is necessary to include many people in each ARA score category, so that group estimates are reasonably precise. Typically, group sizes of 500 are used in social science research (e.g. public opinion surveys); in biomedical research on mortality rates or in the insurance industry, group sizes are in the range of several thousand to tens or even hundreds of thousands. Second, our findings highlight the importance of identifying ARA score categories with extreme estimates of violence risk. An example of ‘extreme’ estimates would be ≤10% v. 50% v. ≥90%. Extreme group estimates may have non-overlapping 95% CIs. Only when both these conditions hold true can ARAs yield potentially useful individual-level risk estimates. (Alternatively, test developers may wish to avoid altogether the concept of ‘groups’ and use statistical procedures that focus on individual predictions, such as logistic regression and event history methods. Of course, large sample sizes are no less important if this is the case.)

Our findings also suggest that people who develop and evaluate ARAs should consider the potential benefits of conceptualising violence risk from a subjectivist perspective, focusing how evaluators do or should form beliefs about an individual’s risk for future violence, especially in the light of uncertain information and decision errors with varying costs (e.g. Hâjkéj, 2003). Although changing the discourse from frequentist to subjectivist will not make predictions of the future any more accurate, it may provide ways of researching and communicating about the problem that are more intuitively understandable to mental health professionals and legal decision-makers alike (for an example, see Mossman, 2006).

We conclude by advising readers that we have addressed only the rather limited issue of the margins of error of group and individual-level risk estimates using ARAs. We did not address other critical issues in construction and forensic use of ARAs (e.g. Hart, 2001, 2003; Litwack, 2001): the questionable representativeness of their construction samples; the absence of calibration or cross-validation research on risk estimates, especially by independent researchers; problems with their legal relevance, owing to a failure to consider the presence of mental disorder and the presence of a causal nexus between mental disorder and violence risk; and their potential prejudicial impact on triers of fact.

**ACKNOWLEDGEMENTS**

We thank John Monahan, Douglas Mossman and David Thornton for their patient, lengthy and helpful critiques of earlier versions of this manuscript. DC received support from the Research and Development Directorate of the Greater Glasgow Primary Care National Health Service Trust while carrying out these analyses.

**REFERENCES**


**ACCURACY OF VIOLENCE PREDICTIONS**

STEPHEN D. HART, PhD, Department of Psychology, Simon Fraser University, Burnaby, British Columbia, Canada and Faculty of Psychology, University of Bergen, Norway; CHRISTINE MICHE, BSc, Department of Psychology, Glasgow Caledonian University; DAVID J. COOKE, PhD, FBPsS, FRSE, Department of Psychology, Glasgow Caledonian University and Forensic Clinical Psychology Services, Greater Glasgow Health Board, Mental Health and Community Trust, Glasgow, Scotland.

Correspondence: Professor Stephen D. Hart, Department of Psychology, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, Canada V5A 1S6. Email: hart@sfu.ca
Assessment and treatment of violence-prone forensic clients: an integrated approach

STEPHEN C. P. WONG, AUDREY GORDON and DEQIANG GU

Background A risk-reduction treatment programme complemented by a focused assessment, both guided by the risk–need–responsivity principles, is suggested as the preferred treatment for violence-prone individuals with personality disorder.

Aims Violence Reduction Programme (VRP) and Violence Risk Scale (VRS) were used to illustrate the design and implementation of such an approach. Participants from a similarly designed Aggressive Behaviour Control Programme were used to illustrate the principles discussed and to test programme efficacy.

Method The VRS was used to assess risk/need and treatment readiness, and DSM–III–IV psychiatric diagnoses of 203 federal offenders.

Results Participants had a high probability of violent recidivism and many violence-linked criminogenic needs, similar to offenders with high PCL–R scores. Most had antisocial personality disorder and substance use disorders; in terms of treatment-readiness, most were in the contemplation stage of change. Outcome evaluation results support the objectives of the VRP.

Conclusions Integrating risk–need–responsivity principles in assessment and treatment can provide useful guidelines for intervention with violence-prone forensic clients with personality disorder.

Declaration of interest S.C.P.W. and A.G. are proprietors of VRS and VRP.

For someone with an entrenched pattern of violent criminal behaviour that is not caused by a major mental illness, intervention to reduce violence risk is one of the few options available for rehabilitation. Assessment and treatment of violence should be based on theoretically sound and empirically validated principles, and should be integrated in their implementation to increase the likelihood of successful outcome. However, common practice often lags behind theory and treatment efficacy suffers. We use the Violence Reduction Programme (VRP) and the Violence Risk Scale (VRS) to illustrate how a theoretically derived and empirically driven treatment programme and assessment process can be integrated in practice. We then describe the participants of a similar programme, Aggressive Behaviour Control Programme, currently being offered to illustrate the principles discussed and to test the efficacy of the programme through outcome evaluations.

Effective correctional treatment

Risk–need–responsivity principles have been identified as useful guidelines for treatment interventions designed to reduce the risk of recidivism. Treatment approaches, often referred to as correctional treatment, that follow the risk–need–responsivity principles are generally more effective in reducing the risk of recidivism in adult and young offenders than those that do not follow such principles (see Andrews et al, 1990; Andrews & Bonta, 2003).

Risk–need–responsivity principles and treatment change

The risk principle states that the intensity of treatment should match the clients’ risk level; clients with ‘high’, ‘medium’ and ‘low’ levels of risk should receive the corresponding intensities of treatment.

The need principle states that the individual’s criminogenic needs (needs that are linked to violence or criminality, such as criminal attitudes, criminal associates etc.) must be assessed, identified and targeted for treatment. Effective correctional treatment should lead to positive changes in the criminogenic needs, resulting in risk reduction. Interventions directed at areas not related to the individual’s recidivism risk.

The responsivity principle states that treatment effectiveness can be maximised if treatment delivery can accommodate the clients’ idiosyncratic characteristics, such as their cognitive and intellectual abilities, level of motivation and readiness for treatment, cultural background, and so forth. Responsivity refers to the individual’s characteristics, which, although not a direct or indirect cause of criminal behaviours, must none the less be taken into account to ensure that treatment and management strategies are effective (see Wong & Hare, 2005: p. 5). One of the most daunting responsivity factors in correctional treatment is to treat the unmotivated, non-adoherent and treatment-resistant client (i.e. dealing with the general issue of treatment readiness). Many individuals with psychopathy or personality disorder are often unmotivated and treatment resistant, at high risk to recidivate and prone to drop out of treatment prematurely (Ogloff et al., 1990). Thus, paradoxically, those who are in need of treatment the most cannot receive the treatment they need. Assessment of treatment readiness, to ensure that treatment delivery matches the clients’ treatment readiness, is therefore essential to reduce treatment drop out, thereby increasing treatment efficacy.

Within this conceptual framework, personality disorder is considered primarily as a responsivity factor. For example, individuals suffering from psychopathy are more likely to be manipulative, lacking in remorse and guilt, self-centred/narcissistic and so forth (Factor 1 characteristics of the Psychopathy Checklist–Revised (PCL–R); Hare, 2003). These interpersonally exploitative and affectively shallow traits are personality traits, and therefore are resistant to change. The behavioural manifestations of these traits and other personality disorder characteristics can significantly interfere with treatment as they impede the formation of a good working alliance with the treatment provider and, therefore, must be appropriately managed for effective correctional treatment and risk reduction to proceed (see Wong & Hare, 2005).
**Treatment readiness**

The Transtheoretical Model of Change or the Stage of Change Model (Prochaska et al, 1992) addresses the issue of treatment readiness, treatment change and the need to match treatment delivery to client readiness. The model postulates that individuals who modify their problem behaviours progress through a series of five stages: the pre-contemplation, contemplation, preparation, action and maintenance stages characterised by specific behaviours.

Those in the pre-contemplation stage have neither insight nor intention to change in the foreseeable future. They are often in denial and externalise blame. Those in the contemplation stage are fence-sitters; they acknowledge their problems but have shown no relevant behavioural change: ‘all talk, no walk’. Those in the preparation stage combine intentions to change with relevant behavioural changes to address problems. However, changes tend to be recent and/or quite unstable. Those in the action stage actively modify their behaviours, attitudes and environment to address their problems; overt behavioural changes are made, commitments followed through and energies expended to change. In the maintenance stage, relapse prevention techniques are used to consolidate, strengthen and generalise the gains made in the action stage.

In progressing through the stages, positive changes become more stable, internalised and sustainable. However, treatment interventions effective for one stage may not be effective or may even be counter-productive for some, at other stages. Lapses or cycling through the stages is considered to be a rule rather than an exception. For example, those in the pre-contemplative stage should be provided with treatment engagement activities such as motivational interviewing (Miller & Rollnick, 1991). Action stage activities such as skill training (e.g. assertiveness training), although appropriate in general for those in the preparation and action stages, are inappropriate for those in the pre-contemplation stage. Prematurely putting unmotivated clients in action-oriented interventions may lead to increased resistance and treatment drop out. Assessment of the client’s treatment readiness, therefore, is critically important in treating resistant clients such as those with psychopathy or personality disorder.

Treatment is a process of change. The primary goal of correctional treatment is to bring about positive changes in crimino-genic needs leading to risk reduction. Treatment changes must be assessed objectively and systematically to determine the amount of risk reduced. Assessment and treatment must be closely integrated: assessments of the clients’ risk, need and responsivity should inform treatment providers of who to treat (risk principle), what to treat (need principle) and how to deliver treatment, in particular to treatment-resistant clients (responsivity principle). Clinicians who provide correctional treatment require the appropriate tools to assess risk, needs, responsivity and treatment readiness, and to measure treatment change.

---

**Assessment to inform treatment**

**Assessing risk—need—responsivity and treatment change**

Many forensic assessment tools are designed primarily for predicting recidivism not complementing treatment. For example, the Violence Risk Appraisal Guide (VRAG; Quinsey et al, 1998) and the Static-99 (Hanson & Thornton, 1999) are designed to predict non-sexual and sexual recidivism respectively. Since these tools use mainly static (unchangeable) predictors, such as criminal history and early behavioural problems, they can predict risk but they cannot assess criminogenic need or responsivity, nor can they measure change in risk. Douglas & Skeem (2005) suggest that development of risk assessment tools with dynamic variables is the next challenge in the field of forensic assessment.

Some assessment tools, such as the Level of Service Inventory (LSI–R; Andrews & Bonta, 1995) are designed to assess risk and needs by incorporating changeable or criminogenic need (dynamic risk) variables together with static variables. The LSI–R uses ten domains to assess risk and need; these include criminal history, education and employment, financial resources, etc. This tool provides useful information on the client’s risk and need but it does not assess the key responsivity issue of treatment readiness. It is also unclear how to link the amount of change observed in treatment with changes in the dynamic need variables (i.e. what behaviours observed in treatment should one use to indicate changes in these domains). For example, within the LSI–R, the financial and employment domain can be measured if the offender was recently employed in the community. However, it is difficult to assess changes in the domain if the individual has been incarcerated for a long time.

**The Violence Risk Scale**

The VRS (Wong & Gordon, 2006) is designed to integrate the assessment of risk, need, responsivity and treatment change into a single tool. It assesses the clients’ level of violence risk, identifies treatment targets linked to violence, assesses the clients’ readiness for change and their post-treatment improvements on the treatment targets. Treatment improvement or lack thereof is linked to quantitative changes in violence risk.

The VRS uses 6 static and 20 dynamic variables derived primarily from and underpinned by the theory of the psychology of criminal conduct and the risk, need and responsivity principles (Andrews & Bonta, 2003; see Fig. 1 for VRS dynamic variables). The linkage between the VRS and the principles of effective correctional treatment is by design such that assessment and treatment are closely integrated theoretically. The VRS static and dynamic variables are rated on a four-point scale (0, 1, 2 or 3), based on a careful review of file information and a semi-structured interview. The VRS static variables can predict general and violent recidivism, but remain unchanged with treatment. Higher ratings on the static variables indicate worse ‘track records’ of dysfunctional and antisocial behaviour. The dynamic variables, such as interpersonal aggression and criminal attitudes, are changeable risk predictors; they can be used as treatment targets and can measure changes in risk. Higher ratings (2 or 3) on dynamic variables indicate that the variables in question are closely linked to violence and are appropriate targets for treatment (need principle). The sum of the ratings of the static and dynamic variables reflects the client’s level of violence risk; the higher the score, the higher the risk. In selecting clients for treatment, those with higher VRS scores should be appropriate candidates for higher intensity intervention (risk principle). The VRS can also be used as a stand alone measure to assess a client’s current risk of violence.

For individuals identified for treatment, the VRS also uses a scheme based on a modified Transtheoretical Model of Change (Prochaska et al, 1992). Each dynamic variable identified as a treatment target (ratings of 2 or 3) is also assessed to...
determine the client’s stage of change (readiness for treatment). The operationalisations of the various stages of change (except pre-contemplation) are designed to measure the extent to which any newly acquired positive attitudes and coping skills are stable, sustainable and generalisable. Progression in treatment from a less advanced to a more advanced stage of change for each treatment target is an indication of improvement, which should lead to risk reduction in that treatment target. The VRS translates the progress from one stage to the next stage into a quantitative risk reduction of 0.5. The only exception is progression from pre-contemplation to contemplation stage, which carries no risk reduction since those in the contemplation stage only ‘talk the talk’ but have not yet ‘walked the talk’ (i.e. they have shown no relevant behavioural change). Positive changes during the treatment programme are reflected as risk reduction measured by the VRS, in other words, integrating treatment change with risk reduction. The pre-treatment risk level (pre-treatment VRS scores) minus the total risk reduction score is a measure of the client’s overall post-treatment risk level. Rating of the VRS variables, the stages of change and the computation of risk scores are provided in detail in the VRS manual (see Wong & Gordon, 1999–2003).

In addition, according to the Trans-theoretical Model of Change, the client’s prototypical behaviours at each stage of change should be matched with appropriate intervention: the responsivity principle. As such, assessment of the client’s stage of change also identifies the most appropriate therapeutic approach to take. A brief summary of therapist tasks that correspond to each stage of change follows.

**Pre-contemplation.** The therapist should: focus on developing a working alliance, enhancing motivation for change and engagement in treatment; raise doubts and create dissonance regarding the client’s current functioning and his hopes of achieving future goals; use cost–benefit analyses to highlight the cost of criminal behaviour.

**Contemplation.** The therapist should: tip decisional balance; evoke reasons to change in order to reduce dissonance; strengthen the client’s confidence to effect change (i.e. increase self-efficacy).

**Preparation.** The therapist should assist the client in: determining the best course of action to change; setting and achieving shorter-term behavioural goals that are planned, observable, measurable and relevant; highlighting successes and emphasising change potential.

**Action.** This is the main skill-teaching and skill-building phase of treatment. The therapist should assist the client in strengthening skills through overpractice and reinforce client’s self-efficacy in problem-solving and achieving treatment goals.

**Maintenance.** The therapist should: assist and encourage the client to practice and generalise learned skills to new and challenging situations by providing access to such situations; identify strategies and interventions to prevent lapses and relapses. Obviously, strengthening and reinforcing the client’s self-efficacy is important whenever the client takes steps to make changes, regardless of the stage of change.

### Integration of assessment and treatment of violence-prone offenders

We will describe the VRP (Gordon & Wong, 2000; Wong, 2000a,b), a risk reduction focused correctional treatment programme for violence prone forensic clients, to illustrate further the integration of treatment and assessment approaches. The design of the VRP is also based on the theory of criminal conduct, the risk–need–responsivity principles and a modified Transtheoretical Model of Change. An integral part of the VRP is the VRS. Treatment services are delivered using a three-phase model described below.

The objectives of the VRP are to reduce the frequency and intensity of violence by first challenging antisocial beliefs, attitudes, schemas and behaviours that support the use of violence and second, assisting programme participants to acquire appropriate skills that can reduce the risk of violence, as well as developing self-efficacy and confidence in using the skills. The VRP is designed to address the treatment needs of high-risk violence prone clients, in particular those who are non-adherent, unmotivated and resistant to treatment. The programme, although structured and goal-oriented, is flexible enough to accommodate the heterogeneity of criminogenic needs and responsivity often found in this client group.

The programme uses cognitive–behavioural therapeutic approaches and social learning principles within a relapse prevention framework to assist participants to make changes and learn new behaviours. It is recognised that learning takes place incrementally (i.e. in small steps) and reinforcement of small incremental improvements is the key. The delivery of the VRP is structured within a three-phase model of treatment delivery (Gordon & Wong, 2000; see Fig. 2). In each of the three phases, participants and those delivering treatment have different tasks and objectives. Phase 1 focuses on helping the client develop
insight into past patterns of violence, on identifying treatment targets and on developing therapeutic or working alliance. Motivational interviewing techniques, which should be used throughout the programme, are particularly important in phase 1 and are essential to engage resistant clients in treatment. Phase 2, which is mainly oriented towards action or skill acquisition, focuses on helping participants to acquire relevant skills to restructure negative thoughts, feelings and behaviours associated with violent and destructive patterns. Phase 3 focuses on relapse prevention strategies and the generalisation of skills across situations and to the community. Phase 3 work consists mainly of consolidation, generalisation and maintenance of phase 2 gains.

The client’s level of readiness for treatment, assessed as one of the five stages of change by the VRS, can be mapped quite readily onto the three phases (see Fig. 2). The pre-contemplation, contemplation and preparation stages are located in phase 1; the preparation and action stages in phase 2 and the action and maintenance stages in phase 3. The preparation stage is located on both phase 1 and 2 and the action stage on both phase 2 and 3 to emphasise the continuity and movement of the stages through the different phases. The three-phase model integrates the treatment readiness of the client with the therapeutic approaches of the staff to form a ‘road map’ as guidance for clients and staff throughout the treatment process. Clients are taught the conceptual meaning of the ‘stages’ and ‘phases’ in order to develop a common language of treatment and change among staff and clients.

With a heterogeneous group of clients, treatment progress is not expected to be smooth or uniform; frequent lapses are the norm. Programmes that are highly scripted, with content that has to be delivered in a specific chronological order and time frame, would not have the flexibility to accommodate the varied needs of the clients. Progress in the three-phase model depends upon the achievement of specific phase objectives (see Gordon & Wong, 2000). Lapses (e.g. regression from action to contemplation stage) would signal staff to allocate additional resources and time to work with the client using phase 1 approaches to help re-engagement in treatment and the process of change. On the other hand, those who progress faster can move on without being held back. The three-phase model provides staff and clients with a road map that has both the structure and flexibility essential for the treatment of a heterogeneous and resistant group of clients. Improvements are quantified and measured using the VRS.

Implementation of the VRP

The VRP is designed so that it can be modified and adapted for use by different organisations to serve different client groups. It can also be adjusted to fit local requirements such as length of treatment, staffing complement, resource availability, security level, management approaches and so forth. A number of treatment programmes based on the conceptual framework of the VRP have been implemented in various sites in the UK. A VRP pilot programme at the Woodhill Close Supervision Centre, a super-maximum security prison, has been completed, and was evaluated by an independent evaluation team (see Fylan & Clarke, 2006). The objectives of this 7-month VRP pilot programme were to reduce the frequency and intensity of violence of very high-risk and violence-prone prisoners, all of whom had committed homicides while incarcerated. In addition, one of the expected outcomes was that by participating in the VRP, the behaviours of the prisoners would improve to the extent that they could be re-integrated into other mainstream prisons or custodial settings. The results of the evaluation should indicate the feasibility of implementing the VRP programme in a super-maximum security prison. Another programme that is conceptually similar to the VRP and has been in operation for over a decade is the Aggressive Behaviour Control (ABC) Programme at the Regional Psychiatric Centre, a secure forensic in-patient facility within the Correctional Service of Canada. Both S.W. and A.G. have been actively involved for many years in the development and modification of the ABC Programme. The design of the ABC programme is similar to that of the VRP; both utilise the three-phase treatment model, adhere to the risk–need–responsivity principles and utilise a cognitive–behavioural therapeutic approach. The design of the ABC programme has to accommodate local requirements such as programme length, staffing complements, resource allocation and management requirements. The ABC programme is about 6–8 months long and, similar to the VRP, is designed for clients that have serious histories of violence, have not had success in past treatment attempts, may belong to gangs and often have significant institutional problems such as episodes of serious violence. Criminogenic factors are addressed in offence cycle groups, psychoeducational groups and individual therapy. Services to address issues of education, work and life skills, relationships with significant others, family dynamics, community support and early abuse are provided where appropriate. Like the VRP, the ABC programme attends to client responsivity such as personality disorders, cognitive and language abilities, cultural background, treatment readiness, and so forth. At the end of the

![Fig. 2 Three-phase treatment delivery model.](image-url)
programme, each participant is required to develop a relapse prevention plan that delineates in detail interventions that can be used to mitigate risks of recidivism.

**METHOD**

**Participants**

Participants included 203 male federal offenders (serving a sentence of ≥2 years), most of whom were referred by the other federal penitentiaries in the provinces of Alberta, Saskatchewan and Manitoba for treatment in the ABC programme over a period of about 7 years (1996–2003). The sample was selected as they were all administered the VRS as a part of the assessment process and they all completed the ABC programme. They were also given a psychiatric diagnosis on or shortly after admission. The sample demographics are given in Table 1.

**Assessment of risk, need and responsivity**

About 1 month after admission, participants were rated using the VRS by staff trained by A.G. Ratings included the static and the dynamic variables and the stages of change for each dynamic variable identified as a treatment target, plus an overall stage of change rating reflecting the predominant stage of change of all the dynamic variables. Most participants tended to show a predominant stage of change for most of the problem areas, but there are exceptions.

**RESULTS**

**Risk ratings**

The mean VRS total score (static plus dynamic variables) for the sample is 55.23 (s.d. = 10.70), which is almost 1.5 s.d. above the mean (35.49; s.d. = 14.97, n = 652; t = 20.76; P < 0.00001) of a sample of randomly selected federal penitentiary inmates from the same three provinces (random sample). In a separate study (Wong & Gordon, 2006), it was found that those who scored 55–60 on the VRS had about 55% and 69% likelihood of recidivating violently and generally, respectively, after 3-year follow-up compared with 25% and 49% for those that scored 35–40, that is, the random sample. Participants in the programme are more than twice as likely to recidivate violently than the general offender sample and have very extensive criminal records (Table 1), with a mean of almost 24 convictions accumulated in an average 8-year criminal career, five of which are violent convictions. The treatment sample comprised violence-prone offenders and it is appropriate to provide them with high-intensity risk reduction treatment.

**Criminogenic needs or dynamic risk ratings**

The VRS dynamic variables rated 2 or 3 are closely linked to violence and can be considered as problem areas or treatment targets. One way to describe the prevalence of problems in the sample is to show the percentages of the sample that rated 2 or 3 on each of the 20 dynamic variables to give a ‘dynamic risk profile’ (Wong & Burt, 2007; see Fig. 1). As a comparison, the dynamic risk profiles of a group with psychopathy (mean PCL–R score = 28.2, s.d. = 2.7; mean VRS score = 58.4, s.d. = 7.7, n = 65; Wong & Burt, 2007) and the random sample are also presented in Fig. 1.

In the ABC sample, all but one of the dynamic variables had prevalence rates of 50% and above, with most variables between 70 and 90% (a very high prevalence of criminogenic problems). The one variable that has very low prevalence is mental illness, which can be linked to violence and can be considered as a problem area. One way to describe the prevalence of problems in the sample is to show the percentages of the sample that rated 2 or 3 on each of the 20 dynamic variables to give a ‘dynamic risk profile’ (Wong & Burt, 2007; see Fig. 1). As a comparison, the dynamic risk profiles of a group with psychopathy (mean PCL–R score = 28.2, s.d. = 2.7; mean VRS score = 58.4, s.d. = 7.7, n = 65; Wong & Burt, 2007) and the random sample are also presented in Fig. 1.

In the ABC sample, all but one of the dynamic variables had prevalence rates of 50% and above, with most variables between 70 and 90% (a very high prevalence of criminogenic problems). The one variable that has very low prevalence is mental disorder, which assesses the presence of associations between Axis I major mental illnesses and violence (not the mere presence of mental illness). The ABC programme is not designed primarily to treat individuals whose violence is the result of Axis I major mental illnesses. Not surprisingly, for the group with psychopathy, 16 out of 20 variables had prevalence rates of 80%. The overall dynamic risk profile of the ABC sample is slightly lower than that of the group with high psychopathy but still indicates a high-risk, high-need profile. The ABC group profile is much higher than the random group on all but the mental disorder variable, further confirming that the ABC group has many more problems and therefore is much higher risk than the average offender population.

**Treatment readiness**

The number of participants in the five stages of change (Lewis, 2004) from a subsample of 191 are shown in Fig. 3. The post-treatment stage of change is also presented to show the advancement in the stages of change as a function of treatment readiness. 

---

1. Because of increasing needs for such services, changes are underway to admit to this programme more patients with major mental illnesses linked to violence.
controlled study was carried out to compare treatment outcomes (about 24 months follow-up) of a treated gang group with a matched control gang group who had received little or no treatment. For members of both groups the mean age was about 24 years and they had about 20 criminal convictions before treatment; they were serving, on average, 6-year sentences. The mean length of treatment was about 8 months. The treated gang group had a significantly lower incidence of recidivism, significantly less major institutional misconduct and committed significantly less serious violent offences than the matched controls (Di Placido et al., 2006). The results suggested that, for a group of high-risk, high-violent and difficult-to-manage offenders, treatment in a risk reduction focus institutional programme, such as the ABC, can reduce both institutional misconduct and violence after release to the community.

Offenders who have committed serious violent acts such as murder or hostage-taking while incarcerated are often housed under extremely restrictive regimes in super-maximum security facilities. Deciding when they are safe enough to be transferred back to regular prisons is difficult, but prison authorities often are required to reintegrate them into the general offender population. Participation in the ABC treatment programme has been used as an alternative strategy to facilitate their reintegration. Within the ABC programme, both their security requirements and treatment needs can be adequately met. Results of an evaluation of such a strategy indicated that over 80% of the offenders (n=31) admitted from the super-maximum institution, the Special Handling Unit in Canada, were successfully reintegrated into a lower-security facility without relapsing (returning to the super-maximum institution) within a 20-month follow-up. They also have significantly lower institutional offence rates after reintegration than before (Wong et al., 2005).

Offenders with high levels of psychopathy were also treated in the ABC programmes (mean length of treatment about 8 months), with the primary treatment objective of reducing their risk for reoffending rather than resolving their personality disorders. In a recent treatment outcome study (Wong et al., 2006), 34 treated offenders with significant levels of psychopathy were matched with 34 untreated controls (mean PCL–R ratings of 28.6 and 28.0 respectively). The two groups were also matched for age (38.5 and 37.9 respectively), past criminal history (17.8 and 19.5 prior convictions respectively), and follow-up time (both 7.4 years). Their PCL scores were 51.1 and 55.2 respectively (P<0.05). They were high-risk, high-need and violent-prone offenders with high psychopathy scores. On follow-up, the treated and matched group did not differ in the number of violent, and non-violent re-convictions and sentencing dates, or the time to first reconviction. However, the treated group had a significantly less violent pattern of re-offence as indicated by the significantly shorter aggregated sentences they received (27.7 vs. 56.4 months respectively, P<0.05). Sentence length has been shown to be a reasonable proxy for the level of violence or severity of offending (Campbell, 1993; Belanger, 2001; Di Placido et al., 2006). Treatment may not prevent offenders with significant levels of psychopathy from reoffending, or even decrease the frequency of reoffending, but it did appear to reduce the degree of violence or severity of reoffending – a harm reduction effect. For offenders with fairly high PCL–R scores, 8 months of treatment is probably not long enough to produce the optimal outcome. Despite the less than optimal treatment ‘dosage’, the results support the contention that risk reduction correctional programmes that use the VRP approach can reduce violent recidivism in forensic clients with high levels of psychopathy.

**Description and treatment outcome of the VRP pilot programme**

The VRP pilot programme is a major part of an overall violence reduction strategy designed for the close supervision centres ‘to reduce physical, emotional and organisational violence, and to provide [prisoners with] an integrated care package . . . which addresses their physical and mental health needs’ (Fylan & Clarke, 2006: p. 6). The other components of the strategy are to provide high-standard mental healthcare to prisoners in close supervision centres and appropriate training to staff to equip them with the necessary skills to manage and care for violence-prone prisoners.

The programme participants were four prisoners with a mean age of 32 years (range 25–36) who spent a mean of 5.25 years (range 1.5–9.0) in the close supervision centre. Three of them murdered a fellow offender and one murdered a member.
of the prison staff while they were incarcerated or in custody. All had dysfunctional or difficult childhoods, long histories of serious violent criminal behaviours, substance misuse, and obviously, very serious institutional violence. One had symptoms of borderline personality disorder and another paranoid schizophrenia. The four treated prisoners were compared with two untreated prisoners (waiting list controls) who were 41 and 24 years of age and spent 8 and 3 years in close supervision centres respectively and had very similar social, criminological and institutional behaviour backgrounds. Data were collected through semi-structured interviews (using questionnaires developed for the purpose) with key staff and the prisoners, behavioural monitoring using Likert-type rating scales for target behaviours, scoring of the VRS and systematically collected behavioural observation narratives. Data were obtained during and after the programme when prisoners were transferred to a new but less supportive environment to test the generalisation of any newly acquired behaviour.

The small sample size precluded quantitative data analyses. We provide a synopsis of the findings taken from the summaries of the report (Fylan & Clarke, 2006; pp. 3 and 47). The authors of the report noted that ‘Data from interviews with staff involved with the program indicate the VRS has produced a marked improvement in prisoner behaviour. While the improvement staff perceived may in part be influenced by their greater insight into the prisoners and their behaviour, there is some independent evidence that violent behaviour has decreased and inter-personal skills have improved. There is also evidence, gained from interviews with staff and prisoners that better insight into prisoner behaviour – the part of both prisoners and staff – has resulted in more effective management of the risk of violence. Prisoners are better able to talk to staff about their emotional reactions and to avoid high-risk situations, and staff are better able to avoid high-risk situations for individual prisoners, and to better interpret and anticipate prisoners’ behaviour. Data collected from the prisoners’ new locations provide evidence that the skills developed during the VRP have been maintained, and that the changes achieved have been maintained in less supportive environments. . . . all three of the prisoners who agreed to be interviewed post-program report that they continue using the skills they developed on the VRP and that it has enabled them to better control their actions and to reduce the frequency and intensity of violent incidents.’

The authors further noted that ‘The VRP is potentially suitable for all prisoners, although their level of motivation to engage with the programme should be sufficiently high’, thus for some prisoners education/ orientation and motivational enhancement strategies should be provided before the programme. Staff do not believe that the VRP would discriminate against any prisoner groups. There is no evidence that (close supervision centre) staff at Woodhill have higher workplace stressors than those at the other (close supervision centre) sites. ‘All four prisoners on the pilot have progressed from or within the (Close Supervision Centre) to a less secure environment which has provided further support for the findings of Wong et al (2005).

Overall, the results suggest that the VRP is efficacious in reducing the frequency and intensity of violent acts by prisoners, and in assisting the reintegration of these prisoners into mainstream custodial settings. The programme staff also felt that they were better equipped to provide more effective management of the risk of violence through staff training, input from the mental health team, and interactions with prisoners within the framework of the VRP. The caveats in the interpretation of the results are the small sample size, the design of the study which limits causal inferences and the lack of statistical testing of the data.

**DISCUSSION**

We suggested that the assessment and treatment of violence-prone forensic clients for the purpose of risk reduction should be theoretically based, empirically driven and closely integrated. Assessment should tell treatment providers about the ‘who’, ‘what’ and ‘how’ of treatment. The goal of treatment is change, and how much positive change has occurred in treatment should be assessed and translated into a measure of risk reduction. Assessment and treatment should be closely integrated. Establishing a clear and psychologically relevant common language between assessment and treatment approaches as well as between staff and clients should increase the chances of achieving the stated goal of risk reduction. The VRP and VRS were described to illustrate how such integration and the use of a common language (i.e. having a common theoretical underpinning) can be achieved in practice and this was illustrated by reference to the existing programmes.

**Risk**

The risk level of the ABC sample indicated a probability of violent recidivism (55%) which is more than two times that of the average offender sample. Although the admission criteria for the ABC programme are not based on VRS scores, the type of clients admitted to the programme did clearly satisfy the violence-prone or ‘high risk’ admission criteria.

**Criminogenic need or dynamic risk**

The criminogenic need or dynamic risk profile of the ABC sample clearly showed that the sample has multiple problem areas linked to violence, and participation in a high-intensity violence-reduction programme would be appropriate. The group had only slightly fewer problems than a group with high levels of psychopathy but many more problems (or higher risk) than a randomly selected group of offenders. The group profile is also useful for planning the treatment programme. Managers and lead clinicians can use the information to decide what types of programmes are needed to address the existing criminogenic needs of a certain population. Overall treatment planning can then be undertaken and resources allocated based on the prevalence of problem areas in the samples of interest.

A similar profile could be constructed for the individual through a comprehensive clinical risk assessment. The ratings (0, 1, 2 or 3) of the 20 dynamic variables, rather than percentages, can be displayed as the individuals’ dynamic risk or problem-strength profile. Ratings of 0 and, to some extent, 1 are the individual’s strengths, and ratings of 2 and 3 are problem/treatment targets. The profile can inform staff of the presence and seriousness of the individual’s problems. Further in-depth investigation may be warranted depending on the presenting problems. Risk reduction interventions could then be formulated based on the individual’s risk profile and stage of change. The level of risk after attending a treatment programme can be re-assessed and similarly presented. The profile is useful for individual treatment planning. Offenders who are prone to violence and those with psychopathy share many similar problems and risk reduction treatment for
both groups should be quite similar. However, management strategies and treatment ‘dosage’ would be different (see Wong & Hare, 2003; Wong & Burt, 2007).

**Treatment readiness**

The majority of the ABC sample was at the contemplation stage on admission to the programme, followed by those in the pre-contemplation stage and the preparation stage, but none in the action or maintenance stage. The majority admitted to having problems but had not done anything about it yet – contemplation. The result is not unexpected as all ABC participants were admitted on a voluntary basis and would have, at least, ‘talked the talk’ by expressing a desire to change. Those in action or maintenance stages do not need such a high-intensity programme. The results suggest that the staff would need to use a lot of phase 1 treatment approaches, such as motivational interviewing and other treatment engagement techniques to try not to encourage and motivate these clients to start taking steps to move forward. Even more so, for the smaller group in the pre-contemplation stage, the first step for staff would be to assist the client in acknowledging their problems and considering the need for treatment. As an illustration, at the end of treatment, the majority of clients moved to the preparation stage, a substantial advance given the relatively short duration of treatment.

**Treatment outcome**

The results of four outcome studies, three with comparison groups, are encouraging. Both the VRP pilot programme and the ABC programme, which is similar in treatment philosophy and design to the VRP, appeared to be effective in reducing the risk and/or the severity of violent recidivism and/or institutional misconduct among perhaps some of the most challenging client groups: violent gang members, prisoners incarcerated in super-maximum security prisons and those with high levels of psychopathy. For those with psychopathy, and probably other high-risk violent offenders, the harm reduction treatment outcome is not unexpected.

In providing treatment to those with psychopathy and other very high-risk, high-need individuals, treatment providers should be realistic in their expectations of changes during treatment and outcomes after treatment. Wong & Hare (2005: p. 9) wrote in the *Guidelines for a Psychopathy Treatment Program*:

> ... it would be a mistake to believe that ... individuals with a history of predatory behavior will become model citizens. Soul will not become Paul, to use a biblical analogy. About the best we can hope for is that psychopaths who have gone through the ... [treatment program] will be significantly (in a practical as well as statistical sense) less prone to engage in violent behavior than they were before the program. Still, even modest reductions in the use of aggression and violence by psychopaths would be of enormous benefit to society.

Statistical analyses of treatment outcomes (criterion variables) using measures of changes in rates of reoffence or time to first reoffence (e.g. using survival analysis) may not be sensitive enough to detect some harm reduction effects, that is, reduction in severity of re offending.

With appropriate modifications, the VRP could be used for the treatment of sex offenders and young offenders, as well as those with Axis I major mental illnesses and co-occurrence of significant antisocial behaviours (personality disorders). Acute symptoms of mental illness have to be appropriately stabilised and staff need to be competent and prepared to deal with the expected periodic decompensations. However, those who are mentally ill and vulnerable should be treated in a separate treatment environment.

The development of the VRP and the VRS is an attempt to integrate correctional treatment and risk assessment for the purpose of providing theoretically derived and empirically driven assessment and interventions to violence-prone and treatment-resistant clients. The VRP and VRS are complementary: each providing the other with information required to fulfill the tasks of assessment, treatment and risk reduction.

**REFERENCES**


Collecting service use data for economic evaluation in DSPD populations
Development of the Secure Facilities Service Use Schedule

BARBARA BARRETT and SARAH BYFORD

Background Economic evaluation of the Dangerous and Severe Personality Disorder Programme is essential to ensure value for money. The collection of individual-level service use information is crucial to any such evaluation, but the best way to collect these data in secure facilities is unclear.

Aims To develop a method for the collection of individual-level service use information for prisoners/patients in secure facilities.

Methods Services provided within secure facilities were identified through examination of facility and policy literature, and discussions with managerial and clinical staff. Appropriate methods of measuring the quantities of services used were then explored and a new research tool capable of capturing all relevant services was developed and pilot tested.

Results The Secure Facilities Service Use Schedule (SF–SUS) records service use information from records and is capable of capturing data on the use of all individual-level services provided within a secure facility plus external services commonly accessed by occupants.

Discussion The SF–SUS is able to collect meaningful individual-level service use information for the economic evaluation of services provided within secure facilities.

Declaration of interest None. Funding detailed in Acknowledgements.

Research in personality disorders has long been neglected and there is little evidence of the clinical effectiveness (Warren et al, 2003) or cost-effectiveness (Swaray et al, 2005) of alternative interventions. The care of people with severe personality disorders, such as dangerous and severe personality disorder (DSPD), is now receiving overdue attention. Substantial funding has been devoted to a programme of new assessment and treatment services for this group and the DSPD Programme is subject to rigorous clinical and criminal evaluation (see Hart et al, 2007; Tyrer et al, 2007, this issue). Scrutiny from an economic perspective is also important because the DSPD Programme needs to be justified in terms of cost-effectiveness as well as effectiveness. Economic evaluation is the comparative analysis of alternative courses of action in terms of both their costs and consequences and is an established practice in the appraisal of health services (Drummond et al, 2005) and mental health services (Knapp, 1995).

A current evaluation of the assessment of DSPD in secure facilities (the IMPALOX study, details available from authors), included an economic component, and provided an opportunity to explore the methodological implications of undertaking an economic evaluation within a secure facility.

Typically, the cost of supporting patients/prisoners in secure facilities is calculated using a top-down approach. Top-down costing involves calculating the total cost of all resources and dividing them by the total population of the secure facility. However, economic evaluation requires individual-level cost data, rather than average costs that do not vary according to prisoner/patient (Drummond et al, 2005). To calculate individual-level costs, individual resource use must first be identified and measured (Beecham & Knapp, 2001). Identification involves drawing up a list of all resources relevant to a particular prisoner/patient. Measurement is the means by which data on the quantity of resources are collected. One obstacle to the successful completion of the IMPALOX economic evaluation was the lack of established methods of measuring resource use in secure settings. This paper explores the characteristics of people referred for DSPD assessment and/or treatment and the nature of secure facilities, in an attempt to develop a method of measuring service use information that is appropriate for research into the economics of secure facilities.

METHOD

Identification of resources
The first step in the calculation of costs for the purpose of undertaking an economic evaluation is the identification of resources. Resources relevant to a DSPD population were identified through a review of relevant literature and policy material. Information on interventions offered as part of the DSPD assessment and treatment programme were collected from a number of policy and service-level documents. (Home Office & Department of Health, 1999; Home Affairs Select Committee, 2000; National Institute for Mental Health in England, 2003; DSPD Programme, 2005; Hart et al, 2007; Tyrer et al, 2007, this issue). General prison and secure hospital services available to all prisoners/patients in secure facilities were located through a review of annual reports of prisons and secure hospitals. In addition, managerial and clinical staff at one prison providing a DSPD assessment and treatment programme and one secure hospital were asked for comments on the services identified through the literature.

Measurement of resources
Once identified, there are a number of methods available to measure the quantity of resources used by participants in an economic evaluation, including questionnaires, diaries or searches of case notes (Byford et al, 2003). In healthcare evaluations, resource use is commonly measured in interviews with patients by using a service use schedule such as the Client Service Receipt Inventory (CSRI; Beecham & Knapp, 2001). Although service use questionnaires often need to be adapted for different participant populations, the principle disadvantage is the need to rely on the memory of interviewees over what can be a significant number of months. Service use diaries are
one method of improving recall and involve asking participants to record their use of services prospectively over the study period. Diaries can be highly structured and involve simple tick boxes. However, the more complex and broad the range of services used, the harder it becomes for diaries to remain manageable. An alternative method is to collect retrospective information from case notes or electronic administrative databases. Records are likely to be more accurate than relying on user recall over a substantial period of time, but record searches can be time consuming, may not record exactly the information needed and will often be hampered by poor completion, missing files and illegible entries.

There are three areas where the DSPD Programme challenges the application of existing methods for collecting service use information. First, DSPD is a complex condition. The Programme involves complex interventions with multiple goals, multiple agencies and a high degree of user involvement (Campbell et al., 2000). Individuals meeting the criteria for DSPD are a heterogeneous group and outcomes are multiple. The complexity of the DSPD intervention has clear implications for the measurement of resource use: individuals are likely to see a wide range of professionals, so data on a broad range of resource use must be collected, making diaries, in particular, a complex and significant burden on respondents.

Second, there may be problems in completing research in individuals with DSPD. By definition, prisoners/patients in the DSPD Programme are challenging individuals. The development of clinical diagnostic interviews for personality disorder has highlighted the difficulties of eliciting accurate responses in interviews (Tyrer & Ferguson, 2000). Therefore collecting service use information directly from prisoners/patients, whether through interview or diaries, might not be appropriate.

Third, there are practical difficulties in carrying out research in secure facilities. For example, access to secure facilities and to prisoners/patients is strictly controlled and requires appropriate permission and/or security clearance. The research must be undertaken alongside the regime of prison/secure hospital, making access to participants more difficult than community-based research, particularly where access is required for long periods of time. Interviews may be interrupted and agreed access may be denied as a result of security alerts.

Despite these difficulties, secure facilities offer one significant advantage for the collection of individual-level service use data compared with community-based research. Resources available in institutional facilities are essentially limited to the services provided on site and although records held in each institution will not record all resource use, they should be sufficient to capture the majority of the services available to prisoners/patients. Although some services are provided by external agencies, the secure nature of the facilities means that such contacts are closely monitored and recorded. Thus, data collection from records is likely to be the most appropriate method of measuring service use in secure facilities.

To ensure that data are collected systematically, a service use schedule is needed, covering the wide and varied range of services provided in secure facilities. None currently exists for the collection of data from records, rather than prisoner/patient interview. We therefore developed a new research tool for the measurement of resource use in secure facilities.

**RESULTS**

**Secure Facilities Service Use Schedule**

The Secure Facilities Service Use Schedule (SF–SUS) was based originally on service use schedules for economic evaluations designed by S.B. – the AD–SUS for adults and the CA–SUS for children and adolescents (see Byford et al., 1999; Barrett et al., 2006) – but was substantially modified on the basis of information collected in the identification phase of the research. Services were separated into three sections: accommodation, service use within the secure facility and service use outside the secure facility. A draft schedule was developed and piloted on 16 patients with personality disorder in a secure hospital and two prisoners in a high-security prison. All participants were being considered for the DSPD Programme.

The pilot study uncovered a number of issues concerning both the type of data and the way in which they are collected. In particular, services that were unique to secure settings were identified, such as travel between institutions requiring security escorts and drivers and non-face-to-face contacts with legal representatives (telephone calls and letters). These services were added to the SF–SUS. The pilot study also enabled us to include suggested data sources in the schedule. We were also able to identify information available electronically and that kept in paper files, and the likely difficulties in gaining access to different data sources. The revised SF–SUS is shown in the data supplement to the online version of this paper. The schedule takes around 2 h to complete, although this varies according to the availability of patient records. Information is recorded for a specific determined period (e.g. 6 months).

**Accommodation**

The accommodation section of the SF–SUS asks for the name of the institution, the wing or the ward and the number of days spent in each location. In prisons, this information can be gathered from the computer database ‘Inmate Information System’ or from the wing record. In secure hospitals, the information can be found in the patient’s medical files. For those in prison, the accommodation section also asks whether the prisoner was subject to any special measures such as a Rule 45 order (segregation) or an open F2052SH form (at risk of self-harm), since such measures result in more intensive staffing and therefore higher costs. Information on special measures is kept in a wing record or inmate personal record. Movement between secure facilities and the number of escorts involved are also recorded in the accommodation section.

**Service use within secure facilities**

Information on service use within secure facilities is held in a number of different locations. In prisons, inmate medical records and wing records provide information on contacts with health and social care professionals and periods spent in the prison hospital. For those in other secure facilities, this information should be available in personal medical records. The SF–SUS gives sections for contacts with 12 categories of professionals and asks for the number of contacts and the average duration of each contact within the specified study period. There is also space to enter the details of contacts with professionals not included in the list. As well as individual contacts with health and social care professionals, prisoners/patients can take part in a range of daily activities. The SF–SUS records the number of hours spent in different types of activities, including therapeutic groups, educational courses, creative activities,
work and sports activities and games, each of which incur a cost. Finally, information is collected on the number of complaints that a prisoner/patient makes during the study period, as some complaints can result in the involvement of the governor and occasionally, external organisations.

**Services external to secure facilities**

Prisoners/patients have access to a range of services external to the institution, particularly those provided by the health and criminal justice sectors. Healthcare received externally in local National Health Service hospitals is recorded in prison/secure hospital medical files and the SF–SUS records the name of the hospital, the type and number of contacts, the length of stay if appropriate and the medical specialty. Contacts with professionals from the criminal justice system, including police officers and legal professionals, are recorded in prisoner’s wing records or secure hospital patient files. Telephone calls and letters in prison are logged, allowing telephone contact with legal professionals and letters from legal professionals to be traced and counted. In addition, parole board hearings and mental health review tribunals are recorded on prisoner wing records and in secure hospital patient files.

**Other resources**

The SF–SUS elicits information on the use of services that can easily be identified, measured and recorded on an individual basis. There are, however, many other resources that cannot be directly allocated on an individual basis, for example security staff, utilities and administration. These resources should be treated as overheads and incorporated using a top-down approach (total cost of all overheads divided by the total population of the secure facility). Overhead costs are then added to the individual level costs, calculated on the basis of data collected with the SF–SUS.

**DISCUSSION**

Economic evaluation of the new DSPD services is crucial to ensure value for money, but existing research methods are not necessarily suitable for use in secure settings. In particular, we identified the lack of established methods for measuring resource use in secure settings. To develop an appropriate research tool to measure resource use for the purpose of economic evaluation, the characteristics of people referred for DSPD assessment and/or treatment and the nature of secure facilities were explored. Prison and secure hospital resources relevant to a population with DSPD were identified and a service use schedule, the SF–SUS, was designed and tested. The SF–SUS obtains service use information from records, rather than interviews, and covers all services relevant to occupants of secure facilities.

The SF–SUS has been used successfully to collect and cost service use information in a secure hospital in a sample of patients being considered for DSPD assessment (Barrett et al, 2005). The average cost per patient was estimated to be £65 545 over 6 months, ranging from £59 119 to £82 709, which indicates that despite the necessity of allocating a large proportion of overhead costs uniformly, the SF–SUS is able to identify substantial variations in service use and cost between individuals. The SF–SUS is currently being used in a number of economic evaluations in secure facilities.

There are a number of problems with the approach to collecting and costing service use information outlined here. First, collecting individual-level service use data in secure facilities is time consuming and can be frustrating. The ease with which data are collected depends upon gaining appropriate permissions and security clearance, the cooperation of the staff and access to all relevant prisoner/patient records. Even with full access, the researcher must abide by the regulations of the secure facilities and may find access denied at times of heightened security. These practical difficulties must be taken into consideration at the planning and funding stages.

Second, detailed costing of individual items of a complex service can never be completely accurate. The researcher is reliant on the accuracy of the records from which the data are collected and the accuracy of the unit costs applied to each item of service. Individual-level costing tends to produce lower estimates than all-inclusive top-down approaches, which involve dividing the total cost of a secure facility by the population of that facility. However, the average individual-level cost calculated in our pilot study was within £1000 of the top-down estimate, suggesting a relatively high degree of accuracy (Barrett et al, 2005). The secure nature of these facilities and the close monitoring of individuals help to produce a level of accuracy that is less likely in community-based research.

Prisoners/patients taking part in the DSPD assessment and treatment programme can have access to all the services and activities provided within secure facilities, and the SF–SUS is designed to capture this full range of resources. The SF–SUS is therefore an appropriate research tool for the collection of service use data for all residents of secure facilities, not just those within the DSPD Programme. Although some categories of prisoners/patients, such as those currently considered appropriate for the DSPD Programme, might be unlikely to return to the community during the course of a typical research project, this might not be the case for all prisoners/patients. For populations in which movement between secure facilities and the community is anticipated, a second version of the SF–SUS is available. This is capable of capturing services in either location.

**ACKNOWLEDGEMENTS**

This work was undertaken as part of the evaluation of the pilot assessment programme at HM Prison Whitemoor funded by the Home Office. We are grateful to Dr Helen Seivewright for her tireless work collecting service use data and for her feedback and input into the SF–SUS.

**REFERENCES**


BARBARA BARRETT, MSc, SARAH BYFORD, MSc, Centre for the Economics of Mental Health, Institute of Psychiatry, King’s College London, De Crespigny Park, London SE5 8AF, UK.

Correspondence: Barbara Barrett, Centre for the Economics of Mental Health, Institute of Psychiatry, King’s College London, Box PO24, De Crespigny Park, London SE5 8AF, UK. Email: b.barrett@kcl.ac.uk


