UNDERSTANDING SELF-REPORT RESPONSE BIAS IN HIGH-FUNCTIONING AUTISM

By

MARILYN ADELE SHER

A thesis submitted to the University of Birmingham

for the degree of

Doctorate in Forensic Psychology (ForenPsyD)

Centre for Applied Psychology

University of Birmingham

July 2020

Abstract

Assessment of self-report response bias, such as random responding, lack of insight/self-reflection, malingering or, conversely, socially desirable responding, should be integral parts of any forensic psychological assessment. However, many of the tools that are used are not designed or specifically validated for use for people who have High Functioning Autism (HFA). For the purposes of this thesis, HFA refers to those people with a diagnosis of Autism who have average to above average cognitive abilities. This thesis aims to address this gap in the evidence base. Chapter 1 provides a general introduction into the assessment of self-report response bias and the current challenges being faced in relation to using existing psychometric measures. Chapter 2 presents a systematic review of the literature on how self-report response bias has been assessed in forensic contexts in the UK over the last 10 years. The findings highlight that the UK seems to favour the Paulhus Deception Scales (PDS: Paulhus, 1998), which is different to the measures used in other parts of the world. Chapter 3 examines the psychometric properties of the PDS and considers its use in forensic contexts, with the focus on UK samples. Chapter 4 presents an empirical study that aimed to establish a normative data set for the PDS and Structured Inventory of Malingered Symptomatology (SIMS: Widows & Smith, 2005) with a High Functioning Autistic community adult sample. This study provides some early evidence that alternative cut-off scores should be used with this population, as part of a wider holistic assessment of response style and bias. Chapter 5 concludes the thesis with a summary of main findings and recommendations for future research and practice.

Acknowledgements

I would like to thank all the wonderful participants who took part in this study, despite the challenging conditions we face in the world at this time. I would also like to thank my supervisor, Dr. Caroline Oliver, for all of her guidance and support from the start, as well as Dr. Christopher Jones for his invaluable assistance and endless patience with my statistics. Thanks to my family, particularly my parents, who were encouraging all the way through and helped me to have the strength and confidence to “get this done”! Finally, I would like to say a special thank you to my loving husband and son for once again being patient whilst I pursued yet another challenging course!

Table of contents

Table of Appendices

List of Tables

List of Figures

Glossary of Terms and Acronyms

CJS	Criminal Justice System
HFA	High Functioning Autism
IM	Impression Management
PDS	Paulhus Deception Scales
SDR	Socially Desirable Responding
SIMS	Structured Inventory of Malingered Symptomology

CHAPTER 1: INTRODUCTION

Introduction

An innate aspect of human personality is the tendency to distort perceptions, favourably or unfavourably, unintentionally or intentionally – to varying degrees (Rogers, 2018a). Most people will exaggerate their skills or qualities, minimise their weaknesses or failings, overestimate their skills or abilities and also have limited awareness of the degree to which they do this (Dufner, Gebauer, Sedikides & Denissen, 2019). We see this type of 'distortion' in every walk of life, whether it is at work, within our relationships, in therapy, and in forensic contexts (Rogers, 2018a). As humans, our instinctual reaction is to protect ourselves, our reputations, and minimise negative consequences (Baron & Byrne, 2000). This tendency to 'deceive' means that whatever the context, when self-report measures are used, clinicians or administrators need to be able to establish the likelihood of the respondent portraying themselves either overly positively or overly negatively (Paulhus, 1998). In contexts involving the criminal justice system, civil law (such as personal injury) or in senior executive occupational roles, determining the personality features of the respondent, and tendencies towards deception and impression management, therefore becomes key. Considering the vast variations in potential populations and contexts, it is a point of contention as to whether or not it is possible to measure or detect response bias tendencies with an acceptable level of accuracy.

According to Rogers (2018a), a detection strategy is *"a conceptually based, empirically validated standardised method for systematically differentiating a specific response style (e.g. malingering or defensiveness) from other response styles (e.g. honest responding and irrelevant responding)"* (p.20). A key element of any assessment is establishing the accuracy of the information included in the assessment.

The ideal approach is to include multiple sources of information to explore areas where information is supportive or inconsistent with other aspects of information. As clinicians, psychologists will draw information from file documentation, such as reports prepared by others, gain information from clinical interviews (of the person and those involved with that person such as carers and family members) and from structured assessments. These assessments may be focused on performance, that is, how the person performs certain tasks or demonstrates certain abilities, such as memory or Intelligence Quotient (IQ) tests. Assessments may also take the form of observational measures completed by those close to them or by those who are caring for them. Finally, assessments may also include self-report measures. These self-report measures may include assessments that explore mental health symptoms, personality, attitudes, beliefs, and a range of other states and traits. All of these approaches to assessment can be utilised to inform decisions-making on whether the person being assessed is engaging honestly or not.

Approaches to Assessing Response Bias: Socially Desirable Responding and Malingering

Response bias describes a range of response styles that include: Giving responses to influence impressions made, to create a more favourable impression (from here-on-in referred to as socially desirable responding: SDR), and the over-reporting of symptomology often referred to as malingering. Self-report measures of SDR are generally either standalone, such as the Paulhus Deception Scales (PDS: Paulhus, 1998) or as a validity subscale that forms a part of a tool that explores other factors, such as the Minnesota Multiphasic Personality Inventory-2 (MMPI-2: Butcher et al., 2009), the

Personality Assessment Inventory (PAI: Morey, 2007) and the Millon Clinical Multiaxial Inventory – Version 4 (MCMI-IV: Millon, Grossman & Millon, 2015) Social Desirability Scale (Y scale).

In terms of malingering, this is often measured using two types of approaches: Firstly, Symptom Validity Tests (SVTs) refer to forced choice assessments where respondents have to provide answers based on multiple choice, yes/no or true/false options to questions (Pella, 2009). They are partly based on real symptomology, but are exaggerated, infrequent, or not entirely real, as well as descriptions of things most people do, but which the respondent is not admitting to or minimises (Pella, 2009). Over time, the meaning of SVTs changed to focusing on faking mental health symptomology rather than neurodevelopmental difficulties (Bender & Frederick, 2018). Symptom Validity Tests (SVTs) were developed to measure over-reporting of symptomology, such as the Portland Digit Recognition Test (PDRT: Binder & Willis, 1991), the Recognition Memory Test (RMT: Warrington, 1984), as well as integrated scales such as the Debasement (Z Scale) on the Millon Clinical Multi-Axial Inventory-IV (Millon et al., 2015). Van Impelen et al (2017) utilised the SIMS (Widows & Smith, 2005). However, assessments have also been developed to measure under-reporting of symptoms (akin to 'faking good') such as the Supernormality Scale (SS), which assesses the under-reporting of symptoms, and a very low score on the Disclosure (x) scale of the MCMI.

In contrast, Performance Validity Tests (PVTs) tend to focus on how people perform on certain tasks and consider atypical patterns in test performance (Rogers, 2018b). One of the earliest assessments developed to explore malingering in memory/feigned memory problems was the Rey Fifteen-Item Test (Rey, 1964). PVTs

were initially designed to explore neurocognitive-biased performance, focusing on drawing information on floor effects (Bender & Frederick, 2018). As such, these approaches consider people's performance on neuropsychological assessments such as the Wechsler Intelligence scales, by looking out for unusually low scores, discrepancies or unlikely styles of responding that are not consistent with the way the test is designed. For instance, the Wechsler scales are designed in a way that items start off easier and become progressively more difficult, so an assessor would expect a person to perform better on earlier items that later items, rather than the reverse (Pella, 2009). Other tests can detect purposeful poor performance by presenting tasks that may seem quite challenging, but are actually quite easy, even for those that have significant organic impairments, such as the Test of Memory Malingering (ToMM: Tombaugh, 1996).

Due to the nature of response bias and what it represents, socially desirable responding and malingering are considered a situational as opposed to a stable factor (Rogers et al., 2010). Irrespective of the direction of response bias, Hart (1995) argues that it falls on a continuum rather than being either present or absent. This acknowledges that some form of bias besets everyone's response styles, but just some more than others, depending on circumstances, incentives, and internal beliefs or personality styles. Whilst it is now accepted that response bias is fluid and situational, some studies have found a relationship between self-report response bias and factors such as IQ (Jobson et al., 2013) and age (Mathie & Wakeling, 2011). For instance, in relation to IQ, Jobson et al.'s (2013) study found that those with lower IQ tended to present with higher levels of social desirability, in line with a number of other studies. Mathie and Wakeling's (2011) study of 1730 sex offenders found that as age increases, impression management increases. Interestingly, the same was not found for self-

deceptive enhancement, suggesting that as offenders' age, they are more likely to deceive others as opposed to themselves (Mathie & Wakeling, 2011). The authors do acknowledge, however, that the correlations were small, and the finding was in contrast to other studies where the reverse was found, in that Self-deceptive Enhancement (SDE) related to age, but not Impression Management (IM) (Mathie & Wakeling, 2011). This therefore means that whilst response bias may be largely dependent on circumstance and context, other factors that may be more stable or fixed, can play a contributory role.

Motivations and Incentives

People can be influenced to respond in a biased way for a multitude of reasons. For instance, they may over-report skills or positive traits, whilst under-reporting psychopathology, to gain something, such as custody of their child (Archer et al., 2016). Conversely they may over-report mental health symptoms to mitigate their role in offences or to remain in a hospital setting as opposed to prison, or they may simply provide responses that are not reflective of their actual functioning due to a general lack of insight (Archer et al., 2016). However, motivation is not always adequately considered in studies of response bias (Ohlsson & Ireland, 2011). Cassano and Grattagliano (2019) suggested that there are three conceptual models which are relevant when considering reasons and evidence for bias in assessment: pathogenic (due to mental health symptomology); criminological (where faking or exaggeration relates to attaining some gain); and adaptation (a strategy devised when a person needs to defend themselves, such as during a trial). This also brings into question whether response bias is a response set that is elicited at a specific time and thus mainly influenced by situational demands, or, if certain personality styles or traits are more prone to response

bias (Tan & Grace, 2008).

There is considerable debate in the literature about whether personality traits are in some way related to different types of response bias styles. Some studies have found that response bias has been associated with certain personality traits, which if present, are considered to increase the likelihood of, for example, malingering, such as in antisocial personality disorder (Hart, 1995). In addition, whilst Tully and Bailey's (2017) study using the Paulhus Deception Scales (PDS: Paulhus, 1998), did not find Impression Management (IM) associated with anti-social personality disorder (APD), they did find that Self-Deceptive Enhancement (SDE), defined by Paulhus (1998, p.9) as *"unconscious favourability bias closely related to narcissism"*, was in fact associated with narcissistic personality traits. In contrast, Impelen et al, (2017) found that whilst those who over-report symptoms tend to exhibit antisocial traits, this was only in the prison context in which their study was undertaken, and there was no evidence that antisocial behaviour in itself was predictive of biased reporting of symptoms. Although Impelen et al's (2017) study had small sample sizes, results were in line with larger reviews. For instance, a systematic review by Niesten et al (2015), also did not find any clear links between psychopathy or antisocial personality disorder, and over or under-reporting of symptoms. Additionally, a meta-analysis exploring similar variables, found a medium association between over-reporting and antisocial lifestyle, but not psychopathic traits (Ray et al., 2013).

The thought of consequences and trying to influence them may also play a role in offenders responding in a biased way. For instance, in a study of 172 violent offenders in the United States of America (USA) by Leite (2015), comparisons were made between pre and post conviction scores on the Minnesota Multiphasic Personality

Inventory (MMPI: Hathaway & McKinley, 1951) Infrequency (Fc) scale and the Personality Assessment Inventory (PAI: Morey, 1991) Negative Distortion Scale (NDS). The study concluded that offenders were more likely to malinger pre as opposed to post conviction, and that scores were higher depending on the level of adjudicated behaviours, with the Fc scale being marginally better at identifying variance depending on adjudications and predicting malingering, than the NDS scale (Leite, 2015).

It is therefore important to establish motivations, as these can sometimes be unintentional, and used 'subconsciously' as a way to preserve a person's "psychological integrity" rather than purposefully deceive (Cassano & Grattagliano, 2019). Some studies have focused on the more positive qualities found to be related to response bias that can serve a person well in day-to-day life, such as conscientiousness, agreeableness, emotional stability and extroversion (Paulhus, 2002; Tan & Grace, 2008). Von Hippel and Trivers (2011, p.13) argue that, from an evolutionary perspective, self-deception is adaptive, useful, and an *"offensive"* measure (e.g. evolved in order to increase the likelihood of an advantageous outcome), as opposed to necessarily a *"defensive"* approach. Yet, most of the research into deception considers it defensive due to a person's difficulty coping with consequences they may face (Von Hippel & Trivers, 2011). A further factor to consider is whether response bias is a dichotomous construct, that is, either present or absent, or whether it is something that appears on a continuum. Young (2017) states, *"in the assessment context, malingering and related negative response biases can be placed on a continuum of dimensionality, from unconscious-based somatic complaints to outright conscious malingering*" (p.83). Von Hippel and Trivers (2011) argue that people can be "*both deceiver and deceived…this is achieved through dissociations of mental processes, including conscious versus unconscious*

memories, conscious versus unconscious attitudes, and automatic versus controlled processes" (p. 1).

It has been argued that tests that explore Socially Desirable Responding (SDR) and malingering cannot in fact state that the person is lying as they do not measure the intention or motivation behind what the person is reporting (Drob et al., 2009). Biased responding is very much dependent on the person and context in which it is being assessed.

Guidance on the Assessment of Response Bias

With the increasing use of measures of response bias, some guidance has developed over the years in order to encourage quality and consistency in how response bias is assessed and interpreted. However, the focus of this guidance has largely been on effort and malingering, as opposed to SDR. In relation to effort and malingering, one could argue that, in general, people are unlikely to exaggerate or, conversely, minimise psychological problems or symptoms, if there is no clear incentive to do so (Salekin, Olley & Hedge, 2010). However, getting an assessment wrong and concluding that a person is purposefully distorting their answers in some way (false positives), when this is not in fact the case, can have far reaching consequences that can have a long term negative effect on them (Rogers, 2010; Walczyk, Sewell & DiBenedetto, 2018). For instance, if a victim of a crime is considered to be feigning their post-traumatic symptoms, when in fact they are genuine, they may not receive the support and therapeutic input that they deserve and require.

Some of the circumstance where false positives may arise relate to factors such as the reliability of measures in clinical contexts, the impact of various mental health

diagnoses, and variability across different countries and cultures. For instance, Rogers, Vitacco and Kurus (2010) argue that common tests such as the MMPI-2 and PAI show only modest test-retest reliability when used in clinical contexts as opposed to research contexts. In relation to mental health diagnoses, people with varying diagnoses can show inconsistencies in measures because their symptoms can wax and wane, as a natural part of their disorder. For instance, people with Axis I disorders have been found to show fluctuating results when assessments are repeated, which is something generally seen in patients with these disorders as it reflects fluctuations in symptomology rather than malingering (Rogers et al., 2010). Finally, in terms of differences between countries and cultures, variations in scores have been found between UK and US/Canadian studies on measures of response bias. For instance, Tully and Bailey's (2017) study of 358 people in the UK found that the UK sample scored significantly higher on the IM scale and Total score of the PDS, compared to the US/Canadian norms in the manual, despite there being no situational demands being placed on them. Thus, it is important to approach such assessments in evidence based and uniform ways that take into consideration other variables, such as diagnoses and cultural/ethnic context.

As mentioned, much of the current available guidance tends to focus on malingering as opposed to socially desirable responding and impression management. For instance, Slick, Sherman and Iverson (1999) made recommendations on the criteria that should be met in order to ascertain whether a person was malingering. The four categories relate to demonstrating that: The person has an external incentive; there must be evidence of actual response bias demonstrated through a neuropsychological assessment/validated response bias test; there is evidence of response bias based on self-

reported information; and the criteria above are not met due to some form of psychiatric, neurological or neurodevelopmental condition (Slick et al., 1999). Slick et al (1999) also provide guidance on establishing the probability of malingering, with varying ranges depending on what criteria are met or not.

Martelli, Nicholson, Zasler and Bender (2012) outlined some further considerations to undertaking assessments of response bias. These included drawing on multiple SVTs that have both built in measures as well as standalone ones, and that these measures are well validated and have the appropriate normative data. They also emphasised the importance of evidence that may go counter to assumptions of response bias, with all of this contributing to an integrated approach where historical, report based, assessment, clinical interviews, and other measures undertaken, are drawn together before making conclusions (Martelli et al., 2012).

Despite the growth and expansion of such tools, there are a number of challenges to consider in relation to how they are used, with whom, and their reliability and validity. As pointed out by Rogers (2018a), numerous terms are used to describe deceptive responding, including but not exclusively, deception, dissimulation, faking, impression management, feigning, social desirability and defensiveness. For the purposes of this thesis, response bias incorporates simulated adjustment (such as defensiveness, social desirability, denial of pathology and impression management) and over-stated pathology (such as malingering, feigning, over-reporting, and suboptimal effort) (Rogers, 2018a).

Alongside the varying terminology, we find differences in approaches across psychological specialisms (Brooks et al., 2016; Cassano & Grattagliano, 2019; McCarter et al., 2009), contexts (Tan & Grace, 2008) and countries (Archer & Wygant,

2012; Dandachi-Fitzgerald et al., 2013; Perinelli & Gremigni, 2016). Finally, many samples used to validate measures rely on normative data from the USA, and use convenience samples that feign response bias, impacting on the validity and reliability of measures (De Marchi & Balboni, 2018; Drob et al., 2009).

Assessing Specialist Populations

Bearing the above in mind, assessing response bias in specialist populations can be challenging. This is because the self-report measures available often do not have the evidence base supporting certain tools' use with certain populations. Yet, in the context of the Criminal Justice System (CJS), response bias needs to be taken into consideration when assessments are undertaken, and this is true for those offenders (or victims) who have neurodevelopmental difficulties. In the context of much of my medico-legal work, for example, I am frequently asked to assess people with an Autism Spectrum Disorder (ASD). Whilst British Psychological Society (BPS) Guidance (BPS, 2009) prompts us to consider malingering and response bias, there is little focus on how this is tackled with an Autistic client group.

How people with ASD cope in day-to-day life can differ depending on their stage of life, context, and level of support, with some difficulties being quite subtle but still leaving them highly vulnerable (Ali, 2018; Underwood et al., 2013). This is particularly true for those considered 'high functioning', who may appear on the surface to manage and be highly skilled, but often mask or learn the 'rules' of social responsivity (Ali, 2018; NAS, 2020b). For the purposes of this thesis, HFA refers to those people with a diagnosis of Autism who have average to above average cognitive abilities. Despite this, they may still struggle to understand and effectively communicate

their own internal state or functioning, which will have implications for how they respond on measures of response bias. Those with High Functioning Autism (HFA) may also demonstrate atypical skills in terms of social reasoning, understanding, and intuitiveness (Lerner, Haque, Northrup, Lawyer & Bursztjan, 2012). For instance, despite their intelligence, they may struggle to make intuitive moral decisions in circumstances they have not experienced before or had the opportunity to learn about. Furthermore, they may have difficulties regulating their emotions and understanding other people's views/feelings (Lerner et al., 2012).

Perspective-taking difficulties and problems understanding others' emotions are often referred to in Autism research as problems with Theory of Mind (ToM). Theory of Mind (ToM) refers to how able a person is at understanding another person's cognitions and views, drawing conclusions on how that person may then experience these mental states and the impact it may have on how they respond (Baron-Cohen, 1992). There is also high co-morbidity of other mental disorders with ASD as well as instances of misdiagnoses (Anckarsater et al., 2008; Murphy, 2003; Murphy et al., 2016). All of these factors may have implications for how people with Autism approach and endorse items on self-report measures of response bias.

Some have argued that people with ASD are less likely to commit offences than the general public due to the way they rigidly follow social and moral rules that are explicit (Grant et al., 2018). In contrast, certain traits may make them more prone to offending including: social communication difficulties and the misunderstandings these may cause; restricted interests that may be 'violent' in nature; obsessive focus on certain structures or routines; and deficits with perspective taking/theory of mind (Dein & Woodbury-Smith, 2010; Grant et al., 2018; Woodbury-Smith et al., 2010). This

becomes more complex in the case of HFA, where it is posited that they draw on cognitive strategies and language skills in order to compensate for deficits, but this ultimately leads to slower or disrupted processing of social information and interactions (Kaland, Smith & Mortensen, 2007; Tager-Flusberg, 2007), thus misinterpreting complex social and moral situations (Grant et al., 2018).

People with HFA can experience a range of difficulties when they come into contact with the Criminal Justice System, which can impact on how they come across and self-report during psychological assessment. This is particularly relevant given that their difficulties are generally not overtly obvious. For instance, they can be naïve in social contexts, struggle in new situations where they are being asked questions by unfamiliar people, misunderstand what is being said or asked of them, become easily agitated or anxious, and find it difficult to understand the impact of their behaviour on others or how they come across (NAS, 2020c). They are also more prone to making false confessions as well as are susceptible to certain types of questioning styles (Lerner et al., 2012; O'Mahony, 2012). A recent meta-analysis (Griego, Datzman, Estrada & Middlebrook, 2019) concluded that those with HFA were not prone to false memory and memory suggestibility. However, the studies included utilised the Gudjonsson Suggestibility Scales (GSS: Gudjonsson, 1997) or similar measures, which rely on verbatim as opposed to gist memory, in response to a story being told, and less on other forms of suggestibility, compliance and acquiescence, outside of interrogative contexts. In addition, understanding any potential links between HFA, co-morbid mental health problems and offending behaviour is central to the role of the assessing clinician (Lerner et al., 2012).

Aims of this Thesis

Assessment of self-report response bias, such as random responding, lack of insight/self-reflection, malingering or, conversely, socially desirable responding, should be integral parts of any forensic psychological assessment. However, many of the tools that are used are not designed or specifically validated for use with a High Functioning Autism population. The current thesis aims to address some of the gaps in the literature by providing a review of self-report measures that can be used in assessing response bias and establishing the evidence base for application in criminal justice contexts.

Having an understanding of response bias in people with HFA is highly relevant in forensic practice. This is because there are potentially many individuals who come into contact with the CJS who have HFA. Standard measures of assessment may not be appropriate as they may erroneously judge someone to be trying to fake or present themselves in a overly positive way, when the person with HFA may not actually be doing that. This is hugely important given the high stakes situations people with HFA may face in the CJS, and who may also not be well-equipped to navigate through the system effectively or cope with the emotional demands placed on them.

Due to the dearth of published research on the assessment of socially desirable responding in people with High Functioning Autism, it was important to consider some of the literature in clinical and neuropsychological contexts as well as forensic contexts. Chapter 2 presents a systematic review of the literature on how self-report response bias has been assessed in forensic contexts in the UK over the last 10 years. The findings highlight that the UK seems to favour the Paulhus Deception Scales/Balanced Inventory of Desirable responding (PDS/BIDR: Paulhus, 1998), which is different to the measures used in other parts of the world. Given the popularity of the PDS in the UK across a

range of contexts, including prison and secure hospital settings, Chapter 3 examines the psychometric properties of the Paulhus Deception Scales (PDS: Paulhus, 1998) and considers its use in forensic contexts, with the focus on UK samples. As a means to establish some evidence base for the use of frequently used self-report measures of response bias, Chapter 4 presents an empirical study that aimed to establish a normative data set for the PDS (Paulhus, 1998) and Structured Inventory of Malingered Symptomology (SIMS: Widows & Smith, 2005) with a High Functioning Autism community adult sample. The reason a non-offending sample was used was to provide an initial understanding of how people with HFA performed on these tools, prior to then applying this information to a narrower offending sample. A community sample may also be relevant in relation to people with HFA who may come into contact with the criminal justice system as victims or witnesses, rather than offenders. This study provides some early evidence that alternative cut-off scores should be used with this population, as part of a wider holistic assessment of response style and bias. Chapter 5 concludes the thesis with a summary of main findings and recommendations for future research and practice, with a consideration of alternative approaches to assessing response bias in a specialist cohort such as HFA.

CHAPTER 2: HOW SELF-REPORT RESPONSE BIAS HAS BEEN ASSESSED IN FORENSIC CONTEXTS IN THE UK OVER THE LAST 10 YEARS: A SYSTEMATIC REVIEW

Abstract

Aims: The current systematic literature review aimed to explore three key areas: identify which self-report measures of response bias were being used by UK psychologists within forensic contexts in the last ten years; establish if there are differences in what measures are used across various forensic contexts and populations; and if the measures used have adequate reliability and validity.

Methods: The review involved a systematic search of published and unpublished literature on self-report response bias measures, limited to UK studies over the last 10 years. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was used to guide the search process. Population, Intervention or Exposure, Comparator, and Outcome (PICO) was used to set the inclusion and exclusion criteria. Published quality review models were applied to structure and score the quality review.

Results: Sixteen studies met the inclusion criteria and were assessed for quality using Critical Appraisal Skills Programme (CASP), with quality scores mostly falling in the 'moderate risk of bias' range. The PDS/BIDR was found to be the most frequently used self-report measure of socially desirable responding (SDR), with it being used across prison, probation and secure mental health settings. The next most frequently used were the Multiphasic Sex Inventory (MSI) Lie scales. The PDS/BIDR and the MSI have adequate validity and reliability but there is a general lack of normative data reported for UK forensic samples.

Conclusions: The UK seems to favour the PDS/BIDR, which is different to the measures used in other parts of the world. Further research using more robust methodologies needs to be undertaken in order to further validate the measures most frequently utilised in the UK.

Introduction

It is rare for people to self-disclose completely and honestly in various domains of life, whether it is at work, in relationships, or therapy, and even less so in contexts involving the criminal justice system (Rogers, 2018a). As a result, people may respond in a style that is biased in some way, depending on the context and what is at stake. When respondents demonstrate bias, whether intentional or not, as clinicians, it is important to establish what this means for the overall assessment.

It has been argued that psychological assessment may be considered incomplete if it does not consider in some way whether presentations and symptomology have been feigned, minimised, or exaggerated (Impelen et al., 2017). This includes being able to establish whether reporting is in fact truthful (Drob et al., 2009). This is particularly relevant in medico-legal contexts. In fact, the British Psychological Society Guidance (BPS, 2009) for psychologists on assessing for effort and malingering highlights the importance of establishing, in medico-legal contexts, the level of effort, deception and potential for malingering. In addition, the National Academy of Neuropsychology guidance stipulates that symptom validity should be assessed, and if the decision is made not to, that this decision will need to be justified (Dandachi-Fitzgerald et al., 2013). Following such guidance on good practice, psychologists often use self-report measures as part of a wider clinical assessment. This information is then used to inform treatment, assess risk and provide valuable expert opinion when used in criminal justice contexts. It is of utmost importance to be as accurate, informed, and evidence based as possible, as well as establish whether respondents are answering questions in a truthful manner or not, and their reasons for responding in that way. Whilst much of the guidance available tends to focus on symptom validity and malingering, there is little

focus on simulated adjustment or socially desirable responding, despite this being frequently assessed in forensic psychological assessments.

The assessment of self-report response bias, most specifically socially desirable responding (SDR), became popular when Edwards (1953, 1957, cited in Uziel, 2010) developed an assessment tool adapted from certain items in the Minnesota Multiphasic Personality Inventory (MMPI: Hathaway & McKinley, 1951). From there came the development of the Marlowe-Crowne Social Desirability Scale (MCSDS: Crowne & Marlowe, 1960), which attempted to improve on earlier scales by helping to differentiate between denial and the absence of traits (Uziel, 2010). However, Paulhus (1998) argued that the measures up to that point in time approached SDR in a one-dimensional manner and so were not sensitive enough to detect the difference between exaggerated response styles and those that were in fact accurate reflections of a person. This was recognised by Sackheim and Gur (1979) who developed the Self-Deception and Other-Deception questionnaires which viewed SDR as two dimensional, and on which Paulhus went on to develop the Balanced Inventory of Desirable Responding (BIDR) and the Paulhus Deception Scales (Paulhus, 1998).

As a result, a number of tools have been developed that explore a range of different response biases or styles, be they minimising, or conversely, exaggerating psychopathology, and depending on situational demand characteristics. These include the Test of Memory Malingering (ToMM: Tombaugh, 1996), the Structured Inventory of Malingered Symptomology (SIMS: Widows & Smith, 2005), the SIRS-2 Rare Symptoms Scale (RS: Rogers et al., 1992), scales of the Minnesota Multiphasic Inventory (MMPI-2: Butcher et al., 2001), and the Paulhus Deception Scales (PDS: Paulhus, 1998), among others. Multiple reviews have been undertaken on a number of

these tools, which Rogers (2018a, p.38) concludes are "*conceptually sound and empirically validated*".

Terminology

Response bias is an umbrella term that refers to a variety of ways that people respond to questionnaires, surveys, and in interviews, that are not accurate or honest (Furnham, 1986). There are many terms that fall under the response bias umbrella, such as socially desirable responding, faking, dissimulating, impression management, response distortion, suboptimal effort, malingering, among others (Furnham, 1986; Paulhus 1998; Rogers, 2018a). Other forms of response bias may also include irrelevant and random responding on questionnaires, when the person simply disengages from the assessment or just randomly responds (Hart, 1995). All of these terms are frequently used interchangeably, but often have important differences.

The International Classification of Diseases version 11 (ICD-11: WHO, 2018, para. 1) defines malingering as "*... the feigning, intentional production or significant exaggeration of physical or psychological symptoms, or intentional misattribution of genuine symptoms to an unrelated event or series of events when this is specifically motivated by external incentives or rewards such as escaping duty or work; mitigating punishment; obtaining medications or drugs; or receiving unmerited recompense such as disability compensation or personal injury damages award*". Overstated pathology or malingering also differs from other disorders, such as fictitious and somatoform disorders, as there is a lack of external incentive and an intent to portray themselves in a certain way, such as mentally ill (Hart, 1995). Rogers (2018a) distinguishes malingering from fictitious and somatoform disorders, as intentional fabrication, and defensiveness,

which relates to purposefully withholding, within the context of extrinsic benefits. Rogers (2018a) also accounts for both these styles being present, referring to 'hybrid responding'.

Developing a two dimensional model, Wiggins (1964) referred to Alpha and Gamma Factors in SDR or simulated adjustment. The Alpha Factor (Wiggins, 1964) is seen as a response-style or trait (Tully & Bailey, 2017) and "*represents an unconscious favourability bias closely associated to narcissism*" (Paulhus, 1998, p.9). Thus, people respond in this way largely because they believe in their inflated qualities (Vispoel, Morris & Kilinc, 2018), and their distorted responses are thus, considered "*less accessible to the conscious mind*" (Uziel, 2010, p.244). On this basis, motivation and awareness are key factors to consider when exploring response bias. In contrast, the Gamma Factor (Wiggins, 1964) represents deliberate and conscious faking or lying due to attempts to manage impressions (Paulhus, 1998; Vispoel et al., 2018), and is considered a response-set or 'state' (Tully & Bailey, 2017). Such individuals' behaviours are purposeful, observed, and thus the person displaying them must be fully aware and conscious of how they are responding, therefore making a choice to lie or be honest (Uziel, 2010, p.244).

Motivations and Incentives

Motivation to deceive is an important variable to consider when assessing response bias. Motivations or gains may include portraying oneself positively in high stake situations, such as child-custody cases, or demonstrating significant mental health problems as a means for gaining a greater settlement in personal injury cases (Archer et al., 2016). Whilst there has been debate about whether response bias is a stable trait or

state (Hart, 1995; Impelen et al., 2017; Niesten et al., 2017; Ray et al., 2013; Tan & Grace, 2008; Tully & Bailey, 2017), there is recognition that situational factors heavily influence whether someone displays deceptive self-reporting and response bias. This has been supported by research that has identified offenders displaying differing levels of response bias pre-post conviction (Henry, Mandeville-Norden, Hayes & Egan, 2010; Leite, 2015) and pre-post treatment (Edwards, Whittaker, Beckett, Bishopp & Bates, 2012; Huntley, Palmer & Wakeling, 2012). It is also recognised that response bias may play a functional or 'positive' role (Cassano & Grattagliano, 2019), such as emotional stability (Paulhus, 2002), and adaptation to challenging circumstances (Von Hippel & Trivers, 2011). For instance, from an evolutionary perspective, people have developed the need to use deception as a strategy to secure resources in their daily lives as well as build a sense of confidence (Von Hippel & Trivers, 2011). In addition, studies have found associations between impression management and 'adjustment approach' interpersonal styles that promote social harmony and shared values, as well as higher levels of positive affect and wellbeing (Uziel, 2010). However, motivation is not always adequately considered in studies of response bias (Ohlsson & Ireland, 2011), and thus assessments cannot conclusively state that a person is being dishonest (Drob, Meehan & Waxman, 2009).

From these perspectives, given that deception can present in so many ways and incorporate a variety of conscious and unconscious processes, can there really be only one type of assessment that would capture such an aspect of the human condition? Should the motivations thus inform the assessment approach?

Approaches

Approaches to assessing response bias need to depend on the type of response bias being examined, such as whether someone is faking good/portraying themselves more positively or faking bad/exaggerating pathology.

There are multiple ways to assess response bias. For instance, some assessments have measures of bias or effort 'built-in' whereas others are stand-alone, either directly administered or self-report. Response bias is generally measured in two ways: through Performance Validity Tests (PVTs) and Symptom Validity Tests (SVTs) (Egeland et al., 2014). PVTs generally examine malingered cognitive impairment and SVTs measure exaggerated psychological and/or somatic symptomology, or conversely, the minimisation of psychological symptomology or 'faking good' (Egeland et al., 2014).

Slick, Sherman, and Iverson (1999) have outlined a structured and evidence-based approach to undertaking a comprehensive and considered assessment of malingering (though not other aspects of response bias). This guidance has featured in much of the literature and validation studies for a range of SVTs and PVTs, and highlights the importance of being able to establish that a test is reliable, valid, and has relevant normative data. This naturally leads to considering what makes a 'good' test.

Characteristics of Good Tests

The standards applied to ensure that a psychometric test is robust, include the following: it should be an interval scale, reliable, valid, discriminate, and have appropriate norms for the population it is being applied to (Kline, 2016, p.1). Archer et al (2016) outlined key factors that should be considered by psychologists when deciding on what test to use in an assessment: that the test is standardised; that there was

adequate evidence of reliability and validity in peer reviewed published papers on the tests; that appropriate normative data existed for the test; and that it was considered appropriate for use in the context it was being used in.

The first area of difficulty relates to how many of these tests are validated. Many studies instruct respondents to 'fake', thus removing the effects of certain variables or traits that may play a role in people responding in a biased way (Tan & Grace, 2008). Drob et al (2009) rightly highlights the problem with generalising studies where participants are asked to feign symptoms and then expect such normative data to be representative of actual clinical populations. This brings us back to the importance of being able to distinguish between actual feigning, 'unconscious' biased reporting, and feigning associated with other conditions, such as factious and conversion disorder, as defined by DSM-V (Cassano & Grattagliano, 2019) as well as Ganser Syndrome, hysteria and dissociation (Drob et al., 2009). As previously mentioned, certain conditions can also lead to elevated scores on response bias scales, such as those with intellectual disabilities (Pollock, 1996, cited in Drob et al., 2009). Alexithymia has also been found to contribute to over-reporting in Post Traumatic Stress Disorder (PTSD), rather than evidence of secondary gain (Young, 2017). In fact, Peters et al (2013) found that patients with schizophrenia scored above cut-off on the Structured Inventory of Malingered Symptomology (SIMS), when in fact their symptomology was genuine. Young (2017) outlines the challenges involved in establishing with certainty, samples that are known to respond in a biased way, and that the reliance on samples purposefully faking good or bad for research studies, can result in making errors about whether bias does in fact exist or not. This highlights the high probability of false positives, where an individual is assessed as displaying biased responding where in

actual fact they are not, particularly if self-report measures are used in isolation when assessing for response bias.

This raises the question of whether it is possible to formally accurately assess the likelihood of someone endorsing questions in biased ways given the variety of populations and contexts in which such assessments are used and applied. This may be particularly relevant in forensic contexts where the stakes are high. For instance, there may be a higher likelihood of response bias being evident to mitigate responsibility at the time of sentencing or when a person is applying for parole or a move to a lower category prison. In terms of populations, are certain groups of people, such as those from different backgrounds or cultures, more likely to display higher levels of response bias on measures, which could in turn have implications on practices in different countries across the world? This is relevant within multicultural countries too, such as the UK, where there are a variety of ethnic and cultural groups in the UK population.

Patterns and Styles Across Countries

There appears to be considerable variability in practice across the world in terms of approaches and the use of assessment tools exploring response bias. The majority of studies are undertaken in North America. In the US, the most frequently used assessments within a battery of forensic assessments are the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) and Personality Assessment Inventory (PAI) (Leite, 2015). A survey of paediatric neuropsychologists in North America found that 92% used a minimum of one PVT (92%) and one SVT (88%) when undertaking assessments, with greater use of standalone PVTs when undertaking forensic assessments (Brooks et al., 2016). A Canadian review of 85 psychologists engaged in child assessments and

130 psychologists in adult assessments found that 67% utilised the MMPI-2/MMPI-A (Minnesota Multiphasic Personality Inventory) and 31% used the Millon Clinical Multiaxial Inventory (MCMI-III) or Millon Adolescent Clinical Inventory (MACI) (Viljoen et al., 2010). An American review of 284 psychologists, reported on by Archer and Wygant (2012) found that 59% used the MMPI-2, followed by the PAI (38%), MMPI-2-RF (29%), MCMI-III (26%), and Rorschach (22%).

A survey of 188 neuropsychologists from the National Academy of Neuropsychology in the United States identified that the majority used the Test of Memory Malingering (ToMM), the Rey Fifteen Item Test (FIT), California Verbal Learning Test and the MMPI as part of assessing for response bias (Sharland & Gfeller, 2007). The most recent review of these types of measures concluded that the most frequently used is the MMPI, and that the PAI is growing in popularity (Cassano & Grattagliano, 2019). Other measures made reference to were the Structured Interview of Reported Symptoms (SIRS), Structured Inventory of Malingered Symptomology (SIMS), and Miller Forensic Assessment of Symptoms Test (M-FAST) (Cassano & Grattagliano, 2019).

Dandachi-Fitzgerald et al (2013) undertook a survey amongst neuropsychologists across six European countries, to understand their process of assessment, and views, in terms of assessing symptom validity. 515 completed the survey, with 55% having undertaken assessments in forensic contexts (Dandachi-Fitzgerald et al., 2013). Variation in practice was evident across countries, with approximately 70% including an SVT in assessment in the Netherlands and Norway, but much lower use in Italy and Finland (22% and 14% respectively) (Dandachi-Fitzgerald et al., 2013). However, in the majority of cases, SVT was reported as looking

for discrepancies, observed behaviour and inconsistencies in cognitive presentations, rather than the use of standalone measures or assessments with embedded validity scales (Dandachi-Fitzgerald et al., 2013). They found that the neuropsychologists tended to place a lot of emphasis on clinical judgments, showed various strategies in terms of how they instructed patients to complete symptom validity tests, and limited uniformity in terms of how they handled outcomes that identified people as over-reporting symptomology (Dandachi-Fitzgerald et al., 2013). Thus, there was more reliance on subjective approaches, especially in comparison to methods used in North America (Dandachi-Fitzgerald et al., 2013). In the European survey conducted by Dandachi-Fitzgerald et al (2013), the most commonly used assessments were the Amsterdam Short-Term Memory Test (ASTM; Schmand et al., 2005), Rey Fifteen-Item Test (FIT), Test of Memory Malingering (ToMM), Word Memory Test (WMT), MMPI-2 and SIMS (Widows & Smith, 2005).

A survey undertaken with neuropsychologists in the UK that explored differences in approaches across forensic and clinical contexts, found that clinicians more often used SVTs in forensic (59%) compared to clinical (15%) contexts (McCarter et al., 2009). The most frequently used measures were the ToMM, FIT, WMT, and in a third of cases it was felt an SVT was not required as over/under reporting of symptoms was obvious based on presentation and performance during the overall assessment (McCarter et al., 2009).

In summary, there are first and foremost wide ranges of terms for response bias, which are frequently used interchangeably. Secondly, there are various factors at play that may motivate people to respond in biased ways in assessments. Thirdly, the number of tools and approaches for assessment are growing, and there is significant variability

in practice between North America and Europe in particular. In addition, the majority of validation studies are based on normative data from North America. Whilst some of this normative data is based on convenience samples, there is a substantial amount of forensic normative data available, but again this is largely based on North American populations. This can be problematic when making inferences about cut-off scores in the UK, with some studies already identifying differences in mean scores, for instance on the PDS, between the UK and American/Canadian samples (Tully & Bailey, 2017).

Finally, there is far less and outdated information on practices in the UK, and limited normative data for forensic populations, which is particularly concerning given the growing use of SVTs within forensic contexts. There are high stakes in ensuring confidence in assessment in forensic contexts– whether it be protecting the public from future harm or ensuring an offender is treated fairly and thus avoiding litigation. This is particularly crucial given the common assumption that offenders will either over- or under-report symptoms or distorted cognitions, or display lack of effort in assessments (Mills & Kroner, 2006). This means that knowing what and how response bias measures are being used in the UK, and establishing that they are robust and reliable, is central to their ongoing effective use and application in the UK.

The Current Review

Preliminary searches were undertaken on 8th January 2020 to establish the originality of the current review. An initial Google search of the terms "response bias" AND foren* AND assess* was undertaken to establish if there was enough information available to warrant a systematic review. This initial search generated 179000 results. An initial search of all University of Birmingham databases using the search term

"response bias" and foren* generated 1950 results. When the search was expanded beyond the university, 2357 results were generated.

It was also important to establish if any systematic or meta-analytic reviews had been done on this topic, and the Cochrane Library (search terms: "response bias" AND Offender OR Court) and Campbell database (search term: response bias) were searched. No relevant results emerged. The Google search identified two reviews, although neither of these reviews focused specifically on UK practice. One did focus on forensic contexts, but this was limited to sexual offenders.

Tan and Grace (2008) undertook a review of socially desirable responding (SDR) in forensic and non-forensic contexts, with the focus on sexual offenders. They found varied and inconsistent practices in how SDR was measured, managed, and controlled for across the studies reviewed (Tan & Grace, 2008). They also found that non-forensic studies relied on convenience samples (such as university students) with less extrinsic demands to drive response bias, whereas forensic studies had samples that mainly consisted of older, predominantly males with less educational exposure, lower cognitive ability, and who potentially had more at stake and so could have more reason to demonstrate socially desirable responding (Tan & Grace, 2008).

Perinelli and Gremigni (2016) undertook a systematic review exploring how clinical psychologists were using SDR scales. Their review focused more specifically on published papers that explored SDR in relation to certain psychological variables, such as personality, over a five-year period (Perinelli & Gremigni, 2016). Their review search only included one database (PsychInfo) and Google Scholar, and excluded dissertations and unpublished research. Of the 35 studies included in their review, most included convenience samples drawn from college and university students, with only

one study in a forensic context. They found that over 70% of included studies explored socially desirable responding as a uni-dimensional concept, and few considered the role of personality. Studies were drawn from North America, Europe and Australia, but differing practices across countries were not overtly considered (Perinelli & Gremigni, 2016).

The current systematic review will attempt to address some of the gaps identified above. Specifically, the review aims to explore up to date information on what response bias measures are being used by psychologists within forensic contexts in the UK by focusing on the last 10 years. This is important as a lot of the American normative data may not be generalisable to the UK population, particularly as a lot of studies in America drew on university samples or prison populations that were predominantly of Hispanic decent, that may not accurately reflect the race and cultures of those involved in the criminal justice process in the UK. In addition, there is variability in practice across the world, so again, focusing on UK practice will be key in establishing a broader sense of where the UK is in terms of response bias compared to other parts of the world.

To summarise, the objectives of the current review were to:

- Identify what self-report response bias questionnaires have been used by psychologists within forensic contexts in the UK over the last 10 years
- Explore any differences in the use of self-report response bias questionnaires across various forensic contexts and populations
- Identify if the assessment measures used have adequate reliability and validity

Method

In order to minimise risk of bias, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-P) 'Introduction' and "Methods' sections (Moher et al., 2009 & 2015; Shamseer et al., 2019) were used to guide the review process.

Sources of Literature

A search of six electronic databases was undertaken on 9th January 2020. The total number of papers identified through this process, prior to removal of duplicates, was 9745. The selection of databases was determined based on a review of other studies on similar topics. The time period was initially unlimited and related to the availability of articles for the specific database. Some bias would have affected the initial searches undertaken: Firstly, it was not possible to search all potential sources of data due to time constraints. Secondly, due to financial constraints and time factors, it was not possible to get papers translated that were in foreign languages, and so only English sources were included in the searches. The database, date range, and number of articles identified were as follows:

- Web of Science Core Collection (1900 to 2019) – 1899 results
- Ovid – PsychInfo (1967 to 2019) – 2814 results
- Medline (1946 to 2019) – 846 results
- Embase (1974 to 2019) – 1290 results
- PubMed (1974 to 2020) – 44 results
- PsycArticles (1912 to 2019) – 2852 results

In order to reduce publication bias, five potential experts in the area of response bias were also contacted (see Appendix B for a list of experts) for any unpublished data. Four out of five experts responded and one provided two additional papers. Experts provided no other unpublished data. Conference presentations and dissertations were included in the initial searches. A further review of reference lists was undertaken, as well as a Google search to minimise the likelihood that other publications were missed. Two additional publications were identified following review of reference lists and the Google search.

Search Strategy

The database searches were undertaken in January 2020. Mapping related terms to words used in the systematic review question helped to generate search terms. The table below summarises the synonyms/related terms used. The search syntax can be found in Appendix A.

Table 1: Search themes and terms

Response Bias	**Forensic**	**Assessments/Evaluations**
Response bias	Forensic	Assessment
Bias	Criminal	Psychometric
Socially Desirable Responding (SDR) / Social desirability	Offender	Questionnaire
Impression management	Mental Health	Evaluation
Lie/lying	Court	Test
Deception		Survey
Deceive		
Dishonest		
Manipulate		
Malingering		

Effort		
Faking good and faking bad		

Study Selection

Following the initial searches, which resulted in 9745 sources, 2500 duplicates were removed on 10th January 2020. Following that, a further 605 secondary sources (e.g. book chapters and commentaries) were removed on 11th January 2020. The balance prior to formal inclusion/exclusion criteria (Population, Intervention or Exposure, Comparator, and Outcome: PICO) being applied was 6640. The two papers provided by experts and two papers added following review of reference lists, brought the total to 6644.

Inclusion Criteria and PICO

The Population, Intervention or Exposure, Comparator, and Outcome (PICO) Framework (Schardt et al., 2007) was applied to identify relevant abstracts and articles, as it is the most widely used tool for addressing clinical issues (Erikson & Frandsen, 2018). The inclusion/exclusion criteria form that was used for the short-listing process can be found in Appendix C, and the PICO summary table can be found in Table 2 below.

Table 2: **Inclusion/Exclusion Criteria (PICO)**

PICO	Inclusion Criteria	Exclusion Criteria
Population	Forensic: • Offenders (all types of 'offenders' defined as receiving criminal charges/convictions for the commission of an offence) • Court referrals Males and Females Any ethnicity	Non-forensic Personal injury/civil
Intervention/ Exposure	Self-report/assessment based response bias measure	No self-report/assessment based response bias measure Computerised response bias measure e.g. response time assessments
Outcome	Response bias measure outcome reported	Response bias measure outcome not reported
Study Design	Experimental; quasi-experimental; observational (e.g. Cohort; Case Control); unpublished theses/dissertations; unpublished data from experts	Qualitative studies; case series studies; reviews, narratives; commentaries; editorials; other types of opinion papers
Other Factors	Year of publication: 2009 to 2019 Language of publication: English UK samples only	Publication prior to 2009 Non-English publications Non-UK

The following rationale for the inclusion and exclusion criteria using the PICO Framework was as follows:

The focus of the review was on the use of self-report response bias measures in forensic contexts, and so the 'population' was focused on all those in contact with the

criminal justice system. These included those with convictions as well as those who did not have convictions as either they were diverted away from the criminal justice context into the mental health system, or were still going through the court process. All males and females were included as it was important to establish if practices differed across sexes. In addition, both adults and adolescents were included as sometimes these tools are used with younger cohorts, and again it was important to establish test use across the age range.

In terms of the intervention, comparator and outcome aspects of PICO, the focus was on the response bias measures reported in studies. The review was focused on the use of response bias measures that were self-report or administered without the use of computers or other technology. Approaches that utilised computers (such as response time assessments) or other equipment (such as polygraph and penile plethysmography: PPG) were excluded, as these measures are less available, can be costly, and less likely to be relied upon in most forensic contexts. It was important to establish that the specific measure used was named in the study, that outcome data was reported on and that the design was quantitative.

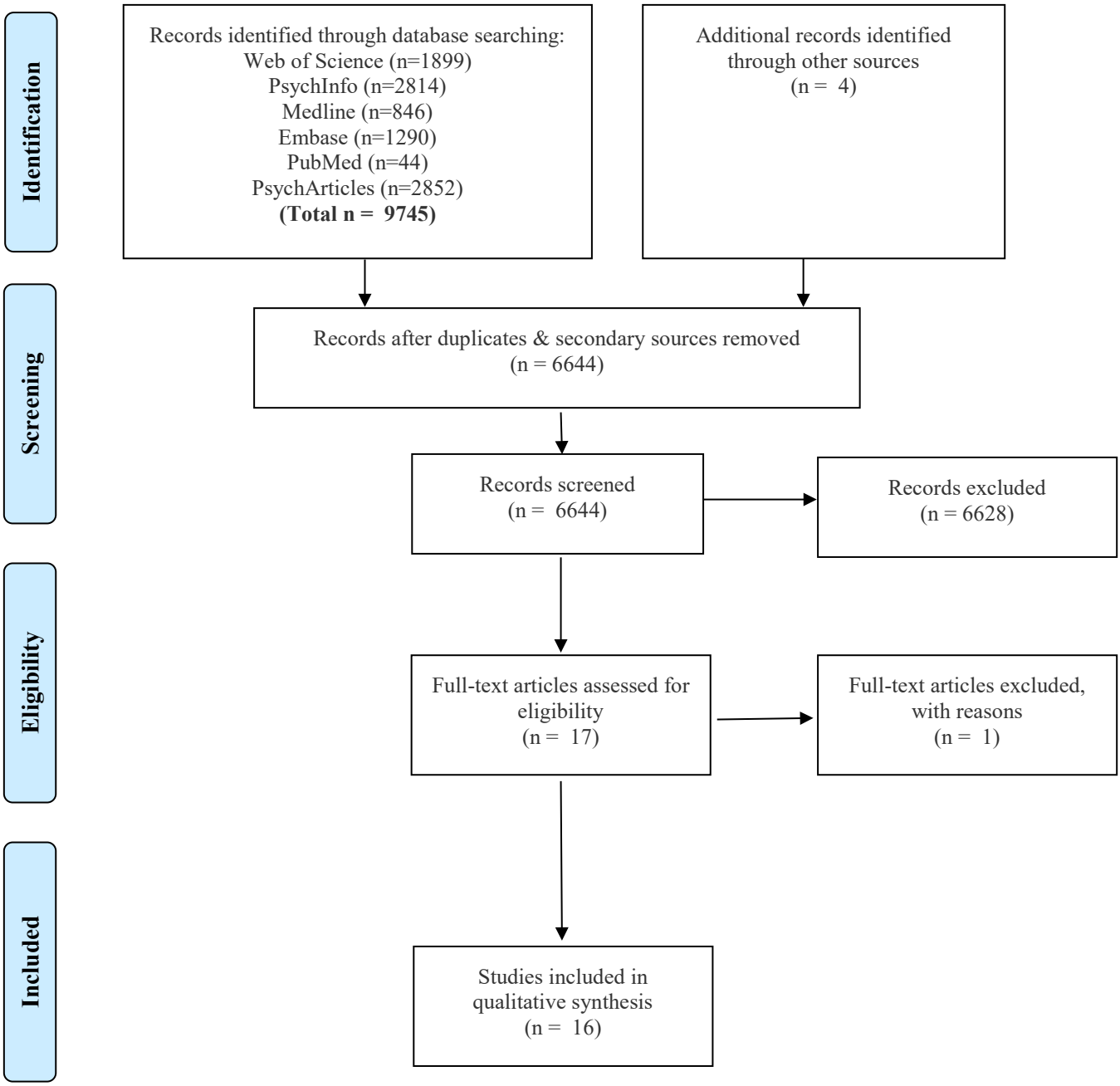
The review included any quantitative studies (such as experimental; quasi-experimental; observational) that were published in peer-reviewed journals. To reduce risk of bias, the review also included any unpublished theses/dissertations as well as unpublished data from experts. Only English publications were included, however, which did introduce some bias, but was necessary due to time and financial factors that would occur through a translation process. However, given the decision to focus exclusively on UK publications, publication language was not likely to cause a problem.

The ten-year timeframe (2009-2019) was introduced to firstly ensure that the

most up-to-date practices and use of tools were reviewed. The decision to focus on UK studies only was important, as the literature has shown that there is variation in practice across the world. Thus it was important to establish how response bias assessment was being approached in the UK compared to other parts of the world, particularly in relation to forensic contexts.

The PRISMA flowchart (Figure 1: Moher et al., 2009) outlines the paper selection process prior to the quality review. Initial searches identified 9745 papers across the six databases. Four further papers were identified through Google, hand searches, and through contact with experts. 3105 duplicates and secondary sources were then removed. When the 10-year timescale was applied, a further 3196 papers were discarded. This left a total of 3444 papers. Following the more stringent PICO inclusion and exclusion criteria being applied, the total number of papers left for quality review was 17.

Figure 1: **PRISMA flowchart**



Quality Assessment

A crucial aspect of undertaking systematic reviews is to establish the quality and credibility of studies to be included in the review, as well as to reduce the risk of bias (Higgins et al., 2019). The Critical Appraisal Skills Programme (CASP, 2018)

checklists for case-control and cohort studies were selected on which to base the quality review on. The CASP tool was selected because the quality criteria were relevant to the review questions being asked in terms of results being valid, available, and beneficial locally (CASP, 2018). In addition, it was developed by a multi-disciplinary working group who aimed to improve the quality assessment of reviews whilst allowing for flexibility/adaptability of its use, and is widely used in the UK (Spittlehouse, 2000). Some adaptations were made to the CASP tools in order to meet the specific review questions. The adapted tool can be found in Appendix D.

Each of the quality items were rated and scored as 'Yes' (2) if fully met the criteria, 'Partially' (1) if there was some evidence the criteria was met, but not fully, 'No' (0) if the criteria was not met, and 'Unclear' (?) if the information was not provided. An 'N/A' option was also available for certain criteria in order to account for that information not being relevant depending on the study design. The 16 items were split into five domains: Study design; Sampling; Analysis; Confounding variables; and Applicability/Value of research. Scores for each item in each domain were averaged to ensure equal/balanced weighting when calculating the overall quality score based on domains, and in line with best practice (Sanderson et al., 2007). Overall quality scores could reach a maximum total of 10, and were converted to percentages. These percentages were then assessed as being at a Low, Moderate or High risk of bias, based on Lees-Warley (2014) bias rating categories. One paper was excluded (See Appendix E) as the Balanced Inventory of Desirable Responding (BIDR) was used, but details not reported on, and those in their sample who had scored above cut-off were excluded from their study.

Data Extraction

Data were extracted for all remaining 16 studies using a Data Extraction Form (Appendix F) to ensure that relevant information was extracted consistently. The data extraction form included information from the Quality Review Form as well as general information, statistics reported, main findings and strengths/limitations. The focus of the majority of studies did not relate to the current reviews questions directly and so only relevant information was extracted for this review. It was important to ensure that data extraction was undertaken in a uniform and consistent way. Schlosser (2007) recommends that inter-rater agreement should be undertaken with 20-30% of selected studies. In order to establish that ratings were reliable and consistent, the author and an independent assessor who had experience of psychological literature and undertaking systematic reviews assessed 20% of the studies. This was achieved by undertaking inter-rater reliability using Cohen's kappa, which yielded a kappa of .84. According to criteria by Fleiss (1971), this level of agreement is "almost perfect". However, some risk of bias remained as not all 16 papers were assessed for inter-rater reliability.

Results

Descriptive Data Synthesis

A summary of the 16 studies included in the quality review can be found in Table 3 below. The majority of studies utilised cross-sectional designs. Ten out of the sixteen studies utilised data that was collected from existing clinical records as opposed to specifically for research. Thus, those that were undertaken in purely research contexts may not be as applicable to everyday clinical practice. Sample sizes ranged from 31 to

2497. Populations included both males and females within probation, prison and secure mental health environments. Offence histories included violent offences, sexual offences (contact and non-contact), and fire setting, and some included both offending and non-offending control groups. Race and mental health diagnoses were not always reported. The majority of studies fell in the 'moderate quality' range, with scores ranging across the review from 42% to 89%. Only three studies looked specifically at the assessment of response bias, utilising the Self and Other Deception Questionnaire (SDQ-ID and ODQ-ID: Jobson et al., 2013), the BIDR (Mathie & Wakeling, 2011) and the MCMI-III (Thomas-Peter, Jones, Campbell & Oliver, 2000). 14 of the studies utilised measures of SDR and IM, with only two exploring over or under-reporting of psychopathology using scales that formed part of symptomology and personality scales. This is unsurprising given that most forensic contexts, particularly those involved in undertaking assessments around risk, release, and pre-post treatment considerations, will be more interested in SDR, IM and faking good, as opposed to malingering or faking of pathology. However, contexts of negligence and at points of assessment for court where criminal responsibility and accountability is considered, assessments of malingering or faking bad, may be more evident. In addition, there may be a tendency to exaggerate symptoms of mental illness as a means to prolong stays in secure hospital settings.

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
Alleyne, Gannon, Mozova, Page & Ciarda (2016) Quality: 65% (moderate risk) Domains: D: 1.5 S: 10.8 A: 1.67 CV: 1 AV: 1.5	Firesetters: 65 female; 128 male Control: randomly selected 63 female offenders Mean age (all prisoners): 34.15 (*SD* = 11.8; range 18-74) 16 Prisons (10 male & 6 female) Race: 88% white; 6% black; 4% Asian; 1% Middle-Eastern Diagnoses: 66% previously engaged with mental health services	Case control – correlational design Research	IM scale of Paulhus Deception Scales (PDS: Paulhus, 1998).	IM of PDS: female firesetter (Mean 7.40; *SD* 4.02); male firesetter (Mean 5.35, *SD* 3.47); Female control (Mean 7.52, *SD* 3.66). MANCOVA: F=9.89, *df* = 2.237, $p < .001$, alpha = .26	Female control scored significantly higher on impression management	Study found unique differences between offender groups Analyses controlled for psychopathology and psychological features Self-selection bias Correlational design, so unclear which variables preceded which

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
Edwards, Whittaker, Beckett, Bishopp & Bates (2012) Quality: 62% (moderate risk) Domains: D: 1.5 S: 0.8 A: 1.67 CV: 0.5 AV: 1.5	34 adolescent males (sexually harmful behaviour) Specialist residential treatment facility Age: 11.6 – 16.3 (mean age 14.3; *SD* 1.2) Race: 79.4% white British; 11.8% black Afro Caribbean; 5.9% mixed; 2.9% white other Diagnoses: 35% Learning Disability; 5.9% ADHD; 5.9% ASD; 52.9% Conduct Disorder	Cohort Comparison done with non-completers, but PRI scores not reported Clinical	Personal Reactions Inventory (PRI: Greenwald & Satow,1970) social desirability scale. Standard scores based on 128 non-offending British adolescent males (alpha .82) Multiphasic Sex Inventory Sexual Multiphasic Sex Inventory Social Desirability subscale (MSI:SSD: Nichols & Molinder, 1984). Standard scores based on 57 adolescent non-sexual male offenders	Non-parametric tests used PRI mean score remained within the normal range (25-41) on pre (mean = 36.62, *SD* = 7.00) and post (mean = 39.47, *SD* = 6.55). Difference between pre/post was not significant (*p = .043*). MSI mean score remained within normal range (17-35) on pre (mean = 18.71, *SD* = 9.09) and post (mean = 22.24, *SD* 6.35). Difference between pre/post was significant (*p <.005*)	PRI and MSI:SSD in normal range pre-and post treatment for whole group. However, when explored, % of those outside norm pre treatment and then % within/below post treatment, sample showed more openness post treatment on both measures.	Study contributes to evidence base for offence specific treatment for adolescent sexual offenders Large heterogeneity within group, impacting on validity of results Small sample size PRI's for non-completers not reported. Details of non-completer group not provided Confounding impact of factors outside group intervention (e.g. therapeutic milieu) could not be controlled for

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
			(alpha .89)			
Elliott, Beech, Mandeville-Norden & Hayes (2009) Quality: 70% (low to moderate risk) Domains: D: 1.5 S: 1 A: 1.5 CV: 2 AV: 1	505 adult male internet offenders (IO's) & 526 adult male contact offenders (CO's) UK Probation Service Mean age: IO's 40.1 (*SD* 11.2); CO's 42.2 (*SD* 14.3) Race and diagnoses not provided	Observational – Cross sectional Comparison of IO's and SO's on various measures to see if the two groups can be distinguishable Clinical	Paulhus Deception Scales (PDS: Paulhus, 1998).	MANOVA identified significant difference between two groups F (1, 1028) = 34.12, $p < .001$ Univariate F test identified significant differences on PDS subscales. IM ($p < .01$; $r = .10$) and SDE ($p < .001$; $r = .25$) Mean scores not provided	Used PDS score to correct socially desirable responses on other measures, using Sanders (1991) technique Contact offenders found to be more likely to display socially desirable responding on measures	Large sample size and able to identify difference between two type of SO's Confounding considered and adjusted for Convenience sample; groups allocated based on index offence rather than previous offence history

Table 3: Studies included in the quality review

Author, year of publication *Quality score* ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	*Population & Location*	*Study design* *Existing clinical versus research data*	*Intervention – outcome measure used*	*Outcome measure internal consistency, reliability & validity*	*Summary of findings*	*Strengths and weaknesses*
Gannon, Alleyne, Butler, Danby, Kapoor, Lovell, Spruin, Tostevin, Tyler & Ciardha (2015) Quality: 60% (moderate risk) Domains: D: 1 S: 1.2 A: 1.3 CV: 1 AV:1.5	99 adult male firesetters (54 offered offence specific treatment (FIPP) – 45 treatment as usual (TAU) Seven English prisons Mean age: FIPP (34.6, *SD* = 13.4); TAU (31.4, *SD* = 11.3) Race: FIPP – 79.7% White European; 5.6% Black Ethnic Minority; 14.7% other. TAU – 82.2% White European; 4.4% Black Ethnic Minority; 13.4% other Diagnoses: FIPP (66.7%) and TAU	Quasi-experimental Clinical	The Impression Management Scale (IM) of the Balanced Inventory of Desirable Responding (BIDR 6; Paulhus, 1991)	FIPP: Pre means 77.9 (21.2); post means 83.1 (21.2) TAU: pre means 73.1 (20.7); post means 72.4 (19.1)	No differences at baseline between FIPP and TAU [$t(97) = 1.12$, $p = .27$, $d=.23$]. No sig. intervention x time interaction [Wilks' Lamda = .97, $F(1.96) = 2.7$, $p = .10$, $\eta_p^2 = 09$]	Power analyses undertaken Compared intervention effectiveness Treatment effects confounded by differences between groups Small sample size due to high attrition

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
	(51.1%) engagement with Mental Health Service					

Table 3: Studies included in the quality review

Author, year of publication *Quality score* ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	*Population & Location*	*Study design* *Existing clinical versus research data*	*Intervention – outcome measure used*	*Outcome measure internal consistency, reliability & validity*	*Summary of findings*	*Strengths and weaknesses*
Gannon, Ciardha, Barnoux, Tyler, Mozova & Alleyne (2013) Quality: 89% (low risk) Domains: D: 2 S: 1.4 A: 2 CV: 2 AV:1.5	136 male prisoners (68 firesetters, 68 non-firesetters) Ten English prisons Mean age: Firesetters (31.93, *SD* 9.33); non-firesetters (35.22, *SD* 12.30) Race: 84% White UK/Irish Diagnoses: Engagement with mental health – Firesetters (57%), Non-firesetters (41.2%)	Case control – matched Research	Impression Management Scale (IM) of Paulhus Deception Scales (PDS: Paulhus, 1998).	Firesetters: mean 4.0 (*SD* = 3.71, CI = 4.01, 5.80) Non-firesetters: mean 5.72 (*SD* = 4.16, CI = 4.72, 6.73) MANOVA (F1, 134) = 1.48, alpha/KR20 = .81	There were no distinguishable differences between firesetters and non-firesetters in relation to impression management (but were distinguishable on other measures	Excellent matching of groups Confounding considered Sample size relatively small
Henry, Mandeville-Norden, Hayes & Egan (2010) Quality: 60% (moderate	422 male internet sex offenders UK National Probation Service	Cross-sectional Clinical	Balanced Inventory of Desirable Responding (BIDR 6; Paulhus, 1991)	ANOVA Mean scores across clusters: Normal – IM 8.77 (3.83); SDE 3.73 (3.02); Total 12.50 (5.47)	Found strong negative correlations between high social desirability and emotional inadequacy	Large sample size Not fully representative of internet offending population as many go

Table 3: Studies included in the quality review

Author, year of publication *Quality score* ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	*Population & Location*	*Study design* *Existing clinical versus research data*	*Intervention – outcome measure used*	*Outcome measure internal consistency, reliability & validity*	*Summary of findings*	*Strengths and weaknesses*
risk) Domains: D: 1 S: 1 A: 2 CV: 0.5 AV:1.5	Mean age: 39.3 (n - 594, *SD* = 12.1, range 17-70) No information on race & diagnoses provided			Inadequate – IM 7.43 (3.63); SDE 1.85 (1.97); Total 9.28 (4.74) Deviant – IM 7.33 (3.39); SDE 2.26 (2.40); Total 9.6 (4.70)	Normal cluster scored most highly for social desirability	undetected Scores were pre-conviction and may change post conviction – influence of high stakes
Huntley, Palmer & Wakeling (2012) Quality: 59% (moderate risk) Domains: D: 1 S: 0.8 A: 1.67 CV: 1 AV:1.5	2497 adult male sexual offenders Small control for test-retest reliability on 26 male sex offenders) 25 UK Prisons Mean age: 43.33 (*SD* = 14.22), range 18-83 Race: 91.1% White; 6.4% Afro-Caribbean, 2.5% other	Cross sectional Clinical	Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984)	Mean BIDR scores not reported Marked correlation between LoC scale and BIDR (r= .42, *p* <.001, one-tailed)	Internal locus of control positively associated with higher levels of SDR Convergent validity - (n=2373) Pearson r – Total (.42); SDE (.48); IM (.27)	Controlled for SDR in analysis of LoC Large sample size Did not use adequate control group No sex offenders who refused or dropped out of SOTP included Test-retest group small and retesting period short (35 days)

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
	No information on diagnoses provided					
Jobson, Stanbury & Langdon (2013) Quality: 55% (moderate risk) Domains: D: 1 S: 0.67 A: 1.3 CV: 1 AV:1.5	100 males - 40 mental health intellectual disability (MHID); 32 intellectual disability (ID); 28 non-ID (NID) MID from medium secure forensic mental health unit. ID and NID from community sample Mean age: MHID 33.03 (*SD* = 12.45); ID 45.88 (*SD* = 15.01); NID 40.64 (*SD* = 10.41) No information on race & diagnoses provided	Cross sectional Research	Self-deception and Other-Deception Questionnaire-Intellectual Disabilities (SDQ-ID & ODQ-ID, Langdon et al, 2010)	Time 1 and Time 2 means: SDQ-ID total - MHID 2.89 (1.62), 2.90 (1.74); ID 3.97 (1.89), 2.85 (1.68); NID 2.32 (.94), 2.5 (1.11) ODQ-ID total – MHID 3.23 (2.49), 3.29 (2.97), ID 4.58 (2.59), 4.11 (2.17); NID 1.64 (1.42), 1.46 (1.97) Internal consistency: SDQ-ID time 1 & 2 (.71, .73); ODQ-ID time 1 & 2 (.81, .84) Test-retest reliability SDQ-ID (.76); ODQ-ID (.61) Sig main group effect for SDQ-ID – F(2.93) = 8.59, *p <.001* and ODQ-ID –	2 Factor structure confirmed and questionnaires shortened AS IQ increases, SDR decreases Offenders scored lower than non offenders for SDR (but difference not evident when IQ controlled for)	Reduced questionnaire length and established psychometric properties of new measure Offender group higher IQ and so may skew results IQ would have benefited from being more closely matched Small sample size

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
				F(2.96) = 12.27, *p <.001* (but not evident when IQ controlled for)		
Mann & Hollin (2010) Quality: 50% (moderate risk) Domains: D: 1 S: 1 A: 1 CV: 0 AV: 2	657 adult male sex offenders Comparison group: 76 non-sexual offenders and 42 non-offenders (male prison officers) UK Prisons Age: SO's - 41.77 (*SD* = 13.00). Non-offenders – 27.95 (*SD* = 6.65) No race or diagnoses information provided	Cross-sectional Clinical	12 items (Lie scale) from Eysenck Personality Questionnaire (EPQ: Eysenck & Eysenck, 1975)	Mean scores not reported for EPQ Lie scale	Those with high SDR showed a low reporting of 13 beliefs on the My Life questionnaire No differences in scores between sex offenders and non-offenders	Provided initial reliability and validity of new tool to assess schemas in sex offenders Assessments were undertaken as part of prison treatment. Lack of confidentiality could impact on response bias Some data not available to compare in relation to non sexual offender group
Mathie & Wakeling (2011) Quality: 65% (moderate	1730 adult male sex offenders UK Prison sample	Before and after study Clinical	Balanced Inventory of Desirable Responding (BIDR 6; Paulhus, 1984,	Total sample - IM pre/post means: 6.34 (4.18); 7.47 (4.35). SDE pre/post means: 5.39 (3.29); 7.21	SDE and IM increased post treatment	Increased debate on role of SDR, whether more trait like, and whether should be used to

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
risk) Domains: D: 1.5 S: 1.3 A: 1.67 CV: 0.5 AV:1.5	Mean age: 43.65 (*SD* = 12.6) Race and diagnoses not provided		1988) Multiphasic Sex Inventory Lie scales (MSI: Nichols & Molinder, 1984).	(3.63) Repeated measures MANOVA: Age significantly correlated with IM: r(1684) = 0.14, *p <0.01* Pre/post treatment sig increase = SDE: f(1.1343) =0.251.31, *p <0.01*; IM: F(1,1343) = 116.86, *p <0.01* ANOVA - High IM and low risk sig associated: F(3,1343) = 4.42, *p <0.01* and less change in scores among low risk (*p <0.01*)	No significant differences found between child molesters and adult sex offenders at either pre or post treatment testing on BIDR Age related to IM but not SDE SDR lower than expected High IM scores sig associated with lower risk of sexual offending	interpret other measures Large sample size Situational pressures influenced ratings as involved in SOTP No non-treatment control Some of the measures the BIDR was examined against have poor psychometric properties
Ohlsson & Ireland (2011) Quality: 72% (low risk)	206 adult male offenders (59% non-violent & 41% violent) CAT C training prison	Cross-sectional Research	Balanced Inventory of Desirable Responding (BIDR 6; Paulhus, 1991)	Means: Overall (12.98, 6.20); Violent (12.81, 5.77); Non-violent (13.10, 6.51)	BIDR used to control for SDR and impact on other measures SDR found to be	Anonymous Good size sample Sample only from one

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
Domains: D: 1.5 S: 1 A: 1.67 CV: 2 AV:1	Age 18- 54: Ranges - 18 to 29 (36.4%), 30 to 41 (32%), 42 to 53 (25.2%), over 54 (6.3%) No information on race and diagnoses provided				related to some aggressive motives	prison
Randall, Carr, Dooley & Rooney (2011) Quality: 48% (moderate risk) Domains: D: 1 S: 1 A: 1.3 CV: 0 AV: 1.5	103 child sex offenders: 30 clerics and 73 laymen. 30 convenience sample control group Community treatment site Mean ages: Clerical (54), lay offenders (44), control (33). Sig difference in age across groups: F(2126)	Cross-sectional Clinical offender sample	Multiphasic Sex Inventory Sexual Social Desirability scale (MSI: SSD, Nichols & Molinder, 1984)	SSD means: Clerical 8.75 (3.72); Lay offenders 9.59 (3.87); Control 12.44 (2.32	Clerical and lay offenders largely similar. Both offending groups had higher levels of SDR than control ANOVA F = 8.43 ANCOVA F = 8.98 Group difference: Clerical – Lay offenders < control	Groups not matched Small sample size

Table 3: Studies included in the quality review

Author, year of publication *Quality score* ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	*Population & Location*	*Study design* *Existing clinical versus research data*	*Intervention – outcome measure used*	*Outcome measure internal consistency, reliability & validity*	*Summary of findings*	*Strengths and weaknesses*
	=26.62, *p <0.01* No race and diagnoses information provided					
Sullivan, Beech, Craig & Gannon (2011) Quality: 64% (moderate risk) Domains: D: 1.5 S: 1.2 A: 1.67 CV: 0.5 AV:1.5	3 groups of male child sex offenders compared – 31 professionals; 31 intra-familial; 31 extra-familial Lucy Faithful Foundation Wolvercote Clinic Age: professionals (mean 45.35, *SD* = 9.13); Intra-familial (mean 43.68, *SD* = 10.04); extra familial (mean 44.71, *SD* = 10.04) No race or diagnoses	Cross-sectional Clinical	Social SSD subscale (MSI:J: Nichols & Molinder, 1984).	MSI:SSD Means – professionals (5.62, *SD* = 0.64); Extra-familial (6.17, *SD* = 0.76); Intra-familial (5.40, *SD* = 0.71)	ANOVA: No significant differences in means scores across 3 groups	Very well matched groups Small sample size Large number of priests in sample may have skewed some aspects on certain measures, particularly those in relation to sexual interest

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
	information reported					
Szlachcic, Fox, Conway, Lord & Christie (2015) Quality: 42% (moderate risk) Domains: D: 1 S: 0.3 A: 1.3 CV: 0.5 AV: 1	31 male inpatients Low to high security mental health units Age range: 23 – 65 (mean 37.32, *SD* = 12.43) Race: 12 Black British, 12 White British, 3 Asian, 3 Mixed, 1 South African Diagnoses: 21 mental health; 2 PD; 8 mental health and PD	Cross-sectional Research	Paulhus Deception Scales (PDS: Paulhus, 1998)	Means: PDS total 9.43, (*SD* = 4.98); IM 6.47, (*SD* = 3.34); SDE 3.34, (*SD* = 3.20) Internal consistency: PDS total (alpha = .74); IM (alpha = .72); SDE (alpha = .81)	IM sig negative correlation with Entitlement schema (Pearson's r = .358, *p* <.*05*) and Entitlement (Pearson's r = .432, *p* <.*05*) PDS scores used to control for responses on other measures	Some initial data on schema modes in forensic mental health/PD sample Small sample size Sample very heterogeneous, increasing risk of confounding
Thomas-Peter, Jones, Campbell & Oliver (2000) Quality: 66% (moderate	94 male (75) and female (19) inpatients and outpatients. Group split into high and low debasement scorers.	Cross-sectional Clinical	MCMI-III (Millon, 1994) – Debasement (z) scale	Means: High Debasement group (85.1, *SD* = 6.86); Low Debasement group (53.06, *SD* = 17.4)	Situational factors impacted on Debasement scores, in support of adaptional model of	Provides some evidence of impact of context on scores Heterogeneous and

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
risk) Domains: D: 1.5 S: 0.6 A: 2 CV: 1 AV: 1.5	Regional secure unit referrals Age range: 17.97 – 57.19 (mean 32.58, *SD* = 9.27) Race not reported Diagnoses: Schizophrenia (22); Personality Disorder (28); Depressive illness (9); Other (11); None (24)			MANOVA (d.f. = 1,2): Comparison of modifying indices: Wilks' Lambda 0.38 (p<.000) Comparison of groups: Wilks' Lambda 0.36 (p<.000)	malingering Differences between low and high scorers groups evident across debasement, disclosure and desirability. Higher debasement also correlated with higher desirability & disclosure scores (d.f. = 1,92, p <.000)	sample small once split into groups and so confounding factors such as diagnosis, offence history, whether inpatient or outpatient, and gender, could not be considered
Wall, Pearce & McGuire (2011) Quality: 77% (low risk) Domains: D: 1.5	83 males. 4 groups comprised of 15 internet child sex offenders (ICSO); 18 contact CSO's (CCSO), 25 non-sexual offenders (NSO), 25	Quasi-experimental Research	Paulhus Deception Scales (PDS: Paulhus, 1998)	Mean scores for PDS not reported ANCOVA: SDE associated with AAQ total F (1.73) 4.70, p <*0.05*, r=0.23. IM/faking good F	Less SDR associated with more emotional avoidance	Confounds well controlled for Confidential study Majority white sample

Table 3: Studies included in the quality review

Author, year of publication ***Quality score*** ***Domain scores:*** *Design (D); Sampling (S); Analysis (A); Confounding variables (CV); and Applicability/Value (AV)*	***Population & Location***	***Study design*** ***Existing clinical versus research data***	***Intervention – outcome measure used***	***Outcome measure internal consistency, reliability & validity***	***Summary of findings***	***Strengths and weaknesses***
S: 1.2 A: 2 CV: 1.5 AV: 1.5	non-offenders (NO) Probation Service community sample Age range: ICSO 22 – 53 (m = 40.80, *SD* = 11.02); CCSO 20 – 70 (m = 43.61, *SD* = 14.61); NSO 19 – 46 (m = 29, *SD* = 8.14); NO 18 – 62 (m = 34.2, *SD* = 13.22) Race: 100% white ICSO and CCSO; 92% white NSO and NO No diagnoses reported			(1.75) 5.73, p <*0.05*, r=0.26		Sample small

Quality Assessment

Quality ratings ranged from 42% to 89%. Most studies fell in the 'moderate quality' range. The highest quality papers (Gannon et al., 2013; Ohlsson & Ireland, 2011; Wall et al., 2011) utilised male prison and probation samples, with excellent matching of groups. For instance, in the Gannon et al (2013) study (89% quality rating), they paid a great deal of attention to matching groups, despite their sample being relatively small (*n*=99). Gannon et al (2013) utilised the IM scale of the PDS, comparing the scores of 68 fire setters and 68 non fire setters, with the sample being taken from ten English prisons. Whilst differences between the two groups were evident on other assessment measures used in their study, there were no significant differences evident in relation to their scores on the IM scale (Gannon et al., 2013).

In contrast, Ohlsson and Ireland (2011) used the BIDR:6 with 206 adult male offenders from a Category (CAT) C prison. Their study used the BIDR:6 to control for SDR on other measures and undertook analyses of SDR as a covariate (Ohlsson & Ireland, 2011). Whilst their study fell just within the 'low risk' quality range (72%), the key limitations of the study were that data were collected from only one prison, and the way they allocated violent versus non-violent offenders into groups meant some with violent histories may have been captured in the non-violent group if they had plea bargained to a lesser offence (Ohlsson & Ireland, 2011).

On the other hand, Wall et al (2011), with a quality rating of 77%, explored a probation service sample comprised of internet and contact sex offenders, non-sexual offenders and non-offenders. Whilst their sample size was quite small, confounds were well controlled for and the participants ratings were treated confidentially, limiting the impact of situational demands. Wall et al (2011) used the PDS, and found that lower

SDR scores were associated with increased emotional avoidance.

Sample Sizes

Sample sizes ranged between 31 (Szlachcic et al., 2015) and 2497 (Huntley et al., 2010). Larger sample studies included Elliot et al (2009) high quality study (70%) with 1031, Henry et al's (2010) moderate quality study (60%) with 422, Huntley et al's (2012) moderate quality study (59%) with 2497, Mann & Hollins (2010) moderate quality study (50%) with 775, and Mathie & Wakeling's (2011) moderate quality study (65%) with 1730. However, some samples were generally quite small, such as Szlachcic et al (2015) with 31, Edwards et al (2012) with 34 and Jobson et al (2103) with 100, impacting on statistical analyses of the SDR measures used.

In most cases mean scores were reported on but it was difficult to descriptively compare scores, mainly due to the variations of the scoring procedures for some versions of the BIDR and how the PDS (a later version of the BIDR) are scored in terms of their likert scales. Whilst Jobson et al's (2013) overall sample was quite small, especially when split into three groups, and there was not adequate matching of groups in terms of IQ (partly contributing to a moderate risk rating of 55%), it provided some promising support for a short and adapted tool for use with intellectual disability samples.

Study Samples

Most studies used prison or probation samples, with a few looking at mental health samples in secure settings (Edwards et al., 2012; Jobson et al., 2013; Sullivan et al., 2011; Szlachcic et al., 2015; Thomas-Peter et al., 2000). The majority of studies

focused on adult offenders, with only Edwards et al's (2012) moderate risk study (62%) utilising an adolescent sample from a specialist residential treatment facility. Most studies focused on males, with only Alleyne et al (2016) including females in the sample of fire setters. In those studies, including those with prison populations, where diagnoses were considered, large proportions had contact with mental health services (Alleyne et al., 2016; Edwards et al., 2012; Gannon et al., 2015; Gannon et al., 2013; Szlachcic et al., 2015). Not all studies reported on race (Elliot et al., 2009; Jobson et al., 2013; Mann & Hollin, 2010; Mathie & Wakeling, 2011; Ohlsson & Ireland, 2011; Randall et al., 2011; Sullivan et al., 2011; Thomas-Peter et al., 2000). Those that did report on race had predominantly white participants (Alleyne et al., 2016; Edwards et al., 2012; Gannon et al., 2015; Gannon et al., 2013; Huntley et al., 2012; Wall et al., 2011), with the exception of Szlachcic et al's (2015) moderate quality study (42%) who had twelve Black British, twelve White British, three Asian, three Mixed and one White African.

Measures Used

A range of measures were administered in the studies in order to assess response bias. In order of frequency of use, measures included the PDS/BIDR (Alleyne et al., 2016; Elliott et al., 2009; Gannon et al., 2015; Gannon et al., 2013; Henry et al., 2010; Huntley et al., 2012; Mathie & Wakelin, 2011; Ohlsson & Ireland, 2011; Szlachcic et al., 2015; Wall et al., 2011) followed by the Multiphasic Sex Inventory (MSI) Lie scales (Edwards et al., 2012; Mathie & Wakeling, 2011; Randall et al., 2011; Sullivan et al., 2011). Other measures used included the Personal Reactions Inventory (Edwards et al., 2012), Eysenck Lie Scale (Mann & Hollin, 2010), the Millon Clinical Multiaxial

Inventory version III (Thomas-Peter et al., 2000), and the Self and Other Deception Questionnaires for Intellectual Disabilities (Jobson et al., 2013). Table 4 below provides a summary of measures used.

Table 4: Summary of measures used

Study	Measure used					
	PDS/BIDR	***MSI***	***PRI***	***ELI***	***MCMI-III***	***SDQ – ID & ODQ-ID***
Alleyne et al., 2016	✓					
Edwards et al., 2012		✓	✓			
Elliot et al., 2009	✓					
Gannon et al., 2015	✓					
Gannon et al., 2013	✓					
Henry et al., 2010	✓					
Huntley et al., 2012	✓					
Jobson et al., 2013						✓
Mann & Hollin, 2010				✓		
Mathie & Wakeling, 2011	✓	✓				
Ohlsson & Ireland, 2011	✓					
Randall et al., 2011		✓				
Sullivan et al., 2011		✓				
Szlachcic et al., 2015	✓					
Thomas-Peter et al., 2000					✓	
Wall et al., 2011	✓					

There was also variation in how the response bias measures were used. Some studies utilised the response bias measure to inform the adjustment of scores on other psychometrics (Elliot et al., 2009; Huntley et al., 2012; Mann & Hollin, 2010; Ohlsson & Ireland, 2013; Szlachcic et al., 2015; Wall et al., 2011). Other studies found relationships between response bias and other variables, such as emotional inadequacy (Henry et al., 2010), locus of control (Huntley et al., 2012), aggressive motives (Ohlsson & Ireland, 2011), entitlement schemas (Szlachcic et al., 2015), to explore the role of situational variables (Thomas-Peter et al., 2000), and emotional avoidance (Wall

et al., 2011). An increase in response bias, most specifically SDR, was also found to be associated with the lowering of IQ (Jobson et al., 2013; Mathie & Wakeling, 2011).

Mathie & Wakeling's (2011) moderate quality study (65%) examined SDR amongst 1730 adult male sex offenders (child sex offenders, adult sex offenders, and offenders with both child and adult victims) across a number of UK prisons where the Sex Offender Treatment Programme (SOTP) was offered. Their study used the BIDR and examined scores before and after treatment, finding that both SDE (Self-Deceptive Enhancement) and IM (Impression Management) scores increased ($p < 0.01$) post treatment (Mathie & Wakeling, 2011). However, the three main groups were not distinguishable in their BIDR scores, and in general were lower than expected (Mathie & Wakeling, 2011). In addition, age was related to IM (r=.14, $p < 0.01$), but not SDE, and high IM related to lower risk for sexual offending (F=4.42, p <0.01) when rated using Thornton et al's (2003) Risk Matrix 2000 (Mathie & Wakeling, 2011). The main drawbacks of the study were that they did not have a control group and some of the offence specific measures did not demonstrate adequate reliability and validity (Mathie & Wakeling, 2011).

Population Scores

There was variation in how populations scored. Interestingly, controls and non-offenders appeared to have higher levels of SDR compared to offending populations (Alleyne et al., 2016; Henry et al., 2010), or no difference at all (Gannon et al., 2013; Mann & Hollin, 2010; Sullivan et al., 2011). An exception was Randall et al's (2011) moderate quality study (48%) that found offenders had higher SDR scores than controls, but this could have related to poor matching of groups, and a number of clerics being

included in their study. Some studies found specific groups of offenders scored higher for SDR, such as contact sex offenders compared to non-contact (Elliot et al., 2009). In addition, little clinically significant change in SDR was evident in those studies that explored treatment effects (Edwards et al., 2012; Gannon et al., 2015; Jobson et al., 2013), though one found an increase in scores post treatment (Mathie & Wakeling, 2011). Jobson et al (2013) did find significant main group effects when the three groups were compared, but these disappeared when IQ was controlled for, and there was a general finding that SDR scores increased as IQ decreased (Jobson et al., 2013).

Psychometric properties

Jobson et al's (2013) moderate quality study (55%) examined the psychometric properties of the Self and Other Deception Questionnaire (SDQ-ID and ODQ-ID) that they had adapted for use with people who had intellectual impairments. They also looked at the relationship between variables such as IQ and criminal offending, and how their scores differed when compared to those without intellectual deficits or offending behaviour, and those with intellectual deficits but no offence history (Jobson et al., 2013). Their sample came from 100 adult males within a medium secure forensic health unit (Jobson et al., 2013). Jobson et al (2013) found that the SDQ-ID and ODQ-ID demonstrated moderate internal consistency for the SDQ-ID (.71 to .73) and good internal consistency for the ODQ-ID (.81 to .84), as well as adequate test-retest reliability (SDQ-ID: .76; ODQ-ID: .61). Szlachcic et al's (2015) moderate quality study (42%) also found adequate internal consistency of the PDS total and SDE / IM subscales (.72 to .81).

Key Limitations

The key limitations of the papers included in the quality review related to sample size, poor matching of groups, large heterogeneity in groups and predominance of male and white samples. There was a general lack of use of adequate controls in most studies. For instance, Huntley et al's (2012) moderate quality study (59%) acknowledged not having included information from offenders who had dropped out, refused, or not completed offence specific treatment. In studies when test-retest designs were used, the time periods between time 1 and time 2 testing being undertaken was often very short. For instance, Huntley et al (2012) reported that the average time between testing was 35 days.

Discussion

Main Findings

The current systematic review aimed to explore up-to-date information on use of self-report response bias measures by psychologists in the UK over the last 10 years. Three key objectives of the current review were: to identify what response bias questionnaires have been used by psychologists in the UK over the last 10 years, whether they were reliable and valid, and if there was any variability of use across forensic contexts and populations. In terms of the first objective, the review identified that the majority of studies utilised the BIDR/PDS. The use of measures is in contrast to North American and European studies, which found greater use of MMPI and PAI (Leite, 2015) in North America, and the FIT, MMPI-2 and SIMS in Europe (Schmand & Lindeboom, 2005). The tendency to use the PDS/BIDR helped to address the second objective of the review, as the PDS has demonstrated adequate reliability and validity,

though could benefit from greater normative data on UK samples, as explored in Chapter 3. Other tools used included the Lie Scales of the MSI, and validity scales of the PRI and MCMI-III. With regards to the third objective of the review, there did not seem to be much variability in terms of what tools were used in various forensic contexts or offending populations. There was, however, one study that attempted to build on an existing measure, in order to validate its use for intellectual disabilities (Jobson et al., 2013).

Strengths and Weaknesses of the Review

Following a structured and rigorous approach when undertaking systematic literature reviews can help to reduce the risk of bias and improve on the information being communicated through the review (Schlosser, 2007). The key strengths of this review was that a clear protocol and process was followed to undertake the review, that being the PRISMA model, which helped to direct the search process to address the main objectives of the review. A range of sources were included in the review, including six databases, contact with experts, Google searches and hand searching through reference lists. Searching of reference lists does introduce a level of bias though as only papers with direct reference to response bias would potentially have been captured. Both published and unpublished studies were considered, to reduce publication bias. However, a comprehensive search of all database and studies could not be undertaken, which inevitably will introduce a level of bias. In addition, it was also not possible to obtain copies of several unpublished dissertations. PICO criteria and inter-rater reliability of quality ratings using the adapted CASP quality review forms, helped to reduce potential inconsistencies in quality ratings, but this was not done on all 16

studies included.

A key limitation is that many studies included in this review did not focus specifically on response bias, though were included as they used response bias tools. This means that other studies that did not have response bias related terms in their title, abstract or key terms would have been missed. Further factors that increased bias in this review related to contacting experts and having more than one of their papers in the review. In addition, quality ratings for the majority of studies included fell in the moderate to high risk of bias category. Finally, the systematic review did not explore all ways of assessing response bias, focusing on self-report questionnaires only. This means that alternative approaches have not been considered.

Conclusion

Applicability of Findings

The current systematic review provided a snapshot of how psychologists use self-report measures to assess for response bias in forensic contexts in the UK. The difference in use of tools across the world highlights the importance of exploring in more depth why the UK are mainly using the PDS, how it differs to the commonly used tools across the world, and establishing whether UK clinicians have chosen appropriately, or if other measures should be considered. The role of understanding the motivation to respond in a biased way on self-report measures should be considered and possibly influence the selection of response bias measure to use, rather than simply using the same one for all. This is also true for the selection of self-report measures exploring exaggeration of symptomology, which is highly relevant for those looking to use mental illness to mitigate their role in offences or to prolong stays in secure hospital

settings.

The review has also identified that these measures are used in a range of forensic settings and contexts. Situational factors will inevitably influence how offenders score, depending on the stakes. Self-report response bias measures are also interpreted and applied differently. However, there is limited normative data available based on UK samples. This means that clinicians may frequently be comparing scores in their assessments with inappropriate norms. In addition, whilst many assume that response bias is a concern and thus adjust scores on other measures as a result, some of the studies highlight that these measures should not necessarily be used as validity checks (Mathie & Wakeling, 2011). In some instances, response bias may even be protective or reflect a lower risk of offending, though focused research in this area is still needed. All of these factors reinforce the approach that self-report measures of response bias should not be used alone, but rather as part of a wider assessment of the person, and applied, where possible, in an individualised way.

Recommendations for Further Research

The review identified that the most commonly used self-report measure to assess response bias in the UK is the PDS/BIDR. It would be useful to undertake a meta-analysis on all UK studies that used the PDS as a means to developing a normative data set that is directly relevant to UK samples. In addition, studies that use and explore the use of self-report measures for response bias should ensure that groups are more specifically defined and matched, perform power calculations and ensure that confounding factors are considered and addressed, in order to improve the quality of studies on these measures. Finally, more effective use of control groups could assist in

gaining a better understanding of the relationship between response bias and other variables, such as personality, age, IQ, risk and motivation.

A large majority of the studies utilised measures of SDR and IM, with few considering the role of malingering of psychopathology. Whilst more often than not forensic psychologists are interested in changes in risk of reoffending and responsivity to treatment, there remain significant problems with offenders using mental health symptoms to mitigate their role in offending behaviour (thus highly relevant to consider in medico-legal assessments for court) or influence their time and stay in secure hospital settings. Further studies exploring the use of measures of malingering in these contexts would be useful.

It would also be useful to undertake further research exploring how various measures of response bias, not only self-report, compare to each other. This is important as a means to determining whether self-report questionnaires are still valuable, or if assessment of response bias is improved with alternative methods, such as attention or reaction time strategies (Wall et al., 2011). Finally, there appears to be a dearth of information on using response bias measures with neurodevelopmental populations, with only one study (Jobson et al., 2013) being identified. However, this may be due to cognitive difficulties, which make rating self-report measures difficult, and reinforces the importance of looking at alternative strategies for assessing biased reporting.

CHAPTER 3

PSYCHOMETRIC PROPERTIES OF THE PAULHUS DECEPTION SCALES

(PDS: PAULHUS, 1998)

Introduction

The interest in measuring response bias gained traction following Edwards' (1953, 1957, as cited in Uziel, 2010) development of a measure that drew on a number of the Minnesota Multiphasic Personality Inventory (MMPI: Butcher, Graham, Ben-Porath, Tellegen & Dahlstrom, 2001) items. From that time there was significant interest in finding ways to measure social desirability and response bias, with the popular Marlowe-Crowne Social Desirability Scale (MCSDS) being developed in the 1960s (Uziel, 2010). Crowne and Marlow attempted to address the issue that arose in the Edwards' measures in that it was difficult to differentiate between respondents who actually did not have symptoms and those that denied them (Uziel, 2010). Whilst the MCSDS became a very popular tool to use in exploring bias (Uziel, 2010), Paulhus (1998) argued that the MCSDS and other similar measures were one-dimensional, assumed that there was only one form of socially desirable way of responding, and failed to differentiate between exaggerations and accurate descriptions people make of themselves. Gur and Sackheim (1979) and Sackheim and Gur (1979) went on to develop the Self-Deception and Other-Deception questionnaire, on which much of Paulhus' scales emerged from (Paulhus, 1998). This review will examine the Paulhus Deception Scales (PDS) by Paulhus (1998), which was previously known as the Balanced Inventory of Desirable Responding (BIDR-7), in relation to its scientific properties and how it is applied, most specifically, in criminal justice contexts.

Paulhus (1998) developed the Paulhus Deception Scales (PDS) in order to distinguish between respondents who portray themselves honestly, those who purposefully manage how they come across, or those that unintentionally portray themselves positively. The PDS explores two specific types of socially desirable

response (SDR) styles: Self-Deceptive Enhancement (SDE) and Impression Management (IM). The Self-Deceptive Enhancement (SDE) scale examines a response-style or trait (Tully & Bailey, 2017) and "*represents an unconscious favourability bias closely associated to narcissism*" (Paulhus, 1998, p.9). Thus, those endorsing items on this scale may largely believe in their inflated qualities (Vispoel, Morris & Kilinc, 2018), and are thus, considered "*less accessible to the conscious mind*" (Uziel, 2010, p.244). In contrast, the Impression Management (IM) scale "*represents a well-known category of social desirability measures aimed at the crude form of dissimulation known as faking or lying*" (Paulhus, 1998, p.9), and is considered a response-set or 'state' (Tully & Bailey, 2017). Endorsements on this scale are seen as more deliberate and conscious (Vispoel et al., 2018), as they explore "*overt behaviours, and as such, their true nature is accessible to the respondent, who has a choice of whether to lie about them or not*" (Uziel, 2010, p.244).

Overview of the PDS

Paulhus' early research was borne out of an attempt to address conflicting views on whether socially desirable responding as a concept existed (Paulhus, 1998). Through his research he concluded that the problems in establishing SDR were because researchers were trying to understand it as one construct, when in fact there were two rather distinct forms – those being SDE and IM (Paulhus, 1998). Paulhus initially developed the BIDR drawing on two unpublished scales that explored conscious and unconscious biases (Paulhus, 1998). He revised the items of the two separate tools into one tool named the Balanced Inventory of Desirable Responding Version 3, and revisions included replacing 'maladjusted' items, including reverse keyed items, and

ensuring items were worded as affirmations (Paulhus, 1998). Exploratory factor analyses helped Paulhus determine item composition for the two scales, followed by confirmatory factor analyses to provide evidence for the separation of the 40 items into two distinct scales (Paulhus, 1984, 1986, 1991).

The PDS is the seventh version of what was previously known as the BIDR. According to the manual (Paulhus, 1998), the PDS is a self-report questionnaire that was developed in order to assess a person's tendency to respond in a socially desirable manner, whether as a result of situational demands or as part of a general personality trait that responds in this way. Most specifically, it assesses self-deception and impression management. On that basis, the author states, "*the PDS captures the two principal forms of socially desirable responding with two (relatively) independent subscales*" (Paulhus, 1998, p.1).

The PDS can be administered using a computerised or Quickscore manual version. The manual states that it takes around five to seven minutes to complete (though there is no time limit) and that people aged 16 or above should be able to read and understand the items. It consists of 40 statements which respondents rate based on a 5-point likert scale that indicates the extent to which a statement applies to them or not (Paulhus, 1998). Quickscore forms include place to administer, score and establish T-scores. Adjustments can be made for missing responses, as long as there are less than five.

The PDS is designed in a way that can assess whether a person has a tendency to respond with one, both or none of the socially desirable response styles, and also only captures scores for extreme responses (Paulhus, 1998). The tool does this by having a 5-point likert scale, but then scores are dichotomised to 0 or 1, with 1s only being

allocated to extremes in scores (Vispoel, Morris & Kilinc, 2018). Vispoel et al (2018) undertook an interesting, but somewhat complicated analysis of the PDS and explored five alternative ways of scoring the PDS, with the aim of improving clinicians' approach to assessing socially desirable responding. They concluded that polytomous scoring of the PDS increased its validity (Vispoel et al., 2018). However, Paulhus (1998) has strongly maintained support for dichotomous scoring as a means to differentiate actual versus exaggerated endorsements. As such, many practitioners tend to follow the manual and use dichotomous scoring, but researchers differ and tend to use polytomous scoring.

The two scales that make up the PDS are very much approached as separate scales, and are interpreted as such. The PDS is also sometimes used alongside a number of other assessments in order to ascertain the likelihood of a person enhancing how they come across on the test battery (Paulhus, 1998). In these cases, a validity check using the IM scale scores can be used to identify if a person 'may be' or 'probably' is, faking good or bad (Paulhus, 1998). The manual provides a formula to calculate the Marlowe-Crowne score from the PDS total raw score. The manual also provides guidance on interpreting the profile of a respondent based on four typologies, depending on whether a person scores high or low on the two scales (Paulhus, 1998; Smith, 2018). The four typologies are: high SDE and high IM; high SDE and low IM; low SDE and high IM; and low SDE and low IM (Paulhus, 1998). However, Smith (2018) argues that many forensic studies did not find it possible to fit into these typologies, possibly because the PDS is not sensitive to specific conditions and due to there being an overlap in what the SDE and IM measure.

The PDS is accompanied by a manual, which outlines how the tool was

developed, instructions for administration and interpretation, and evidence base for the validity and reliability of the assessment. The test authors emphasise that the PDS alone should not be used to determine whether a person is responding honestly or not, and should include information gained from other sources, such as direct contact with the person and information from other assessments (Paulhus, 1998). Thus, the PDS is a stand-alone measure often used alongside other measures, but there are options to use alternative tools that have built in deception measures, such as the Marlowe Crowne and Eysenck Personality Inventory.

The PDS has been used in a range of contexts. For instance, in exploring treatment readiness (Casey, Day, Howells & Ward, 2007; Day, Howells, Casey, Ward, Chambers & Birgden, 2009), police officer applicants (Detrick & Chibnall, 2008), time banditry in employees (Baskin, McKee & Buckley, 2017), child molester reactions (Mitchell, Keylock, Campbell, Beech & Kogan, 2012), the moral reasoning of students (Hren, Vujaklija, Ivanisevi, Knezevic, Marusic, & Marusic, 2006), those who cheat in educational contexts (Williams, Craig & Paulhus, 2010) and public attitudes towards sex offenders (Olver & Barlow, 2010). The majority of the validation data for the PDS were collected using American and Canadian samples (Tully & Bailey, 2017), and so this raises issues as to whether the normative data can be generalised to UK samples. Next, I will consider the characteristics of the PDS.

Characteristics of the PDS

The PDS is a self-report measure. That is, it is a measure completed by the respondent himself or herself. It is generally used as a companion measure, in that it is used alongside other measures to establish the potential for response biases (Vispoel et

al., 2018). Whilst widely implemented, it is important to objectively examine whether or not the PDS is reliable and valid for use with the various populations it is being used with. Kline (1986) argues that in order for a psychometric test to be considered robust, it should be reliable, valid, discriminate and have appropriate norms for the population it is being used for.

Reliability

Reliability refers to the consistency and stability of a test – that is, how accurate a test is at measuring something rather than the result being due to random error (Hammond, 2002). Internal reliability considers whether each item that makes up the scale is measuring the same concept, and test-retest reliability explores whether the same score is achieved when administered over time to the same set of participants (Hammond, 2002).

Internal reliability is measured using Cronbach's alpha, and a criterion of .7 is usually considered an acceptable level of internal reliability (Field, 2005). The manual provides a table (see Table 5) summarising the Cronbach's Alpha coefficients for the PDS, which was used to demonstrate internal reliability (Paulhus, 1998). As illustrated in Table 5 the PDS scores demonstrated good internal reliability, with Cronbach alpha's of between .70 and .86. Other independent researchers have also found evidence of good internal consistency. For example, Cronbach Alpha's of .71 were found in an independent study of 143 undergraduate students in the Netherlands (Sweldens, Puntoni, Paolacci & Vissers, 2014). A study undertaken in Australia (Gignac, 2013) also established good internal consistency for the SDE and IM scales as well as the PDS as a whole, reporting Cronbach alpha's over .80 (n=466). Finally, a UK study (Tully &

Bailey, 2017) identified high Cronbach alpha's for the SDE (0.90) but much lower findings for IM (0.68; *n*=321).

Table 5*:* Internal Reliability (Cronbach's alpha) for the PDS (Reproduced from the PDS User's Manual; Paulhus, 1998)

Sample	*N*	*SDE*	*IM*	*Total*
General population	441	.75	.84	.85
College students	289	.70	.81	.83
Prison entrants	603	.72	.84	.86
Military recruits	124	.72	.83	.85

Despite the adequate Cronbach alpha's found in the above studies (and others), Vispoel et al (2018) argue that reducing ratings on the PDS from polytomous to dichotomous scores has a negative impact on reliability and validity. They state that dichotomous scores result in strong floor effects and small standard deviations for low and high scores, resulting in skewed distributions (Vispoel et al., 2018). This is problematic because it makes it harder to detect and discriminate between smaller differences (Vispoel et al., 2018). However, when their study undertook analysis on polytomous scores, these issues were addressed. The authors supported the use of G-coefficients over alpha, stating that this alternative is more effective in taking into consideration measurement error (Vispoel et al., 2018). Using polytomous scores, the authors reported that the G-coefficients indicated .80 levels of reliability.

In terms of test-retest reliability, according to Cohen (2009), Pearson's r of .50 or above are considered large, .30 to .49 medium, and below .30 is small. In terms of test-retest reliability, no data could be found in relation to the PDS. However, when Lanyon and Carle (2007) did their comparison between the BIDR and PDS, they reported correlations of .78 (IM) and .69 (SDE), concluding that the SDE scale for the

PDS and BIDR were not acceptable in terms of short-term test-retest reliability. Meanwhile, Egerton, Rees, Bose, Lappin, Stokes, Turkheimer and Reeves (2010) reported that a number of measures of social desirability, including the BIDR showed excellent test-retest reliability for periods between one and two months. However, as highlighted by Mathie and Wakeling (2011), variables being measured by the BIDR (and therefore the PDS too) are likely to vary across time and context, and thus test retest reliability may not be an important area of reliability for the PDS to meet.

Validity

Validity refers to how well a test measures what it is meant to (Hammond, 2002). In relation to the PDS, concurrent and construct (convergent and discriminant) validity are the most relevant aspects of validity to be considered. Before these terms are defined and explored, it is important to note that one area of concern that has been raised, is that a lot of the validation data on the PDS is based on studies undertaken with the BIDR, and whether the two scales are actually that similar. This question was largely addressed when Lanyon and Carle (2007) undertook a comparison of the two tools. The only key area of discrepancy was the large difference in mean scores reported for the PDS and BIDR scales (Lanyon and Carle, 2007). However, they related this to the differing scoring systems of the PDS versus the BIDR (Lanyon & Carle, 2007). As such, in the literature it is generally assumed that findings from the BIDR are also relevant for the PDS.

Concurrent Validity

Concurrent validity refers to how well a test measures the same thing as similar

tests available that explore the same construct (Hammond, 2002) – in this case, how the PDS compares to other psychometrics examining elements of social desirability and impression management.

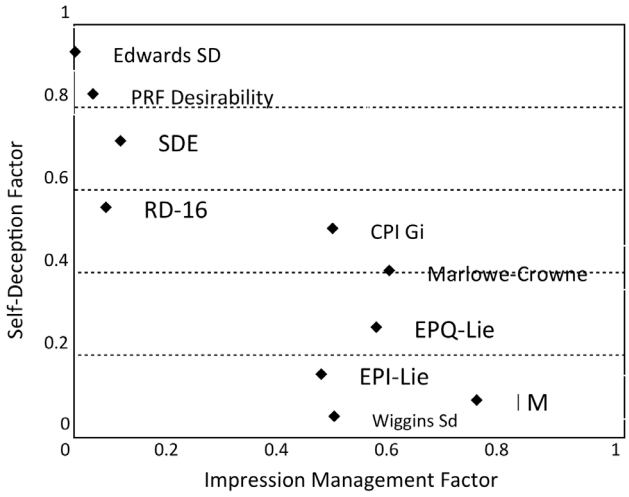
According to the manual, a sample of 320 undergraduates completed ten measures of social desirability, including the PDS. The ten measures/variables were then factored using principal-components extraction (Paulhus, 1998). When varimax rotation was undertaken, the factor loadings appeared which were consistent with the factor analyses previously undertaken on the PDS. Figure 1 below replicates the representation in the manual, and reflects a clustering of scales around the SDE subscale, and a separate clustering around the IM subscale. This highlights the distinctiveness of the two scales that make up the PDS.

The SDE scale has been found to correlate with measures that assess peoples' tendency to confirm or deny problems, such as the ERS scale of Block (1957, cited in Lanyon & Carle, 2007). The SDE scale is also considered to 'detect' narcissistic, arrogant and self-inflated traits (Lanyon & Carle, 2007). The manual (Paulhus, 1998) reports (and as per Figure 1 above) that the SDE shows associations with measures such as Edwards Social Desirability (SD) scale and, IM associates with Wiggins Social Desirability (Sd) Scale and Eysenck's EPI Lie scale. The PDS Total Score also shows high correlations (.73) with the Marlowe-Crowne Scale (Crowne & Marlowe, 1960; Paulhus, 1998) and the Eysenck Personality Inventory (EPI and EPQ) Lie scales (Lanyon & Carle, 2007; Paulhus, 1998). A full review of the factor analyses can be found in Paulhus (1991) paper. Lanyon and Carle (2007) state that whilst the two subscales do differ from one another in terms of divergent validity, there is still quite a lot of overlap between them, which suggests "*a single concept of favorable self-*

presentation is also salient" (p.874). Taken together, this evidence suggests that the PDS demonstrates adequate concurrent validity.

Figure 2

Comparison of Common SDR measures (Reproduced from the PDS User's Manual; Paulhus, 1998)



Construct Validity

Construct validity considers how well items (in uni-dimensional scales) or groupings of items (in multidimensional scales) correlate to each other (Hammond, 2002). In order to demonstrate construct validity, measures need to demonstrate convergent and discriminant validity (Carlson & Herdman, 2019).

Paulhus (1998) argues that the IM scale remained largely the same as the version in the BIDR, and thus previous validation studies were still relevant in determining the convergent validity for the IM scale of the PDS. A review by Thomas, Lanyon and Millsap (2009) found high specificity (.86) and sensitivity (.91) estimates for the IM scale. This means the IM was found to be highly accurate in identifying those engaging

in impression management and those that do not. In terms of convergent validity, the manual reports convergent validity for the IM scale in terms of it being more accurate at detecting shifts in responses depending on situational demands (Paulhus, 1998), with it citing a number of studies where this was demonstrated. According to the author, the PDS Total Score is reportedly highly correlated with the Marlowe-Crowne Scale (Crowne & Marlowe, 1960) as well as the Eysenck Personality Inventory Lie scale (Lanyon & Carle, 2007). The manual also provides convergent validity for the SDE in a number of areas, such as "Hindsight and Overconfidence"; "Overclaiming"; and poor awareness of how a person is viewed by other people (Paulhus, 1998).

With regards to discriminant validity, this refers to a psychometric's ability to identify variables that are theoretically unrelated are in fact, unrelated (Hubley, 2014). The PDS manual reports that this is met through factor analyses, which show that the two scales load as separate factors, further evidenced by low inter-correlations and different patterns in the various separate studies that they undertook, and which are summarised in the manual (Paulhus, 1998). When comparing the BIDR and PDS, Lanyon and Carle (2007) found that discriminant validity was adequate for both SDE and IM scales, but slightly less so for the SDE scale of the PDS. They explain that the stronger IM results reflect the focused and narrow definition of impression management as a concept, compared to the broader and less unified definition of self-deceptive enhancement (Lanyon & Carle, 2007). In addition, G-theory (Generalizability Theory) approaches served to improve convergent and discriminant validity (Vispoel et al., 2018). This is because G-Theory *"is a comprehensive framework for representing reliability of scores in relation to multiple sources of measurement error"* (Vispoel et al., 2018, p.69).

A further way to explore construct validity is by using a criterion comparison group. According to Dufner et al (2019, p.49), "*The most direct and straightforward operationalization of SE is in terms of criterion-discrepancy measures, which compare one's self-views with an objectively assessed external benchmark*." However, there is an absence of criterion group research in the studies validating the PDS (Smith, 2018). One of the difficulties in validating tools such as the PDS is the difficulty of not having a criterion comparison group, due to the challenge of being able to fully ascertain a person's honesty, particularly in forensic assessment (Thomas et al., 2009). This was acknowledged by Tully and Bailey (2017) in their study too – that criterion and predictor ratings were obtained from the same respondents. As a result, often validation studies rely on people simulating dishonest or inflated response styles (Thomas et al., 2009). Thomas et al (2009) proposed utilising Latent Class Analysis (LCA), such as by looking at whether multiple assessments that measure the same/similar thing come up with comparable results, to determine a tools validity when there is a lack of a true criterion comparison group. However, this does not appear to have been done with the PDS.

Structural Validity

Structural validity concerns how well the dimensional constructs of a psychometric are reflected in the scales scores and the overall construct being measured (Mokkink et al., 2010). In the case of the PDS, structural validity has been determined through factor analysis. As mentioned, Lanyon and Carle (2007) undertook a study to explore whether the PDS and BIDR were comparably valid, and whether the validity data could be applied to forensic contexts. They collected data on 519 participants,

consisting of three groups of forensic participants, who varied between criminal justice cases with a range of offenders and patients with personal injury related concerns (Lanyon & Carle, 2007). The other two groups were made up of undergraduate students who had to complete the scales as part of a requirement for their course (Lanyon & Carle, 2007). Participants were also required to complete a range of other measures as a means to establishing concurrent validity of the PDS and BIDR (Lanyon & Carle, 2007). The study concluded that whilst the SDE and IM scales were found to be more strongly positively related for the PDS when used with a forensic sample, the 2-factor model fit was not confirmed for the BIDR or PDS (Lanyon & Carle, 2007). This was not surprising given similar findings in other studies (Uziel, 2010; Vispoel et al., 2018).

Some studies have concluded that the factor analyses undertaken on the PDS could only demonstrate adequate model fit when all items of the PDS were grouped together, thus meaning that the IM and SDE as separate constructs, do not achieve discriminant validity (Tully & Bailey, 2017). Tully and Bailey (2017) argued that the factor analysis approach used to validate the PDS was somewhat restrictive, and may explain the problems in validating the constructs that make it up. When an alternative approach - Exploratory Structural Equation Modeling (ESEM) was utilised, the two scales of the PDS were considered distinct and valid, reporting a Root Mean Square Error of Approximation (RMSEA) of 0.026 (90% CI's=0.019-0.032) (Tully & Bailey, 2017). Therefore, taken together, there seems to be evidence that structural validity is evident for the PDS, but the extent is dependent on the factor analytic approach taken, with more recent studies favouring the ESEM approach.

Normative Data on the PDS

The PDS manual provides normative data for the general population, college students, prison entrants and military recruits, drawn from an American and Canadian sample (Paulhus, 1998). Whilst much of the validation data of the PDS has been on college students, and some prison inmates or military, the PDS is used extensively in forensic and occupational contexts (Lanyon & Carle, 2007). In an analysis of the uses of the PDS, Vispoel and colleagues (2018) found that *"35.4% were used as contamination checks, 25% for statistical controls, 20.8% as dependent variables, 16.7% as flags for invalid responses, 16.7% as explanatory variable, 8.3% as part of construct validation, and 2.1% as indexes of situational effects on SDR"* (p.71). The variety of contexts in which the PDS is used raises the question of whether the normative data in the manual is adequate to generalise to other populations, most specifically with the range of offenders and contexts in which they are currently being assessed in the UK.

As noted previously, another factor is that a large number of published research utilised Paulhus' earlier tool – the BDIR, rather than the PDS. In addition, there are very few studies on the PDS in the UK. One of the few studies providing normative data for the PDS on a general population UK sample was that undertaken by Tully and Bailey (2017) in which they assessed 358 people from the general population. They found that the UK sample were statistically in line with American samples for SDE, but scored significantly higher on the IM scale and Total Score (Tully & Bailey, 2017). What was notable in the Tully and Bailey (2017) study though, was the disproportionally high female sample (80.7%), which may have an implication on generalisation of their validation data. Vispoel et al's (2018) study drew on a student sample in America, which also consisted of majority females (74.2%) and Caucasians (89.2%). This

suggests that it may be worthwhile exploring if there are any differences between how males and females score on the PDS, and if different cut-offs may be required.

No normative data was found for the use of the PDS with people who have neurodevelopmental difficulties, such as learning disability or autism. However, Langdon, Clare and Murphy (2010) modified the Seikheim and Gur SDQ and ODQ for use with people with intellectual disabilities. Their study used a very small sample (32 men with intellectual disability and 28 without intellectual disability). However, they found some preliminary evidence that people with intellectual disability showed a tendency to score more highly on the measures of self and other deception (Langdon et al., 2010). This suggests that it might be important to identify PDS normative data for people who have neurodevelopmental difficulties.

Conclusion

One of the biggest challenges in reviewing the PDS has been determining which versions of the PDS/BIDR were used in what studies, particularly the older studies, and how the various versions correlate. Validation data seems to have been combined for each version with the next, possibly disguising some of the flaws of the tool. However, Lanyon and Carle's study (2007) provides some reassurance that the last version of the BIDR (version 6) and the PDS are comparable.

On the whole, the PDS has demonstrated reliability in a number of countries, with military, forensic and college/undergraduate samples. In addition, there seems to be good evidence for concurrent and construct validity, with the exception of criterion validity. As noted by Thomas et al (2009, p. 228), "*when no external criterion exists by which to classify participants in the study, there is no absolute way to evaluate the*

accuracy of this technology here". However, the debate appears mainly to be around the PDS's structural validity. This is because the reliability and validity of the PDS, based on the reviews of the literature, very much depended on the way in which the factors were statistically explored. When alternative approaches to factor analysis were considered, such as ESEM, the IM and SDE scales were found to be more distinct.

Normative data exists, but like with many other tools, and more specifically with the PDS given the impact the outcome of the results can have for an individual, adequate selection of norms needs to be specific rather than generalised. Using ESEM, Tully and Bailey (2017) found that the PDS was a valid tool to utilise in the UK, using UK based normative data. However, their study had a disproportionate number of females, and this has implications for comparing data to males. No studies could be identified that had used the PDS with people who have neurodevelopmental difficulties, such as learning disability or autism. This could be problematic given possible prevalence rates of people with ASD in UK prisons and secure hospitals falls between 7% and 15% (Hare et al., 1999; NAS, 2020; Scragg & Shah, 1994; Siponmaa et al., 2001; Talbot, 2009).

Clinicians' ability to assess for response bias is very important, particularly in forensic contexts. A tool such as the PDS may be helpful in that regard, and there is adequate evidence for its use. However, as suggested by Paulhus (1998), it should be used to inform rather than determine response bias. Understanding and exploring the function and reasoning for elevated scores on the PDS should be a key part of the forensic assessment, rather than simply reporting deception tendency or likelihood. Furthermore, there is clear scope for increasing normative data for the use of the PDS in the UK, particularly in relation to males, offenders, and people with mental health or

developmental disabilities. This type of research will help gain a greater understanding of the role deception plays, conscious or not, thus informing a holistic understanding of a person's response style.

CHAPTER 4

UNDERSTANDING SELF-REPORT RESPONSE BIAS IN HIGH-FUNCTIONING AUTISM

Abstract

The current study aimed to establish a normative data set for the PDS and SIMS in a community adult sample of people with High Functioning Autism (HFA). Assessments were administered via an online platform and were anonymous. 70 surveys were completed, with respondents contributing from 16 countries, including the UK, USA, a number of European Countries, as well as South America and Australia. The majority of subscales and total scores for the PDS and SIMS fell above cut-off for self-report response bias. In addition, some relationships were evident between high scores and education level as well as psychological distress. This study provides some early evidence that alternative cut-off scores could be used with this cohort, as part of a wider holistic assessment of response style and bias.

Terms: response bias, impression management, socially desirable responding, malingering, autism

Introduction

In prisons, hospitals, probation services and medico-legal assessment contexts across the UK, psychometrics are routinely utilised. These can be used to explore and evaluate treatment efficacy as well as to determine an individual's risk factors at various points of their offence pathway through the court, detention and community (Wakeling & Barnett, 2014). When faced with offenders or victims who have neurodevelopmental difficulties, psychologists will typically be asked by the court to assess their specific deficits and establish how those relate to their abilities in response to specific legal questions (Salekin, Olley & Hedge, 2010), such as their understanding of the trial process, their ability to instruct solicitors or take part in a trial.

In forensic psychology, we often consider risk factors in terms of static (those that are generally stable) and those that are dynamic (variable/changeable), recognising that multiple variables can impact on dynamic risk (Wakeling & Barnett, 2014). Many psychometrics used in forensic psychology to evaluate individual risk factors, beliefs, personality and other factors, rely on self-report measures, which naturally introduces the risk of response bias. Many of the psychometrics have been criticised for being so transparent that people can easily lie, fake, present themselves in a biased way and introduce further bias because of variations or limitations in insight, shame and sense of guilt (Wakeling & Barnett, 2014). One way this has been addressed is by utilising questionnaires that explore typical response biases, such as socially desirable responding (SDR), impression management (IM), and malingering.

In terms of self-report measures of response bias, these tend to fall into two main categories: those that explore malingering, that being faking/over-reporting of the existence of psychopathology (Furnham, 1986), and those that explore socially

desirable responding, whereby the person tries to portray themselves in a more favourable light and manage impressions others may form of them (Rogers, 2018a). As such, various tools have been developed that focus on these two key areas. Popular tools exploring self-reported malingering include the Structured Inventory of Malingered Symptomology (SIMS: Widows & Smith, 2005), which is a standalone measure. Measures that explore deception and socially desirable response styles by including validity scales within the full assessment include: the Minnesota Multiphasic Personality Inventory (MMPI: Butcher, Graham, Ben-Porath, Tellegen & Dahlstrom, 2001), the Millon Clinical Multiaxial Inventory – Version 4 (MCMI-IV: Millon, Grossman & Millon, 2015) and the Multiphasic Sex Inventory (MSI: Nichols & Molinder, 1984). In contrast, the Supernormality Scale (SS) and a very low score on the Disclosure (x) scale of the MCMI assess the under-reporting of symptoms. The Paulhus Deception Scales (PDS: Paulhus, 1998) is an example of a standalone measure of socially desirable responding and impression management (Paulhus, 1998).

Response bias is recognised as situational and context driven, falling on a continuum (Hart, 1995; Rogers et al., 2010, Tan & Grace, 2008; Young, 2017). People may over or under-report abilities, traits or symptoms, to gain something, such as parole, to mitigate risk or to obtain a job (Archer et al., 2016; Cassano and Grattagliano, 2019; Leite, 2015). Thus, understanding motivation and incentives that may contribute to a person displaying response bias, is important to consider (Ohlsson & Ireland, 2011). At the same time, factors such as IQ (Jobson et al., 2013), personality (Hart, 1995; Impelen et al., 2017; Paulhus, 1998; Tully & Bailey, 2017) and age (Mathie & Wakeling, 2011) have been found to play a contributory role, and are important to understand.

Despite response bias measures, that is, measures of SDR or conversely, malingering, being around for many decades now, they are not without their shortcomings. The various self-report response bias questionnaires that are frequently used by clinicians as part of their assessments have not been developed with specialist populations in mind. Whilst the British Psychological Society (BPS) does acknowledge the need to consider the impact, for instance, of neurodevelopmental difficulties in relation to response bias, effort and malingering (BPS, 2009), the guidance provides little information on how to do this. Some conditions, such as learning difficulties and social communication disorders, may render people particularly susceptible to produce raised scores on assessments of malingering and effort (Lerner et al., 2012), or conversely come across in a socially desirable way (Langden, Clare & Murphy, 2010), and so this remains an important area to address.

Therefore, it is important to develop normative data, bearing in mind different contexts and different populations, so that it is possible to contextualise an individual's scores with comparable population norms. This may be particularly relevant for those with neurodevelopmental disorders who may inadvertently produce distorted responses on these measures, due to factors such as cognitive inflexibility, suggestibility, and social communication deficits, leading assessors to conclude that they are trying to 'fake good' or 'fake bad', when this is not actually the case.

As outlined in previous chapters, varying base rates and differences in cut-off scores for different populations, settings, contexts and even clinicians, create problems with the meaningfulness of results from such measures (Drob et al., 2009). Attempting to generalise results from purely clinical contexts to forensic contexts can also have a seriously adverse effect on the quality and accuracy of such assessment (Cassano &

Grattagliano, 2019). A significant factor to consider is that response bias may look different in forensic groups as opposed to non-forensic populations (Tan and Grace, 2008). This picture may be further complicated when considered in the context of those who have pre-existing conditions that may render them more likely to appear as if they are malingering or engaging in SDR when they are not, for example, those with Learning Disabilities and Autism.

Autism and the Criminal Justice System

Autistic Spectrum Disorder (ASD) is a lifelong condition and refers to a spectrum or dyad of neurodevelopmental difficulties that can vary considerably between people, but encompasses impairments with social communication, repetitive behaviours, cognitive flexibility/rigidity, restrictive/specialist interests, and problems understanding and making sense of other people's feelings, thoughts and behaviours (Ali, 2018; Attwood, 2015; Baron-Cohen, 2008; National Autistic Society [NAS], 2019; Wing, 1997). Alongside these, they can also have difficulties with executive function, language, sensory sensitivities, and problems with emotion recognition and regulation (Lai, Lombardo & Baron-Cohen, 2014). A type of Autism, often described as High Functioning Autism (HFA) or Asperger's, refers to those who are of average or above average intelligence, and may also have better language skills than others with Autism, but may have certain learning difficulties or problems with processing and making sense of communication and social information (NAS, 2019). This can mean that their difficulties may not be as obvious, particularly when considering those who find themselves in trouble with the law (Browning & Caulfield, 2011).

In terms of prevalence, according to the NAS (NAS, 2020a), approximately 1.1% of the UK population has a diagnosis of ASD. However, many go undiagnosed. Lai & Baron-Cohen (2016) highlighted that many females with ASD may also have been missed due to variations in how they present. Only with improvements in diagnostic approaches over the last five years have we begun to see an increase in the prevalence of ASD amongst the female population (NAS, 2020b). However, there has been very little research on the prevalence of ASD in criminal justice contexts, which means that many may not have been identified (Ali, 2018; Underwood et al., 2013). Population differences do seem to exist based on the small pockets of research conducted. For instance, Scragg and Shah (1994) found between 1.5% and 2.3% of inpatients in a high secure hospital in the UK had ASD, and Siponmaa et al (2001) found a prevalence rate of 3%, but a further 15% were identified as having a pervasive developmental disorder. A further study of all UK high secure hospitals conducted by Hare et al (1999) found a prevalence rate of 1.6%, and an additional finding that those with a diagnosis of ASD tended to be detained over 10 years longer that those with other diagnoses (Hare et al., 1999). This suggests that prevalence rates of those with Autism in forensic contexts may be higher than what appears on the surface. As a result, it becomes important to have appropriate normative data for response bias measures that can be used in forensic assessments of Autistic offenders across community, secure hospital and prison contexts.

In relation to response bias specifically, some conditions (such as learning difficulties and social communication disorders, such as Autism) may render people particularly susceptible to inadvertently produce raised scores on assessments of malingering and may also be seen as displaying less effort (Lerner et al., 2012). This is

because people with Autism are often described as being 'black and white' in their thinking (and so may endorse extreme ends of a likert scale when rating questionnaires) and can find it hard to perspective take, possibly causing difficulty when endorsing scenario based questions (Mazefsky & White, 2014), both of which could potentially impact on how they respond to questionnaires.

People with ASD also show a greater tendency to be compliant and/or suggestible (Chandler, Russell & Maras, 2018; Lerner et al., 2012; O'Mahony, 2012) – a further factor than can impact on their response styles when completing questionnaires. Gudjonsson and Clark (1986, p.84) define suggestibility as "*The extent to which, within a closed social interaction, people come to accept messages communicated during formal questioning, as the result of which their subsequent behavioural response is affected*. As Gudjonsson (2003) goes on to explain, the definition considers that, in the context of an assessment or interview, five factors are playing a role: *"A social interaction; a questioning procedure; a suggestive stimulus; acceptance of the stimulus; and a behavioural response*" (p.345). This demonstrates how many stages can be relevant and impact on a person with ASD being suggestible in forensic contexts. For instance, by asking if they experience something (e.g. felt criticised by other people), they may endorse this as if they had, even if they had not thought that prior to being faced with the question. They may also over endorse symptoms (Lerner et al., 2012). This is because how suggestible a person is, very much relates to situational demand characteristics (Gudjonsson, 2003). That is, if an ASD individual is suspected of malingering and is informed during assessment that they are being asked about their symptoms of psychosis (using the Symptom Inventory of Malingered Symptomology:

SIMS) they may feel that they should be reporting symptoms, and therefore endorse psychosis related items, due to a tendency to be suggestible.

NHS Guidance (2019) aims to ensure that those with ASD can be held responsible for criminal behaviours whilst still ensuring fair and just treatment by the Criminal Justice System (CJS), in line with the Equality Act (2010) and the NHS Long-term Plan (2019). Advocates of ASD emphasise the importance of the courts having a better understanding of ASD so that a person's journey through the CJS, from arrest to sentencing, takes into account the person's particular needs, vulnerabilities and difficulties (Ali, 2018; Mouridsen, 2012). This is particularly key when considering sentencing, risk management and treatment options (Ali, 2018; Freckleton, 2013; Murphy, 2010) and when ensuring that approaches are consistent and fair (Browning & Caulfield, 2011). On that basis, it is crucial that clinicians can approach assessments with relative confidence, and that the information they gain is accurate and reliable. Thus, it is crucial that when examining honesty and accuracy of self-report information, this is done in line with good practice and includes the use of valid and reliable assessment measures that have appropriate normative data.

Aims of the Current Research

The current study aims to address some of the gaps in the understanding of self-report response bias amongst an Autistic client group. A key focus of the research is to obtain normative data on the SIMS and PDS and explore any differences that emerge between the scores for the general population in the manuals and the scores of people with High Functioning Autism (HFA).

On this basis, it is hypothesised that people with HFA will endorse items differently to the general population on measures that assess malingering (such as the SIMS) and those that assess SDR (such as the PDS), in a context where there is no motivation to display response bias. In addition, it would be useful to explore how levels of psychological distress (measured using the CORE:OM) in people with HFA may relate to how they then score on a relevant subscale (Affective disorders) of the SIMS bearing in mind their tendency to be suggestible. Therefore, the research is exploratory in nature.

Method

Sample

The final study sample consisted of 70 adults who reported having a diagnosis of HFA. 79 respondents consented to complete the survey. Five were removed from the data set as they only completed the demographics questions and did not complete any of the questionnaires. One respondent was excluded as they disclosed a diagnosis of Learning Disability, and so did not meet the criterion for High Functioning Autism. Three respondents only completed the first part of the survey and the CORE:OM, omitting the PDS and SIMS. Data were collected anonymously via an on-line survey from a community sample of males and females aged 18 or over who confirmed that they had an established diagnosis of High Functioning Autism (HFA). Confirmation of a diagnosis of Autism was established by asking respondents to confirm this when they emailed for the survey link, and again when consenting to participate upon entering the online survey. A section was added to the survey for them to disclose any other diagnoses they had. This enabled a decision to be made as to whether these other

diagnoses would have had a marked impact on their responses and required their data to be excluded from the analysis.

The demographics of the sample are summarised in Appendix K. The mean age of the sample was 34.01 (range = 18-69; SD = 13.19). There were slightly more males (53.4%) than females (45.1%) in the sample, and one respondent (1.4%) identified themselves as 'other'. The majority were employed (43.8%) and had an undergraduate or postgraduate qualification (54.8%). Respondents came from a variety of different countries, with the majority residing in the UK (46.6%) followed by the USA (26%). In relation to diagnoses other than HFA, the majority (57.5%) did not have other diagnoses, followed by 21.9% having a Depressive Disorder, Anxiety Disorder (6.8%), ADHD (6.8%), Personality Disorder (2.7%), Eating Disorder (2.7%) and a Trauma related disorder (1.4%). Although there is high comorbidity of mental health difficulties in people with Autism, particularly low mood and anxiety, not all would have sought a formal diagnosis or had symptoms severe enough to meet diagnostic criteria. This may account for the high percentage that did not disclose any other diagnoses in the current community sample.

Procedure/data collection

The study collected data anonymously from respondents in 16 countries, including the UK, USA, a number of European Countries, as well as South America and Australia, between 25th February 2020 and 13th April 2020. The decision was made to collect data beyond the UK, for two reasons. Firstly, this study was the first to explore self-report response bias in people with HFA. As a result, it was important to establish a general sense of how people with HFA approach such measures, with the secondary aim

of narrowing the focus to look at specific groups, whether they are country specific, by gender/sex, or any other specific variable. Secondly, expanding the study beyond the UK would generate a larger sample size. . Social media platforms were intentionally targeted to try and maximize sample sizes and these platforms are not country-specific (though limited to English-speaking). Additionally, at the planning stage of the research it was unclear what the response rate would be like. It was hoped that the sample size would be large enough to split respondents by country to explore any normative differences.

It is important to note that the data were collected during the COVID-19 pandemic, though this was unintentional. Arrangements had been made to visit certain Autism support groups in the UK in order to inform them about the research and allay any concerns potential respondents had, but due to COVID-19 lockdown, this could not go ahead.

Respondents were approached via social media platforms (such as Twitter, LinkedIn, Facebook, Reddit) and Autism support groups (via organisations such as the National Autistic Society; Scottish Autism and Project Aspie) and asked to express an interest in taking part. An incentive to participate was offered whereby respondents could opt into a prize draw for a gift voucher once they had completed the study. The advert directed them to email the principal researcher, at which point they were given a link and password to enter the online survey. This was a requirement of the test publishers as the SIMS and PDS are restricted tests.

Once respondents arrived at the website, they were asked some preliminary questions to confirm they were over 18 and had a diagnosis of HFA. Once confirmed, they were directed to the next page where they were provided with an information sheet

giving details of the study and asking for consent to participate. Once consented, the survey opened.

As soon as respondents completed the questionnaires, they were asked to create a unique identifying code. They could use this code, rather than any personally identifiable information, to withdraw their data from the study, should they subsequently wish to do so. Respondents were given the option to withdraw from the research, but opportunity to withdraw was limited to a specific date, at which point their data were merged with all the other respondents' data, and so it was not possible to remove their specific responses. None of the respondents requested for their data to be withdrawn.

The survey consisted of collecting demographic data that is outlined above as well as the completion of three questionnaires: CORE:OM, PDS and SIMS, details of which are summarised below. The CORE:OM was selected to gauge the level of respondents' psychological distress as a means to determine whether the SIMS Affective subscale was able to distinguish genuine from malingered mental health symptomology, particularly given people with HFA are prone to being suggestible.

Measures

CORE-OM

The Clinical Outcomes in Routine Evaluation - Outcome Measure (CORE-OM; Evans et al., 2000) is a 34-item self-report questionnaire of psychological distress. The measure explores subjective well-being (4 items), commonly experienced problems or symptoms (12 items), and life/social functioning (12 items). The measure includes six items on risk to self and to others. For the purposes of this study, the risk items were

excluded. It has a five-level response choice, ranging from *"not at all"* to *"most or all of the time"*. The time frame rated is the preceding seven days. Normative data is available for clinical and non-clinical populations as well as males and females separately.

The CORE:OM has been used extensively with a range of populations. Evans et al (2002) undertook an extensive evaluation of the psychometric properties of the CORE:OM on a UK sample. According to the CORE:OM user manual (CORE, 1998), the non-clinical norms were based on 1106 convenience sample and a clinical population of 890 people receiving a range of psychological interventions across the UK.

Paulhus Deception Scales (PDS)

The PDS (Paulhus, 1998) is a 40-item self-report questionnaire consisting of two scales – the Self-Deceptive Enhancement Scale (SDE) and the Impression Management Scale (IM). The author states, "*the PDS captures the two principal forms of socially desirable responding with two (relatively) independent subscales*" (Paulhus, 1998 p. 1). A critique of the PDS outlined in Chapter 3 concluded that the PDS has demonstrated adequate reliability and validity. The systematic review outlined in Chapter 2 identified that the PDS is one of the most widely used socially desirable response bias measures in the UK.

Structured Inventory of Malingered Symptomology (SIMS)

The SIMS (Widows & Smith, 2005) consists of 75 items that are rated as true or false. It screens for malingered psychopathology across five scales: Psychosis, Low

Intelligence, Neurologic Impairment, Affective Disorders, and Amnestic Disorders, as well as providing an overall score of malingering. Impelen et al (2014) undertook a meta-analysis of the SIMS. Their main findings were that the SIMS could distinguish between honest and instructed feigners, could detect elevated scores in groups predicted to be high scorers, showed false positives in those who had diagnoses of schizophrenia, cognitive deficits and psychogenic seizures, and cut-off scores showed adequate sensitivity but poor specificity when the manualised cut-off scores were applied.

Ethical considerations

Ethics approval for the current study was received from the University of Birmingham Research Ethics Committee (Ref: ERN_19-1770) on 25.02.2020. Respondents consented to participation after reviewing a detailed information sheet. In terms of confidentiality, all surveys were anonymous. As soon as interested participants emailed for the link and password to the online platform, and the details were sent back to them, their emails were then immediately deleted. On completion of the questionnaires, respondents were asked to generate a code that they could then use should they decide to withdraw from the study. No personally identifiable information was collected throughout data collection and data analysis. At various points in the survey respondents were signposted to various organisations should they require emotional support due to the questions being asked.

Data analysis

All data were analysed using SPSS version 25. As several of the subscales deviated from normality it was not possible to use parametric tests (with asymptomatic

significance) of association and difference. Instead, bootstrap tests of association and difference were employed (as these do not rely upon distributional assumptions) and significance testing was undertaken using bootstrap probability estimates, standard errors and confidence intervals. The bootstrap confidence intervals and significance test have been shown to be robust for small samples, do not depend on normal distribution assumptions and are appropriate for use with this data. When comparing groups on multiple occasions there is the risk of finding differences by chance, and thus a false positive result (Ranganathan, Pramesh & Buyse, 2016). To control for this, analyses were undertaken using the Bonferroni correction. The Bonferroni correction adjusts the alpha value by the number of comparisons made (Ranganathan et al., 2016).

Results

Distribution of CORE, PDS and SIMS total scores and subscales

Table 6 below provides descriptive statistics and tests of normality of distribution for each of the CORE:OM, PDS and SIMS subscales. CORE:OM mean scores for all the subscales and total score, fell in the 'clinical range' when compared to the norms provided in the CORE:OM manual. Normative data in the manual for the CORE:OM corresponded more with the 5th and 25th percentile scores found in the current study.

Table 6: Mean and percentile scores for CORE:OM, PDS and SIMS

Scale	*Valid N*	*Manual means*	*Mean*	*Median*	*SD*	*Percentile*					*One-Sample Kolmogorov-Smirnov Test*	
						5th	*25th*	*75th*	*95th*	*99th*	*Test Statistic*	*Asymp. Sig. (2-tailed)*
CORE Wellbeing	73	*3.64*	8	8	4	1	4	11	14	15	0.102	0.06
CORE Psych distress	73	*10.80*	24	23	10	8	18	33	39	42	0.085	0.20
CORE Feelings	73	*10.20*	22	23	8	8	16	28	37	39	0.066	0.20
CORE total	73	*24.64*[1]	54	55	20	21	42	70	86	91	0.061	0.20
PDS IM (t-score)	70	*50*	62	62	9	44	57	67	78	80	0.112	0.03*
PDS SDE (t-score)	70	*50*	53	49	10	42	46	57	72	76	0.22	0.01*
PDS total (t-score)	70	*50*[2]	64	62	12	46	54	70	86	97	0.127	0.01*
SIMS psychosis	70	*0.82*	3	1	4	0	0	5	12	13	0.279	0.01*
SIMS Neurologic Impairment	70	*1*	5	3	4	0	2	7	13	15	0.191	0.02*
SIMS Amnestic	70	*1.15*	5	4	4	0	1	8	12	15	0.17	0.03*
SIMS low intelligence	70	*1.42*	2	2	2	0	1	4	6	7	0.214	0.04*
SIMS affective	70	*3.27*	6	7	2	3	4	8	11	12	0.101	0.08
SIMS Total	70	*7.67*[3]	21	18	13	6	11	28	53	56	0.146	0.01*

[1] *Based on 1084 non-clinical population (CORE System Group, 1998)*
[2] *Based on 441 general population (Paulhus, 1998)*
[3] *Based on 238 honest responders (Widows & Smith, 2005)*
** Significant for p <.05*

One of the primary aims in the study was to generate normative data for the SIMS and PDS for a HFA population. The above scores were therefore compared with the norms provided in each manual respectively.

The normative data reported in the SIMS manual (Widows & Smith, 2005) consisted of 476 undergraduate students from four universities in the USA, who participated in an "analogue simulation study" (p.12). Ages ranged from 17 to 66 (mean age 24.43 years), and comprised of predominantly females (71%). Compared to the normative data reported in the SIMS manual, the mean score of the respondents in this study fell above cut-off on all subscales and total scores of the SIMS, which indicates that the HFA respondents as a whole would be considered to be malingering their symptomology in relation to psychosis, neurologic impairment, amnestic disorders and affective disorders. An exception was found with the Low Intelligence sub-scale, with the mean score attained being 2, which fell on the cut-off (>2).

The normative data in the PDS manual (Paulhus, 1998) consisted of 441 American and Canadian urban and rural respondents between the ages of 21 and 75. Compared to the normative data in the PDS manual, mean scores for the PDS SDE subscale and Total PDS scores for the current sample fell in the 'average' and 'above average' range respectively. This indicates that people with HFA are no more likely that the general population to demonstrate SDE, as their IM scores most likely impacted on the Total score 'elevation'. In relation to the PDS IM score, the current samples mean scores fell in the 'above average' range. This indicates that people with HFA are more likely to display impression management, when compared to the general population. In terms of the invalidity cut-off score for the IM scale, the mean of 10.86 (*SD = 3,432)*

was identified according to the manual as “may be faking good”, suggesting that people with HFA as a whole show a tendency towards presenting themselves in a positive light.

Individual Difference Factors

Age

Correlations between age and the CORE:OM, PDS and SIMS subscale scores are presented in Table 7.

Table 7: Pearson r and bootstrap correlation coefficients

Scale	*Age*	*Bootstrap (with 1000 repetitions)*		
	Pearson r	*Bias*	*Lower 95% CI*	*Upper 95% CI*
CORE Wellbeing	-0.061	0.001	-0.284	0.168
CORE Psych distress	0.023	0.000	-0.228	0.263
CORE Feelings	-0.078	-0.001	-0.327	0.168
CORE Total	-0.033	0.001	-0.287	0.207
PDS IM (T-score)	0.002	-0.002	-0.217	0.251
PDS SDE (T-score)	-0.040	-0.005	-0.272	0.194
PDS Total (T-score)	-0.034	-0.002	-0.267	0.219
SIMS Psychosis	-0.304*	0.000	-0.454	-0.155
SIMS Neurologic Impairment	-0.084	0.002	-0.286	0.143
SIMS Amnestic	-0.028	0.000	-0.245	0.196
SIMS Low intelligence	-0.153	0.004	-0.372	0.104
SIMS Affective	-0.121	0.000	-0.376	0.135
SIMS Total	-0.167	0.003	-0.361	0.058

** Significant for p<.01*

Correlations were undertaken to explore any differences between age and scores on the CORE:OM, PDS and SIMS. Only the SIMS Psychosis subscale evidenced a significant correlation with age ($r=-0.304$, $p<.01$), indicating that scores on this subscale go down as age increases. Age differences are not provided in any of the measure manuals, but the finding that symptoms of psychosis reduce with age is in line with other studies, as previously discussed, and where the SIMS produced false positives with those who had a diagnosis of schizophrenia (Impelen et al., 2014).

Gender differences

In order to explore differences in gender, a series of bootstrapped t-tests between males and females were undertaken (see Table 8).

Table 8: Bootstrap t-tests between male and female respondents

	Sex						*Bootstrap*				
	Male			*Female*						*95% Confidence Interval*	
	Mean	*Standard Deviation*	*Valid N*	*Mean*	*Standard Deviation*	*Valid N*	*t*	*Bias*	*Std. Error*	*Sig. (2-tailed)*	*Lower*	*Upper*
CORE Wellbeing	7.54	4.12	39	7.82	3.82	33	-0.418	0.008	0.958	0.664	-2.334	1.383
CORE Psych Distress	24.46	10.33	39	23.7	8.96	33	0.271	0.007	2.337	0.787	-4.35	4.982
CORE Feelings	23.82	8.23	39	21.03	8.39	33	1.403	0.005	2.007	0.165	-1.517	6.619
CORE total	55.82	21.04	39	52.55	18.9	33	0.622	0.019	4.821	0.535	-7.082	12.223
PDS IM (T-score)	60.46	8.11	37	63.09	8.94	32	-1.284	-0.149	2.012	0.195	-6.693	1.259
PDS SDE (T-score)	52.3	8.66	37	53.16	10.85	32	-0.365	-0.033	2.421		-5.754	3.822
PDS total (T-score)	62.49	12.05	37	65.37	12.38	32	-0.980	-0.168	2.901	0.33	-8.783	2.441
SIMS Psychosis	4.14	4.33	37	1.84	2.78	32	2.566	0.001	0.853	0.014**	0.587	3.985
SIMS Neurologic Impairment	5.92	4.4	37	3.25	2.58	32	3.012	-0.001	0.853	0.007***	0.958	4.22
SIMS Amnestic	5.62	4.64	37	3.94	3.55	32	1.673	-0.001	0.944	0.082	-0.089	3.633
SIMS Low Intelligence	2.46	2.21	37	1.69	1.67	32	1.617	-0.003	0.466	0.106	-0.136	1.691
SIMS Affective	6.51	2.64	37	6.31	2.25	32	0.337	0.003	0.58	0.737	-0.893	1.358
SIMS Total	24.65	15.33	37	17.03	9.16	32	2.455	-0.001	2.965	0.02*	1.773	13.5

** Significant for p<.05*
*** Significant for p<.01*
**** Significant for p<.001*

The only significant differences found between males and females were on the Psychosis (*p<.01*) and Neurological Impairment (*p<.001*) subscales of the SIMS and the Total score (*p<.05*) of the SIMS. Differences in scores across males and females are not reported in the PDS and SIMS manuals. In relation to the CORE:OM, the manual reports some small significant differences between the non-clinical male and female sample, with females scoring higher than males. However, differences were not evident between males and females in their clinical sample, with exception of the Wellbeing subscale (p <.001).

Differences in country of origin

Only the UK (n=34) and the USA (n=19) had more than three valid responses. Accordingly, it was not possible to assess differences between respondents from other countries with less than three valid responses. Table 9 summarises differences between UK and USA samples on the CORE:OM, PDS and SIMS subscale and total scores.

Table 9: Bootstrap t-test of differences between the UK and the USA samples

Scale	UK			USA			Bootstrap based on 1000 repetitions						
	Mean	SD	N	Mean	SD	N	Mean Diff	t	Bias	Std. Error	Sig. (2-tailed)	Lower 95% CI	Upper 95% CI
CORE: Wellbeing	8.41	3.47	34	5.53	3.78	19	2.716	2.677	-0.003	1.063	0.007*	0.531	4.606
CORE Psych distress	26.24	9.65	34	20.47	9.83	19	5.526	1.969	-0.001	2.792	0.05*	-0.077	10.755
CORE Feelings	24.41	7.65	34	19.74	8.42	19	4.233	1.902	0.067	2.285	0.077	-0.415	8.696
CORE Total	59.06	18.59	34	45.74	19.84	19	12.475	2.303	0.064	5.542	0.024*	1.428	22.914
PDS IM T-score	59.91	7.82	33	66.05	8.87	19	-6.144	-2.597	-0.063	2.431	0.018*	-10.914	-1.308
PDS SDE T-score	50.24	7.42	33	54.32	10.41	19	-4.073	-1.641	-0.024	2.746	0.135	-9.904	1.025
PDS Total T-score	60.27	8.84	33	69.53	13.69	19	-9.254	-2.964	-0.061	3.512	0.017*	-15.965	-1.873
SIMS Psychosis	3.55	4.19	33	3.63	4.37	19	-0.086	-0.07	-0.015	1.23	0.005*	-2.504	2.199
SIMS Neurologic Impairment	4.88	4.34	33	5.32	4.18	19	-0.437	-0.355	-0.023	1.206	0.944	-2.802	1.897
SIMS Amnestic	5.24	4.47	33	6.21	4.44	19	-0.968	-0.753	-0.013	1.285	0.724	-3.61	1.564
SIMS Low Intelligence	3	2.18	33	1.26	1.37	19	1.737	3.13	0.006	0.498	0.003*	0.834	2.683
SIMS Affective	7.03	2.42	33	5.58	2.78	19	1.451	1.975	-0.017	0.766	0.054	-0.12	2.988
SIMS Total	23.7	14.37	33	22	14.66	19	1.697	0.407	-0.06	4.186	0.708	-6.545	9.757

** Significant for p<.05*

Seven of the tests measures (CORE Wellbeing, CORE Psych distress, CORE total, PDS IM, PDS total, SIMS Psychosis, and SIMS Low Intelligence) evidenced significant differences between the UK and the USA respondents. In relation to the CORE:OM Wellbeing and Distress subscales and CORE:OM Total score, UK mean scores were significantly higher than USA mean scores ($p<.05$), indicating higher levels of psychological distress and lower levels of wellbeing. The PDS IM mean subscale score was lower in the UK compared to the USA ($p<.05$), with a similar finding on the PDS Total score ($p<.01$), indicative of the UK sample being less likely to impression manage compared to USA samples, with USA scores falling in the "above average"

range, according to the PDS manual. The SIMS Psychosis mean subscale score was also higher in the USA compared to the UK ($p<.01$), indicating that USA respondents were considered to report more malingered psychotic symptoms that UK respondents, though according to the SIMS manual, both groups still scored above cut-off for malingering psychotic symptoms. In contrast, the Low Intelligence mean subscale scores were higher in the UK compared to the USA ($p<.01$), suggesting that UK respondents reported more malingered symptoms associated with cognitive deficits and also scored above cut-off, indicative of malingering of learning problems. A t-test was undertaken to explore whether age was a factor in the differences between scores between the USA and UK, but no significant differences were found. In addition, Chi-Square analyses between gender and UK and USA scores did not identify any significant differences.

Secondary diagnosis

Differences between secondary diagnosis and the CORE:OM, PDS and SIMS subscale and total scores were undertaken using Analysis of Variance (ANOVA). No significant differences were found.

Work Status

Differences between work status and the CORE:OM, PDS and SIMS subscale and total scores were undertaken using Analysis of Variance (ANOVA). No significant differences were found.

Education status

Differences between educational status and the CORE:OM, PDS and SIMS subscale and total scores were undertaken using Analysis of Variance (ANOVA). Table 10 shows that a significant difference was found for education level in relation to the CORE:OM Wellbeing subscale [$F(6,66) = 3.271$, $p<.05$], the CORE:OM Feelings

subscale [F(6,66) = 3.844, p <.01]; and the CORE:OM Total score [F(6,66) = 2.875, p<.05].

Table 10: Differences between educational categories

Scale	*Education level*																			
	No formal qualifications			*GCSE/O Level*			*A Level / Senior High*			*Vocational*			*Undergraduate degree*			*Postgraduate degree*				
	Mean	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*F*	*Sig.*
CORE Wellbeing	7.43	4.5	7	11.6	2.55	10	6.63	4.14	8	9.17	2.64	6	6.59	3.92	29	6.55	3.56	11	3.271	0.011*
CORE Psych distress	23.86	8.84	7	29.9	10.26	10	21	10.1	8	28.5	9.54	6	22.62	9.61	29	22.18	9.25	11	1.366	0.249
CORE Feelings	25.57	7.93	7	30.3	6.45	10	19.25	12.41	8	25	5.76	6	20.9	6.6	29	17.36	8.38	11	3.844	0.004*
CORE Total	56.86	19.75	7	71.8	15.98	10	46.88	24.33	8	62.67	16.05	6	50.1	18.24	29	46.09	20.43	11	2.875	0.021*
T PDS IM	57.29	10.27	7	60.33	8.43	9	64.63	7.63	8	63.67	4.97	6	62.1	7.76	29	62.78	14.18	9	0.635	0.674
T PDS SDE	52.57	6.53	7	51.11	8.1	9	50.88	11.22	8	54.5	11.08	6	51.07	9.23	29	61.22	10.11	9	1.806	0.125
T PDS Total	59.29	14.43	7	61.44	10.53	9	65.63	13.31	8	67.17	11.02	6	63.24	10.99	29	70.33	17.17	9	0.857	0.515
SIMS Psychosis	2.57	3.05	7	3.67	4.03	9	1	1.07	8	4.17	5.46	6	3.72	4.25	29	2.22	3.23	9	0.867	0.508
SIMS Neurologic Impairment	4.86	3.93	7	5.56	5.03	9	3.13	1.64	8	7.33	4.72	6	4.34	3.77	29	3.89	3.86	9	1.017	0.415
SIMS Amnestic	5.86	5.84	7	4.44	3.71	9	3.5	2.62	8	5.83	5.71	6	5.45	4.31	29	3.11	3.06	9	0.736	0.600
SIMS Low Intelligence	1.71	2.21	7	3.22	1.79	9	2	1.51	8	2.33	2.25	6	2.28	2.1	29	1.33	1.87	9	0.92	0.474
SIMS Affective	7	2.24	7	7.33	2.5	9	5	2.73	8	8.17	2.14	6	6.17	2.27	29	5.33	2.78	9	1.932	0.102
SIMS Total	22	15.66	7	24.22	13.38	9	14.63	7.54	8	27.83	18.95	6	21.97	13.09	29	15.89	11.77	9	1.089	0.375

**Significant for p<.05*

Post hoc comparisons using Games-Howell test indicated that, in relation to the CORE:OM Total score, respondents with up to a GCSE/O-Level qualification had higher scores (possibly reflecting higher psychological distress) compared to undergraduates (p <.05). In addition, in relation to the CORE:OM Wellbeing subscale, GCSE/O-Level respondents showed higher scores than Undergraduates (p <.01) and postgraduates (p <.05), which may imply lower levels of wellbeing amongst the GCSE/O-level respondents. A similar finding was evident in relation to the CORE:OM Feelings subscale, with GCSE/O-Level respondents showing higher scores than

Undergraduates ($p < .01$) and postgraduates ($p < .01$), potentially indicating higher levels of difficult feelings amongst the GCSE/O-level respondents. However, these results should be interpreted with caution given the small sample sizes.

Reliability of the Measures

In order to establish whether the PDS, SIMS and CORE:OM are psychometrically sound for use with a HFA population, internal consistency of all measures, and convergent validity of the SIMS was explored.

Internal consistency

The internal consistency of the scales was calculated using Cronbach's Alpha. Table 11 provides a summary of the internal reliability of the CORE:OM, PDS and SIMS total and subscale scores.

Table 11: Internal Reliability of the CORE, PDS and SIMS total scores and subscales

Scale	*Number of items*	*Cronbach's Alpha Coefficient*	*Cronbach's Alpha per manual*
CORE Total	28	0.929	0.94
CORE Wellbeing	4	0.822	0.77
CORE Psych distress	12	0.889	0.90
CORE Feelings	12	0.803	0.86
PDS Total (T-score)	40	0.712	0.85
PDS IM (T-score)	20	0.668	0.84
PDS SDE (T-score)	20	0.654	0.75
SIMS Total	75	0.936	0.88
SIMS Psychosis	15	0.899	0.82
SIMS Neurologic Impairment	15	0.853	0.83
SIMS Amnestic	15	0.882	0.83
SIMS Low Intelligence	15	0.624	0.85
SIMS Affective	15	0.498	0.86

All total scores were well within acceptable ranges (above .7): CORE:OM (n=73) Cronbach alpha of .929; the PDS (n=70) Cronbach alpha of .712; and the SIMS (n=70) Cronbach alpha of .936. However, at a subscale level, the Cronbach alphas were below the acceptable range on the PDS SDE scale (.654), the SIMS Low Intelligence subscale (.624) and the SIMS Affective subscale (.498), indicating that there are items on those subscales which are not correlating well and thus not measuring the same thing. This would raise concerns about the use of these subscales with people who have HFA.

Convergent validity

Given that the SIMS explores malingering of symptomology, it was considered useful to explore how actual levels of psychological distress (measured using the CORE:OM) in people with HFA may relate to how they then score on the relevant SIMS subscale (Affective disorders), possibly due to suggestibility. Convergent validity between the CORE:OM Total and SIMS Affective scale was undertaken using Pearson correlation. Table 12 provides a summary of scores. A significant relationship was found with high total scores on the CORE:OM being correlated with high scores on the Affective scale of the SIMS ($r = .604$, $p < .000$). This would suggest, firstly, that the SIMS Affective subscale is measuring a similar construct as the CORE:OM. Secondly, this may also suggest that genuinely experienced psychological distress would be identified as malingering of affective symptoms, when using the SIMS with a HFA population.

Table 12: Correlations between the SIMS affective subscale and the total score and subscales of the CORE:OM

CORE:OM	*SIMS: Affective*	*Bootstrap (with 1000 repetitions)*		
	Pearson r	*Bias*	*Lower 95% CI*	*Upper 95% CI*
CORE Total	0.604*	-0.004	0.442	0.730
CORE Wellbeing	0.569	-0.002	0.436	0.681
CORE Psych distress	0.547	-0.003	0.378	0.682
CORE Feelings	0.542	-0.007	0.331	0.692

**Significant*

The CORE:OM Cronbach Alphas in the current study were largely in line with those reported in the manual for the non-clinical population. In terms of the PDS, the Cronbach Alpha scores in the current study were lower than the manual for the general population, on both the subscales and total score. The SIMS faired similarly in the current study compared to the manual's general population, with the exception of the Low Intelligence and Affective subscales, with both scales, particularly the affective subscale, demonstrating poor internal consistency in the current study. These findings suggest that the CORE:OM and PDS could be used with a HFA population, but that certain scales on the SIMS would not be appropriate to use with people who have HFA.

Following completing the surveys, some respondents spontaneously emailed some informal feedback, highlighting certain difficulties they experienced when completing the measures. For instance, respondents felt that some items were ambiguous. At times they were unsure how to endorse an item as they expressed that their feelings may differ in different situations, and so, in order to answer a question accurately, felt they needed a scenario or context around it. Additionally, some experienced frustration at not knowing what something meant, having some initial thoughts and then doubting themselves. All of these factors are likely to have impacted on how they approached their responses to items on the survey.

Discussion

People with High Functioning Autism (HFA) sometimes find themselves involved in the criminal justice system, whether as a victim, witness or offender. As part of this process, they are likely to be assessed in some form, such as their ability to give evidence or participate in a trial, exploring risk and responsibility, or gauge treatment responsivity. BPS (2009) guidance recommends that assessments in criminal justice contexts should include consideration of response bias, more specifically socially desirable responding, impression management and malingering. However, none of the self-report measures frequently used by psychologists have been designed or validated for use with an Autistic population. This means that conclusions may be drawn on the role of response bias in assessments that are inaccurate and lack empirical support. This study aimed to address this gap in the evidence base.

The study hypothesised that people with HFA, who had little or no motivation to respond in a socially desirable way nor malinger symptomology, would score differently on two commonly used tools in the UK – the PDS and the SIMS, when compared to the normative data in the published manuals. This was hypothesised because the literature has identified that people with HFA could display response biases to certain types of questioning styles (Lerner et al., 2012; O'Mahony, 2012). This hypothesis has largely been proven in the current study. In relation to the SIMS, HFA respondents scored above cut-off on the Psychosis, Neurologic Impairment, Amnestic Disorders and Affective Disorders sub-scales of the SIMS, as well as the Total score. The only subscale in which people with HFA did not score above cut-off on was the Low Intelligence subscale. In relation to the PDS, the scores of those people with HFA were in line with the general population for Self-Deceptive Enhancement (SDE), but

fell in the above average ranges on the Total score and Impression Management (IM) subscale. Their elevated scores for IM were reflective of possible 'faking good' as measured by this test. This is an important finding as it may suggest that the IM on the PDS, when used with people with HFA, can lead to a false positive for impression management, thus having significant negative implications for how their reporting is viewed by assessing clinicians. It is also important to consider that respondents had complete anonymity in the current study and still had elevated scores, and what the impact would be if they were assessed in person.

This general pattern of elevated scores on a measure of malingering and socially desirable responding is important to consider in the context of respondents having no clear motive for displaying response bias, as in the current study, where their responses were anonymous and outside of a 'high stakes' context. The current study's findings are in line with Lerner et al (2012) who found an increased likelihood of elevated scores on assessments of effort and malingering amongst those with neurodevelopmental conditions, including Autism. The reason for this is likely related to some of the key attributes of people with HFA, which include the inability to cognitively process in a flexible way, perspective taking, processing social and emotional information and general Theory of Mind difficulties (Ali, 2018; Baron-Cohen, 2008; Lerner et al., 2012; NAS, 2019; Wing, 1997). The tendency for those with neurodevelopmental conditions such as Learning Disability and Autism to be suggestible is also a likely contributory factor to these findings (Chandler et al., 2018, Lerner et al., 2012; O'Mahony, 2012).

It is important, however, to consider a closer inspection of the two measures explored. In terms of the SIMS, this self-report measure explores malingered psychopathology (Widows & Smith, 2005). One explanation for elevated scores

amongst an HFA population is that they may actually be experiencing higher levels of co-morbid mental health disorders, which fits with previous research (Murphy et al., 2016). This would suggest that the SIMS is not effectively discriminating between genuine and faked symptomology. A further explanation may relate to people with HFA being found to be more suggestible than the general population (Chandler et al., 2018). Thus, the way in which 'symptoms' are presented in the SIMS, may serve to 'prime' people with HFA to endorse the symptom as something they have experienced. These factors may explain the elevated scores on most of the subscales, but do not explain the finding that the Low Intelligence subscale was not above cut-off in this population. This is likely related to the nature of the Low Intelligence questions at an individual level. The items that make up the Low Intelligence subscale consist mainly of basic general knowledge or mathematical calculations that are very obviously either correct or incorrect. Thus, this subscale is tapping into an alternative cognitive structure – that being factually based as opposed to somatic or psychological experiences that the other subscales are measuring. Thus, in the context of not having any clear incentive or motive to malinger, they correctly identified which items were true and false.

The patterns of scores for the PDS were interesting and also deserve closer inspection. In contrast to the SIMS, the PDS does not assess malingering, but rather socially desirable responding, and separately considers SDE and IM (Paulhus, 1998). The people with HFA in the current study scored in line with the general population for SDE, which indicates that they are no more likely to display rigid over-confidence. When considered in this way, this finding makes sense if we consider that whilst people with HFA may effectively 'mask' to compensate for social and communication difficulties (Kaland et al., 2007; Tager-Flusberg, 2007), they largely lack confidence

and experience anxiety about their social selves. This also fits with Baron-Cohen's (1992) findings that people with Autism struggle to be deceptive in more complex contexts.

This may go some way in explaining the elevated IM scores, and is suggestive of people with HFA portraying themselves in a more positive light. However, a further reason for the elevated IM scores may relate to the way the PDS is designed and scored. The PDS presents a statement, which is scored on a scale of one to five, with the extreme ends of the continuum attracting a score. This lends itself to higher item endorsement, due to people with HFA being more prone to rigid or black and white thinking (Mazefsky & White, 2014) and propensity to adhere to social and moral rules that are explicitly outlined (Grant et al., 2018), particularly when worded in the way they are in the PDS. As a result, they may be mistakenly seen as 'faking good' when in fact they are responding honestly, or at least are unaware that they are endorsing distorted responses.

The analysis of scores also identified some individual differences amongst the HFA sample, which needs further consideration. Respondents in this study did not differ significantly on the PDS and SIMS with regards to any secondary diagnoses, level of education or work status. In terms of age, the only significant finding was that scores on the SIMS Psychosis subscale went down as people aged. This is unsurprising given that the experience of psychotic symptomology in the general population decreases with age (Auslander and Jeste, 2014), but that the SIMS Psychosis scale produces false positives with those diagnosed with Schizophrenia (Impelen et al., 2014). However, the lack of variability depending on age across all other subscales of the SIMS as well as the PDS, is in contrast to the finding in the forensic study by Mathie

and Wakeling (2011), that age correlated with high IM scores on the PDS. This again may be explained by the differing motives in this context compared to offending populations, where there may be extrinsic factors contributing to respondents displaying self-report response bias.

Despite the growing evidence that there are key differences between males and females with Autism (NAS, 2020), the only significant differences in scores in this study were found on the SIMS Total score and the Psychosis and Neurological Impairment subscales. This generally suggests that males and females with HFA are no more or less likely to over endorse psychopathology on self-report measures. However, when using the SIMS, some alternative means and percentiles should be used for the specific scales where differences between genders were identified.

Differences in response styles for the PDS and SIMS were evident when comparisons were made between the UK and USA. The directionality of differences was in contrast to previous findings amongst the general population for the PDS, undertaken by Tully and Bailey (2017). Their study found slightly lower SDE scores but higher IM scores in the UK compared to USA/Canadian norms. However, the current study found lower UK IM scores compared to the USA respondents. Despite this, the UK norms for the current sample were more closely matched to Tully and Bailey's (2017) mean scores that they found in their study. This may suggest that at least on the PDS, UK norms may differ from the manual, but HFA scores in the UK are relatively similar to the UK general population. The only significant differences evident on the SIMS were for the Psychosis and Low Intelligence scales. These differences may reflect actual differences, but may also be confounded by other variables, such as age and gender. However, t-tests in relation to age, and chi-square exploring gender

differences, did not identify anything significant, suggesting that some other factor is contributing to the differences on the Psychosis and Low Intelligence subscales of the SIMS.

Given the variation in norms with the current HFA sample compared to normative data in the respective manuals of the PDS and SIMS, this does suggest that these measures can be used with an HFA population, but that higher cut-off scores need to be utilised. The measures themselves demonstrated good internal consistency, with the exception of the SIMS Low Intelligence and Affective subscales, and to some extent, the PDS SDE subscale. This introduces some caution when interpreting these subscales with an HFA population.

In relation to ratings on the CORE:OM, whilst the respondents' scores all fell well within the clinical range indicating high levels of psychological distress, the impact of data collection for the research taking place during the COVID-19 pandemic cannot be ignored. This may be particularly relevant given how change in routine and structure can increase anxiety in people with Autism (NAS, 2019). As a result, scores potentially may be more elevated as a result.

Other variations in CORE:OM scores may also be worth noting, but again, with some caution due to the period of time when data collection took place. If the assumption is that SIMS scores were elevated amongst people with HFA compared to the general population, this may be due to them genuinely experiencing higher levels of psychological distress than the average population, in line with previous studies (Anckarsater et al., 2008; Murphy et al., 2016). Their scores on the CORE:OM falling in the clinical ranges may be evidence of this. However, this would mean that the SIMS

was producing false positives – that it was not able to distinguish between real and malingered symptomology.

No differences were evident in relation to age, secondary diagnosis, or work status. However, differences were evident between the UK and USA scores, with UK means being higher. In addition, CORE:OM Total scores appeared significantly lower amongst those with higher levels of education. However, this may be an artifact of the COVID pandemic (such as greater financial difficulties for those with lower levels of education due to the economic impact of the pandemic), thus confounding the results. Other potential reasons for these differences are beyond the scope of this study and would require larger samples.

Conclusions and Limitations

The use of measures of malingering and socially desirable responding has its place as part of wider clinical assessments in medico-legal contexts. However, it is imperative that we utilise measures that are valid and reliable in order to have confidence in the conclusions that we draw. This study provides evidence that scores on the PDS and SIMS do differ amongst a HFA population, and that relying purely on normative data in the respective manuals may increase the likelihood of false positive outcomes. Whilst the PDS and SIMS would not be used alone to determine malingering or SDR, their contribution to the process must be evidence based or not included at all. This study has provided a starting point for establishing the evidence base for the use of these tools with HFA. However, interpretation of certain subscales, particularly the SIMS and PDS SDE subscale, may need to be approached with caution due to poor internal consistency when used with HFA. Whilst the current study provides enough

evidence that these measures operate differently in a HFA population, the data collected in the current study is not sufficient to provide alternative normative data for a HFA population. The findings of this study should therefore provide a starting point for further research. However, in forensic and clinical contexts, the information gained from the current study should provide clinicians with some supporting evidence that these measures should ideally not be used with people with HFA. This is especially important as we are still in the early stages of understanding how and why these measures operate differently amongst those people with HFA.

Furthermore, whilst the PDS and SIMS are relatively quick and cost effective psychometrics for adding to an assessment, there is growth in alternative performance based measures, particularly in relation to assessing malingering, which may be more useful to rely on than self-report measures such as the SIMS. In relation to SDR, alternatives to the PDS are limited, and the general concept of SDE and IM are difficult to accurately measure, and so in these cases, exploring motive, context, and drawing on collateral information will go a long way in substantiating or disproving these types of response biases.

The study is not without its limitations. Small sample sizes impacted on the extent to which meaningful analyses could be undertaken in terms of exploring individual differences within the current sample. Information on ethnicity was also not collected. Although this ideally should have been, given the small sample size, meaningful analysis on different response styles based on ethnic background would not have been possible. Cultural differences may also have made a difference to respondents' scores, but this could be explored in future research.

It was also not possible to ascertain with certainty that a respondent did in fact have a formal diagnosis that was made in line with the rigorous approach recommended by the National Institute of Clinical Excellence (NICE). Diagnostic approaches of ASD/HFA may also differ in different countries, hence some may be considered to have HFA in one country and not in another. These factors mean that some respondents may have completed the survey even though they did not have the appropriate diagnosis, impacting on the validity of the results. Future research on those with a confirmed diagnosis is needed.

The study was also undertaken during the COVID-19 pandemic, which arguably had an impact on data collection and sample size. Arrangements had been made to visit certain Autism support groups in the UK in order to inform them about the research and allay any concerns potential respondents had. This was important because early feedback that I had been given was that potential respondents were wary of unknown researchers and worried about websites or links being computer scams. Thus meeting me and hearing about the research might have allayed these concerns and increased the likelihood of participation. However, due to travel and in person contacts being restricted due to the pandemic, these arrangements were cancelled.

In addition, the impact of the COVID-19 pandemic on how respondents approached certain measures, such as the CORE:OM requires consideration. Whilst the most obvious impact was on the CORE:OM (as it is a measure of psychological distress), there was likely less impact on the PDS and SIMS given the nature of the items and content of those measures. Further research exploring differences in CORE:OM scores before, during, and after the pandemic, may shed light on the level of impact that the pandemic had on the current data. In addition, expanding the normative

data set for the PDS and SIMS with a HFA sample would increase the applicability of the results of this study.

CHAPTER 5

DISCUSSION

The thesis aimed to address some of the gaps in the literature on response bias by providing a review of self-report measures that can be used in assessing response bias and establishing the evidence base for application in criminal justice contexts. In addition, it aimed to collect normative data for commonly used self-report measures (the PDS and SIMS) and consider their value and applicability in HFA.

Summary of Findings

Chapter 2 presented a systematic review of the literature on how self-report response bias has been assessed in forensic contexts in the UK over the last 10 years. The findings highlighted that across all contexts, UK clinicians predominantly use the Paulhus Deception Scales/Balanced Inventory of Desirable responding (PDS/BIDR: Paulhus, 1998). However, the measures used in the studies tended to consider socially desirable responding, with few papers that explored over-reporting of psychopathology/malingering. In addition, little emphasis was placed on motives that drive dishonest responding, resulting in scores on these measures alone being relied upon to make assumptions on respondents' validity of self-report. This goes against guidance by Slick et al (1999) and Martelli et al (2012) about the comprehensive assessment of response bias, as well as the failure to recognise the protective function response bias may have on a persons functioning (Paulhus, 2002; Tan & Grace, 2008; Von Hippel & Trivers, 2011). Furthermore, there was little variation in the use of self-report response bias measures across contexts, thus failing to recognise the implications of referring to normative data mainly from North America on substantially heterogeneous populations. This becomes particularly relevant for specialist populations such as those with neurodevelopmental disabilities, with only one UK study actively

attempting to validate a measure for those with Intellectual Disabilities (Jobson et al., 2013).

Due to the predominance of the Paulhus Deception Scales (PDS: Paulhus, 1998). being utilised in the UK in a variety of forensic contexts, including prison and hospital settings, Chapter 3 examined the psychometric properties of this measure. Whilst the PDS is used in multiple contexts with a vast range of populations, normative data drawn on is based on US and Canadian samples (Tully & Bailey, 2017). Internal reliability of the PDS was evident with adequate Cronbach Alpha scores, but some differences were evident in the strength of this when examining the two subscales that make up the PDS – the SDE and IM scales. For instance, when examined using a UK sample, IM alpha scores were lower, with an alpha of .68 (Tully & Bailey, 2017). In terms of validity, challenges were faced in determining this with certainty given the seven variations of the BIDR/PDS as the scale developed, as well as differences in how the measure is scored. Bearing this in mind, the PDS demonstrated adequate concurrent, construct, and structural validity, though structural validity outcomes varied depending on the statistical methods used to establish it. In terms of normative data, this exists but relies heavily on North American samples as well as convenience samples, with a lack of normative data available for specialist populations, such as those with neurodevelopmental conditions. Whilst the PDS is frequently used in the UK, and data is available for analysis, no meta-analytic studies have as yet been undertaken on the PDS, in the UK or elsewhere in the world.

Chapter 4 moves on to present an empirical study aimed at establishing a normative data set for the PDS (Paulhus, 1998) and Structured Inventory of Malingered Symptomology (SIMS: Widows & Smith, 2005) with a High Functioning Autistic

community adult sample. The study hypothesised that people with HFA, who had no obvious motivation to be deceptive in their responses, would nevertheless score differently to the normative data in the respective manuals of the PDS and the SIMS. This is because, we already know from the literature that people with ASD have problems with understanding their emotional states (Lai et al., 2014), problems with intuition (Lerner et al., 2012), can be prone to suggestibility and acquiescence (Chandler et al., 2018; Lerner et al., 2012; O'Mahony, 2012), and may have limited self-insight and display concrete thinking (Ali, 2018; Attwood, 2015; Baron-Cohen, 2009; Mazefsky & White, 2014; NAS, 2019). Elevations in scores on the SIMS were found, in line with Lerner et al (2012), suggesting that people with HFA over-report or malinger symptomology. However, the standard measures of psychological distress indicated that the HFA sample did in fact experience elevated levels of psychological distress, which would suggest that the SIMS was unable to differentiate between actual and feigned symptomology in this client group. This is important to consider, particularly as the evidence base has established higher levels of co-morbidity of mental health problems in those with HFA (Anckarsater et al., 2008; Murphy, 2003; Murphy et al., 2016). Alternatively, is it that people with HFA are more likely to endorse greater pathology due to suggestibility, compliance or concrete thinking, and thus the SIMS is in fact detecting over-reporting of symptomology, particularly given the SIMS includes unusual symptoms that do not accurately reflect real symptoms? The meta-analysis by Griego et al (2019) concluded that people with HFA did not display false memory and memory suggestibility. However, the studies included drew on specific measures (such as the GSS: Gudjonsson, 1997) that rely on a person recalling events read out in a story format. Measures such as the GSS rely on verbatim memory – memory structures that

are not evident to being effected in those with HFA given their average or above average cognitive functioning (Griego et al., 2019). This is very different to being presented with a list of symptoms and being asked if this is something they experienced. It is not fully clear though, at this stage, based on the research evidence, to say with certainty which factor is at play and contributing to elevated scores on measures like the SIMS or those measures that assess mental health difficulties. However, given the high rates of comorbidity in Autism generally (Anckarsater et al., 2008; Murphy, 2003; Murphy et al., 2016), and tendencies to be prone to certain types of questioning styles (Lerner et al., 2012; O'Mahoney, 2012) it is likely that the SIMS is prone to producing false positives with people with HFA.

In terms of the PDS, SDE scores were in line with the general population. However, IM responses fell in the "possibly faking good" range, with people with HFA having a tendency to impression manage and portray themselves positively. This may on the surface appear contradictory, given how people with HFA performed on the SIMS. However, a likely reason for the elevated IM scores may relate to the way the PDS is designed and scored, as extreme ends of the scoring system attracts a score, making it prone to extreme scoring in those with concrete thinking styles. This fits with the literature in that people with HFA are more prone to concrete, rigid or black and white thinking (Mazefsky & White, 2014) and have a propensity to adhere to explicit social and moral rules (Grant et al., 2018). As a result, the PDS IM scale may be producing false positives in people with HFA, as was evident in the SIMS. An alternative explanation, however, may relate to differences in scoring patterns evident in different countries. This is because, a recent study undertaken by Tully and Bailey (2017) found that the UK norms for the PDS IM scale were higher amongst the UK

general population compared to the normative data based on USA samples, included in the PDS manual. The study outlined in Chapter 4 found that the HFA IM scores in the UK were relatively similar for HFA and the general population in Tully and Bailey's (2017) study. This suggests that the people with HFA were not in fact engaging in impression management, and rather higher cut-off scores should be used that are more representative of the UK population.

When exploring individual differences, there appeared to be a main effect of age on the SIMS psychosis scale, where older participants scored lower, suggesting less reporting of bogus psychotic symptoms. If the assumption is that the SIMS is not able to distinguish real from feigned symptomology in people with HFA, then this would make sense, given the literature suggests that reporting of psychotic symptoms decreases with age (Auslander and Jeste, 2014), though more age differences have been evident when explored in forensic samples (Mathie & Wakeling, 2011). No clinically significant differences were evident between males and females, with the exception of SIMS total and two of the SIMS scales, though the sample size may have been a factor in this finding. Differences in scores were also evident when comparisons were made between USA and UK respondents.

The measures demonstrated adequate to poor internal consistency. Concurrent validity was evident when the CORE:OM was compared to the SIMS Affective subscale, yet the SIMS is supposedly measuring faked symptoms whereas the CORE is designed to measure actual symptoms. Thus, theoretically, if someone is scoring highly on the CORE (i.e. is genuinely distressed) then they should not be scoring highly on the SIMS, which is supposed to only be identifying faked symptoms. This highlights the need for further research with a larger sample that may enable low scores on the

CORE:OM to be compared to above cut-off scores on the SIMS. In addition, a more in-depth evaluation of a series of single cases may help to better understand the mechanisms behind the response styles of people with HFA on all of the measures used in the empirical study. Furthermore, elevations in CORE:OM scores were evident, but the confounding effect of data being collected during the COVID-19 pandemic, could have played a significant role in this. Overall, the study provided some initial evidence that alternative cut-off scores should be used with those who have HFA, as part of a wider holistic assessment of response bias.

Spontaneous informal feedback from some respondents highlighted certain difficulties they experienced when completing the measures, such as finding some items were ambiguous, experiencing uncertainty about how to endorse an item given how they felt differed across situations, and difficulties understanding some items. These factors are likely to have impacted on their response styles, and should be considered if developing new response bias measures for this client group. Finally, a major limitation of the study was the small sample size, meaning generalisability of the results and analysis of group differences, might have been impacted upon.

Moving forward, the relevance of the findings outlined above to undertaking assessments with people with HFA in forensic contexts, requires consideration. Given that people with HFA look like they are engaging in response bias where there are *no* situational demands, it raises the question about what might their responses look like where there are. This is particularly pertinent given that those with Autism are known to experience higher levels of anxiety, tend to try and mask their difficulties, and struggle in new contexts (Ali, 2018; Langden et al., 2010; Lerner et al., 2012; NAS, 2019; NAS, 2020c; Underwood et al., 2013) – all highly relevant when faced with a clinician they

do not know and where the stakes may be high, such as in a forensic context. In order to understand the relevance of these findings, it will be necessary to undertake further research in forensic contexts to establish comparisons between HFA/ASD and non-HFA/ASD forensic populations. In addition, undertaking research in a similar fashion to how many of these response bias tools are normed, that being, asking those with HFA to feign or display SDR, and see how that compares to the current normative data, may be useful.

Conclusions and Recommendations

As already mentioned, incorporating measures of malingering and socially desirable responding has its place as part of wider clinical assessments in medico-legal, risk assessment, and forensic treatment contexts. However, it is imperative that we utilise measures that meet standards of reliability and validity for the populations we use them on, in order to have confidence in the conclusions that we draw. There is already recognition that witnesses and defendants on the Autistic Spectrum are vulnerable in police and court contexts, with the growing use of Registered Intermediaries to support with communication and other psychological needs (O'Mahony, 2012). However, there is limited guidance on how to adjust assessment processes for this cohort. There is variability in the use of response bias measures across the world, coupled with, in many cases, the failure to recognise the limitations of normative data being available for specialist populations, in this case, those with HFA. There is also limited evidence of response bias assessments being undertaken holistically, with sufficient attention being placed on motivation and varying presentations, again highly relevant with HFA.

Understanding and exploring the function of and reason for elevated scores on any response bias measure should form an integral part of a wider assessment, rather than simply concluding a tendency to deceive or be dishonest. Given the detrimental impact false positives on measures of response bias can have for an individual with HFA, at all stages of their contact with the CJS, selection of appropriate norms is essential. This is especially important given that the prevalence rates of Autism in UK forensic institutions has been shown to range from between 1.5% and 15% (Hare et al., 1999; NAS, 2020; Scragg & Shah, 1994; Siponmaa et al., 2001; Talbot, 2009). A further important consideration is that this figure does not capture the number of witnesses and victims who are assessed as part of the criminal justice process. At present, there are no estimates available here.

This thesis provides some preliminary evidence that scores on the PDS and SIMS do differ amongst a HFA population, and that relying on normative data cited in the manuals means an increased likelihood of false positive outcomes. As a result, assessment may conclude that the individual is purposefully distorting their responses when they may not be, and thus negatively impacting on the overall assessment of their psychological functioning. Whilst the PDS and SIMS would not be used alone to determine malingering or SDR, their contribution to the process must be valuable and reliable, or not included at all. Additionally, whilst it is recognised that the PDS and SIMS are a simple, fast, and cost effective means of adding to an assessment, there is growth in alternative performance based measures, which may be more useful to rely on than self-report measures such as the SIMS. For instance, Sadek, Daniel and Langden (2020) used the dot-probe task that utilises pictures and measures attentional bias, indicating preferences or otherwise, for certain behaviours, such as violence. In their

study of offenders with intellectual disabilities, it was possible to distinguish between offenders and non-offenders, with offenders showing significantly more attentional bias towards negative images. This procedure could remove problems associated with reading difficulties and misinterpretation of language used (Sadek et al., 2020). Thus, utilising attentional and response time approaches to assess offence supportive cognitions could ultimately negate the need for response bias measures (like the PDS) with offenders, particularly in relation to assessing the role of offence supportive beliefs/interests in risk of re-offending or treatment efficacy, but more thought is required on how this could be applied in other areas of forensic assessments. In relation to SDR, exploring a person's particular traits of HFA, motive, context, and collateral information, is more important in understanding why they may portray themselves in a certain light as a means to substantiate or disprove response bias.

In forensic and clinical contexts, including when assessing people with HFA for court (as witnesses, victims or defendants), this research should provide some evidence and justification for excluding the use of self-report response bias measures with people who have a diagnosis (or suspected diagnosis) of HFA. Instead, information gained via observations of the specific individual's effort and performance during assessment, file information and discussions with those close to or working with the person, will help highlight any inconsistencies that may suggest dishonest engagement in the assessment process, or in contrast, provide evidence for consistent and honest engagement. A central part of establishing motive or intent to deceive or be honest requires an understanding of the person with Autism in an individualised way, particularly as people with Autism differ in the way that their particular traits may present.

Recommendations for Future Research

The systematic review on response bias measures used in the UK identified that the most commonly used self-report measure is the PDS/BIDR. Some of the samples sizes were quite large, and so it would be valuable to undertake a PDS meta-analysis with the aim of providing a normative data set that is directly relevant to a UK forensic cohort. Future studies should also aim to assess the reliability and validity of the PDS with different populations in order to gain clarity on how factors such as: personality, cognitive ability, age, context, and other motivating factors, may be influencing response style. It would also be valuable to ensure information on ethnicity is included in future studies, as there may be variations in how people from various ethnic backgrounds endorse questionnaires. UK research should also pay more attention to how malingering is assessed in forensic contexts, as there are ongoing challenges in offenders attempting to mitigate responsibility for offending or extend support in mental health environments by over-reporting psychopathology (Rogers et al., 2018; Thomas-Peter et al., 2000). Studies exploring medico-legal assessments undertaken with victims, witnesses, and defendants for court, would be useful to add to the evidence base.

Research examining the full range of response bias approaches and comparing their efficacy in identifying response bias may be helpful, particularly in specialist populations. This is key, as it may be that alternative approaches, such as attention or reaction time approaches may be a way of dispensing the need for assessing self-report response bias, in certain contexts, in line with Sadek et al (2020). In addition, algorithms may be used to adjust scores on other measures (based on the PDS score), such as undertaken in the study by Elliott et al (2009).

In relation to HFA, it would be useful to build on the empirical study outlined in Chapter 4 by gathering larger HFA samples in low or no stakes situations. It is also important to consider the impact of complete anonymity of online ratings of self-report response bias measures compared to if they are completed in person, outside of clinical or high stakes situations (as a witness, victim or offender). This is particularly relevant for people with HFA due to social, communication and emotional difficulties that are likely to be exacerbated in face-to-face contexts, but may provide a more accurate baseline for low-stakes normative scores.

It will also be useful to collect CORE:OM data on people with HFA outside of a COVID context to see how differently they may score. It may be useful to consider other or additional means of establishing actual psychological functioning alongside response bias measures, such as through semi-structured interviews or by using observational measures. These approaches may provide a more accurate way of evaluating the extent to which people with HFA may be producing scores on measures of response bias that indicate intentional or unintentional distortion. In addition, it may be useful to replicate the common methodology for obtaining normative data with measures such as the PDS and SIMS by conducting laboratory experiments whereby individuals with HFA are instructed to fake good or fake bad. This data could then be compared to both of the relevant manuals non-HFA normative data as well as to no or low stakes normative data, to explore any key differences. Undertaking a detailed analysis of a series of single cases within criminal justice contexts may help inform a more detailed, qualitative and holistic understanding of the role that specific HFA traits, situational and contextual factors may have on response styles. Finally, gathering data from people with HFA in forensic contexts, such as prison, secure hospital, and those

involved in trials (as defendants, victims or witnesses), would provide the opportunity to further understand response bias in criminal justice contexts.

Recommendations for Future Practice

Whilst Slick et al (1999) and Martelli et al (2012) have developed guidance on the holistic assessment of malingering, similar guidance should be available for other forms of response bias, such as socially desirable responding and impression management. Drawing on the principles of Slick et al (1999) and Martelli et al (2012), the following is proposed for undertaking assessments of biased responding, relevant with all populations, but particularly in cases where Autism is suspected or with a confirmed HFA cohort.

Firstly, multiple sources of information should be available and drawn upon when considering whether response bias is a factor. These include, but are not limited to, information gained from: clinical interviews with the individual; clinical interviews with other sources (e.g. carers, professionals, family); observational data/information; performance on measures that form part of the wider assessment (e.g. cognitive assessment; memory, personality; mental health symptomology; measures of functioning); self-report measures of psychopathology and functioning; collateral information (such as previous professional assessments/reports or case related documentation, transcripts of interviews, such as Achieving Best Evidence (ABE) interviews). People with HFA or those suspected of having Autism should have assessment information available on cognitive functioning, adaptive behaviour, and suggestibility as part of this information reviewed.

Once all information from these multiple sources is considered, if no response bias is suspected, then assessment results can be reported and formulated. If response bias is suspected, then it is important to explore motive (situational and context specific) as well as any personality, diagnostic, or cultural factors that may influence the *individual* person's response style. This is particularly relevant in HFA, as Autism consists of a spectrum of traits that can vary, sometimes in unique ways, across each person.

It is then important to select an appropriate tool or approach to explore response bias further, whether it be looking more closely at performance on measures such as the WAIS-IV or validity scales on the MCMI-IV, attentional bias techniques, as well as identifying which specific response bias measures (malingering versus SDR measures, and performance versus self-report approaches) should be administered. Once a tool/s are selected, it is crucial that the clinician establish whether the measure/tool demonstrates adequate reliability and validity. If it does not, then it should not be used.

If it does, then it is necessary to identify whether or not there is normative data available that is relevant to the individual person being assessed. Appropriate normative data should include data for those with neurodevelopmental conditions (ideally Autism if confirmed or suspected), based on country specific populations, relevant context and appropriate sex. In cases where appropriate normative data is not available, the clinician must reflect on what will be added by formally assessing response bias further including the implications of findings, and consider an alternative or qualitative approach. Alternatively, the measure can be used, but the assessing clinician must ensure limitations are transparently and clearly reported on. If appropriate normative data is

available, the measure can be used, reporting on literature of the normative data drawn upon, and being clear about any limitations relevant to relying on that normative data.

If a person has been identified as highly suggestible or has poor language skills, self-report measures exploring malingering should ideally not be used. Alternatively, the assessing clinician may help the person rate the measure, for instance, by reading items to them. The clinician should then ask for examples to support each rating, increasing the threshold for endorsements and to mitigate the effects of suggestibility playing a role.

This process should be followed each time response bias is assessed with an individual, as their motivations, context, and presentation can change and thus influence each of the steps. In relation to the PDS specifically, this self-report measure can be used with people with HFA or those suspected to have Autism, as long as it is utilised as part of the above steps. The same is true for the SIMS, though this should be used very cautiously with Autistic/HFA populations, and the Psychosis and Low Intelligence subscales should be omitted.

It is important to have an understanding of why an individual may have elevated scores in assessments. This requires an individualised approach, particularly as Autism consists of a spectrum of traits that can vary in presence and severity from one person to the next. In forensic and clinical contexts this means taking into consideration many factors, such as (but not exclusively): the environment the person is being assessed in, whether they are familiar with the assessing clinician or not, levels of cognitive flexibility in terms of social rules and norms; and degree of masking they may engage in as a way to fit in as opposed to being deceitful.

This thesis has highlighted the importance of identifying those within the CJS who have Autism, particularly HFA, as they are often 'missed' due to them being high functioning, masking difficulties well, or being misdiagnosed (Lai et al., 2016; NAS, 2020a). At this time, there are limited accurate prevalence rates of Autism in the CJS available (Ali et al., 2018; Underwood et al., 2013), particularly in relation to witnesses and victims. Improved identification of HFA is key so that appropriate assessment processes can be utilised and relevant normative data drawn on for accurate interpretation of scores. There is also clear scope for the development of assessment tools that are appropriate for use with people with HFA, and which have relevant normative data available. This is imperative given the 'high stakes' situations those with HFA in criminal justice contexts find themselves being assessed in, and the dangers of erroneously concluding that they are faking good or faking bad. Finally, guidance for clinicians who are involved in assessing people with HFA in forensic contexts needs to be improved by being more specific on what is currently available in the form of measurement tools, how assessments should take place with what is currently available, and by highlighting the gaps in evidence base.

REFERENCES

Ali, S. (2018). Autistic spectrum disorder and offending behaviour – a brief review of the literature. *Advances in Autism*, *4*(3), 109–121. doi: 10.1108/aia-05-2018-0015

Alleyne, E., Gannon, T. A., Mozova, K., Page, T. E., & Ciardha, C. Ó. (2016). Female Fire-Setters: Gender-Associated Psychological and Psychopathological Features. Psychiatry, 79(4), 364–378. doi: 10.1080/00332747.2016.1185892

Anckarsäter, H., Nilsson, T., Saury, J.-M., Råstam, M., & Gillberg, C. (2008). Autism spectrum disorders in institutionalized subjects. *Nordic Journal of Psychiatry, 62*(2), 160-167. doi:10.1080/08039480801957269

Archer, R. P., Wheeler, E. M. A., & Vauter, R. A. (2016). Empirically Supported Forensic Assessment. Clinical Psychology: Science and Practice, 23(4), 348-364. doi:10.1111/cpsp.12171

Archer, R.P. & Wygant, D. (2012). Child Custody Evaluations: Ethical, Scientific, and Practice Considerations. Psychology Faculty and Staff Research. 5. https://encompass.eku.edu/psychology_fsresearch/5

Attwood, T. (2015). *The complete guide to Asperger's syndrome*. Jessica Kingsley Publ.

Auslander, L. A., & Jeste, D. V. (2004). Sustained remission of schizophrenia among community-dwelling older outpatients. *Am J Psychiatry, 161*(8), 1490-1493. doi:10.1176/appi.ajp.161.8.1490

Baron, R. A., & Byrne, D. (2000). *Social Psychology* (9th ed.). Boston: Pearson, Allyn and Bacon.

Baron-Cohen, S. (1992). Out of Sight or Out of Mind? Another Look at Deception in Autism. *Journal of Child Psychology and Psychiatry*, *33*(7), 1141–1155. doi: 10.1111/j.1469-7610.1992.tb00934.x

Baron-Cohen, S. (2008). Autism. *British Journal Of Psychiatry*, *193*(4), 321-321. https://doi.org/10.1192/bjp.193.4.321

Bender, S.D. & Frederick, E. (2018). Neuropsychological Models of Feigned Cognitive Deficits. In R. Rogers & S. Bender, Clinical Assessment of Malingering and Deception (4th ed., pp. 42-60). London: Guildford Press.

Binder, L. M., & Willis, S. C. (1991). Assesment of motivation after financially compensable minor head trauma. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3(2), 175–181. https://doi.org/10.1037/1040-3590.3.2.175

British Psychological Society. (2009). Assessment of Effort in Clinical Testing of Cognitive Functioning for Adults. British Psychological Society (BPS).

Brock Baskin, M., McKee, V., & Buckley, M. (2017). Time Banditry and Impression Management Behavior. *Journal Of Leadership & Organizational Studies*, *24*(1), 39-54. https://doi.org/10.1177/1548051816661479

Brooks, B. L., Ploetz, D. M., & Kirkwood, M. W. (2016). A survey of neuropsychologists' use of validity tests with children and adolescents. Child Neuropsychology, 22(8), 1001-1020. doi:10.1080/09297049.2015.1075491

Browning, A., & Caulfield, L. (2011). The prevalence and treatment of people with Asperger's Syndrome in the criminal justice system. *Criminology & Criminal Justice*, *11*(2), 165–180. doi: 10.1177/1748895811398455

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & Dahlstrom, W. G. (2001). *Mmpi-2: Minnesota multiphasic personality inventory - 2*. Minneapolis: University of Minnesota Press.

Carlson & Herdman. (2019). Convergent Validity and Discriminant Validity: Definitions, Examples. Retrieved 1 Dec 2019 from www.management.pamplin.vt.edu/directory/Articles/Carlton1.pdf

Casey, S., Day, A., Howells, K., & Ward, T. (2007). Assessing Suitability for Offender Rehabilitation. *Criminal Justice And Behavior*, *34*(11), 1427-1440. https://doi.org/10.1177/0093854807305827

CASP Checklists - CASP - Critical Appraisal Skills Programme. (2018, May 1). Retrieved from https://casp-uk.net/casp-tools-checklists/

Cassano, A. & Grattagliano, I. (2019). Lying in the medicolegal field: Malingering and psychodiagnostic assessment. Clinical Therapeutics, 170(2), 134-141. doi: 10.7417/CT.2019.2123

Chandler, R., Russell, A., & Maras, K. (2018). Compliance in autism: Self-report in action. *Autism*, *23*(4), 1005-1017. https://doi.org/10.1177/1362361318795479

Cohen, J. (2009). *Statistical power analysis for the behavioral sciences*. New York, NY: Psychology Press, Taylor & Francis Group.

Core System Group (1998). *CORE System (Information Management) Handbook*. Leeds, Core System Group. Retrieved from http://www.coreims.co.uk

Dandachi-FitzGerald, B., Ponds, R. W., & Merten, T. (2013). Symptom Validity and Neuropsychological Assessment: A Survey of Practices and Beliefs of Neuropsychologists in Six European Countries. Archives of Clinical Neuropsychology, 28(8), 771-783. https://doi.org/10.1093/arclin/act073

Day, A., Howells, K., Casey, S., Ward, T., Chambers, J., & Birgden, A. (2009). Assessing Treatment Readiness in Violent Offenders. *Journal Of Interpersonal Violence*, *24*(4), 618-635. https://doi.org/10.1177/0886260508317200

Dein, K., & Woodbury-Smith, M. (2010). Asperger syndrome and criminal behaviour. *Advances In Psychiatric Treatment*, *16*(1), 37-43. https://doi.org/10.1192/apt.bp.107.005082

De Marchi, B. & Balboni, G. (2018). Detecting malingering mental illness in forensics: Known-Group Comparisons and Simulation Design with MMPI-2, SIMS and NIM. PeerJ, 6:e5259; DOI 10.7717

Detrick, P., & Chibnall, J. (2008). Positive Response Distortion by Police Officer Applicants. *Assessment*, *15*(1), 87-96. https://doi.org/10.1177/1073191107306082

Drob, S.L., Meehan, K.B. & Waxman, S.E. (2009). Clinical and Conceptual Problems in the Attribution of Malingering in Forensic Evaluations. Journal of the American Academy of Psychiatry and Law, 37, 98-106. PMID:19297641

Dufner, M., Gebauer, J., Sedikides, C., & Denissen, J. (2019). Self-Enhancement and Psychological Adjustment: A Meta-Analytic Review. *Personality And Social Psychology Review*, *23*(1), 48-72. https://doi.org/10.1177/1088868318756467

Edwards, R., Whittaker, M. K., Beckett, R., Bishopp, D., & Bates, A. (2012). Adolescents who have sexually harmed: An evaluation of a specialist treatment programme. Journal of Sexual Aggression, 18(1), 91-111. https://doi.org/10.1080/13552600.2011.635317

Egeland, J.; Andresson, S.; Sundseth, O.O. & Schanke, A.K. (2014). Types of Modes of Malingering? A Confirmatory Factor Analysis of performance and Symptom

Validity Tests. Applied Neuropsychology: Adult, 0: 1-12. https://doi.org/10.1080/23279095.2014.910212

Egerton, A., Rees, E., Bose, S., Lappin, J., Stokes, P., Turkheimer, F., & Reeves, S. (2010). Truth, lies or self-deception? Striatal D2/3 receptor availability predicts individual differences in social conformity. *Neuroimage*, *53*(2), 777-781. https://doi.org/10.1016/j.neuroimage.2010.06.031

Elliott, I. A., Beech, A. R., Mandeville-Norden, R., & Hayes, E. (2009). Psychological Profiles of Internet Sexual Offenders. Sexual Abuse: A Journal of Research and Treatment, 21(1), 76-92. https://doi.org/10.1177/1079063208326929

Equality Act 2010: Guidance - GOV.UK. (n.d.). Retrieved May 8, 2020, from https://www.gov.uk/guidance/equality-act-2010-guidance

Eriksen, M. B., & Frandsen, T. F. (2018). The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. Journal of the Medical Library Association : JMLA, 106(4), 420–431. https://doi.org/10.5195/jmla.2018.345

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, *180*(JAN.), 51–60. http://doi.org/10.1192/bjp.180.1.51

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J. and McGrath, G (2000). CORE: Clinical Outcomes in Routine Evaluation. *Journal of Mental Health, 9, 3, 247-255.*

Field, A. P. (2005). *Discovering statistics using SPSS: (and sex, drugs and rocknroll)*. Second Edition. London: SAGE.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378–382. https://doi.org/10.1037/h0031619

Freckleton, G. (2013). Education at the Royal Courts of Justice. *The Law Teacher*, *47*(2), 269-270. https://doi.org/10.1080/03069400.2013.790151

Furnham, A. (1986). Response bias, social desirability and dissimulation. Personality and Individual Differences, 7(3), 385–400. doi: 10.1016/0191-8869(86)90014-0

Gannon, T. A., Alleyne, E., Butler, H., Danby, H., Kapoor, A., Lovell, T., … Ciardha, C. Ó. (2015). Specialist group therapy for psychological factors associated with firesetting: Evidence of a treatment effect from a non-randomized trial with male prisoners. Behaviour Research and Therapy, 73, 42–51. doi: 10.1016/j.brat.2015.07.007

Gannon, T. A., Ciardha, C. Ó., Barnoux, M. F., Tyler, N., Mozova, K., & Alleyne, E. K. (2013). Male Imprisoned Firesetters Have Different Characteristics Than Other Imprisoned Offenders and Require Specialist Treatment. Psychiatry: Interpersonal and Biological Processes, 76(4), 349-364. https://doi.org/10.1521/psyc.2013.76.4.349

Gignac, G. (2013). Modeling the Balanced Inventory of Desirable Responding: Evidence in Favor of a Revised Model of Socially Desirable Responding. *Journal Of Personality Assessment*, *95*(6), 645-656. https://doi.org/10.1080/00223891.2013.816717

Grant, T., Furlano, R., Hall, L., & Kelley, E. (2018). Criminal responsibility in autism spectrum disorder: A critical review examining empathy and moral

reasoning. *Canadian Psychology/Psychologie Canadienne*, *59*(1), 65–75. doi: 10.1037/cap0000124

Griego, A. W., Datzman, J. N., Estrada, S. M., & Middlebrook, S. S. (2019). Suggestibility and false memories in relation to intellectual disability and autism spectrum disorder: a meta-analytic review. *Journal of Intellectual Disability Research, 63,* 1464-1474. https://doi.org/10.1111/jir.12668.

Gudjonsson, G. H. (1997). *The Gudjonsson Suggestibility Scales*. Hove: Psychology Press.

Gudjonsson, G. H. (2003). *Wiley series in the psychology of crime, policing and law. The psychology of interrogations and confessions: A handbook.* John Wiley & Sons Ltd.

Gudjonsson, G. H., & Clark, N. K. (1986). Suggestibility in police interrogation: A social psychological model. *Social Behaviour, 1*(2), 83–104.

Gur, R., & Sackeim, H. (1979). Self-deception: A concept in search of a phenomenon. *Journal Of Personality And Social Psychology*, *37*(2), 147-169. https://doi.org/10.1037/0022-3514.37.2.147

Hammond, S. (2002). *Using Psychometric Tests*. In: Breakwell, G.M.; Hammond, S. & Fife-Shaw, C. (Eds.), *Research Methods in Psychology* (pp.175-193). London:Sage

Hare, D.J., Gould, J., Mills, R., and Wing, L. (1999) 'A preliminary study of individual with autistic spectrum disorders in three special hospitals in England', *The National Autistic Society*. Available at: www.aspires-relaationships.com/3hospitals.pdf.

Hart, K. J. (1995). The assessment of malingering in neuropsychological evaluations: Research-based concepts and methods for consultants. Consulting Psychology Journal: Practice and Research, 47(4), 246–254. https://doi.org/10.1037/1061-4087.47.4.246

Henry, O., Mandeville-Norden, R., Hayes, E., & Egan, V. (2010). Do internet-based sexual offenders reduce to normal, inadequate and deviant groups? Journal of Sexual Aggression, 16(1), 33-46. https://doi.org/10.1080/13552600903454132

Higgins, J.P.T., Savovic, J., Page, M.J., Elbers, R.G. & Sterne, J.A.C. (July 2019). Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J. & Welch, V.A. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions version 6.0, Cochrane, 2019. Retrieved February, 10, 2020 from www.training.cochrane.org/handbook

Hren, D., Vujaklija, A., Ivanisevic, R., Knezevic, J., Marusic, M., & Marusic, A. (2006). Students moral reasoning, Machiavellianism and socially desirable responding: implications for teaching ethics and research integrity. *Medical Education*, *40*(3), 269–277. doi: 10.1111/j.1365-2929.2006.02391.x

Hubley, A. M. (2014). Discriminant Validity. *Encyclopedia of Quality of Life and Well-Being Research*, 1664–1667. doi: 10.1007/978-94-007-0753-5_751

Huntley, F. L., Palmer, E. J., & Wakeling, H. C. (2011). Validation of an Adaptation of Levenson's Locus of Control Scale With Adult Male Incarcerated Sexual Offenders. Sexual Abuse: A Journal of Research and Treatment, 24(1), 46-63. https://doi.org/10.1177/1079063211403163

Impelen, A. V., Merckelbach, H., Jelicic, M., & Merten, T. (2014). The Structured Inventory of Malingered Symptomatology (SIMS): A Systematic Review and Meta-Analysis. *The Clinical Neuropsychologist*, *28*(8), 1336–1365. doi: 10.1080/13854046.2014.984763

Jobson, L., Stanbury, A., & Langdon, P. E. (2013). The Self- and Other-Deception Questionnaires-Intellectual Disabilities (SDQ-ID and ODQ-ID): Component analysis and reliability. Research in Developmental Disabilities, 34(10), 3576-3582. https://doi.org/10.1016/j.ridd.2013.07.004

Kaland, N., Mortensen, E., & Smith, L. (2007). Disembedding performance in children and adolescents with Asperger syndrome or high-functioning autism. *Autism*, *11*(1), 81-92. https://doi.org/10.1177/1362361307070988

Kline, P. (1986). *A Handbook of Test Construction: Introduction to Psychometric Design.* London:Routledge

Kline, P. (2016). A handbook of test construction: introduction to psychometric design. London: Routledge.

Lai, M. C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. *Lancet, 383*(9920), 896-910. doi:10.1016/s0140-6736(13)61539-1

Langdon, P. E., Clare, I. C., & Murphy, G. H. (2010). Measuring social desirability amongst men with intellectual disabilities: The psychometric properties of the Self- and Other-Deception Questionnaire—Intellectual Disabilities. *Research in Developmental Disabilities*, *31*(6), 1601–1608. doi: 10.1016/j.ridd.2010.05.001

Lanyon, R. I., & Carle, A. C. (2007). Internal and External Validity of Scores on the Balanced Inventory of Desirable Responding and the Paulhus Deception

Scales. *Educational and Psychological Measurement*, *67*(5), 859–876. doi: 10.1177/0013164406299104

Lees-Warley, G.T. (2014). Deliberate fire-setting by adults with developmental disabilities. [Unpublished doctoral dissertation]. University of Birmingham.

Leite, V. (2015). The MMPI-2 Criminal Offender Infrequency Scale And PAI Negative Distortion Scale: A Comparison Study Of Malingering Scales Within A Forensic Sample [Unpublished doctoral dissertation]. Alliant International University

Lerner, M.D.; Haque, Q.S.; Northrup, E.C.; Lawer, L. & Bursztajn, H.J. (2012). Emerging Perspectives on Adolescents and Young Adults With High-Functioning Autism Spectrum Disorders, Violence, and Criminal Law. Journal of the American Academy of Psychiatry and the Law, 40:177–90. PMID: 22635288

Mandeville-Norden, R., & Beech, A. R. (2009). Development of a Psychometric Typology of Child Molesters. Journal of Interpersonal Violence, 24(2), 307-325. https://doi.org/10.1177/0886260508316479

Mann, R. & Hollin, C. (2010). Self-reported schemas in sexual offenders, The Journal of Forensic Psychiatry & Psychology, 21:6, 834-851, DOI:10.1080/14789949.2010.511240

Martelli, M. F., Nicholson, K., Zasler, N. D., & Bender, M. C. (2012). Assessing and Addressing Response Bias. *Brain Injury Medicine*. doi: 10.1891/9781617050572.0085

Mathie, N. L., & Wakeling, H. C. (2011). Assessing socially desirable responding and its impact on self-report measures among sexual offenders. Psychology, Crime & Law, 17(3), 215-237. https://doi.org/10.1080/10683160903113681

Mazefsky, C. A., & White, S. W. (2014). Emotion regulation: concepts & practice in autism spectrum disorder. *Child and adolescent psychiatric clinics of North America*, *23*(1), 15–24. https://doi.org/10.1016/j.chc.2013.07.002

McCarter, R. J., Walton, N. H., Brooks, D. N., & Powell, G. E. (2009). Effort Testing in Contemporary UK Neuropsychological Practice. The Clinical neuropsychologist, 23(6), 1050-1066. doi:10.1080/13854040802665790

Millon, T. (1994). Millon Clinical Multiaxial Inventory Manual-III manual. Minneapolis, MN: National Computer Systems.

Millon, T., Grossman, S., & Millon, C. (2015). *MCMI-IV*. Bloomington, MN: Pearson.

Mitchell, I. J., Keylock, H., Campbell, N., Beech, A. R., & Kogan, D. (2012). Do Child Molesters Show Abnormal Disgust and Fear of Contamination Reactions? *Psychiatry, Psychology and Law*, *19*(2), 282–294. doi: 10.1080/13218719.2011.561768

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4(1):1. doi: 10.1186/2046-4053-4-1

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … Vet, H. C. D. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. doi: 10.1016/j.jclinepi.2010.02.006

Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Mouridsen, S. E. (2012). Current status of research on autism spectrum disorders and offending. *Research in Autism Spectrum Disorders, 6*(1), 79–86. https://doi.org/10.1016/j.rasd.2011.09.003

Murphy, D. (2003) Admission and cognitive details of male patients diagnosed with Asperger's Syndrome detained in a Special Hospital: comparison with a schizophrenia and personality disorder sample. *The Journal of Forensic Psychiatry & Psychology, 14*(3), 506-524. doi: 10.1080/1478994031000152736

Murphy, B.P. (2010) Beyond the first episode: Candidate factors for a risk prediction model of schizophrenia. *International Review of Psychiatry, 22*(2), 202-223. doi: 10.3109/09540261003661833

Murphy, C., Wilson, C. E., Robertson, D. M., Ecker, C., Daly, E. M., Hammond, N., Mcalonan, G. M. (2016). Autism spectrum disorder in adults: diagnosis, management, and health services development. *Neuropsychiatric Disease and Treatment*, *Volume 12*, 1669–1686. doi: 10.2147/ndt.s65455

National Autistic Society (2019). *Asperger Syndrome*. Retrieved from https://www.autism.org.uk/about/what-is/asperger.aspx

National Autistic Society:NAS (2020a). Autism facts and history. (n.d.). Retrieved from https://www.autism.org.uk/about/what-is/myths-facts-stats.aspx

National Autistic Society (2020b). Women and Girls. (n.d.). Retrieved from https://www.autism.org.uk/professionals/training-consultancy/online/women-and-girls.aspx

National Autistic Society (2020c). Criminal Justice. Retrieved from https://www.autism.org.uk/professionals/others/criminal-justice.aspx

NHS England and NHS Improvement (2019). People with a learning disability, autism or both: Liaison and Diversion Managers and Practitioner resources (2019). Publishing number 000948

NHS England: *NHS Long Term Plan*. (2019). Retrieved May 8, 2020, from https://www.england.nhs.uk/long-term-plan/

Niesten, I. J., Nentjes, L., Merckelbach, H., & Bernstein, D. P. (2015). Antisocial features and "faking bad": A critical note. International Journal of Law and Psychiatry, 41, 34–42. doi: 10.1016/j.ijlp.2015.03.005

Ohlsson, I. M., & Ireland, J. L. (2011). Aggression and offence motivation in prisoners: exploring the components of motivation in an adult male sample. Aggressive Behavior, 37(3), 278-288. https://doi.org/10.1002/ab.20386

Olver, M. E., & Barlow, A. A. (2010). Public attitudes toward sex offenders and their relationship to personality traits and demographic characteristics. *Behavioral Sciences & the Law*, *28*(6), 832–849. doi: 10.1002/bsl.959

O'Mahony, B. M. (2012). Accused of murder: Supporting the communication needs of a vulnerable defendant at court and at the police station. *Journal of Learning*

Disabilities and Offending Behaviour,3(2), 77-84. doi:10.1108/20420921211280060

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. doi: 10.1037/0022-3514.46.3.598

Paulhus D.L. (1986). *Self-Deception and Impression Management in Test Responses.* In: Angleitner A., & Wiggins J.S. (Eds.), *Personality Assessment via Questionnaires*. Heidelberg: Springer

Paulhus, D.L. (1991). *Measurement and control of response bias.* In Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (2010). *Measures of personality and social psychological attitudes*. San Diego: Academic.

Paulhus, D.L. (1998). Paulhus Deception Scales: User Manual. United States: MHS.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement (p. 49–69). Lawrence Erlbaum Associates Publishers.

Pella, R.D. (2009). Evaluation of embedded malingering indices in a non-litigating, relief seeking sample: a partial cross-validation using control, clinical, and derived groups. LSU Doctoral Dissertations. 4024. https://digitalcommons.lsu.edu/gradschool_dissertations/4024

Perinelli, E., & Gremigni, P. (2016). Use of Social Desirability Scales in Clinical Psychology: A Systematic Review. Journal of Clinical Psychology, 72(6), 534-551. doi:10.1002/jclp.22284

Peters, M. J. V., Jelicic, M., Moritz, S., Hauschildt, M., & Jelinek, L. (2013). Assessing the Boundaries of Symptom Over-Reporting Using the Structured Inventory of Malingered Symptomatology in a Clinical Schizophrenia Sample: Its Relation to Symptomatology and Neurocognitive Dysfunctions. Journal of Experimental Psychopathology, 64–77. https://doi.org/10.5127/jep.023811

Randall, P., Carr, A., Dooley, B., & Rooney, B. (2011). Psychological characteristics of Irish clerical sexual offenders. The Irish Journal of Psychology, 32(1-2), 4-13. https://doi.org/10.1080/03033910.2011.610191

Ranganathan, P., Pramesh, C. S., & Buyse, M. (2016). Common pitfalls in statistical analysis: The perils of multiple testing. *Perspectives in clinical research*, *7*(2), 106–107. https://doi.org/10.4103/2229-3485.179436

Ray, J. V., Hall, J., Rivera-Hudson, N., Poythress, N. G., Lilienfeld, S. O., & Morano, M. (2013). The relation between self-reported psychopathic traits and distorted response styles: A meta-analytic review. Personality Disorders: Theory, Research, and Treatment, 4(1), 1–14. doi: 10.1037/a0026482

Rey, A. 1964. *L'examen clinique en psychologie [The clinical examination in psychology]*, Paris: Presses Universitaires de France.

Rogers, R. (2018a). An Introduction to Response Styles. In R. Rogers & S. Bender, Clinical Assessment of Malingering and Deception (4th ed., pp. 3-17). London: Guildford Press.

Rogers, R. (2018b). Detection Strategies for Malingering and Defensiveness. In R. Rogers & S. Bender, Clinical Assessment of Malingering and Deception (4th ed., pp. 18-41). London: Guildford Press.

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). Structured Interview of Reported Symptoms (SIRS) and professional manual. Odessa, FL: Psychological Assessment Resources.

Rogers, R., Vitacco, M. J., & Kurus, S. J. (2010). Assessment of malingering with repeat forensic evaluations: Patient variability and possible misclassification on the SIRS and other feigning measures. *Journal of the American Academy of Psychiatry and the Law*, *38*(1), 108-114.

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. Journal of Consulting and Clinical Psychology, 47(1), 213–215. doi: 10.1037/0022-006x.47.1.213

Sadek, S. A., Daniel, M. R., & Langdon, P. E. Attentional bias toward negative and positive pictorial stimuli and its relationship with distorted cognitions, empathy, and moral reasoning among men with intellectual disabilities who have committed crimes. *Aggressive Behavior,* 1-11. doi:10.1002/ab.21908

Salekin, K. L., Olley, J. G., & Hedge, K. A. (2010). Offenders With Intellectual Disability: Characteristics, Prevalence, and Issues in Forensic Assessment. *Journal of Mental Health Research in Intellectual Disabilities*, *3*(2), 97–116. doi: 10.1080/19315861003695769

Sanderson, S., Tatt, I. D., & Higgins, J. P. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. International Journal of Epidemiology, 36(3), 666–676. doi: 10.1093/ije/dym018

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC

Medical Informatics and Decision Making, 7, 16. doi: http://dx.doi.org/10.1186/1472-6947-7-1

Schlosser, R. W. (2007). Appraising Systematic Reviews. Focus, Technical Brief 17, 1-8. https://doi.org/10.1002/9780470750605.ch5

Schmand, B., Lindeboom, J., Merten, T., & Millis, S. (2005). Amsterdam Short-Term Memory Test. PsycTESTS Dataset. doi: 10.1037/t12622-000

Scragg, P., & Shah, A. (1994). Prevalence of Asperger's Syndrome in a Secure Hospital. *British Journal of Psychiatry,165*(5), 679-682. doi:10.1192/bjp.165.5.679

Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA, the PRISMA-P Group. Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015.349:g7647. doi: 10.1136/bmj.g7647

Sharland, M., & Gfeller, J. (2007). A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. Archives of Clinical Neuropsychology, 22(2), 213–223. doi: 10.1016/j.acn.2006.12.004

Siponmaa L, Kristiansson M, Jonson C, Nydén A, Gillberg C. (2001). Juvenile and young adult mentally disordered offenders: the role of child neuropsychiatric disorders. *J Am Acad Psychiatry Law 29*(4):420-6. PMID:11785613

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic Criteria for Malingered Neurocognitive Dysfunction: Proposed Standards for Clinical Practice and Research. The Clinical neuropsychologist, 13(4), 545-561. doi:10.1076/1385-4046

Smith, G. (2018). *Brief Measures for the Detection of Feigning and Impression Management.* In Rogers, R., & Bender, S. D. (2018). *Clinical assessment of malingering and deception*. New York, NY: The Guilford Press.

Spittlehouse, M. A. K. E. C. (2000). Introducing critical appraisal skills training in UK social services: another link between health and social care? Journal of Interprofessional Care, 14(4), 397-404. doi:10.1080/13561820020003946

Sullivan, J., Beech, A. R., Craig, L. A., & Gannon, T. A. (2010). Comparing Intra-Familial and Extra-Familial Child Sexual Abusers With Professionals Who Have Sexually Abused Children With Whom They Work. International Journal of Offender Therapy and Comparative Criminology, 55(1), 56–74. doi: 10.1177/0306624x09359194

Sweldens, S., Puntoni, S., Paolacci, G., & Vissers, M. (2014). The bias in the bias: Comparative optimism as a function of event social undesirability. *Organizational Behavior and Human Decision Processes*, *124*(2), 229–244. doi: 10.1016/j.obhdp.2014.03.007

Szlachcic, R., Fox, S., Conway, C., Lord, A., & Christie, A. (2015). The relationship between schemas and offence supportive attitudes in mentally disordered sexual offenders. Journal of Sexual Aggression, 21(3), 318-336. https://doi.org/10.1080/13552600.2014.966166

Tager-Flusberg, H. (2007). Evaluating the Theory-of-Mind Hypothesis of Autism. *Current Directions in Psychological Science, 16*(6), 311–315. https://doi.org/10.1111/j.1467-8721.2007.00527.x

Talbot, J. (2009). No One Knows: offenders with learning disabilities and learning difficulties. *Tizard Learning Disability Review*, *14*(1), 18–26. doi: 10.1108/13595474200900004

Tan, L., & Grace, R. C. (2008). Social Desirability and Sexual Offenders: A Review. Sexual Abuse, 20(1), 61–87. https://doi.org/10.1177/1079063208314820

Thomas, M. L., Lanyon, R. I., & Millsap, R. E. (2009). Validation of diagnostic measures based on latent class analysis: A step forward in response bias research. *Psychological Assessment*, *21*(2), 227–230. doi: 10.1037/a0015693

Thomas-Peter, B. A., Jones, J., Campbell, S., & Oliver, C. (2000). Debasement and faking bad on the Millon Clinical Multi-axial Inventory III: An examination of characteristics, circumstances and motives of forensic patients. *Legal and Criminological Psychology,5*(1), 71-81. doi:10.1348/135532500167985

Tombaugh, T.N. (1996). TOMM: Test of Memory Malingering. Pearson Publishers.

Tully, R. J., & Bailey, T. (2017). Validation of the Paulhus Deception Scales (PDS) in the UK and examination of the links between PDS and personality. Journal of Criminological Research, Policy and Practice, 3(1), 38–50. doi: 10.1108/jcrpp-10-2016-0027

Underwood, L., Forrester, A., Chaplin, E., & Mccarthy, J. (2013). Prisoners with neurodevelopmental disorders. *Journal of Intellectual Disabilities and Offending Behaviour,4*(1/2), 17-23. doi:10.1108/jidob-05-2013-0011

Uziel, L. (2010). Rethinking Social Desirability Scales. Perspectives on Psychological Science, 5(3), 243–262. doi: 10.1177/1745691610369465

Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing Violence Risk and Psychopathy in Juvenile and Adult Offenders: A Survey of Clinical Practices. Assessment, 17(3), 377–395. https://doi.org/10.1177/1073191109359587

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Using G-Theory to Enhance Evidence of Reliability and Validity for Common Uses of the Paulhus Deception Scales. Assessment, 25(1), 69–83. doi: 10.1177/1073191116641182

Von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. Behavioral and Brain Sciences, 34(1), 1-16. doi:10.1017/S0140525X10001354

Wakeling, H., & Barnett, G. (2014). The relationship between psychometric test scores and reconviction in sexual offenders undertaking treatment. *Aggression and Violent Behavior*, *19*(2), 138–145. doi: 10.1016/j.avb.2014.01.002

Walczyk, J. J., Sewell, N., & DiBenedetto, M. B. (2018). A Review of Approaches to Detecting Malingering in Forensic Contexts and Promising Cognitive Load-Inducing Lie Detection Techniques. *Front Psychiatry, 9*, 700. doi:10.3389/fpsyt.2018.00700

Wall, G. K., Pearce, E., & McGuire, J. (2011). Are Internet offenders emotionally avoidant? Psychology, Crime & Law, 17(5), 381-401. https://doi.org/10.1080/10683160903292246

Warrington, E. K. (1984). Recognition Memory Test: RMT. Windsor: NFER-NELSON Publ. Co.

Widows, M. R., & Smith, G. P. (2005). Sims: Structured Inventory of Malingered Symptomatology: professional manual. Lutz: Psychological Assessment Resources.

Wiggins, J. S. (1964). Convergences among stylistic response measures from objective personality tests. Educational and Psychological Measurement, 24(3), 551–562. https://doi.org/10.1177/001316446402400310

Williams, K. M., Nathanson, C., & Paulhus, D. L. (2010). Identifying and profiling scholastic cheaters: Their personality, cognitive ability, and motivation. *Journal of Experimental Psychology: Applied*, *16*(3), 293–307. doi: 10.1037/a0020773

Wing, L. (1997). The autistic spectrum. *Lancet, 350*(9093), 1761-1766. doi:10.1016/s0140-6736(97)09218-0

Woodbury-Smith, M., Clare, I., Holland, A. J., Watson, P. C., Bambrick, M., Kearns, A., & Staufenberg, E. (2010). Circumscribed interests and 'offenders' with autism spectrum disorders: a case-control study. *The Journal of Forensic Psychiatry & Psychology, 21*(3), 366-377. doi:10.1080/14789940903426877

World Health Organization. (2018). International classification of diseases for mortality and morbidity statistics (11th Revision).Retrieved 16 February 2020, from https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1136473465

Young, G. (2017). PTSD in Court III: Malingering, assessment, and the law. International Journal Of Law And Psychiatry, 52, 81-102. doi: 10.1016/j.ijlp.2017.03.001

APPENDICES

Appendix A: Search Terms

"response bias*" or bias* or "social* desirab*" or "impression manag*" or lie or lying or decep* or deceiv* or dishonest* or liar or manipulat* or maling* or effort* or fak*

(assess* or psychometric* or questionnaire* or evaluat* or test* or survey*)

(forensic or criminal* or offend* or "mental health" or court)

selected all years but English only

Web of science core collection: - 1900 to 2020

"response bias*" or bias* or "social* desirab*" or "impression manag*" or lie or lying or decep* or deceiv* or dishonest* or liar or manipulat* or maling* or effort* or fak*
AND
(assess* or psychometric* or questionnaire* or evaluat* or test* or survey*) near/3
(forensic or criminal* or offend* or "mental health" or court)

1899 results

Ovid – Psychinfo – 1967 to 2019

Response bias* or bias* or social* desirab* or impression manag* or lie or lying or decep* or deceiv* or dishonest* or liar or manipulat* or maling* or effort* or fak*
AND
(assess* or psychometric* or questionnaire* or evaluat* or test* or survey*) adj3
(forensic or criminal* or offend* or mental health or court)

2814 results

Medline – 1946 to 2020

Response bias* or bias* or social* desirab* or impression manag* or lie or lying or decep* or deceiv* or dishonest* or liar or manipulat* or maling* or effort* or fak*
AND
(assess* or psychometric* or questionnaire* or evaluat* or test* or survey*) adj3
(forensic or criminal* or offend* or mental health or court)

846 results

Embase – 1974 to 2020

Response bias* or bias* or social* desirab* or impression manag* or lie or lying or decep* or deceiv* or dishonest* or liar or manipulat* or maling* or effort* or fak*
AND

(assess* or psychometric* or questionnaire* or evaluat* or test* or survey*) adj3 (forensic or criminal* or offend* or mental health or court)

results – 1290

PubMed – 1974 to 2020

Response bias* or bias* or social* desirab* or impression manag* or lie or lying or decep* or deceiv* or dishonest* or liar or manipulat* or maling* or effort* or fak*
AND
(assess* or psychometric* or questionnaire* or evaluat* or test* or survey*) near/3 (forensic or criminal* or offend* or mental health or court)

results – 44

PsycArticles – 1912 to 2020

Response bias* or bias* or social* desirab* or impression manag* or lie or lying or decep* or deceiv* or dishonest* or liar or manipulat* or maling* or effort* or fak*
AND
(assess* or psychometric* or questionnaire* or evaluat* or test* or survey*) adj3 (forensic or criminal* or offend* or court)

results – 2852

Appendix B: Experts Contacted

Ruth Tully – response received

Theresa Gannon – response received

Peter Langdon – response received

Helen Wakeling – response received

Rebecca Szlachcic – no response

Sample of email sent:

Dear Dr _____,

I am currently undertaking a CPD Doctorate in Forensic Psychology at the University of Birmingham, England. I currently also work as a Consultant Clinical and Forensic Psychologist with young people and adults as well as undertake medico-legal reports for court.

As part of my thesis I am undertaking a systematic literature review on the assessment of response bias in forensic contexts over the last 10 years. As part of my systematic search I came across your *articles/research* _____. I am emailing you, as I have identified that you are an expert in this field and wondered whether you may have any articles or studies, either published or unpublished, that you would be so kind as to forward me. I am hoping to include all relevant research, fitting my inclusion criteria, on this topic. If this is possible, I would be very grateful.

Kind regards,

Marilyn Sher

Appendix C: Pre-defined Inclusion / Exclusion Criteria Form

Study characteristics	Inclusion criteria	Criteria met?
Population	Forensic / court referrals (e.g. all those with convictions in prison or hospital; probation samples; court referrals eg pre-trial assessment/assessment for sentencing Male and female Adult and adolescent	Yes ☐ No ☐ Unclear ☐
Intervention	Self-report response bias measure used	Yes ☐ No ☐ Unclear ☐
Comparator	Type of response bias measure named	Yes ☐ No ☐ Unclear ☐
Outcome	Response bias measure outcome reported	Yes ☐ No ☐ Unclear ☐
Study design	Experimental Quasi-experimental Observational	Yes ☐ No ☐ Unclear ☐
	Published journal article Theses/dissertations Conference presentations Unpublished data from experts	Yes ☐ No ☐ Unclear ☐
Other	UK	Yes ☐ No ☐ Unclear ☐
	Published/reported 2009-2019	Yes ☐ No ☐ Unclear ☐
	Published in English	Yes ☐ No ☐
Criteria for Quality review met?	***Yes*** ☐ ***No*** ☐	***Reasons:***

Appendix D: Quality Review Form (adapted from CASP templates)

Author: **Year:** **Title:** **Source:**			
Question		**Quality score/information**	**Comments**
1.	What is the study design?	☐ Cohort ☐ Case Control ☐ Experimental ☐ Quasi-experimental	
A) Study design – Domain average (score / 2) =			
2.	Did the authors use an appropriate design and method to answer their questions?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
3.	Are the cases precisely defined?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
B) Sampling – Domain average for all items (score / 5) = **Domain average excluding Q7&Q8 (score / 3) =**			
4.	Was there an established reliable system for selecting all the cases?	☐ Yes (2) ☐ Partially (1) ☐ No (0)	

		☐ Unclear (?) ☐ N/A (na)	
5.	Was there a sufficient number of cases selected/appropriate to the aims of the study?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
6.	Was there a power calculation?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?) ☐ N/A (na)	
Is there a control/comparison group?		☐ Yes ☐ No	If yes, go to Q7. If no, go to Q9
7.	Were the control group representative of a defined population?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?) ☐ N/A (na)	
8.	Was there a sufficient number of controls selected?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?) ☐ N/A (na)	
C) Analysis – Domain average (score / 4) =			
9.	Were the measurement methods similar for all groups?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	

		☐ N/A (na)	
10.	Do the measures truly reflect what they are supposed to measure? (validated)	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
11.	Was the inter-rater reliability of the measures ascertained and reported?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
12.	Is the analysis appropriate to the design?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
D) Confounding variables – Domain average (score / 2) =			
13.	Have the authors taken account of the potential confounding factors in the design and/or analysis?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
14.	Are the results adjusted for confounding?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
E) Applicability/value of research – Domain average (score / 2) =			
15.	Are the study participants sufficiently representative of the local population?	☐ Yes (2) ☐ Partially (1) ☐ No (0) ☐ Unclear (?)	
16.	Do the results of this study	☐ Yes (2)	

	fit with other available evidence?	☐ Partially (1) ☐ No (0) ☐ Unclear (?)	

Global domain score: **(Maximum 10)**

Quality Score (Percentage)	Methodological Quality	Risk of Bias Rating
70%- 100%	Strong Methodological Quality	Low risk of bias
40% - 70%	Intermediate Methodological Quality	Moderate risk of bias
0% to 40%	Weak Methodological Quality	High risk of bias

Above risk of bias rating categories adapted from Lees-Warley (2014)

Appendix E: Excluded Paper

Mandeville-Norden, R., & Beech, A. R. (2009). Development of a Psychometric Typology of Child Molesters. *Journal of Interpersonal Violence*, *24*(2), 307-325. https://doi.org/10.1177/0886260508316479

The above paper was excluded as the required data was not reported on. The authors used the BIDR, but did not provide scores. In addition, those in the sample who had scored above cut-off were excluded from the study.

Appendix F: Data Extraction Form for Quality review

Author:	
Year:	
Title:	
Source:	
Questions:	
1. What is the study design?	
2. Did the authors use an appropriate design and method to answer their questions?	
3. Are the cases precisely defined?	
4. Was there an established reliable system for selecting all the cases?	
5. Was there a sufficient number of cases selected/appropriate to the aims of the study?	
6. Was there a power calculation?	
7. Were the control group representative of a defined population?	
8. Was there a sufficient number of controls selected?	
9. Were the measurement methods similar for all groups?	
10. Do the measures truly reflect what they are supposed to measure? (validated)	
11. Was the inter-rater reliability of the measures ascertained and reported?	
12. Is the analysis appropriate to the design?	
13. Have the authors taken	

account of the potential confounding factors in the design and/or analysis?	
14. Are the results adjusted for confounding?	
15. Are the study participants sufficiently representative of the local population?	
16. Do the results of this study fit with other available evidence? In what way?	
17. Quality review score:	
18. Samples	
19. Sample size	
20. Population characteristics	
21. Recruitment procedures	
22. Measures used	
23. Internal consistency, reliability and validity of measures	
24. Details of comparative conditions	
25. Method of statistical analysis	
26. Magnitude and direction of results	
27. Summary of findings	
28. Strengths	
29. Limitations	

APPENDIX G: Research Flyer

Do you or someone you know have a diagnosis of High Functioning Autism / Asperger's? Please help with some important research!

I am a Psychologist undertaking doctoral research at the University of Birmingham and am specifically looking at the response styles of adults with a diagnosis of High Functioning Autism/Asperger's. I am asking those with this diagnosis to fill out some simple questionnaires on-line. An example question is: "*talking to people has felt to much for me*". You would rate this question on a scale from 0 to 4, with 0 being '*not at all*' and 4 being '*most or all of the time*'.

The current study aims to address some of the gaps in the understanding of response styles amongst an Autistic client group, as well as provide evidence based data for certain assessments commonly used. This is important because some questionnaires used as part of assessments are not designed for people with ASD.

If you are 18 or over and have a diagnosis of High Functioning Autism / Asperger's, please email me to express an interest in participating. The survey will take no more than 30 minutes. My email address is: [redacted]. When you email me you will be given a link and password to undertake the survey on-line. I will then delete your email from my system to ensure confidentiality.

All eligible respondents who complete the survey will be entered into a prize draw for a £50 Amazon voucher.

I appreciate you taking the time to read this and I look forward to hearing from you. Feel free to circulate this message to anyone you feel may be eligible.

Kind regards
Marilyn Sher
University of Birmingham, United Kingdom

APPENDIX H: List of Demographic Information Collected

1. Age

2. Sex
Male
Female
Other (please specify)

3. Gender

4. Work status
Employed
Unemployed
Student
Retired

5. Education level
No formal education
GCSE/O-Level
A Level/Senior high
Vocational (equivalent to A level)
Undergraduate degree
Postgraduate degree

6. Country you are currently living in

7. Any other diagnoses (other than ASD/Asperger's)

APPENDIX I: Information Sheet Web Page

Thank you for following the link.

You are being invited to take part in a research study. Before you decide to participate, it is important for you to understand the purpose of the research and what your participation will involve. Please take your time to read through the following information carefully, and discuss it with others if you wish.

The purpose of the research is to understand the ways people on the Autistic Spectrum respond to certain types of questionnaires. The research is important because it will give people using the questionnaires a more accurate understanding of what scores mean when used with people who have Autism. The research may also influence certain guidance that has been drawn up by the British Psychological Society on use of certain questionnaires.

If you choose to participate in the research you will be required to electronically complete 3 questionnaires that are officially published. The manuals for the questionnaires suggest that they should not take any longer than 30 minutes all together. The questionnaires will be active from December 2019 to the end of March 2020, giving participants three months to complete the questionnaires. Once the data has been collected it will be analysed using a computer package known as SPSS and may be used for publication.

Participation in this research is entirely voluntary. If you decide to take part you will need to click 'continue' to the next page where you will be presented with a consent form. By pressing 'accept' you are consenting to take part in the research. Before you are automatically directed to the questionnaires, you will need to enter the password. During completion of the questionnaire, if at any time you wish to withdraw, on each page there will be an option to press the 'quit' button. Any information will be automatically erased. Once you have completed the questionnaires, there will be a final page, which will display details of relevant support services. In addition, a unique user ID number will be generated. Please keep a note of this user ID number as you may need it if you decide to withdraw. You will also be asked if you would like to receive a summary of the results by email and whether you want to be entered into the prize draw for a £50 Amazon voucher. All your data will be kept confidential and anonymous, in line with the Data Protection Act 2018. If you have any questions about the study and your participation, prior to giving consent, feel free to email the researcher (Marilyn Sher) using the following email address: [redacted]. You can also contact Marilyn Sher if you have any queries or issues relating to the conduct of the study. If you would prefer to contact someone independent of the study, you can email Dr Caroline Oliver using the following email address: [redacted]

Should you wish to withdraw after completion of the questionnaires, you will have up until 15th April 2020, two weeks after the questionnaires close, to decide. A reason for withdrawal is not required and there are no penalties. The procedure for withdrawal is to contact the researcher via email and quote you user ID number. Your data will then be immediately erased.

A Research Ethics Committee reviews all proposals for research using human participants before they can proceed. The University of Birmingham Research Ethics Committee has reviewed this proposal.

The questionnaires will require you to think about your day-to-day experiences and feelings. If you experience any difficulties as a result of this, please refer to the list below showing relevant organisations that can provide support and advice:

Samaritans

Call: 116 123
Website: https://www.samaritans.org/how-we-can-help/contact-samaritan/

MIND

Call: 0300 123 3393
Text: 86463
Website: https://www.mind.org.uk

111

Call: 111

National Autistic Society (NAS)

Website: https://www.autism.org.uk/services/helplines/main/contact.aspx

I would like to take this opportunity to thank you for taking the time to read the information sheet.

QUIT

CONTINUE

APPENDIX J: Consent Form Web Page

I have understood the details of the research as explained to me by the researcher, and confirm that I have consented to act as a participant. I further confirm that I am 18 or older and have a diagnosis of Autism (high functioning/Asperger's).

I understand that my participation is entirely voluntary, the data collected during the research will not be identifiable, and I have the right to withdraw from the study up to two weeks after the questionnaires close without any obligation to explain my reasons for doing so. I am aware any data collected will be destroyed.

I further understand that the data I provide may be used for analysis and subsequent publication.

APPENDIX K: Demographic Characteristics of Participants

Characteristic		
Age (n=72; range = 18-69)	34.01 (mean)	13.19 (SD)
Sex (n=73)	***Frequency***	***Percentage***
Male	39	53.4%
Female	33	45.1%
Other	1	1.4%
Employment Status (n=71)		
Employed	32	43.8%
Unemployed	21	28.8%
Student	14	19.2%
Retired	4	5.5%
Level of Education (n=71)		
No formal qualifications	7	9.6%
GCSE or equivalent	10	13.7%
A Level / Senior High School	8	11%
Vocational	6	8.2%
Undergraduate Degree	29	39.7%
Postgraduate Degree	11	15.1%
Country currently living in (n=72)		
UK	34	46.6%
Germany	2	2.7%
Switzerland	1	1.4%
Sweden	2	2.7%
Norway	1	1.4%
USA	19	26%
Australia	2	2.7%
Canada	3	4.1%
Estonia	1	1.4%
Portugal	1	1.4%
Belgium	1	1.4%
Finland	1	1.4%
France	1	1.4%
El Salvador	1	1.4%
Austria	1	1.4%
Spain	1	1.4%
Any other diagnoses (n=73)		
Anxiety Disorder	5	6.8%

Characteristic		
Depressive Disorder	16	21.9%
ADHD	5	6.8%
Trauma related	1	1.4%
Personality Disorder	2	2.7%
Eating Disorder	2	2.7%
No other disorder	42	57.5%