The reliability and validity of mental state reasoning tasks for use in adolescents and adults

A systematic review

Bremner, Georgie

Awarding institution:
King's College London

Volume I:

Systematic Review

Empirical Research Project

Service Related Project

Georgie Bremner

Thesis submitted in partial fulfilment of the degree of

Doctorate in Clinical Psychology

Institute of Psychiatry, Psychology & Neuroscience

Kings College London

May 2016

Acknowledgements

I would like to thank my supervisors Dr Jennifer Lau and Dr Victoria Pile for their consistent and thoughtful support with this project. They were always a pleasure to work with and I am very grateful to have had their guidance. Thank you to the amazing teachers Kate Tapper and Joe Robinson, who were incredibly generous in giving me their time as well as their AS Psychology students. I am also so appreciative to each and every student who completed the study – I doubt it fulfilled their expectations of having a psychologist come to visit…!

I would also like to thank Dr Rumina Taylor for her supervision of the service evaluation project, as well as to all my placement supervisors who reviewed a case study.

Thank you to Carole and Mark for being so supportive during that final push in the ASB and to everyone in our lovely cohort who have continuously blown me away in showing so much genuine kindness and support.

Finally thank you to my wonderful family for always providing a happy base for recharging, and to Ben for always being there and believing in me.

Contents

Systematic Review

The reliability and validity of mental state reasoning tasks for use in adolescents and adults: A systematic review 3

Empirical Research Project

Mental state reasoning in adolescent social anxiety 59

Service Related Project

Facilitating an acceptance and commitment therapy group in an acute inpatient setting: a service evaluation 128

Systematic Review

The reliability and validity of mental state reasoning tasks for use in adolescents and adults: A systematic review

Supervised by Dr Jennifer Lau

& Dr Victoria Pile

Abstract

Mental state reasoning (MSR) is a skill that enables individuals to understand other people's mental states and is a key aspect of social cognition. There has been increasing interest in this construct, as subtle MSR impairments have been associated with number of different clinical populations. However, little is known about the reliability or validity of the instruments used to assess this skill, which limits their utility in both research and clinical settings. This review set out to identify and evaluate the measurement properties of MSR tasks for use in populations beyond childhood. A systematic search was performed using four databases, for articles concerning the development or evaluation of the measurement properties of a MSR task. Both the methodological quality of the studies and the quality of the instrument measurement properties were systematically evaluated using validated criteria. The search strategy returned a total of 4523 articles, of which 18 studies, evaluating 16 different MSR instruments, were included in the review. The majority of studies were found to be of poor or fair methodological quality, which means that evidence regarding the majority of the instrument measurement properties is limited or indeterminate. Only three instruments, the Social Attribution Task (SAT), the Virtual Assessment of Mentalising Ability (VAMA) and the Reading the Mind in Films task (RMIF) demonstrated adequate results on more than one measurement property. This review therefore highlights the need for larger, well-designed studies to assess the measurement properties of currently available MSR tasks.

Contents

1. Introduction 7

1.1 Aims and Objectives 11

2. Methods 11

2.1 Search Strategy 11

2.2 Selection criteria 12

Inclusion Criteria 12

Exclusion criteria 12

2.3 Assessment of methodological quality of the studies 12

2.4 Measurement properties 13

2.5 Quality assessment of measurement properties 14

2.6 Best evidence synthesis 16

3. Results 16

3.1 General description of identified studies 18

3.2 General description of instruments 18

3.3 Findings of the review 19

3.3.1 The Awkward Moments Test (AMT) 29

3.3.2 Bell Lysaker Emotion Recognition Test (BLERT) 30

3.3.3 Child and Adolescent Social Perception Measure (CASP) 31

3.3.4 Empathic accuracy test (EAT) 32

3.3.5 False Belief Sequencing Task (FBST) 33

3.3.6 Frith-Happe Animations Task (FH-A) 33

3.3.7 Hinting Task 34

3.3.8 Movie for the Assessment of Social Cognition (MASC) 35

3.3.9 Reading the Mind in the Films Task (RMIF) 36

3.3.10 Social Attribution Task: original and multiple choice version (SAT and SAT-MC) 37

3.3.11 Strange Stories (SS) 39

3.3.12 Short Story Task (SST) 39

3.3.13 The Awareness of Social Inference Test (TASIT) 40

3.3.14 ToM Stories 41

3.3.15 Virtual Assessment of Mentalising Ability (VAMA) 42

4. Discussion 43

4.1 Content Validity and Ecological Validity 44

4.2 Construct Validity **45**
4.3 Reliability **46**
4.5 Responsiveness **47**
4.6 Interpretability **47**
4.7 Limitations **47**
4.8 Clinical applications **48**
4.9 Future directions **49**
4.10 Conclusions **50**
5. References **51**
6. Appendix **56**

List of Tables

Table 1 Measurement property domains assessed in the current review 14
Table 2: Quality criteria for measurement properties (adapted from Terwee et al., 2007). 15
Table 3: Levels of evidence for the quality of the measurement property. 16
Table 4: Study characteristics 20
Table 5: Measure characteristics 22
Table 6: Methodological quality and quality criteria ratings per study and instrument property 26
Table 7: Best evidence synthesis of the measurement properties per task 28

List of Figures

Figure 1: Flowchart of search results and paper selection 17

1. Introduction

Mental state reasoning (MSR) is an aspect of social cognition that refers to the capacity to recognise one's own, as well as other peoples' mental states, including thoughts, desires, intentions and emotions. This ability is critical for predicting and making sense of others' behaviour (and therefore modulating one's own behaviour accordingly), as well as for efficient communication, social learning, and empathic concern (Baimel, Severson, Baron, & Birch, 2015). MSR is a multidimensional construct, with a number of inter-related terms used to describe this skill, including 'theory of mind', 'mindreading', 'perspective-taking' and 'mentalising', and is part of a wider concept of 'metacognition' (Kuhn, 2000). There is growing understanding that impaired MSR may underpin the social difficulties seen in several clinical populations, which can result in reduced community functioning and poor quality of life (Maat, Fett, & Derks, 2012; Van Donkersgoed et al., 2014). Indeed, because of its potential role in the development or maintenance of mental health conditions, MSR has become an important target for treatment (e.g. Bateman & Fonagy, 2010; Moritz, Veckenstedt, Randjbar, Vitzthum, & Woodward, 2011; Penn, Roberts, Combs, & Sterne, 2007). In this way, valid and reliable measures of MSR are needed to both empirically test the hypothesised role of MSR impairments, as well as to assess the outcome of such interventions. This review sets out to investigate what tasks are currently being used to assess MSR ability and to evaluate their reliability and validity for use in populations beyond childhood.

Both the conceptualisation and measurement of MSR has developed significantly over recent decades. Early measures investigated the development of MSR in the form of 'false belief' understanding – that is, the understanding that other people can have beliefs that are different from our own. These tasks have been valuable in demonstrating the early development of MSR, which begins with first-order understanding (identifying the mental state of another person), around 3-5 years of age, followed by second-order understanding (identifying what one person thinks about another person's thoughts), around 6-7 years of age (Callaghan et al., 2005; Wellman, Cross, & Watson). These false belief tasks typically focus on the attribution of beliefs and intentions to characters in a story (e.g. Wimmer & Perner, 1983), and because they were originally designed for use in children, they tend to suffer from ceiling effects in older, more able individuals. Therefore, in order to assess MSR abilities in individuals beyond childhood, more sophisticated measures are needed.

More recently, researchers have developed 'advanced' measures of MSR, which have demonstrated that the more difficult aspects of MSR, such as understanding irony, sarcasm and metaphor, continue to develop into adolescence and early adulthood (Kaland, Smith, & Mortensen, 2007; Keulers, Evers, Stiers, & Jolles, 2010; Sebastian et al., 2012). Additionally, they have been able to demonstrate more subtle MSR impairments in a number of clinical populations beyond the autism spectrum disorders (ASDs), including schizophrenia (Brune, 2005) conduct disorder (Happé & Frith, 1996) borderline personality disorder (BPD; Sharp & Vanwoerden, 2015) traumatic brain injury (TBI; McDonald et al., 2003) and depression (Harkness, Jacobson, Duong, & Sabbagh, 2010). Different types of MSR impairment have been associated with different disorders; for example, reduced MSR in ASD, but 'hyper-MSR' impairments in schizophrenia (e.g. Langdon & Brock, 2008) and BPD (e.g. Sharp & Vanwoerden, 2015). Thus, more sophisticated MSR assessments enable researchers to understand what specific aspects of MSR might be impaired in individuals with different mental health problems, which may then provide insight into their particular social functioning deficits. This research can then inform the development of interventions that target these MSR errors.

As already mentioned, MSR reasoning is a multi-dimensional construct, which can be differentiated with respect to types of mental states (desire, belief, knowledge, intention and emotion), which can be further differentiated in terms of valence (positive, neutral or negative). MSR can also be controlled and explicit, or automatic and implicit, as well as self or other focused. This multi-dimensionality is reflected more in the newer more advanced measures, which assess MSR ability beyond understanding of only desire and belief. Indeed, some researchers, particularly in the area of cognitive neuroscience, now make the distinction between 'cognitive MSR' (understanding thoughts and intentions) and 'affective MSR' (understanding emotions). Affective MSR can be further distinguished from affective *empathy*, which is considered as the ability to 're-feel' the emotions of others, rather than to understand/identify them per se (Shamay-Tsoory, 2011). Such distinctions in terminology have arisen from individuals demonstrating differential abilities across these different domains, in addition to evidence of at least partly different neural correlates mediating the different subcomponents (Kalbe et al., 2010).

Tasks not only differ with respect to what aspects of MSR they are measuring, but they also vary greatly with respect to their mode of presentation. For instance, a number of measures assess understanding of concepts such as white lies, irony, misunderstandings and faux pas, which often involve short stories or cartoons that are presented visually and read by either the

participant or the experimenter. Participants are then asked about what the characters were thinking, intending or feeling - examples include the Strange Stories task (Happe, 1994) and the Faux Pas Test (Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 1999). Other tasks have used video stimuli involving the attribution of mental states to animated objects, such as the Social Attribution Task (Klin, 2000) and the Frith-Happe Animations task (White, Coniston, Rogers, & Frith, 2011). In more recent years, a number of 'dynamic' tasks have been developed to be more ecologically valid, which have used natural interactions (e.g. Roeyers, Buysse, Ponnet, & Pichal, 2001) acted video stimuli (e.g. Dziobek et al., 2006), as well as virtual reality technology (e.g. Canty, Neumann, Fleming, & Shum, 2015).

MSR tasks also vary with respect to what, and how much, information they provide to participants, in order for them to make an accurate MSR judgement. Most tasks tend to provide a degree of contextual information, alongside basic information regarding the character(s) to which the participant has to attribute a mental state (Achim, Guitton, Jackson, Boutin, & Monetta, 2013). This distinguishes MSR tasks from social cue recognition tasks, which typically present acontextual information (e.g. only facial expression), or from social knowledge tasks, which usually present contextual information without relating to a particular person or character (e.g. a prototypical situation). Therefore, for the purpose of this review a MSR task was considered only if it measured the ability to attribute mental states to at least one character or person, drawing upon both person-based and context-based information. For this reason, tasks such as the Reading the Mind in the Eyes task (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) were not included, as this presents only perceptual information (photographs of the eye region), without contextual information, and would therefore be considered a social cue recognition test by the above criteria. We were also specifically interested in performance-based tasks, which have a degree of objectivity in terms of having a 'correct' answer. For this reason, self-report questionnaires and interview-based assessments (e.g. that considered the number of MSR terms used during a dialogue) were not considered.

Despite the varied use MSR tasks in clinical and developmental research, little is known about their validity and reliability (Ahmadi, Jalaie, & Ashayeri, 2016; Blijd-Hoogewys, van Geert, Serra, & Minderaa, 2008; Pinkham, Penn, Green, & Harvey, 2015). This is problematic in terms of the strength of the conclusions that can be drawn from studies that use these tasks. Moreover, clinicians and researchers should be making evidence based decisions about what measures to use for any particular population and purpose. Ideally, for a task to be considered as having adequate measurement properties, it should be assessed on three different quality domains:

reliability, validity and responsiveness. Reliability is the extent to which scores for individuals (who have not changed) are the same for repeated measurement, which includes across different items (internal consistency), over time (test-retest), or across different raters (inter-rater). Reliability also includes measurement error, which is the systematic and random error of a score that is not attributed to true changes in the construct to be measured. Validity is the degree to which an instrument measures the construct(s) that it purports to measure, and has three domains: content validity, construct validity (which includes structural validity, hypothesis testing and cross-cultural validity) and criterion validity (the degree to which scores are an adequate reflection of a 'gold-standard') (Mokkink et al., 2010c). Responsiveness is the third measurement property domain, which is the ability of an instrument to detect changes over time, and is thus a useful property for intervention research. The current review will be assessing the included MSR tasks on these separate domains, with the exception of cross-cultural validity (as an exclusion criteria is studies using non-English versions of measures), criterion validity (as there is currently no consensus as to what the 'gold standard' instrument for MSR assessment is) and measurement error (as there is no parameter of measurement error for nominal scales, which is the format of the included MSR tasks).

The few reviews to date that have considered the measurement properties of MSR instruments (e.g. Henry, Cowan, Lee, & Sachdev, 2015 and Ahmadi et al., 2016) have lacked an adequate rating tool to critically evaluate the methodological quality of the included studies. This is important, as only studies of adequate methodological quality can lead to appropriate conclusions about the measurement properties of an instrument. Therefore, this review used the only validated quality assessment tool that has been specifically developed for the methodological evaluation of measurement properties - the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN; Mokkink et al., 2010c). Additionally, to assess the quality of the measurement properties of the included MSR tasks, criteria established by Terwee et al., (2007) was used. Both the instrument measurement property criteria (Terwee et al., 2007) and the methodological quality criteria (COSMIN checklist) also considered the 'interpretability' of measures, which is the degree to which performance on the measure can be ascribed qualitative meaning. This requires authors to provide information about clinically meaningful differences in scores between subgroups as well as floor and ceiling effects. This is particularly relevant for measures of MSR, which may be prone to ceiling effects in adolescent and adult populations. Finally, although not a property assessed by either the measurement property or methodological quality criteria, this review also considered the ecological validity of the included MSR tasks. This is because it has been suggested that problems with MSR are most

pronounced in complex everyday-life settings (Dziobek, 2012) and so how closely a task approximates the demands of real-life is important.

In summary, there has been increasing evidence in recent years that MSR is a skill that continues to develop beyond childhood, and may be subtly impaired in a number of mental health conditions. Indeed, interventions have already been developed to specifically target impaired MSR (e.g. Fonagy & Bateman, 2010). Thus, reliable and valid MSR instruments have an important role in shedding light on how MSR may be impaired in different mental health problems (and how this may impact social functioning) as well as in the evaluation of interventions.

1.1 Aims and Objectives

Since there is little information about the reliability and validity of MSR instruments, particularly for use in older individuals, this review aims to identify what measures are currently being used to assess MSR ability in adolescents and adults, and to systematically evaluate the available evidence regarding their measurement properties.

2. Methods

2.1 Search Strategy

A systematic literature search (11th November 2015) was conducted to identify eligible studies, using search terms related to MSR (mentali?ing OR mentali?ation OR "theory of mind" OR mind$reading OR "mental state attribution" OR "mental state decoding" OR "mental state reasoning" OR "perspective taking") combined with ('AND') terms related to psychometric properties ("coefficient of variation" OR "ceiling effect" OR " internal consistency" OR reproducib* OR valid* OR reliab* OR psychometric OR specificity OR sensitivity). The psychometric property terms were obtained from a highly sensitive search filter developed by Terwee et al., (2009) specifically for finding studies on measurement properties. The databases PsychINFO, Medline, Web of Science (Core collection) and PubMed were used, limited to journal articles with adolescent or adult human subjects, where such limits were available. Reference lists of included studies were screened to identify additional relevant studies. The search was re-run on 24th April 2016 to incorporate any relevant newly published papers.

2.2 Selection criteria

Inclusion Criteria

Studies were included if they satisfied all of the following inclusion criteria:

- The study is published in a peer-reviewed journal article in English, after 1980.
- The mean age of study participants is over 12.
- The study participants are of normative IQ (IQ>70).
- The study aim is to develop or evaluate the measurement properties of a MSR instrument (which does not include alternative forms designed for re-test).
- The MSR measure fits the following definition: The instrument tests the ability to attribute mental states (e.g. thoughts or feelings) to a character whereby information is presented about at least one character and some information about the interpersonal and physical context, for example, the presence of other individuals or the specific location the character is in.

Exclusion criteria

Studies were excluded if they fulfilled one of the following exclusion criteria:

- The study uses a measure that was not delivered in the English language.
- The study is not primary research (e.g. a review).
- The study uses a measure developed for a highly specific population group in a way that would limit its applicability to any other population.

Any articles where the decision to include them was uncertain were discussed with two other experienced researchers in the field (JL and VP) and an agreement was reached whether or not to include the article.

2.3 Assessment of methodological quality of the studies

The methodological quality of the studies was evaluated using the COSMIN checklist (Mokkink et al., 2010c), which evaluates the methodological quality of studies on validity (content validity, construct validity and cross-cultural validity), reliability (internal consistency, reliability and measurement error) and responsiveness. A methodological quality score is determined per measurement property, and is rated on a four-point scale (poor, fair, good or excellent), considering aspects such as the adequacy of the sample size used, whether the most appropriate

statistical methods were used and whether data on missing items was described. The overall methodological quality of each measurement property is determined by taking the lowest rating achieved from any of the items ('worst score counts') (Terwee et al., 2012). In the current review, five of the included articles (28%) were randomly selected for double rating by an independent assessor. Inter-rater reliability using the COSMIN checklist, in terms of scoring 'poor', 'fair', 'good' or 'excellent' was shown to be excellent (Cohen's weighted kappa = .91). The full COSMIN checklist is available online (www.cosmin.nl).

2.4 Measurement properties

The current review investigated the methodological quality of the following measurement properties: internal consistency, reliability (test-retest, inter-rater), content validity and construct validity (including structural validity and hypothesis testing). Content validity was assessed if the study referred to how the instrument was developed, which was usually only the case for the original development articles. Responsiveness was not included, as this was not assessed by any of the studies. Although not measurement properties, both the interpretability and ecological validity were also evaluated, as these are considered important characteristics of MSR tasks. Definitions of all the instrument characteristics evaluated in the current review are presented in Table 1.

Table 1 Instrument characteristics assessed in the current review

Validity	**Reliability**	**Interpretability**
1) Content: The degree to which the content of the instrument is an adequate reflection of the construct to be measured. 2) Construct: *Structural validity:* Refers to the instrument's structure, usually investigated via factor analysis. *Hypothesis testing:* The degree to which the scores of an instrument are consistent with theoretically derived hypotheses (for instance with regard to relationships to scores of other instruments, or differences between relevant groups). 3) Ecological validity: The extent to which results can be applied to real-life situations outside of research settings (or how closely the task approximates the demands of real-life MSR)	1) Internal Consistency: The extent to which items in a measure are correlated (expressed by Cronbach's α) 2) Reliability: The degree to which repeated measures in stable subjects provide similar results both over time (re-test reliability) and across raters (inter-rater)	The degree to which qualitative meaning can be ascribed to scores on a measure: includes information about clinically meaningful differences in scores between subgroups as well as floor and ceiling effects

2.5 Quality assessment of measurement properties

In order to determine the quality of measurement properties of the MSR instruments, the rating system proposed by Terwee et al., (2007) was used (see Table 2). For each instrument property a criterion is defined for a positive (+), negative (-) or indeterminate (?) rating, depending on the outcome of the study under review.

Table 2: Quality criteria for measurement properties (adapted from Terwee et al., 2007).

Property	**Rating**	**Quality criteria**
Reliability		
Internal consistency	+	(Sub)scale unidimensional AND Cronbach's α(s) ≥ .70 and < .95
	?	Dimensionality not known OR Cronbach's α not determined
	−	(Sub)scale not unidimensional OR Cronbach's α(s) < .70 or > .95
Reliability	+	ICC/weighted Kappa ≥ .70 OR Pearson's r ≥ .80
	?	Neither ICC/weighted Kappa, nor Pearson's r determined
	−	ICC/weighted Kappa < .70 OR Pearson's r < .80
Validity		
Content validity	+	A clear description is provided of the measurement aim, the target population, the concept(s) being measured, and the item selection. Target population and experts in the field were involved in the development process.
	?	Not enough information available
	−	No clear description is provided of the measurement aim, the target population, the concept(s) being measured, and the item selection, or target population and experts in the field were NOT involved in the development process.
Construct validity		
Structural validity	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	−	Factors explain < 50% of the variance
Hypothesis testing	+	At least 75% of the results are in accordance with the hypotheses AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	−	< 75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
Interpretability		
Interpretability	+	+ Means and SD scores presented for subgroups
	?	Doubtful design of method
Floor and ceiling effects	+	≤15% of the respondents achieved the highest or lowest possible scores
	?	Doubtful design or method
	-	> 15% of the respondents achieved the highest or lowest possible scores

Note: ICC=intraclass correlation coefficient; + = positive rating, ? =indeterminate rating, - = negative rating.

2.6 Best evidence synthesis

To summarise all the evidence on the different measurement properties of the different instruments, the results were combined to give one overall 'best evidence' rating, taking into account the number of studies, the methodological quality (according to COSMIN criteria) and the consistency of the results across different studies (where applicable). How the best evidence ratings were determined is summarised in Table 3.

Table 3: Levels of evidence for the quality of the measurement property.

Level	Rating	Criteria
Strong	+++ or − − −	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++ or − −	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or −	One study of fair methodological quality
Conflicting	+/−	Conflicting findings
Unknown	?	Only studies of poor methodological quality

3. Results

The search process can be seen in the PRISMA Flowchart in Figure 1. The two searches resulted in 2,410 articles in PsychINFO, 1,183 articles in Medline and 930 articles from Pubmed. Therefore, the total number of articles returned was 4,523, which was reduced to 3277 after duplicates were removed. Titles and abstracts were then screened for relevance and 122 papers were retained for further inspection of full-text versions. Seventy-six articles were then removed for not meeting inclusion criteria: the sample did not have a mean age of above 12 (n = 16), the sample were not all of normative IQ (n = 3), the study was not explicitly a development or validation study (n = 15) or the MSR task did not fit the required definition (n = 42). Twenty-one were removed as they met exclusion criteria: measures were delivered in a non-English language (n = 14), the study was a review article (n= 6) and one study used a measure developed for a highly specific population group[1]. Eleven studies met multiple inclusion /exclusion criteria. A reference search of the remaining articles (n = 14) identified four relevant studies, bringing the total number of included papers to 18.

[1] One paper was excluded as it used a measure specifically designed for medical students, depicting doctor/patient consultations.

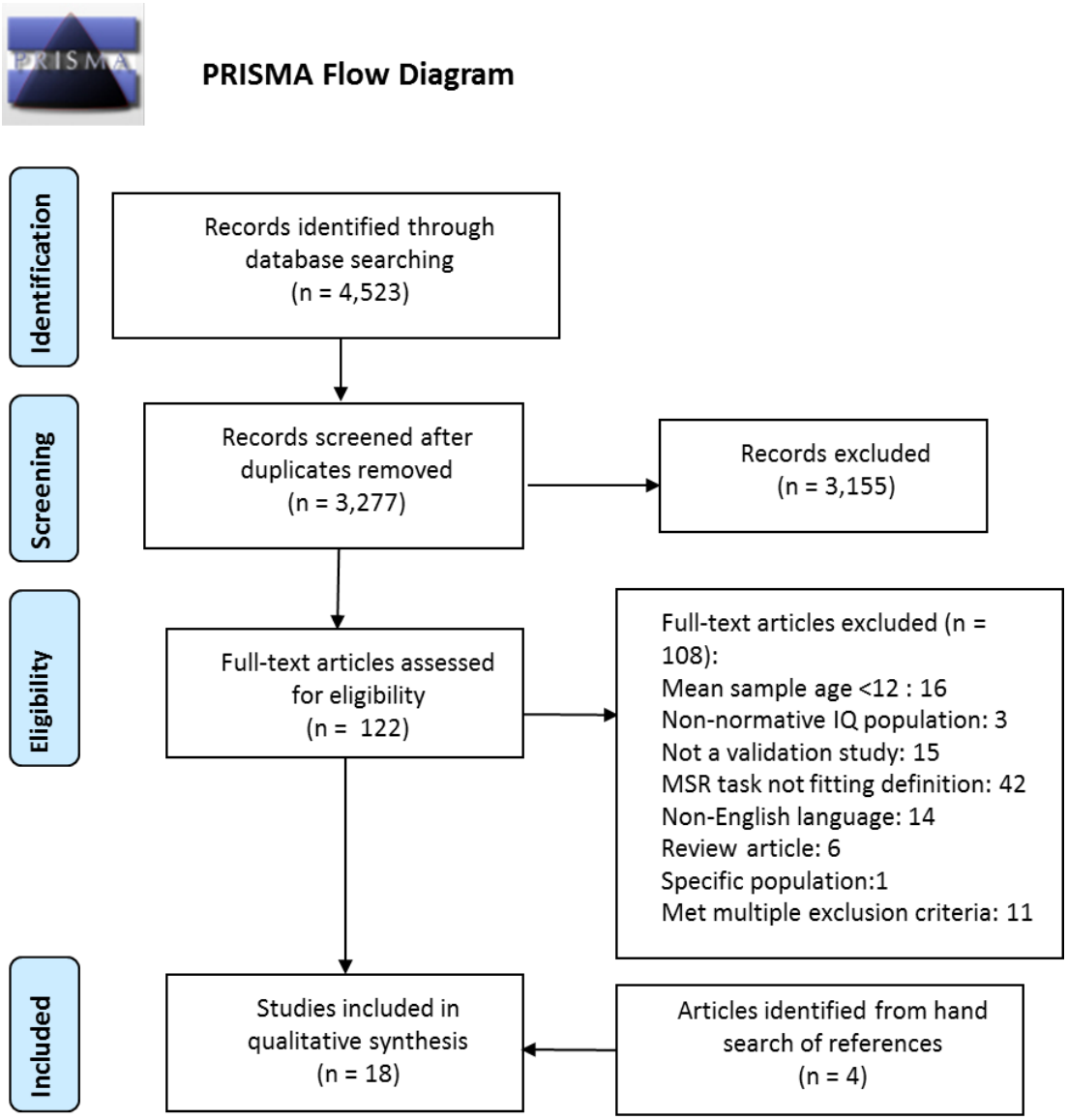


Figure 1: Flowchart of search results and paper selection

3.1 General description of identified studies

Eighteen articles evaluating 16 different measures were included in the study. The characteristics of the 18 studies are presented in Table 4. One study evaluated the measurement properties of multiple instruments (Pinkham et al., 2015) and some measures were evaluated across multiple studies. Seven studies were conducted in the UK, seven in the US, three in Australia and one in Canada. The majority of the studies were comparison studies, with eight investigating differences between individuals with an ASD[2] and a control group (Klin, 2000; Dziobek et al., 2006; Golan, Baron-Cohen, Hill, & Golan, 2006; Happé, 1994; Heavey, Phillips, Baron-Cohen, & Rutter, 2000; Koning & Magill-Evans, 2001; Roeyers et al., 2001; White et al., 2011), three comparing participants with a diagnosis of a schizophrenia spectrum disorder (SCZ) with a control group (Bell, Fiszdon, Greig, & Wexler, 2010; Bell, Bryson & Lysaker, 1997; Pinkham et al., 2015) and one study comparing individuals with traumatic brain injury (TBI) with a control group (McDonald et al., 2011). Two studies investigated the measurement properties of an instrument on a diverse heterogeneous clinical and non-clinical sample (Corcoran et al., 2011; Shryane et al., 2008) and four studies used a single group of participants from the healthy population (Canty et al., 2015; Dodell-Feder, Lincoln, Coulson, & Hooker, 2013; Johannesen, Lurie, Fiszdon, & Bell, 2013, McDonald et al., 2003). Sixteen of the studies used an adult sample, and two used an adolescent sample.

3.2 General description of instruments

A summary of the 16 MSR measures is displayed in Table 5. The majority (seven) of the measures were in the format of audio-visual film clips presenting social interactions and characters, to which participants are required to attribute mental states, using both verbal and non-verbal information. Four tasks required participants to infer the mental states of characters in a short story or vignette (using only verbal information), either read by the participant or the experimenter. Three tasks were in the form of silent film clips, requiring participants to attribute mental states to animated objects using only non-verbal information. One task used picture (non-verbal) stimuli requiring the participant to put the pictures in order to tell a coherent story. One task used virtual reality technology, whereby participants inferred the mental states of the virtual characters (using verbal and non-verbal information) with whom they were interacting. Seven tasks assessed both cognitive and affective aspects of MSR, five assessed only affective

[2] An Autism Spectrum Disorder (ASD) refers to a range of conditions characterised by social and communication difficulties. Different studies included in the review refer to different subgroups on this spectrum, including pervasive developmental disorder (PDD), high functioning autism (HFA) and Asperger's Syndrome (AS).

aspects, and four only cognitive aspects. All tasks assessed MSR from a second-person perspective, with the exception of one task that assessed MSR from a first person perspective. The majority of tasks (nine) used open-ended questions, four used multiple-choice questions, two tasks had both question types, and one task used picture sequencing as its response format.

3.3 Findings of the review

The methodological quality per study (poor, fair, good, or excellent) and the quality of the measurement properties of the instruments (negative, indeterminate, or positive) are presented in Table 6. The best evidence synthesis of the results per instrument is presented in Table 7. The results per MSR task are also described below. The mean scores obtained by study participants on the MSR tasks are presented in the Appendix.

Table 4: Study characteristics

Study	MSR Task(s)	Population	Design	Country	Setting	Mean age (SD)	Gender (% female)
Bell et al., 2010	SAT-MC	SCZ (n = 66) CON (n = 85)	Comparison	US	Health care	SCZ = 42.73 (10.4) CON = 31.72 (8.58)	SCZ = 39.4% CON = 88%
Bell et al., 1997	BLERT	SCZ (n = 50) SUB (n = 25) CON (n = 81)	Comparison	US	Health care; College	SCZ = 42.85 (7.99) SUB = 41.0 (6.54) CON = 20.56 (1.63)	SCZ = 2% SUB = 0% CON = 51%
Canty et al., 2015	VAMA	HP (n = 65)	Validation on single heathy population group	AUS	University	HP =25.98 (7.01)	65%
Corcoron et al., 2011	FBST	1 (n = 39) 2 (n = 29) 3 (n = 33) 4 (n = 29) 5 (n = 20) 6 (n = 27) 7 (n = 29) 8 (n = 31)	Validation on heterogenous clinical and non-clinical group	UK	Not reported	1 = 33.95 (8.4) 2 = 34.7 (10.3) 3 = 39 (14) 4 = 76.9 (6) 5 = 36 (10) 6 = 48.4 (11) 7 = 77.6 (8.1) 8 = 75.6 (5.5)	1 = 33% 2 = 38% 3 = 54% 4 = 66% 5 = 45% 6 = 67% 7= 52% 8 = 71%
Dodell-Feder et al., 2013	SST	HP (n = 74)	Validation on single heathy population group	US	University	HP = 27.8 (9.6)	64%
Dziobek et al., 2006	MASC	AS (n = 19) CON (n = 20)	Comparison	US	University	AS = 41.6 (10.4) CON = 39.9 (12.6)	AS = 10% CON = 10%
Golan et al., 2006	RMIF	AS/HFA (n = 22) CON (n = 22)	Comparison	UK	University	AS/HFA = 29 (9.8) CON = 25.4 (9.6)	ASD = 23% CON = 18%
Happe, 1994	SS	ASD (n = 24) MH (n = 11) Child CON (n = 69) Adult CON (n = 10)	Comparison	UK	via internet link	ASD = 20.6 MH = 19.4 Child CON = 8.6 Adult CON = 20.5	Figures not provided but study reports: "*Both the able autistic group and the MH*

							control group showed a preponderance of males, whereas the normal adult and child groups were made up of equal numbers of male and female subjects."
Heavey et al., 2000	AMT	HFA (n = 16) CON (n = 15)	Comparison	UK	Health care; Work-place; Home	HFA = 34.7 (9.5) 22-51 CON = 30.7 (8.1) 22-45	ASD = 6% CON = 0%
Johannesen et al., 2013	SAT-MC	HP (n = 51)	Validation on single heathy population group	UK	School; Home	HP = 19.27 (1.41)	60.8%
Klin., 2000	SAT	HFA (n = 20) AS (n = 20) CON (n = 20)	Comparison	US	University	HFA = 20.5 (5) AS = 18.9 (11.8) CON = 20.2 (7.4)	Not reported
Koning & Magill-Evans, 2001	CASP	SSD (n = 32) CON (n = 61)	Comparison	CAN	Varied: "at a location convenient for participants"	SSD = 13.91 (1.10) CON = 13.96 (0.85) Range = 12 - 15	O%
McDonald et al., 2003	TASIT	CON (n = 129) TBI (n = 7)	Content Development	AUS	University	"young adults"	Not reported
McDonald et al., 2013	TASIT	TBI (n = 16) CON (n = 16)	Comparison	AUS	University	TBI (13 - 19); CON (13-19)	TBI: 31% CON: 31%
Pinkham et al., 2015	BLERT TASIT Hinting task	SCZ (n = 179) CON (n = 104)	Comparison	US	University	SCZ = 42.11 (12.32) CON = 39.20 (13.70)	SCHZ = 35% CON = 53%

Roeyers et al 2001	EAT	PDD (n = 24) CON (n = 24)	Comparison	US	University	PDD = 23.8 (6.4) 17 – 46 CON = 23.1 (3.8) 19 - 31	8%
Shryane et al., 2008	Theory of Mind Stories	1 (n = 39) 2 (n = 29) 3 (n = 33) 4 (n = 29) 5 (n = 20) 6 (n = 27) 7 (n = 29) 8 (n = 31)	Validation on heterogenous clinical and non-clinical group	UK	Not reported	1 = 33.95 (8.4) 2 = 34.7 (10.3) 3 = 39 (14) 4 = 76.9 (6) 5 = 36 (10) 6 = 48.4 (11) 7 = 77.6 (8.1) 8 = 75.6 (5.5)	1 = 33% 2 = 38% 3 = 54% 4 = 66% 5 = 45% 6 = 67% 7= 52% 8 = 71%
White et al 2011	FH-A	ASD (n = 16) CON (n = 15)	Comparison	UK	University	ASD = 33 (10.31) CON = 36.53 (9.86)	ASD = 25% CON = 27%

Note: SAT-MC = Social attribution test--multiple choice; BLERT = The Bell Lysaker Emotion Recognition Test; VAMA = Virtual assessment of mentalising ability; FBST = False Belief Sequencing Task; SST = Short Stories Task; MASC = The movie for the assessment of social cognition; RMIF = Reading the Mind in Films Task; SS =Strange Stories; AMT = Awkward Moments Test; SAT = Social attribution test; CASP = Child and Adolescent Social Perception Measure; TASIT = The Awareness of Social Inference Test; EAT = Empathic accuracy test; FH-A = Frith-Happe Animation; 1 = schizophrenia spectrum with current persecutory delusions; 2 = schizophrenia spectrum with persecutory delusions in remission; 3 = Healthy adults; 4 = very late onset schizophrenia-like psychosis with current persecutory delusions; 5 = affective psychosis with current persecutory delusions; 6 = unipolar depression; 7 = older age unipolar depression; 8 = older age healthy adults; HFA = High functioning autism; AS = Asperger's syndrome; PDD = Pervasive developmental disorder; SSD = social skills deficits; HP = healthy population; CON = control group; SCZ = schizophrenia; TBI = traumatic brain injury

Table 5: Measure characteristics

Task	**Stimuli**	**Task requirements**	**Number of items & response format**	**Score range**	**Aspects of MSR assessed**
Awkward Moments Test	Audio-visual film clips UK advertisements /TV series clips	Participants answer test questions about characters' mental states e.g. "how do you think the young man was feeling at the end of the clip?" and control questions (e.g. about a visual feature or something that was said in the dialogue). Participants then engage in a semi-structured interview about the intentions of the characters.	7 MC (4 options) test questions 7 MC (4 options) control questions Open-ended interview questions regarding intentions of characters	Test questions: 0-7. Intentions score: 0-42	Affective Verbal / Non-verbal

Bell Lysaker Emotion Recognition Test	Audio-visual film clips	Participants identify the expressed emotion (happiness, sadness, fear, disgust, surprise, anger, or no emotion) using facial, vocal-tonal and upper body movement cues of a male actor.	21 MC (7 options) test questions	Test questions: 0-21	Affective Verbal / Non-verbal
Child and Adolescent Social Perception Measure	Audio-visual video film clips	Questions about characters' mental states (emotions) from film clips based on facial, vocal-tonal and body movement cues (emotion score). Participants are also asked what cues they used to make this judgement (nonverbal cues score).	10 film clips: 10 emotion questions; 10 non-verbal cue questions Open ended questions	Emotion score (ES): 0 – 10 Nonverbal cue (NC) score: 0 -10	Affective Verbal / Non-verbal
Empathic accuracy test	Audio-visual video film clips	Participants answer questions about real peoples' mental states (thoughts and feelings) from film clips of real interactions.	Video 1: 36 questions Video 2: 36 questions Open ended questions	Video 1: 0-36 Video 2: 0-36	Affective / Cognitive Verbal / Non-verbal
False Belief Sequencing Task	Pictorial	Participants put four picture cards in order so that they told a coherent story, involving false beliefs.	Four stories with 4 cards to order Control non-mental state stories	Test score: 0-8	Cognitive Non-verbal
Hinting task	Verbal	Participants explain the intended meaning of statements in stories.	10 stories Open ended questions	Total scores range from 0 - 20.	Cognitive Verbal
Frith-Happe Animations	Silent film clips	Participants decide what type of interaction is being displayed between two animated triangles: no interaction (NI), physical interaction (PI) OR mental interaction (MI). For correctly identified MI clips participants are additionally asked what emotion they believe to have been felt by the two different triangles.	12 clips Subjective scoring (from open ended questions) Objective MC scoring: 3 MC options for interaction questions; 5 MC options for emotion questions	<u>Subjective scoring</u> Intentionality: 0 -15 Appropriateness: 0 -15 <u>Objective scoring:</u> MCQ- Categorisation scores: 0-12, divided into 4 for each of the 3 animation types. MCQ-Feelings scores: 0-8	Affective Non-verbal

The movie for the assessment of social cognition	Audio-visual 15 minute film of 4 characters at a dinner party.	Participants answer questions about four characters' mental states from a 15-minute film. Control questions ask about non-mental state related questions.	45 questions open ended questions 6 control questions	Test score: 0-45 Control question score: 0-6	Affective / Cognitive Verbal / Non-verbal
Reading the Mind in Films Task	Audio-visual film clips	Participants identify characters' mental states from clips taken from movies.	22 MC (4 options) questions No control clips / questions	Test score: 0-22	Affective Verbal / Non-verbal
Social attribution test	Silent film clip	Participants describe the interaction between moving shapes and responding to questions such as "what happened here?" or "what kind of a person is big / little triangle?"	1 film clip – 17 questions to provide 17 narratives Open ended questions No control clips / questions	Total of 17 narratives coded to produce 7 index scores: 1) pertinence index 2) salience index 3) ToM cognitive index 4) ToM affective index 5) animation index 6) person index 7) problem solving index	Affective / Cognitive Non-verbal
Social attribution test--multiple choice	Silent video	Participants select the best explanation of the behaviour of moving shapes. The animation is shown twice and then short segments are presented followed by multiple-choice questions about the actions depicted.	19 MC test questions No control questions	Test questions: 0-19	Affective / Cognitive Non-verbal
Strange Stories	Verbal (accompanying pictures)	Participants answer questions about the intended meaning of statements in stories, which contain concepts such as lies, white lies, jokes, pretence, misunderstandings & sarcasm (e.g. "Was it true what X said?") and justification questions (such as "'Why did X say that?").	24 vignettes Open ended questions 6 Control vignettes 'physical stories' – to control for domain	Test scores: 0-24 Control scores: 0-6	Cognitive Verbal

			general processes to be accounted for		
Short Stories Task	Verbal (written) stimulus	Participants answer questions about the mental states of characters in 'The End of Something', a short story by Ernest Hemingway (such as "Why is Nick afraid to look at Marjorie?").	8 mental state questions 5 comprehension questions Open ended questions	Total scores for MSR questions: 0 – 16 (max score of 2 per question) Comprehension scores: 0-10	Affective / Cognitive Verbal
The Awareness of Social Inference Test	Audio-visual film clips	Participants answer questions about the mental states of characters in film clips of everyday social interactions. Part 1 assesses emotion recognition; Part 2 assesses the ability to detect sarcastic vs literal meanings in speech; Part 3 assesses the ability to detect sarcasm vs lies in speech.	64 questions (16 vignettes: 4 questions per vignette) Open ended questions No control clips / questions	Test scores: 0 - 64.	Affective / Cognitive Verbal / Non-verbal
ToM Stories	Verbal (accompanying pictures)	Participants are read short stories involving false belief and deception. Test questions relate to 1st order and 2nd order MSR. Memory questions are also included to test comprehension.	6 stories – 2 questions per story Open ended questions Control memory questions	Test scores: 0-12	Cognitive Verbal
Virtual assessment of mentalising ability	Audio-visual virtual reality environment	Participants navigate through a virtual reality environment, and respond to questions that relate to a series of 10 interactions that occur between the test-taker and virtual 'friends'. Answers correspond to 1st order cognitive, 2nd order affective, 2nd order cognitive, 2nd order affective ToM. Control scenarios require integration of film information without requiring social understanding.	40 items (10 interactions each with 4 MC questions) - max 2 points per item. 2 Control scenarios	Test scores: 0 - 80.	Affective / Cognitive Verbal / Non-verbal

Note: MC = Multiple choice; ToM – Theory of Mind

Table 6: Methodological quality and quality criteria ratings per study and instrument property

Task	Study	Reliability						Validity						Interpretability[3]		
								Construct Validity								
		Internal Consistency		Test-retest		Inter-rater		Hypothesis Testing		Structural Validity		Content Validity		Interpret-ability	Floor effects	Ceiling effects
				Q	M	Q	M	Q	M	Q	M	Q	M	Q	Q	Q
AMT	Heavey et al., 2000					+	Fair	?	Poor			?	Poor	+	+	+ (ASD group)
BLERT	Bell et al., 1997			-	Fair			+	Poor					+		
	Pinkham et al., 2015	?	Poor	-	Fair			+	Poor					+	+	+
CASP	Koning & Magill-Evans, 2001							+	Fair					+		
EAT	Roeyers et al., 2001							?	Poor			?	Poor	+		
FBST	Corcoron et al., 2011							+	Fair	?	Fair					
FH-A	White et al., 2011					?	Poor	?	Poor					+		
Hinting task	Pinkham et al., 2015	?	Poor	-	Fair			+	Poor					+	+	+
MASC	Dziobek et al., 2006	?	Poor	+	Poor	+	Poor	?	Poor			+	Fair	+	+	+
RMIF	Golan et al., 2006							?	Fair			+	Excellent	+		
SAT	Klin, 2000					+	Fair	+	Fair			+	Fair	+		
SAT-MC	Bell et al., 2010	?	Poor	?	Poor			+	Poor					+	+	+
	Johannesen et al. 2013	?	Poor	?	Poor			+	Fair					+	+	+

[3] As interpretability is not considered to be a measurement property, there is no methodological quality rating (using COSMIN)

SST	Dodell-Feder et al., 2013	?	Poor	?	Poor	+	Poor	+	Fair			?	Fair	+	+	+
SS	Happe, 1994					?	Poor	?	Poor			?	Poor	+	+	+ (ASD group)
TASIT	McDonald et al., 2003											+	Excellent			
	McDonald et al., 2013							?	Poor					+		
	Pinkham et al., 2015	?	Poor	-	Fair			+	Poor					+	+	+
ToM Stories	Shryane et al., 2008							+	Fair	?	Fair			+		
VAMA	Canty et al., 2015	?	Poor	+	Fair			-	Fair			+	Excellent			

Note: Q = Instrument measurement property quality criteria; M = Methodological quality

Table 7: Best evidence synthesis of the measurement properties per task

Task	Reliability			Validity		
	Internal consistency	Reliability		Construct validity		Content validity
		Test-retest	Inter-rater	Hypothesis testing	Structural validity	
AMT			+	?		?
BLERT	?	- -		?		
CASP				+		
EAT				?		?
FBST				+	?	
Hinting task	?	-		?		
FH-A			?	?		
MASC	?	?	?	?		+
RMIF				+		+++
SAT			+	+		?
SAT-MC	?			+		
SS			?	?		?
SST	?		?	+		?
TASIT	?	-		?		+++
ToM Stories				+	?	
VAMA	?	+		+		+++

3.3.1 The Awkward Moments Test (AMT)

Content and Ecological Validity. The AMT was developed in order to create a more sensitive tool for exploring the mentalising difficulties in adults with ASD (Heavey et al., 2000). The measure consists of suitable clips from TV advertisements, which were selected by experts on the basis of depicting socially awkward moments between characters of various ages and relationships, and in a variety of contexts. However, the study does not provide information about how items were selected or whether the items were piloted in the target population. The study was therefore rated as having poor methodological validity, and so evidence for the content validity is unknown. Although the task attempts to approximate the demands of real life in its use of actors, the use of TV advertisements is likely to increase the dramatic nature of the interactions, raising questions about the task's ecological validity.

Construct Validity. Heavey et al. (2000) also performed hypothesis testing, with some formation of hypotheses a priori. As expected, the ASD group performed significantly worse than controls on both the test and intention questions of the AMT, with the latter more sensitive in differentiating groups. This remained significant after removing two participants from each group who had the lowest IQ scores (in order to better match groups in terms of IQ scores) and after controlling for comprehension and reading ability. The AMT test question scores correlated significantly with performance on the Strange Stories task (Happe, 1994) in the control group (r = .60) but not in the ASD group (r = .48). The AMT intention scores were not significantly correlated with the Strange Stories in either group. Performance on the AMT correlated with ability subtests and Full Scale IQ (r = .57 – .78) in control group only, and AMT was not related to verbal IQ in ASD group. Since the measurement properties of the comparator tasks were not described and the sample size was small, the methodological quality was rated as poor. Therefore, the evidence for the construct validity (as assessed by hypothesis testing) of the AMT has been rated as indeterminate.

Reliability. Heavey et al., (2000) also evaluated inter-rater reliability, where they calculated a high intraclass correlation coefficient (ICC) of .99, with fair methodological quality, providing some limited positive evidence for the task's reliability.

Interpretability. The study reported the means, SDs and ranges of scores obtained for both the ASD and control group, which demonstrated that 0% of participants in either group scored the lowest score possible, and none of the participants in the ASD group scored above five

(maximum = 7). The range of scores for the control group was 3-7 (mean = 5.27), indicating possible ceiling effects.

3.3.2 Bell Lysaker Emotion Recognition Test (BLERT)

Content and Ecological Validity. The BLERT consists of short video vignettes, depicting the same male actor speaking about a work related topic (as the measure was developed in the context of a work rehabilitation programme for people with schizophrenia). Content validity was not formally assessed due to inadequate information provided about the development of the measure. However, as the BLERT only assesses the recognition of six emotions, its content validity (with respect to comprehensiveness) is questionable. The task has good ecological validity in that is uses dynamic video stimuli providing vocal, upper body and facial cue information.

Construct Validity. The Bell et al., (1997) study carried out hypothesis testing, and as expected they found that controls performed better on the BLERT than the SCZ patient group. However, the study was of poor methodological quality, due to groups not being matched on age, gender or ethnicity, and it was likely that the mean IQ of the control group of college students (which was not collected) would have been significantly higher than that of the SCZ patient sample (mean IQ = 90.32). The second study (Pinkham et al., 2015) did not form a priori hypotheses, although it was still possible to deduce what was expected. This study also demonstrated poorer performance in the SCZ patient sample compared with controls, as well as significant correlations between BLERT scores and measures of social functioning: UCSD Performance-Based Skills Assessment, Brief (UPSA-B; Mausbach, Harvey, Goldman, Jeste, & Patterson, 2007) Total (*r*=.32), Social Skills Performance Assessment (SSPA; Patterson, Moscona, McKibbin, Davidson, & Jeste, 2001) Average (*r*=.26) and Specific Level of Functioning Scale (SLOF; Schneider & Struening, 1983) Total (*r*= .31) in the SCZ group. They found that BLERT continued to predict social functional outcome (SLOF Total: β = .353) over and above the variance accounted for by performance on neurocognitive measures. However, this study did not include information about the psychometric properties of comparator instruments, and so the study was rated as having poor methodological quality. Additionally, the control group had significantly more years in education than the patient groups, which was not controlled for in the analyses. The evidence for the construct validity of the BLERT is therefore unknown as only studies of poor methodology are available.

Reliability. One study (Pinkham et al., 2015) evaluated the internal consistency of the BLERT, which was found to be adequate for the SCZ patient group (Cronbach α=.74) but questionable for the control group (Cronbach α=.63). However, the study was of poor methodological quality with respect to this measurement property, since factor analysis was not performed in order to check the unidimensionality of the measure. Best evidence synthesis for internal-consistency is therefore unknown.

Two studies investigated the test-retest reliability (Bell et al., 1997; Pinkham et al., 2015), which were both of fair methodological quality. Pinkman et al. (2015) found Pearson's *r*s of .70 for the SCZ patient group, and .68 for the control group, over a 2-4 week interval. Bell et al., (1997), reported a Pearson's coefficient of .76 in their SCZ patient group, and a weighted Kappa of .93 for disease severity category stability, over a five-month interval. Both studies therefore scored a negative rating for test-retest reliability, as the correlation coefficients did not meet minimum criteria proposed by Terwee (2011), indicating a moderate level of negative evidence for the test-retest reliability of the BLERT.

Interpretability. Bell et al., (2010) only reported average percentage correct with regards to scores, with no information on SDs or ranges, although the control group's average score of 92.3% suggests a risk of ceiling effects in these college students. Pinkman et al., (2015) reported both means and SDs of scores for both groups and found that 0.6% of patients, and 0% of the controls, were scoring at floor, and 0% of the patient group, and 2% of the control group were scoring at ceiling.

3.3.3 Child and Adolescent Social Perception Measure (CASP)

Content and Ecological Validity. The CASP was developed for use in the clinical setting for young people with mental health problems. The content validity was not formally assessed due to lack of information about the development of the measure (the original development study was not included in the review). However, the film clips show young people interacting in a range of everyday settings and characters depict a range of emotional intensities, indicating good ecological validity.

Construct Validity. Koning & Magill-Evans, 2001 conducted hypothesis testing, with hypotheses formulated a priori and was rated as having fair methodological quality. Significant correlations were demonstrated between CASP emotion scores (CASP-ES) and a measure of social skill (the

Social Skills Rating System; SSRS; Gresham & Elliot, 1990) as rated by parents (*r* =.63) teachers (*r* =.56) and students (*r*=.39). CASP non-verbal cue (CASP-NC) scores also significantly correlated with parent (*r* = .52) teacher (*r* = .50) and student (*r*=.34) rated versions of the SSRS. CASP scores were also significantly associated with problem behaviours, whilst not being significantly related to an estimate of IQ (*r*=.13). Both CASP ES and NS scores revealed significantly poorer performance in the SSD group compared with controls, who were matched in terms of vocabulary performance, and demonstrated good classification accuracy for the clinical group (96.9%) as well as controls (86.2%), suggesting some limited positive evidence for the CASP's construct validity.

Reliability. No study in the current review investigated the reliability of the CASP.

Interpretability. The study provided information on means and SDs for the SSD and control groups, although no additional information was provided to ascertain whether participants were scoring at floor or ceiling.

3.3.4 Empathic accuracy test (EAT)

Content and Ecological Validity. The EAT was developed to assess participants' 'online' inferences of other peoples' successive thoughts and feelings, to assess the specific deficits of adults with ASC (Roeyers et al., 2001). The task uses footage of real people interacting and uses their actual thoughts and feelings to ascertain the correct responses, suggesting good ecological validity. However, as the two videos depict just one narrow situation (two volunteers waiting to take part in a research study) it is likely that the comprehensiveness of the item content is compromised. As there was no indication of piloting the task in the target population, or about how items were selected, the methodological quality was rated as poor, and so overall evidence is indeterminate.

Construct Validity. The authors conducted hypothesis testing on the EAT, which found significant correlations between Video 2 of the EAT and the Eyes task (*r* = .57), in the control group, but not for the ASD group. They also found a significant correlation between Video 1 and the Strange Stories task (*r* = .38) in the ASD group, but not for controls. Performance on Video 1 of EAT did not correlate with performance on the Strange Stories task in either group. Although no IQ data was collected for the control sample, task performance was not associated with IQ in the ASD group. The study was of poor methodological quality due to not providing psychometric

information about comparator measures and had a small sample size. Evidence is therefore unknown regarding the construct validity of this measure.

Reliability. No study in the current review investigated the reliability of the EAT measure.

Interpretability. Roeyers et al. (2001) provided means and SDs per group for performance on the EAT, but did not report any information on floor or ceiling effects.

3.3.5 False Belief Sequencing Task (FBST)

Content and Ecological validity. The FBST task involves ordering four cards in order to tell a coherent story, with characters acting on false beliefs. As there was no information about the development of the measure, the content validity was not assessed. However, this picture task lacks ecological validity in terms of approximating the demands of real-life MSR.

Construct Validity. Corcoran et al. (2011) investigated the structural validity of the FBST, and found a unidimensional IRT model to fit the data well. However, there is no information about how much variance is explained by this single factor, and so the structural validity is unknown. Hypothesis testing was also conducted, and as expected no association was found between clinical diagnosis or symptoms and the FBST, after controlling for age and IQ. With respect to convergence validity, the authors found significant correlations between the picture sequencing factor and the three factors from the ToM Stories task (Shryane et al., 2008): the false belief factor (r=.63), the 1st order deception factor (r=0.55) and the 2nd order deception factor (r=.60), which remained significant after controlling for age and IQ. As the study was rated as having fair methodological quality, there is limited positive evidence for the construct validity of the FBSQ.

Reliability. No study in the current review investigated the reliability of the FBST.

Interpretability. Means and SDs are not reported, and there is no information about floor or ceiling effects.

3.3.6 Frith-Happe Animations Task (FH-A)

Content and Ecological Validity. The task was originally developed to assess the theory of mind difficulties in individuals with ASC. The task assesses attribution of mental states to animated

triangles, and assesses thoughts, feelings and intentions. The content validity was not formally assessed, as no study referred to the development of the measure. However, the use of animated shapes reduces the task's ecological validity.

Construct Validity. White et al., (2011) conducted hypothesis testing, where they found that as expected the ASD group performed significantly worse than the control group in terms of ToM intentionality scores, and the ToM appropriateness scores (but not on the non-mental state goal-directed or random interaction items). There were also group differences on the objectively scored multiple-choice categorisation questions (MCQ-Categorisation) and multiple-choice feelings questions (MDQ-Feelings). The study also found positive but non-significant correlations between the MCQ-categorisation scores and performance on false belief tasks (r=.49; p=.07) and between the objective MCQ-feelings scores and the subjective intentionality scores on the ToM animations (r=.48, p=.06) in the ASD group. This suggests a degree of convergence between the subjective and objective scoring methods, as well as with other false belief tasks. Although the study matched the groups with respect to age, gender, verbal and performance IQ, the sample size was small, resulting in a poor rating for the study's methodological quality. Therefore, evidence for the construct validity of the FH-A is unknown.

Reliability. White et al. (2011) investigated the inter-rater reliability of the FH-A task, which was calculated for the intentionality scale (ICC = .83) and for the appropriateness scale (ICC = .70). Despite high ICCs, because the study sample was small, it received an indeterminate rating for this measurement property.

Interpretability. Means and SDs of the two groups, on both subjective and objective scoring methods, were provided. However, no information was provided to determine floor or ceiling effects.

3.3.7 Hinting Task

Content and Ecological Validity. The Hinting task involves an examiner reading aloud vignettes, in which characters drop a hint, and the participant is required to infer the true meaning of what was said. Due to the verbal format of presentation, the ecological validity is reduced. Content validity was not formally assessed as there was no information about the development of the measure.

Construct Validity. Pinkham et al., (2015) investigated hypothesis testing, and as expected the SCZ patient group performed significantly worse on the task than controls. The study also found significant positive correlations between the Hinting task and three different measures of social functioning: USPA Total (*r*= .46), SSPA Average (*r*= .39), SLOF Total (*r* = .20). Moreover, performance on the Hinting task continued to predict UPSA-B (β = .242) and SSPA (β = .258) over and above that accounted for by performance on the neurocognitive tasks. However, the study did not comment on the psychometric properties of comparator instruments and was therefore rated as poor in terms of methodological quality. Therefore, evidence for the task's construct validity is unknown.

Reliability. Pinkman et al., (2015) also investigated the internal consistency of the Hinting task, which found Cronbach's αs of .73 (adequate) and .56 (inadequate) for the SCZ patient and control groups respectively. However, because the study was rated as having poor methodological quality (due to not performing factor analysis to confirm the dimensionality of the scale) the evidence for this measurement property is unknown. The same study also investigated the test-retest reliability, which was rated as having fair methodological quality for this measurement property. They found reliability coefficients of .64 for the SCZ patient sample and .42 for the control sample, suggesting inadequate test-retest reliability. Best evidence synthesis is therefore limited negative evidence for the test-retest reliability of the Hinting task.

Interpretability. The study provided means and SDs for the two groups. Neither the SCZ patient or control group scored the lowest possible score. However, 1.17% of the SCZ patient group and 7% of the control group scored the highest possible score, indicating a risk of ceiling effects particularly for control populations.

3.3.8 Movie for the Assessment of Social Cognition (MASC)

Content Validity and Ecological Validity. Dziobek et al (2006) refer to the development of the MASC, which was rated as having limited positive evidence for content validity. Experts ensured that the MASC assessed a range of different ToM aspects (including thoughts, feelings, intentions) that require reading both verbal (literal and figurative) and non-verbal (facial and body) communications. The study also refers to pilot work (although a sample size for the pilot data is not provided) and so is rated as having fair methodological quality. The MASC also has good ecological validity in that it aims of approximate a real life situation, and the relatively long video clip (15 minutes) allows for the development of characters and scenarios.

Construct Validity. Dziobek et al. (2006) also performed hypothesis testing, and as expected they found that the AS group performed significantly worse on the MASC than the control group, but that this was not the case for the control questions, which might add to the construct validity of the task. The MASC was also shown to be accurate at discriminating between groups using ROC curves. The area under the curve was .98 for the MASC, which was higher than the comparator instruments, indicating the measure's superior sensitivity. The study also demonstrated that the MASC significantly positively correlated with the Strange Stories task (*r*=.47) in the ASD group, but not in the control group. For the control group, performance on the MASC significantly positively correlated with performance on facial affect recognition (*r*=.72), providing some degree of convergent validity. The MASC did not correlate significantly with IQ in either group. Due to the small sample size, the study was of poor methodological quality, which means that evidence for the construct validity of the MASC is indeterminate.

Reliability. Dziobek et al. (2006) reported a good internal consistency (Cronbach's α = .84), however because no factor analysis was performed to check the unidimensionality of the scale, the evidence is indeterminate. The study also investigated inter-rater reliability, which was found to be excellent (ICC = .99 for whole sample; .98 for the AS group; .94 for the control group). Test-retest reliability was also good (*r* = .97 for whole sample; .92 for the AS group; .89 for the control group). However, the study was of poor methodological study with respect to reliability, as the sample sizes were small. Thus, evidence for the reliability of the MASC is also unknown.

Interpretability. Means, SDs and ranges were provided for each group, which demonstrated no floor or ceiling effects in either group. However, ceiling effects were demonstrated on control items, with both groups obtaining a mean score of 3.9 (max = 4), suggesting that these items were not of comparable difficulty.

3.3.9 Reading the Mind in the Films Task (RMIF)

Content and Ecological Validity. The RMIF task was developed to assess the MSR difficulties associated with ASD. Golan et al. (2006) described the development of the measure, and refer to experts selecting clips from films and TV series that displayed suitable emotional content. The study also refers to the careful selection of multiple-choice responses using independent judges and pilot data from the general population. Moreover, the task assesses 22 different complex emotions, which suggests a comprehensive assessment of the construct being measured. The

study was rated as having excellent methodological quality for assessing content validity, indicating strong level of positive evidence. However, the measure may have reduced ecological validity due to clips being chosen specifically for their dramatic nature.

Construct Validity. Golan et al. (2006) conducted hypothesis testing on the RMIF task and as expected found that the ASD group performed significantly worse than the control group, after controlling for age, verbal and performance IQ. The measure was also found to be significantly negatively correlated (*r*=-.52) with a measure of ASD traits (the Autism Quotient; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) and significantly positively correlated with a measure of complex emotion recognition, the Cambridge (CAM) Face-Voice Battery (Golan, Baron-Cohen, & Hill, 2006) (*r*=.63 for CAM faces; *r*=.62 for CAM voices and *r*=.61 for emotional concepts recognised) demonstrating convergent validity. As expected, RMIF performance also correlated with verbal IQ (*r*=.48), but not performance IQ. The measure was also able to successfully classify 90% of participants, demonstrating its sensitivity. The study was rated as being of fair methodological quality, and so best evidence synthesis is that there is limited positive evidence for the RMIF task's construct validity.

Reliability. No study in the current review investigated the reliability of the RMIF measure.

Interpretability. Means and SDs are reported but no information is provided on floor or ceiling effects.

3.3.10 Social Attribution Task: original and multiple choice version (SAT and SAT-MC)

Content and Ecological Validity. The SAT was originally developed to detect mental state attribution difficulties in individuals with ASD, in a format that aimed to reduce the confound of verbal IQ. The development study (Klin, 2000) selected the clip based on previous research by Heider and Simmel (1944) that found that the healthy population naturally attributed mental states to moving animated shapes. Questions relate to thoughts, feelings and intentions (indicating comprehensiveness) however, it is unclear how test questions were selected. Thus, the content validity of the SAT and SAT-MC tasks is indeterminate, and they lack ecological validity due to the use of animated shapes.

Construct Validity. The study by Klin (2000), is rated as having fair methodological quality for hypothesis testing. As predicted it was found that the HFA and AS groups performed significantly

worse than the control group, and performance on the SAT was found not to be related to verbal IQ, suggesting limited positive evidence for construct validity. Bell et al. (2010) demonstrated that the SAT-MC was also able to discriminate between a patient (SCZ) and control group, whereby the patient group performed significantly worse. The SAT-MC also predicted group membership with 60.6% sensitivity and 77% specificity (Bell et al., 2010). The SAT-MC was also shown to correlate modestly with the BLERT (r = .47) and the Hinting task (r = .37) in the SCZ patient group. However, they did not describe the psychometric properties of comparator instruments, resulting in a poor rating of methodological quality. Positive correlations between the SAT-MC and the BLERT were also found by Johannesen et al., (2013) (r = .28), which was of fair methodological quality in their healthy population sample, indicating further evidence of convergent validity. A best evidence synthesis suggests limited positive evidence for the construct validity of the SAT-MC and the SAT.

Reliability. None of the studies in the current review investigated the internal consistency of the original version of the SAT, although two studies investigated the internal consistency of the multiple-choice version (Bell 2010; Johannesen et al., 2013). Bell et al. (2010) found the internal consistency of the SAT-MC to be good (Cronbach α = .83) with a split half reliability using Spearman Brown of .75. However, Johannessen et al. (2013) found evidence of poor internal consistency (Cronbach α = .56) in their healthy population sample. Neither of the studies conducted factor analysis in order to assess the dimensionality of the scale, indicating poor methodological quality. Evidence regarding the internal consistency of the SAT-MC is therefore unknown.

Klin (2000) investigated the inter-rater reliability of the original version of the SAT (which is not a relevant measurement property for the multiple choice version) and found evidence of good reliability for the different indices (ICCs ranging from .76 – .91). The study was rated as being of fair methodological quality and so there is limited positive evidence for the SAT's inter-rater reliability.

Interpretability. All three studies (Bell et al., 2010; Klin, 2000; Johannessen et al., 2013) provided means and SDs of scores. Bell et al., (2010) demonstrated that neither the patient nor the control group scored at floor on the SAT-MC, and no more than 3% of the patient sample, and no more than 5% of the control group, scored the highest possible score. Johannessen et al. (2013) demonstrated that none of their healthy sample scored the lowest possible score on the SAT-MC, although 9.8% scored the highest possible score, indicating risk of ceiling effects especially in the healthy population.

3.3.11 Strange Stories (SS)

Content and Ecological Validity. The Strange Stories task explores a wide range of mental states such as lying, persuasion, white lies and jokes, across the different vignettes, implying good content validity. However, the study did not pilot the items, which gives it a poor methodological rating. Thus, evidence for the content validity of the task is indeterminate. The ecological validity of the SS task is limited due to its pencil and paper format.

Construct Validity. Happe (1994) conducted hypothesis testing and found that as expected the ASD group performed significantly worse than healthy adults with respect to providing accurate mental state justifications. The ASD group's performance was also significantly worse than a group matched on verbal IQ, and scores were also related to performance on false belief tasks. However, because of the study's poor methodological quality rating (due to the small sample size) the construct validity is unknown.

Reliability. Happe (1994) found excellent agreement between two raters (92-100% agreement) when coding the SS task. However, the study was of poor methodological quality due to the use of inappropriate statistics according to both COSMIN and Terwee et al (2007) criteria (percentage agreement rather than ICCs or weighted kappa). Therefore, the evidence for inter-rater reliability is unknown.

Interpretability. Means, SDs and ranges of scores were provided for each subgroup, which demonstrated that within the adult control group all participants scored in top three possible scores (see Appendix), suggesting ceiling effects.

3.3.12 Short Story Task (SST)

Content and Ecological Validity. The SST was developed to assess individual differences in implicit and explicit MSR (Dodell-Feder et al., 2013). The task assesses a range of different mental states (thoughts, feelings and intentions) at both first and second order levels of inference. Content validity was rated indeterminate, as although the study refers to pilot data, it is unclear how items were selected, and how comprehensiveness was assessed. The ecological validity is limited due to the verbal text format of the task.

Construct Validity. Dodell-Feder et al. (2013) also conducted hypothesis testing, where they found significant positive correlations between performance on the SST and performance on the Eyes task (Baron-Cohen et al., 2001) (*r*=.59), after controlling for IQ (*r*=.42). SST performance was also correlated with the fantasy subscale of the Interpersonal Reactivity Index (Davis, 1980; *r*=.45) as well as with IQ (*r*=.24). The methodological quality for the study for this measurement property was fair, and so there is limited positive evidence for the convergent validity of the SST.

Reliability. Dodell-Feder et al., (2013) assessed the internal consistency of the SST, which was found to be poor for both the mental state questions (Cronbach α=.54) as well as the comprehension questions (Cronbach α=.31). However, because no factor analysis was performed to check the dimensionality of the measure, the methodological quality was poor - thus evidence is indeterminate. The inter-rater reliability was found to be excellent (ICC=.98), although again due to the poor methodological quality of the study (for having a small sample size in test-retest reliability analysis) the strength of this evidence is unknown.

Interpretability. Means, SDs, ranges were reported and a histogram demonstrated a normal distribution of scores. No participants scored at floor or ceiling on the test questions, although 48.6% of participants scored at ceiling on the control comprehension questions, suggesting that these items were not of comparative difficulty.

3.3.13 The Awareness of Social Inference Test (TASIT)

Content and Ecological Validity. The TASIT was developed as a criterion referenced test to detect difficulties in social cognition in individuals with TBI. Video vignettes depict everyday social interactions that probe understanding of the emotions, intentions, beliefs, and meanings (e.g. lies or sarcasm) of the speakers in social exchanges. The development study (McDonald et al., 2003) refers to the selection of items by experts as well as pilot data in both the healthy population and TBI population. The study was therefore rated as having excellent methodological quality, providing strong positive evidence for the task's content validity. The task also has good ecological validity due to its dynamic format.

Construct Validity. Two studies (Pinkman et al., 2015; McDonald et al., 2011) performed hypothesis testing on aspects of the TASIT. Pinkman et al (2015) found that the SCZ patient group performed significantly worse on Part 3 of the TASIT than controls, and McDonald et al. (2011) found significantly worse performance on Part 2 of the TASIT in their IQ-matched TBI group. The

Pinkman study also demonstrated a positive significant correlation between the TASIT Part 3 and three measures of social functioning: USPA Total (*r*=.44), SSPA Average (*r*=.31), SLOF Total (*r*=.30). However, because the psychometric properties of the comparator measures were not described, the methodological quality was rated as poor. The McDonald et al. (2011) study also demonstrated that performance on both TASIT parts 1 and 2 was related to relative-reported (but not self-reported) difficulties in understanding of both literal (TASIT 1: *r*=.66; TASIT 2: *r*=.53) and non-literal communication (TASIT 1: *r*=.74, TASIT 2: *r*=.58.). TASIT 3 performance was related to both relative- and self-reported difficulties in both literal and non-literal language (*r*s=.51 - .72). They also found that TASIT Part 3 was significantly correlated to a face recognition task (*r*=.5). None of the above correlations were significant in the control group. All three parts of the TASIT correlated with IQ in the TBI group (*r*s=.63 – .79), but only TASIT Part 3 significantly correlated with IQ (*r*=.58) in the control group. Therefore, there is some mixed evidence for the construct validity of the TASIT, but due to poor methodological quality of studies the strength of this evidence is unknown.

Reliability. One study (Pinkman et al. 2015) investigated the internal consistency of the TASIT Part 3, which was found to be adequate for the control sample (Cronbach α = .76) and good in the SCZ patient sample (Cronbach α = .81). However, once again due to lack of factor analysis, the study was rated as having poor methodological quality, and so the strength of this evidence is unknown. Pinkman et al. (2015) also investigated the test-retest reliability of the TASIT Part 3, which they found to be poor for the control sample (*r* = .54) and questionable in the SCZ patient sample (*r* = .60). The quality of this study was rated as fair, suggesting limited negative evidence of the test-retest reliability of the TASIT.

Interpretability. Means and SDs were provided by all three studies. Pinkman et al. (2015) also reported that 7% of the SCZ patient sample (but 0% of the control sample) scored at floor on this task. Neither group scored the maximum possible score.

3.3.14 ToM Stories

Content and Ecological Validity. The ToM stories consist of three false belief and three deception stories, with both first and second order questions per story. No study in the review reports on the development of the task, and so content validity was not assessed. Due to the presentation format of the ToM stories, the task lacks ecological validity.

Construct Validity. Shyane et al., (2008) investigated the structural validity of the ToM stories task, where they found that a three-factor Item–Response Theory (IRT; Birnbaum, 1968; cf. Rasch, 1960) model provided the best fit. However, it is not clear how much variance is explained by the factors and so the structural validity is indeterminate. In terms of hypothesis testing, current paranoia was found to be significantly associated with lower scores on the first-order deception factor (β = -.58) and the false-belief factor (β = -.37), after controlling for the effects of age and IQ. Additionally, the depressed with paranoid delusions group had significantly lower average deception first-order (β = -.27) scores than the controls, but this effect was not seen for false beliefs (p<.1). The other diagnostic groups (unipolar depression and schizophrenia with persecutory delusions in remission) did not perform significantly differently from controls on any of the factors. In terms of sensitivity, the authors found that a both the first-order deception and the single false belief factor were found to measure the ToM skill well in samples where the skill was compromised, and these two factors proved sensitive to the presence of persecutory delusions. IQ scores related to false belief and second order deception factors, but not first order deception factor. As the study was of fair methodological quality, there is limited positive evidence for the construct validity of the ToM Stories task.

Interpretability. Means and SDs were presented for each of the eight diagnostic subgroups, although information to determine floor or ceiling effects was not reported.

3.3.15 Virtual Assessment of Mentalising Ability (VAMA)

Content and Ecological Validity. The VAMA was developed to detect individual differences in MSR ability, and was designed to assess a range of mental states, covering first-order and second-order, cognitive and affective ToM, including concepts such as false belief, deception, faux pas, humour, sarcasm and persuasion. The study refers to the use of expert judges and pilot work in the development of the measure and selection of items, indicating excellent methodological quality and strong positive evidence for content validity. Additionally, with its use of virtual technology, which provides individuals with a more immersive experience in a social interaction, the task has good ecological validity.

Construct Validity. Hypothesis testing demonstrated some moderate relationships between the VAMA subscales and the Faux Pas Recognition Test subscales (11/45 correlations were significant) as well as the Hinting task (3/9 correlations significant), although there were no significant correlations between performance on the VAMA and Yoni task (Shamay-Tsoory, Tibi-

Elhanany, & Aharon-Peretz, 2006). Strong correlations were demonstrated with measures of social functioning: SSPA Total scores (*r*= .91) and for Social Functioning Scale (SFS; Birchwood, Smith, Cochrane, Wetton, & Copestake, 1990) subscales, where 5/8 correlations were significant (r=.27–.60). There were also moderate correlations between IQ and VAMA total score (*r*=.41). As the study was of fair methodological quality, there is some limited positive evidence for hypothesis testing.

Reliability. Canty et al. (2015) investigated the internal consistency of the VAMA, which they found to be generally good for the separate subscales (Cronbach αs = .69 – .84). However, due to lack of factor analysis the study had poor methodological quality, thus overall evidence for internal consistency is unknown. Test-retest reliability was found to be excellent for the different subscales (ICCs: .93 – .99), and as the study was of fair methodological quality, this provides limited positive evidence for the VAMA's test re-test reliability.

Interpretability. Means of the four VAMA subscales were presented graphically, with no information provided on SDs or the range of scores, and so the presence of floor or ceiling effects could not be determined.

4. Discussion

This systematic review was conducted to identify the measurement properties of MSR tasks for use in adolescent and adult populations, taking the methodological quality of the studies into account. Eighteen studies reporting on the measurement properties of 16 different MSR tasks were included. No one MSR measure achieved positive ratings across all the measurement properties. The MASC, SST, TASIT and VAMA tasks had the most measurement property domains assessed by the included studies, although the RMIF, SAT and VAMA tasks were the only tasks to achieve positive ratings on more than one measurement property. The VAMA achieved the highest number of positive ratings, with positive evidence for content validity, test-retest reliability and construct validity (as assessed by hypothesis testing). For the remaining tasks, future research is clearly needed to provide more conclusive evidence about the adequacy of their measurement properties.

4.1 Content Validity and Ecological Validity

Nine of the included studies provided some information about the how the measure was developed, which indicated positive evidence of content validity for the MASC, RMIF, TASIT and VAMA tasks. A difficulty was that many studies did not make clear what dimensions of the MSR construct the task was aiming to measure (e.g. only cognitive or affective aspects, or which types of non-literal utterances) which made comprehensiveness difficult to assess. Additionally, a criteria set out by Terwee et al. (2007), is whether there was involvement of the target population in the measure's development. This was also difficult to assess, as although a number of studies referred to pilot work, very little information was provided with respect to the sample size and demographics of the pilot participants, or what information was used from this initial work in terms of developing the measure. It is possible that target population involvement in the development of objective performance based tasks with 'correct' answers, is less well defined than for self-report health outcome questionnaires (for which the measurement property and COSMIN criteria were initially developed). For example, it might be harder to see how target populations (such as those with ASD) might be used in terms of consultation about the relevance and comprehensiveness of task items, which may require insight and expertise regarding social communication difficulties. However, target population involvement is still likely to be an important process in selecting the most relevant or sensitive items, as well as for ensuring participant understanding (as done by McDonald et al., (2003) in the development of the TASIT). Thus, to improve the content validity of future MSR tasks, researchers should aim to carry out pilot work in relevant populations, as well as to provide details about this work in their published studies.

In terms of ecological validity, the tasks varied significantly in their presentation format. The tasks that employed story and picture-based stimuli (the FBST, Hinting task, SS, SST and ToM Stories) and those using animated shape stimuli (the FH-A, SAT and SAT-MC) lacked ecological validity, in that they did not approximate the demands of real-life MSR. Most of tasks attempted to improve ecological validity by using audio-visual dynamic stimuli (the BLERT, TASIT, MASC, RMIF, AMT, CASP and VAMA), which integrate audio and visual social (perceptual and linguistic) information, within a social context. However, these measures have a limitation in that they require the test taker to adopt a second person perspective upon a pre-recorded social interaction. The VAMA task is the exception to this, as it uses virtual reality technology that enables the test taker to adopt a first person perspective in real-time interactions. As virtual reality technology continues to improve and become more widely available, more MSR tasks can make use of this technology to improve ecological validity. However, as more ecologically valid tasks often place greater demands on domain-general processes such as language and executive

functioning, for some researchers ecological validity will be less important. There is therefore a trade-off between ecological validity and a task's 'purity' in assessing separate aspects of mental state attribution (e.g. verbal versus non-verbal, or cognitive versus affective) that needs to be considered when selecting tasks.

4.2 Construct Validity

With respect to construct validity, only two studies investigated structural validity, which demonstrated a single factor for the FBSQ and three factors for the ToM Stories task, however neither study reported how much variance was explained by each factor. Assessing structural validity is important in terms of demonstrating what dimensions of MSR (and indeed other abilities) a task might be tapping into. This in turn can inform what specific abilities might be impaired in different individuals. Thus, improved understanding of the structural validity will be a worthwhile goal of future research.

All studies assessed construct validity by means of hypothesis testing (comparison of groups and correlations with other constructs). The majority of the studies were rated as having poor methodological quality, usually due to small sample sizes in analyses (especially in subgroup analyses) and not reporting on the psychometric properties of comparator instruments. This means that for the majority of the studies the construct validity of the tasks assessed has not been adequately determined. A few studies of fair methodological quality resulted in limited positive evidence for the construct validity of the CASP, FBST, RMIF, SAT, SAT-MC, SST and ToM Stories tasks.

An important issue when assessing construct validity is whether studies controlled for the general intellectual ability of participants, which serves as a potential confound in MSR performance. Indeed, a number of studies demonstrated significant associations between measures of intellectual ability and MSR task performance (on the AMT, FBST, RMIF, SST, TASIT, ToM Stories and VAMA). This relationship is unsurprising given that these tasks are all verbal in nature (perhaps with the exception of the FBST), which increases reliance on language based abilities, including memory, comprehension and executive functioning. This highlights the importance of matching groups on intellectual ability in comparison studies, which although was attempted in the majority of studies, some did not (e.g. Bell et al. 2010 and Bell et al. 1997). Additionally, although Pinkman et al., (2015) did not match groups (individuals with schizophrenia versus controls) in terms of ability, they investigated whether MSR measures were able to predict social functioning over and above neurocognitive measures. Future studies would

benefit from including similar analyses, as this is provides more convincing evidence of a measure's construct validity that merely demonstrating group differences.

In addition to studies controlling for general ability, a number of the instruments incorporated control items, with the aim of accounting for more domain general abilities, such as comprehension and memory. Eight of the tasks in the current review had such control items (MASC, FH-A, SS, SST, AMT, FBST, ToM Stories and VAMA) although there is evidence (e.g. from inspections of means, SDs and ranges of scores; see Appendix) that performance on these items may suffer from ceiling effects, which was the case for the AMT, FH-A and MASC. This highlights the difficulty of having control questions that are as equally cognitively demanding as test questions, which is important for ensuring an instrument's specificity in measuring MSR ability.

4.3 Reliability

Internal consistency was investigated in nine studies, for seven MSR tasks. For the MASC, TASIT and VAMA the internal consistency was found to be adequate (Cronbach's αs > .7). The BLERT, Hinting task and SAT-MC demonstrated adequate internal consistencies in SCZ patient groups, but not in control samples. The SST and the FBST demonstrated inadequate internal consistencies. However, as none of the studies performed factor analysis in order to check the (uni)dimensionality of the measures, the Cronbach's αs cannot be properly interpreted. Thus, researchers are encouraged to check a measure's dimensionality before reporting internal consistencies.

Test-retest reliability was studied in six studies for the BLERT, Hinting task, MASC, TASIT and VAMA tasks. The MASC and VAMA demonstrated adequate test-retest reliability (ICCs > .7, or Pearson's *r*s > .8), although for the MASC this finding is inconclusive due to the study's poor methodological quality rating. The BLERT, Hinting task and TASIT were found to have inadequate test-retest reliability, in studies of fair methodological quality. Inter-rater reliability was studied in six studies for the EAT, FH-A, MASC, SAT, SS and SST tasks, all of which demonstrated excellent reliability. However, due to poor methodological quality of the majority of studies (mainly due to small samples), the findings are inconclusive for the FH-A, MASC, SST and SS tasks. Thus the review found limited positive evidence for only the AMT and SAT tasks, in terms of inter-rater reliability. Therefore, future research would benefit from using larger samples to assess reliability.

4.5 Responsiveness

None of the studies evaluated the measurement property of responsiveness, which as mentioned previously is likely due to MSR tasks not being primarily used as outcome measures. However, this is likely to become an increasingly relevant property of MSR assessments, as psychological interventions that target MSR have been developed (e.g. mentalisation-based therapies for BPD; Bateman & Fonagy, 2010), which will need reliable, valid and responsive measures of MSR for their evaluation. This is an important focus for the future development and evaluation of MSR measures.

4.6 Interpretability

The majority of studies provided means and SDs of scores of subgroups. For seven studies enough information was provided to determine the percentage of participants scoring the highest possible score (where more than 15% indicates ceiling effects). These studies demonstrated no evidence of ceiling effects for the BLERT, Hinting task, MASC, SAT-MC, SST, and TASIT. The AMT and SS tasks demonstrated no ceiling effects in their patient (ASD) samples, although ranges showed that control participants were scoring the highest score, although the number who did so was not reported. No data was reported on the CASP, EAT, FBST, FH-A, RMIF, SAT, ToM Stories and VAMA, and so whether these tasks suffer from ceiling effects is unknown. It would be greatly beneficial if future studies reported more information on ceiling effects, as this is critical for ensuring a task's utility in more able populations.

4.7 Limitations

This review has a number of limitations. The exclusion criteria limited our selection to only studies that were explicit in their aim to develop a new, or assess the psychometric properties of an existing, MSR task. This means that the review has likely missed many papers that did investigate some aspects of validity or reliability of the included MSR tasks. Additionally, the exclusion of tasks that were delivered in a non-English language meant that the cross-cultural validity of tasks could not be assessed. Another limitation was that the applicability of the COSMIN and Terwee et al., (2007) criteria in the assessment of MSR tasks (i.e. not self-report health-related outcome questionnaires) was at times questioned. For instance, many studies were penalised for not describing how missing data was handled (only able to score 'fair' on methodological quality), which may have proved more stringent criteria for studies assessing MSR tasks. This is because MSR tasks are usually administered 1:1 by an experimenter, thus reducing the possibility of missing items and the likelihood of studies to report on this. Additionally, as previously mentioned, when assessing content validity it was less clear what role

a target population should have in developing a performance-based measure. Another limitation of COSMIN was that it required some degree of subjectivity in the rating process, for example, in determining methodological quality of test-retest reliability, the rater decides whether the time interval is “appropriate”. Whilst the COSMIN tool could be ‘customised’ to make it more applicable to the particular type of measures under review, this would impact the tool’s validity. Finally, as ecological validity was not a property evaluated by either COSMIN or Terwee et al (2007) measurement property criteria, this was not assessed in a standardised way. Future studies evaluating the ecological validity of tasks would benefit from the use of guidelines for this, to reduce the degree of subjectivity.

4.8 Clinical applications

The findings of this review suggest that the VAMA task has the best available evidence with respect to measurement properties, with positive evidence for content validity, hypothesis testing and test-retest reliability. The RMIF and SAT tasks also demonstrated positive evidence with respect to two measurement property domains. This does not mean that the remaining tasks are inadequate, but rather that their reliability and validity, without further research, is questionable. Ultimately, which task is likely to be of most value to clinicians and researchers, will depend on the population being assessed and the purpose of the assessment.

For researchers seeking instruments that do not suffer from ceiling effects in populations beyond childhood (e.g. for tracking developmental trajectories of MSR into adulthood), this review found a number of tasks to be suitable. However, it was noted that only two studies (Koning & Magill-Evans, 2001 using the CASP and McDonald et al., 2011 using the TASIT) used adolescent samples, which highlights the lack of MSR instruments that have evaluated in this population. It is possible that studies tend to use mixed child and adolescent participants, which (if the mean age was below 12) would have been excluded from this review. With the growing interest of the continued development of the social brain during adolescence (e.g. Blakemore, 2012; Fett et al., 2014) and the cognitive implications of this in terms of social and emotional development, validated tools for this age group are needed.

Many of the tasks in this review demonstrated sensitivity to MSR deficits in different clinical populations, including ASD (SAT, MASC, RMIF, SS, AMT, CASP, EAT, and FH-A), schizophrenia (SAT-MC, BLERT, Hinting task, TASIT) and traumatic brain injury (TASIT). These tasks might therefore be useful in shedding light on the social difficulties associated with these clinical

disorders. Furthermore, depending on what aspects of MSR are of interest (e.g. verbal or nonverbal, cognitive or affective) different tasks will be more appropriate. The VAMA, EAT, TASIT and the MASC appear to be the most comprehensive in terms of assessing multiple aspects of MSR over a large number of test items. However, simpler, quicker tasks that measure a narrower range of MSR abilities (such as the BLERT or FH-A task) may be more suitable for some types of study, such as neuro-cognitive research that seeks to investigate different aspects of MSR at a neuronal level.

For clinicians, MSR tasks are likely to be most useful for the purposes of clinical assessment and as intervention outcome measures. Therefore, tasks that are more closely related to real-world functioning may be more relevant. A number of tasks in the current review were shown to predict social functioning, including the BLERT, Hinting task, CASP and VAMA. Such associations can also help to delineate what aspects of MSR might be most important for social functioning, which may in turn inform the development of specialised treatments.

4.9 Future directions

This review has highlighted that for the vast majority of included MSR tasks, many measurement properties had not been investigated. In particular, few studies investigated the structural validity of measures and none of the included studies considered the responsiveness of tasks. Furthermore, for those measurement properties that have been assessed, such as internal consistency, hypothesis testing and reliability, the evidence is largely indeterminate due to the poor methodological quality of studies. Thus, larger, well-designed studies using appropriate statistical methods are greatly needed, to provide more conclusive evidence about the of measurement properties of MSR tasks. Of clinical relevance, more information about the construct validity of tasks, in terms of predicting real-world functioning, will be an important avenue for future research.

With respect to the development of future MSR tasks, it is appreciated that there are inherent challenges with this, as MSR is a complex multi-dimensional construct that is difficult to operationalise. It would be helpful if more researchers provided details about the aim of their measure, as well as the concept(s) they seek to assess with it. For those wishing to develop highly ecologically valid instruments, which enable the test taker to engage in life-like MSR from a first person perspective, virtual reality technology may offer a fruitful option. Another challenge is the many trade-offs between different aspects of reliability and validity, such as

tensions between greater task comprehensiveness versus lower internal consistency, as well as higher ecological validity versus lack of task specificity. Some of these difficulties may be overcome, for instance, by the inclusion of well-developed control items to account for non-specific task demands.

4.10 Conclusions

The last three decades have seen a number of MSR tests be developed, which have been widely used in clinical, developmental, and neuro-cognitive research. However, evidence regarding their measurement properties is currently limited, or unknown, which may impact their current value in these fields. This review highlights the need for future research to address the inconclusive findings; to enable researchers and clinicians to make evidence based decisions when selecting MSR instruments.

5. References

Achim, A. M., Guitton, M., Jackson, P. L., Boutin, A., & Monetta, L. (2013). On what ground do we mentalize? Characteristics of current tasks and sources of information that contribute to mentalizing judgments.

Ahmadi, S. Z. Z., Jalaie, S., & Ashayeri, H. (2016). Validity and Reliability of Published Comprehensive Theory of Mind Tests for Normal Preschool Children: A Systematic Review. *Iranian Journal of Psychiatry*.

Klin, (2000). Attributing Social Meaning to Ambiguous Visual Stimuli in Higher-functioning Autism and Asperger Syndrome: The Social Attribution Task. *Journal of Child Psychology and Psychiatry*, *41*(07), 831–846.

Baimel, A., Severson, R. L., Baron, A. S., & Birch, S. A. J. (2015). Enhancing "theory of mind" through behavioral synchrony. *Frontiers in Psychology*, *6*, 870.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–18.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–51.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (n.d.). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17.

Bateman, A., & Fonagy, P. (2010). Mentalization based treatment for borderline personality disorder. *World Psychiatry : Official Journal of the World Psychiatric Association (WPA)*, *9*(1), 11–5.

Bell, M. D., Fiszdon, J. M., Greig, T. C., & Wexler, B. E. (2010). Social attribution test--multiple choice (SAT-MC) in schizophrenia: comparison with community sample and relationship to neurocognitive, social cognitive and symptom measures. *Schizophrenia Research*, *122*(1-3), 164–71.

Birchwood, M., Smith, J., Cochrane, R., Wetton, S., & Copestake, S. (1990). The Social Functioning Scale. The development and validation of a new scale of social adjustment for use in family intervention programmes with schizophrenic patients. *The British Journal of Psychiatry*, *157*(6), 853–859.

Blakemore, S. J. (2012). Development of the social brain in adolescence. *Journal of the Royal Society of Medicine*, *105*(3), 111-116.

Blijd-Hoogewys, E. M. A., van Geert, P. L. C., Serra, M., & Minderaa, R. B. (2008). Measuring theory of mind in children. Psychometric properties of the ToM Storybooks. *Journal of Autism and Developmental Disorders*, *38*(10), 1907–30.

Brune, M. (2005). "Theory of Mind" in Schizophrenia: A Review of the Literature. *Schizophrenia Bulletin*, *31*(1), 21–42.

Bell, M., Bryson, G., & Lysaker, P. (1997). Positive and negative affect recognition in schizophrenia: a comparison with substance abuse and normal control subjects. *Psychiatry research*, *73*(1), 73-82.

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., … Singh, S. (2005). Synchrony in the onset of mental-state reasoning: evidence from five cultures. *Psychological Science*, *16*(5), 378–84.

Canty, A. L., Neumann, D. L., Fleming, J., & Shum, D. H. K. (2015). Evaluation of a newly developed measure of theory of mind: The virtual assessment of mentalising ability. *Neuropsychological Rehabilitation*, 1–37.

Corcoran, R., Bentall, R. P., Rowse, G., Moore, R., Cummins, S., Blackwood, N., … Shryane, N. M. (2011). Making sense of theory of mind and paranoia: the psychometric properties and reasoning requirements of a false belief sequencing task. *Cognitive Neuropsychiatry*, *16*(6), 547–61.

Davis, M. H. (n.d.). Measuring individual differences in empathy: Evidence for a multidimensional approach.

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: a new task for assessing theory of mind in adults. *PloS One*, *8*(11).

Dziobek, I. (2012). Comment: Towards a more ecologically valid assessment of empathy. *Emotion Review*, *4*(1), 18-19.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., … Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*(5), 623–36.

Elliott, S. N., & Gresham, F. M. (1993). Social skills interventions for children. *Behavior modification*, *17*(3), 287-313.

Fett, A. K. J., Shergill, S. S., Gromann, P. M., Dumontheil, I., Blakemore, S. J., Yakub, F., & Krabbendam, L. (2014). Trust and social reciprocity in adolescence–a matter of perspective-taking. *Journal of adolescence*, *37*(2), 175-184.

Golan, O., Baron-Cohen, S., & Hill, J. (2006). The Cambridge Mindreading (CAM) Face-Voice

Battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *Journal of Autism and Developmental Disorders*, *36*(2), 169–83.

Golan, O., Baron-Cohen, S., Hill, J. J., & Golan, Y. (2006). The "reading the mind in films" task: complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience*, *1*(2), 111–23.

Happé, F., & Frith, U. (1996). Theory of mind and social impairment in children with conduct disorder. *British Journal of Developmental Psychology*, *14*(4), 385–398.

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154.

Harkness, K. L., Jacobson, J. A., Duong, D., & Sabbagh, M. A. (2010). Mental state decoding in past major depression: Effect of sad versus happy mood induction.

Heavey, L., Phillips, W., Baron-Cohen, S., & Rutter, M. (2000). The Awkward Moments Test: a naturalistic measure of social understanding in autism. *Journal of Autism and Developmental Disorders*, *30*(3), 225–36.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243-259.

Henry, J. D., Cowan, D. G., Lee, T., & Sachdev, P. S. (2015). Recent trends in testing social cognition.*Current opinion in psychiatry*, *28*(2), 133-140.

Johannesen, J. K., Lurie, J. B., Fiszdon, J. M., & Bell, M. D. (2013). The Social Attribution Task-Multiple Choice (SAT-MC): A Psychometric and Equivalence Study of an Alternate Form. *ISRN Psychiatry*, *2013*(4), 830825.

Kaland, N., Smith, L., & Mortensen, E. L. (2007). Response times of children and adolescents with Asperger syndrome on an "advanced" test of theory of mind. *Journal of Autism and Developmental Disorders*, *37*(2), 197–209.

Keulers, E. H. H., Evers, E. A. T., Stiers, P., & Jolles, J. (2010). Age, sex, and pubertal phase influence mentalizing about emotions and actions in adolescents. *Developmental Neuropsychology*, *35*(5), 555–69.

Koning, C., & Magill-Evans, J. (2001). Social and language skills in adolescent boys with Asperger syndrome. *Autism : The International Journal of Research and Practice*, *5*(1), 23–36.

Langdon, R., & Brock, J. (2008). Hypo- or hyper-mentalizing: It all depends upon what one means by "mentalizing." *Behavioral and Brain Sciences*, *31*(03), 274–275.

Maat, A., Fett, A.-K., & Derks, E. (2012). Social cognition and quality of life in schizophrenia. *Schizophrenia Research*, *137*(1-3), 212–8.

Mausbach, B. T., Harvey, P. D., Goldman, S. R., Jeste, D. V, & Patterson, T. L. (2007). Development of a brief scale of everyday functioning in persons with serious mental illness. *Schizophrenia*

Bulletin, *33*(6), 1364–72.

McDonald, S., Flanagan, S., Rollins, J., & Kinch, J. (2003). TASIT: A new clinical tool for assessing social perception after traumatic brain injury. *The Journal of head trauma rehabilitation*,*18*(3), 219-238.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (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), 737–45.

Moritz, S., Veckenstedt, R., Randjbar, S., Vitzthum, F., & Woodward, T. S. (2011). Antipsychotic treatment beyond antipsychotics: metacognitive intervention for schizophrenia patients improves delusional symptoms. *Psychological Medicine*, *41*(9), 1823–32.

Patterson, T. L., Moscona, S., McKibbin, C. L., Davidson, K., & Jeste, D. V. (2001). Social skills performance assessment among older patients with schizophrenia. *Schizophrenia Research*, *48*(2-3), 351–60.

Penn, D. L., Roberts, D. L., Combs, D., & Sterne, A. (2007). Best practices: The development of the Social Cognition and Interaction Training program for schizophrenia spectrum disorders. *Psychiatric Services (Washington, D.C.)*, *58*(4), 449–51.

Pinkham, A. E., Penn, D. L., Green, M. F., & Harvey, P. D. (2015). Social Cognition Psychometric Evaluation: Results of the Initial Psychometric Study. *Schizophrenia Bulletin*, sbv056–.

Roeyers, H., Buysse, A., Ponnet, K., & Pichal, B. (2001). Advancing advanced mind-reading tests: empathic accuracy in adults with a pervasive developmental disorder. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 271–8.

Schneider, L. C., & Struening, E. L. (1983). SLOF: a behavioral rating scale for assessing the mentally ill. *Social Work Research & Abstracts*, *19*(3), 9–21.

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., Brito, S. A. De, McCrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective Theory of Mind in adolescents and adults. *Social Cognitive and Affective Neuroscience*, *7*(1), 53–63.

Shamay-Tsoory, S. G. (2011). The neural bases for empathy. The Neuroscientist, 17(1), 18-24.

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, *1*(3-4), 149–66.

Sharp, C., & Vanwoerden, S. (2015). Hypermentalizing in Borderline Personality Disorder: A Model and Data. *Journal of Infant, Child, and Adolescent Psychotherapy*, *14*(1), 33–45.

Shryane, N. M., Corcoran, R., Rowse, G., Moore, R., Cummins, S., Blackwood, N., … Bentall, R. P. (2008). Deception and false belief in paranoia: modelling theory of mind stories. *Cognitive Neuropsychiatry*, *13*(1), 8–32.

Smeets, T., Dziobek, I., & Wolf, O. T. (2009). Social cognition under stress: differential effects of stress-induced cortisol elevations in healthy young men and women. *Hormones and Behavior*, *55*(4), 507-513.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., … de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42.

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & de Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *21*(4), 651–7.

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Quality of Life Research, 18(8), 1115-1123.

Van Donkersgoed, R. J. M., De Jong, S., Van der Gaag, M., Aleman, A., Lysaker, P. H., Wunderink, L., & Pijnenborg, G. H. M. (2014). A manual-based individual therapy to improve metacognition in schizophrenia: protocol of a multi-center RCT. *BMC Psychiatry*, *14*(1), 27.

Wellman, H. M., Cross, D., & Watson, J. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, *72*(3), 655–84.

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: a quick and objective test of Theory of Mind for adults with autism. *Autism Research : Official Journal of the International Society for Autism Research*, *4*(2), 149–54.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–28.

6. Appendix

Task and study	Study population	Mean (SD); Range
AMT Heavey et al., 2000	ASD group: N = 16 Control group: N = 15	Test questions: ASD: 3.13 (1.2); 1 – 5 CON: 5.27 (1.22); 3 – 7 Control questions: ASD: 5.94 (0.77); 5-7 CON: 6.67 (0.49); 6-7 Intention questions: ASD: 8.31 (3.74); 2 – 14 CON: 22.6 (7.41); 9-37
BLERT Bell et al., 1997	SCZ group: N = 50 SUB group: N = 25 CON group: N = 81	Total percentage correct: SUB: 77.2% SCZ: 64.8% College sample: 92.3%
BLERT Pinkham et al., 2015:	SCZ group: N = 179 CON group: N = 104	SCZ: 13.24 (3.82) Controls: 15.74 (2.89)
CASP Koning & Magill-Evans, 2001	SSD: N = 32 CON: N = 61	CASP ES scores: SSD: 29.53 (7.18) CON: 51.31 (8.17) CASP NS scores: SSD: 34.09 (14.07) CON: 65.93 (14.39)
EAT Roeyers et al 2001	PDD: N = 24 CON: N = 24	Video 1 PDD: 32.71 (13.49) CON: 33.31 (14.40) Video 2: PDD: 14.98 (8.82) CON: 23.26 (10.25)
FBST Corcoron et al., 2011	1: N = 39 2: N = 29 3: N = 33 4: N = 29 5: N = 20 6: N = 27 7: N = 29 8: N = 31	Not reported
FH-A White et al., 2011	ASD: N = 16 CON: N = 15	**Objective Scoring**: MCQ-categorisation total: ASD: 7.94 (1.95) CON: 10.00 (0.93) MCQ-feelings: ASD: 2.38 (2.00) CON: 5.47 (1.30) **Subjective Scoring:** Intentionality *Random:* ASD: 0.26 (0.57)

		CON: 0.47 (0.66) *Goal-directed:* ASD: 2.24 (0.41) CON: 2.21 (0.39) *ToM:* ASD: 3.02 (0.49) CON: 3.62 (0.60) <u>Appropriateness</u> *Random:* ASD: 4.51 (0.59) CON: 4.48 (0.54) *Goal-directed:* ASD: 3.04 (0.90) CON: 3.45 (0.66) *ToM:* ASD: 1.71 (0.70) CON: 2.66 (0.9)
Hinting task Pinkham et al., 2015:	SCZ group: N = 179 CON group: N = 104	SCZ: 13.65 (3.8) CON: 16.85 (2.01)
MASC Dziobek et al., 2006	AS: N = 19 CON:N = 20	MASC test questions: AS: 24.4 (5.9); 13-33 CON: 34.8 (2.7); 30 – 39 Control questions (4): AS: 3.9 (0.2); 3 – 4 CON: 3.9 (0.1); 3.5 – 4
RMIF Golan et al., 2006	ASD/HFA: N = 22 CON: N = 22	ASD: 14.96 (3.28) CON: 18.77 (2.41)
SAT Klin, 2000	HFA: N = 20 AS: N = 20 CON: N = 20	Controls: 1) Pertinence index: 6.6% (11.1) 2) Salience index: 78.4% (22.0) 3) ToM cognitive index 13.6% (10.7) 4) ToM affective index: 11.5% (7.4) 5) Animation index: 3.5 (1.4) 6) Person index: 4.2 (1.5) 7) Problem solving index: 79.3% (22.0) AS: 1) Pertinence index: 36.0% (23.8) 2) Salience index: 23.5% (20.6) 3) ToM cognitive index 4.5% (2.9) 4) ToM affective index: 2.9% (2.2) 5) Animation index: 1.9 (1.2) 6) Person index: 2.1 (1.2) 7) Problem solving index: 39.5% (27.0) HFA: 1) Pertinence index: 26.2% (18.0) 2) Salience index: 20.5% (20.4) 3) ToM cognitive index 4.3% (3.1) 4) ToM affective index: 2.5% (2.0) 5) Animation index: 1.6 (1.1) 6) Person index: 1.9 (1.0) 7) Problem solving index: 29.2% (24.2)

SAT-MC Bell et al., 2010	SCZ: N = 66 CON: N = 85	SCZ: 11(4); 2 – 19 CON: 15(3); 5/6 – 19
Johannesen et al., 2013	HP: N = 51	16.53 (2.23)
SST Dodell-Feder et al., 2013	HP: N = 74	MSR scores: 8.6 (2.6); 0-16 Control scores: 9 (1.2); 0-10
SS Happe, 1994	Able ASD: N = 18 MH: N = 13 Child CON: N = 26 Adult CON: N = 10	Mental test questions: Able ASD: 15.7 (4); 9 – 21 CON: 23.7 (0.7); 22-24 Verbal IQ control group: 21.4 (2); 17-24
TASIT McDonald et al., 2013	TBI adolescents: N = 16 CON adolescents: N = 16	TASIT 1: TBI: 23.81 (2.74) CON: 24.06 (2.41) TASIT 2: TBI: 50.44 (6.96) CON: 54.50 (3.14) TASIT 3: TBI: 52.88 (7.31) CON: 55.56 (4.37)
Pinkham et al., 2015:	SCZ group: N = 179 CON group: N = 104	SCZ:44.55 (7.55) CON: 51.44 (5.68)
ToM Stories Shryane et al., 2008	1: N = 39 2: N = 29 3: N = 33 4: N = 29 5: N = 20 6: N = 27 7: N = 29 8: N = 31	ToM total % correct: 1: 78.4 (17.3) 2: 86 (15.05) 3: 90.15 (11.1) 4: 65.2 (18.3) 5: 82.9 (11.6) 6: 90.7 (11.1) 7: 74.0 (15.0) 8: 85.2 (12.3)
VAMA Canty et al., 2015	HP: N = 65	Not reported

Empirical Research Project

Mental state reasoning in adolescent social anxiety

Supervised by Dr Jennifer Lau

and Dr Victoria Pile

Abstract

Adolescence is a developmentally sensitive period for the emergence of social anxiety disorder (SAD), a disorder characterised by fear of negative evaluation from others. This study investigated whether socially anxious youth have difficulties with understanding the mental states of other people, a key aspect of social cognition, which may play a role in the development and maintenance of SAD. A total of 159 students (aged 16-18) were recruited to investigate whether social anxiety symptoms were related to performance on two mental state reasoning (MSR) tasks. Sixty-five students performed the MSR tasks under conditions of 'social-stress' to investigate whether any impairments were due to the effects of anxiety or self-focused attention. Measures of social competence (via questionnaires and behavioral assessments) were also included, to explore whether social anxiety was associated with poorer social skill, and if so, whether dysfunctional MSR skills could explain this. Social anxiety was found to significantly predict 'hyper-MSR' on one task, and superior MSR performance on another task. The social stress condition did not significantly impact performance, even for those who were highly socially anxious, suggesting that these are more habitual patterns of MSR. However, mediation analyses indicated that state anxiety and a form of self-focused attention ('public awareness') may mediate increased hyper-MSR in those with higher social anxiety, indicating a role for state factors. Finally, although social anxiety was related to lower ratings of social competence, this was not the case for assessor-rated measures. Overall, these findings suggest that social anxiety may be characterised by MSR abilities at the upper end of the normal continuum. If found to play a role in the development or maintenance of the disorder, then these findings offer preliminary support for the use of interventions that target hyper-MSR in SAD, which have already been developed for other clinical populations.

Contents

1. Introduction 64

1.1 Cognitive models of Social Anxiety 64

1.2 Social Competence Theories 66

1.3 Mental State Reasoning and Social Anxiety 66

1.4 Clinical Significance of MSR in Adolescent Social Anxiety 68

1.5 Current Study Aims 69

1.6 Primary Hypotheses 69

1.7 Secondary Hypotheses 70

2. Method 70

2.1 Design 70

2.2 Power Analysis 70

2.3 Participants 70

2.4 Procedure 71

2.5 Materials 73

2.5.1 Baseline measures 73

2.5.2 MSR tasks 75

2.5.3 Manipulation checks 76

2.5.4 Social Behaviour Assessment 77

2.6 Statistical Analysis 77

2.6.1 Missing data 77

2.6.2 Data analysis over-view 78

2.6.3 Data assumptions 79

3. Results 79

3.1 Group characteristics 79

3.1.1 Demographics 79

3.1.2 Baseline measures 80

3.1.3 Measures of self-focus and state anxiety 80

3.2 MSR task validity and ceiling effects 82

3.3 Primary analyses 82

3.3.1 Bivariate correlations between baseline measures and MSR variables 82

3.3.2 Hypothesis 1: Social anxiety will predict MSR ability 85

3.3.3 Hypothesis 2: Social Anxiety will be more strongly related to MSR under conditions of social stress 88

3.3.4 Hypothesis 3: State anxiety and / or self-focused attention will mediate the relationship between SAS-A and MSR Variables 89
3.4 Secondary analyses **92**
3.4.1 Hypothesis 4: Social Anxiety will be related to poorer social competence 92
3.4.2 Hypothesis 5: Measures of MSR performance will explain any relationship between social anxiety and social competence 93
4. Discussion **93**
4.1 Clinical implications **97**
4.2 Limitations **97**
4.3 Conclusions **98**
5. References **99**
6. Appendices **105**
6.1 Appendix A: Ethical approval letters **105**
6.2 Appendix B: Information sheets and consent forms **108**
6.3 Appendix C: Study Baseline Measure **116**
6.4 Appendix D: Instructions and description of the MSR tasks **121**
6.5 Appendix E: State Anxiety and Self-focused attention measure **123**
6.6 Appendix F: Behavioural task checklist for assessors **125**
6.7 Appendix G: Demographic and baseline characteristics of students who completed the behaviour task **127**

List of Tables

Table 1: Group Characteristics..... 81
Table 2: Bivariate correlations between study measures 83
Table 3: Output for hierarchical regression with MASC total score as the DV..... 85
Table 4: Output for hierarchical regression with MASC hyper-MSR error score as the DV 86
Table 5: Output for hierarchical regression with FH-A total score as the DV 86
Table 6 Output for hierarchical regression with FH-A MI score as the DV..... 87
Table 7: Pearson's bivariate correlations between SAS-A and the state measures..... 88
Table 8: Pearson's partial correlations between SAS-A and the MSR variables..... 89

List of Figures

Figure 1 The study procedure..... 72
Figure 2: Model of social anxiety as a predictor of MASC total score mediated by state anxiety 90
Figure 3: Model of social anxiety as a predictor of MASC hyper-MSR errors mediated by state anxiety..... 91
Figure 4: Model of social anxiety as a predictor of MASC hyper-MSR errors mediated public self-awareness 92

1. Introduction

Social anxiety disorder (SAD) involves fear of negative evaluation by others in social or performance situations (American Psychiatric Association, 2013). Exposure to such situations consistently leads to marked anxiety, and so such situations tend to be avoided or endured with extreme distress (Veale, 2003). SAD is one of the most persistent of the anxiety disorders (NICE, 2013) and individuals with SAD will have an increased likelihood of future comorbid anxiety disorders, depression and substance abuse (Brunello et al., 2000). The transition into adolescence has been marked as a developmentally sensitive period for the emergence of SAD, with the majority of cases developing around mid-adolescence, and almost all cases developing before the age of 20 (Beesdo et al., 2007; NICE, 2013). Young sufferers are likely to have poorer outcomes in terms of academic achievements, as well as social and romantic relationships, which can have long standing effects that are hard to overcome in later life (Leigh & Clark, 2016).

The period of adolescence is characterised by a myriad of physical, neurobiological, social-emotional and cognitive changes. Cognitively, adolescents engage in increased metacognition (thinking about their own thoughts) as well as consideration of the thoughts of others (Elkind & Bowen, 1979; Selman, 1980), which happens in combination with increasing importance placed on the opinions and perceptions of peers. Adolescents are also more likely to experience enhanced 'public self-awareness', a form of self-focused attention on the outwardly observable aspects of the self and the impression one may be giving to others (Higa, Phillips, Chorpita, & Daleiden, 2008; McMillan and Ebesutani, 2011) and they are more likely to use social comparison as a method of self-evaluation (Pfeifer et al. 2009). Some of these cognitive changes are thought to be underpinned by maturing neural circuits and are likely to be adaptive as adolescents begin to engage in increasingly complex social situations and peer networks (Haller, Cohen Kadosh, Scerif, & Lau, 2015). However, it is possible that for some vulnerable individuals, this developmentally-typical heightening of social concerns may highlight risk factors associated with the development of SAD. Understanding how symptoms arise during this unique period is important for devising and implementing early interventions.

1.1 Cognitive models of Social Anxiety

According to cognitive models, biased information processing of social information is key to the development and maintenance of SAD (e.g. Clark & Wells, 1995). As social information is inherently ambiguous, individuals constantly have to track and interpret complex social cues.

When this is done in a negatively biased way, for instance when a person with SAD interprets somebody laughing as evidence for having done or said something foolish, this can lead to significant distress and avoidance of similar situations. Indeed, experimental paradigms have demonstrated that socially anxious individuals tend to evaluate themselves more negatively, and mistakenly believe that others judge them negatively during a social interaction (Niels Christensen, Stein, & Means-Christensen, 2003).

In addition to cognitive biases, which have been well demonstrated in adult (Clark & McManus, 2002) and, albeit to a lesser extent, in adolescent SAD (e.g. Miers, Blöte, de Rooij, Bokhorst, & Westenberg, 2013) excessive self-focused attention is also considered to be a key maintaining factor. Self-focused attention can be divided into two different types: 1. enhanced awareness of self-relevant thoughts, feelings, and internal sensations (*private self-awareness*) and 2. enhanced awareness of the externally observable aspects of the self and the impression one is giving to others (*public self-awareness*). Both private and public forms are hypothesised to lead to further anxiety, as awareness of negative thoughts, feelings and self-impressions is intensified, and so reduces the possibility of noticing positive or neutral social feedback (George & Stopa, 2006). Indeed, excessive self-focus, particularly in the form of heightened public self-awareness, has been shown to characterise highly socially anxious youth (Miers et al., 2013; Higa et al., 2008).

Arguably the best support for cognitive models comes from the demonstrated effectiveness of Cognitive Behavioural Therapy (CBT) in treating SAD, which targets both biased cognitions and self-focused attention. The success of CBT in adults has been demonstrated in a number of meta-analyses (e.g. Fedoroff & Taylor, 2001; Gould, Buckminster, Pollack, Otto, & Massachusetts, 1997; Taylor, 1996) and is currently a first line treatment for SAD recommended by NICE (NICE, 2013). However, the evidence base of CBT with young people is less strong (NICE, 2013) and generic CBT interventions for anxiety appear to be less effective with social anxiety than other anxiety disorders (Ginsburg et al., 2011; Hudson et al., 2015). A recent case series using specialised CBT with socially anxious adolescents highlighted the need for adaptations to the adult treatment protocol, such as liaising with schools to support friendship building (Leigh & Clark, 2016). The difficulty of ongoing peer victimisation was also emphasised, as this can serve to reinforce negative cognitions and low self-esteem. Thus, it is possible that additional factors that are not targeted in existing CBT packages, such as additional support with peer relationships, might be more relevant in the maintenance of social anxiety in young people.

1.2 Social Competence Theories

Unlike cognitive accounts of SAD, social competence theories suggest that an individual's negative expectations and evaluations regarding social ability may in fact be the result of previous poor performance and negative outcomes in social situations (Banerjee, 2008). In line with this, young people with social anxiety are more likely to be disliked and rejected by peers, and have fewer friendships (Greca & Lopez, 1998; Ranta, Kaltiala-Heino, Fröjd, & Marttunen, 2013; Ranta, Kaltiala-Heino, Pelkonen, & Marttunen, 2009), implying actual deficits in social competence. Additionally, researchers have found that young people with SAD exhibit significantly poorer social skills than non-anxious peers, as rated by self- and parent- report, as well as behavioral assessments rated by both adults and peers (Alfano, Beidel, & Turner, 2006; Inderbitzen-Nolan, Anderson, & Johnson, 2007; Miers, Blöte, & Westenberg, 2010; Scharfstein, Beidel, Sims, & Rendon Finnell, 2011; Spence, Donovan, & Brechman-Toussaint, 1999), although this finding not always been replicated (Cartwright-Hatton, Tschernitz, & Gomersall, 2005; Erath, Flanagan, & Bierman, 2007). Further evidence in favour of problematic social skills is that interventions targeting social skills can achieve significant therapeutic benefits for adolescents with social anxiety (Spence, 2003). However, what is less clear is whether socially anxious adolescents also have difficulties with understanding social situations - that is whether they struggle with core social cognitive abilities such as mental state reasoning.

1.3 Mental State Reasoning and Social Anxiety

Mental state reasoning (MSR) is the ability to understand the mental states (e.g. thoughts, intentions, and desires) of other people, which can be used to make sense of others' actions as well as help guide our own behaviour and responses. A number of closely related terms have been used to describe this skill, including theory of mind (ToM), mentalising, mindreading, social understanding, socio-cognitive ability and perspective-taking.

One can see how difficulties with MSR could contribute to the onset and maintenance of social anxiety, in that not being able to fully or accurately understand the beliefs, desires and behaviours of other people, could produce a threatening social world. This is in line with studies which have found that the cognitive error of 'mind reading' (making assumptions about what others are thinking about you) is the strongest predictor of anxiety symptoms in young people (Schwartz & Maric, 2014). This might be especially apparent in adolescence, a time when the

social environment is perhaps particularly intimidating, as concerns about peer rejection become paramount (Petersen & Leffert, 1995).

As well as MSR difficulties possibly increasing an individual's vulnerability to experiencing social anxiety, cognitive-behavioural explanations of social anxiety would also predict that high levels of social anxiety could negatively impact the development of MSR. This could occur, for instance, via mechanisms such as long-term avoidant behavior (reducing social experience), negative information processing biases (distorting social experience) and excessive self-focus (reducing attention to important features of social interactions). Indeed, allocating attentional resources to self-monitoring processes in SAD has been shown to be detrimental to task performance on unrelated, complex social tasks (Hope, Heimberg, & Klein, 1989). Finally, it is also possible that a reciprocal relationship exists between MSR difficulties and increased social anxiety.

Some studies have shown that higher levels of social anxiety in young people is associated with dysfunctional MSR abilities, in terms of significantly longer reaction times (Melfsen & Florin, 2002) and reduced accuracy on facial affect recognition tasks (Simonian, Beidel, Turner, Berkes, & Long, 2001), which require MSR based on a static facial expression. Additionally, on a more advanced MSR task, requiring higher-order skills such as understanding false beliefs, deception and irony, Banerjee and Henderson (2001) found that social anxiety was associated with poorer understanding of faux-pas scenarios in a sample of 6-11 year olds.

MSR impairments have also been demonstrated in socially anxious adults. Hezel and McNally (2014) found that individuals with SAD were significantly impaired relative to healthy controls on the Movie for the Assessment of Social Cognition (MASC; Dziobek et al., 2006), a video-based, advanced measure of MSR. Those with SAD were more likely to make 'hyper-MSR' errors in their judgements of the characters' thoughts and feelings in the film clips, suggesting an over-interpretive style of MSR. This was replicated by Washburn et al., (2016) using the same task, and is in keeping with findings from Tibi-Elhanany and Shamay-Tsoory (2011) who found that compared to adults with low levels of social anxiety, the high socially anxious group demonstrated an elevated tendency for identifying emotions but relatively decreased accuracy on identifying others' thoughts and beliefs. The authors attributed this discrepancy to anxious participants' tendency to 'over-mentalise', which may result in inaccurate social perceptions.

What is less clear from the above studies, is why individuals with social anxiety might display dysfunctional MSR abilities. In particular, whether habitual trait-level difficulties in social understanding act as a vulnerability factor in the development of social anxiety, or whether state-level factors, involving established cognitive-behavioural mechanisms, such as state anxiety or self-focused attention, have a negative impact on MSR ability.

Hezel and McNally (2014) explored whether the cognitive process of self-monitoring could explain MSR impairments in social anxiety, by employing a 'cognitive load' condition to mimic the effects of self-focused attention, which involved participants memorising a sequence of symbols before completing the MSR tasks. Although this manipulation significantly impacted MSR ability in both groups, those with SAD performed significantly worse even with no cognitive load. Thus, the authors concluded that poorer MSR in SAD was not due to the cognitive effects of self-focus, implying that trait-level factors could be at play. However, it may have been that even under conditions of no experimental load, those with SAD were still engaging in self-monitoring processes, especially if the experimental situation itself was found to be anxiety provoking. Moreover, using a non-social memory task may not have been an ecologically valid or effective means of simulating the effects of self-focus in social anxiety. Therefore, in order to test the hypothesis that state factors, such as anxiety and self-focus, may explain MSR impairments in SAD, a manipulation that induces a sense of social evaluation or 'social threat' might be a more valid and powerful way of testing this.

1.4 Clinical Significance of MSR in Adolescent Social Anxiety

If MSR is found to be impaired in social anxiety, this would have implications for both the theoretical models regarding the development and maintenance of social anxiety, as well as treatments. Since there are a number of ways in which MSR can be disrupted (such as insufficient MSR or hyper-MSR), it is important to uncover the precise nature of any impairment in social anxiety. Impairments in MSR skills could underpin some of the social difficulties seen in individuals with social anxiety, such as reduced social competence and increased social fears. MSR errors could therefore become important targets for treatment, which could be delivered by extending current treatment programmes such as CBT, cognitive bias modification and social skills training. Given that adolescence reflects a protracted period of social learning and development (as well as a possible critical period for development of the 'social brain'; Blakemore, 2008), it is feasible that interventions delivered during this period may yield stronger and longer-lasting benefits (Lau, 2013).

1.5 Current Study Aims

The primary aim of the study was to investigate whether social anxiety is related to dysfunctional MSR in adolescence. Age-appropriate MSR tasks were selected, which included the Frith-Happe Animations task (FH-A; White, Coniston, Rogers, & Frith, 2011) and the MASC task (Dziobek et al., 2006), both of which have demonstrated group differences between socially anxious and non-socially anxious individuals in previous studies (e.g. Runchman et al, in prep; Hazel & McNally, 2014). The MASC is particularly informative due to enabling the qualitative analysis of errors (from no-MSR to hyper-MSR), which enables the exploration of whether social anxiety is related to hyper-MSR in adolescents, as has been demonstrated in adults (Hazel & McNally, 2014; Washburn et al., 2016).

Secondly, the study extends previous research by including a 'social stress' condition, involving the use of video cameras and an upcoming social performance task, with the aim of increasing self-focused attention and state anxiety in participants. This manipulation has been commonly used in social anxiety research (e.g. Chen et al., 2010; Mansell & Clark, 1999), and enabled the exploration of whether anomalies in MSR are due to the effects of induced state factors (anxiety and self-focus), or whether the difficulties are more consistent with habitual patterns of MSR.

Finally, measures of social competence (self-reported, teacher-reported, as well as observer and peer-rated behavioral assessments) were included, to explore whether social anxiety was indeed associated with poorer social skill, and if so whether MSR anomalies could explain this.

1.6 Primary Hypotheses

1) Social anxiety will predict poorer overall performance on the FH-A task and the MASC task. On the MASC task the reduced total score will be due to a higher number of hyper-MSR errors.

2) The relationship between social anxiety and MSR performance will be stronger under conditions of social stress.

3) Measures of self-focused attention and state anxiety will explain any relationship between social anxiety and impaired MSR.

1.7 Secondary Hypotheses

4) Social anxiety will be related to poorer social competence.

5) Impaired MSR performance will explain any relationship between social anxiety and poorer social competence.

2. Method

2.1 Design

The study employed a cross-sectional design to investigate MSR in social anxiety. All participants were tested in classroom settings where they completed a series of self-report questionnaires, computerised MSR tasks, and ratings of state anxiety and self-focused attention. Half of the classes were randomly allocated to a social stress induction condition, to investigate the impact of state anxiety and self-focused attention on the relationship between social anxiety and MSR ability. A subset of randomly selected students from the social stress condition went on to complete a conversation and speech task, which served as a behavioural assessment of social competence.

2.2 Power Analysis

A power analysis using G-power software calculated that a sample size of 69 would be needed to detect a relationship between social anxiety and MSR ability, with an effect size of r=.33 (based on Hezel and McNally's (2014) study) with 80% power, at the two-tailed 5% significance level. In order to detect a more conservative effect of r=.25, which was expected since the current study used an unselected non-clinical sample, a sample of 123 would be needed. The current study therefore aimed to recruit between 69–123 participants to test the primary hypothesis that social anxiety is related to MSR ability and a further 69-123 participants to form the social threat condition, to explore the role of state anxiety and self-focused attention on MSR ability.

2.3 Participants

Participants were typically developing adolescents aged between 16-18 years old, attending a large sixth form college in London. A total of 159 participants, from 11 mixed ability AS

Psychology classes, took place in the study. Five of the classes were randomly allocated to the social stress induction condition.

2.4 Procedure

Ethical issues and approval. Ethical approval for the study was granted from the King's College London Psychiatry, Nursing and Midwifery Research Ethics Sub-committee on 10th December 2014 and approval was also granted for a subsequent modification on 27th April 2015 (see Appendix A for documents). Information about the study was approved by the school head teacher and was provided to participants by the main researcher at least 48 hours prior to participation in the study. Written consent was obtained from all participants at the experimental stage of the study. Information sheets and consent forms can be found in Appendix B.

Testing procedure. Testing sessions were constructed to fit within a double lesson period of one hour and 30 minutes. An overview of the study procedure is illustrated in Figure 1.

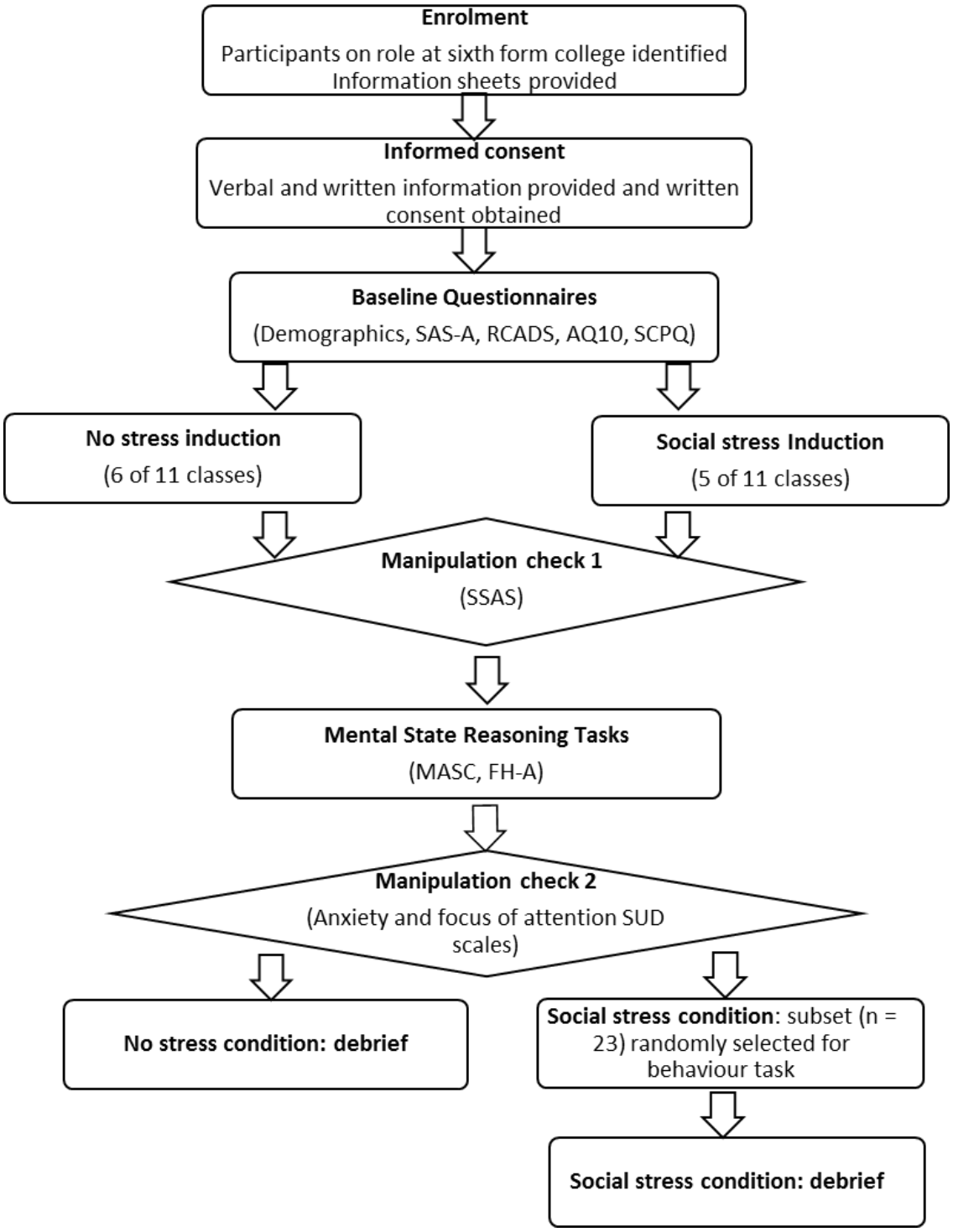


Figure 1 The study procedure

Note. SAS-A = Social Anxiety Scale-Adolescents; RCADS = The Revised Child Anxiety and Depression Scale; AQ-10 = Autism Quotient – 10 item version; SCPQ = Social Competence with Peers Questionnaires; SSAS = Situation Self Awareness Scale

Procedure for the 'social stress' condition

After written consent was received from each participant, the baseline questionnaires were completed in the classroom setting. Students were then informed that a random selection of them would be asked at the end of the testing session to have a conversation with some strangers, as well as give a speech in front of an audience and camera. They were told that they would be assessed on their social skills and public speaking ability. Participants then completed ratings of self-focused attention, before completing the two computerised MSR tasks. Additionally, the computer room was set up with multiple cameras on tripods, with the aim of enhancing self-focus during completion of the MSR tasks. After the MSR tasks participants then re-rated measures of self-focused attention and state anxiety before the researcher randomly selected 4-5 students from each class to complete the behavior assessment. After the behavioural task, three assessors (two adult and one unfamiliar peer) rated their performance. All participants were then debriefed about the study.

Procedure for the 'no-stress' condition

The procedure for the no-stress condition was identical to that of the social stress condition in terms of the questionnaires, rating scales and MSR tasks completed. However, students in the no-stress condition were not told about, nor asked to take part in, the conversation and speech task, nor did the computer room have any cameras set up.

2.5 Materials

2.5.1 Baseline measures

Copies of all baseline measures used in the current study can be found in Appendix C.

Social Anxiety Scale for Adolescents (SAS-A; La Greca & Lopez, 1998)

The SAS-A is a 22-item self-report measure of social anxiety symptoms, with 18 statements (such as 'I worry about what others say about me') and four filler items (such as 'I like to play sports'). Each item is rated on a five-point Likert-scale according to how much the item is 'true for you', ranging from 0 ('not at all') to 5 ('all the time'). Studies have found the scale to be a valid and reliable measure of social anxiety for both clinical and community samples in seven to 18 year-olds (Storch, Masia-Warner, Dent, Roberti, & Fisher, 2004) with a suggested clinical cut-off score of 50 (La Greca & Lopez, 1998). Test-retest reliabilities range from .47 to .75 for a six-month

interval (Vernberg, Abwender, Ewell, & Beery, 1992). Internal consistency for the total scale has been shown to be adequate (α = .87) (La Greca & Harrison, 2005) and was also .87 in the current sample.

The Revised Child Anxiety and Depression Scale (RCADS; Chorpita et al., 2000)

The Revised Child Anxiety and Depression Scale (RCADS) was used in the study to ensure equivalence of anxiety and depression symptoms between groups, as well as to assess the specificity of the relationship between MSR and social anxiety. The RCADS is a 47-item, youth self-report questionnaire with subscales including separation anxiety disorder (SEP), SAD, generalised anxiety disorder (GAD), panic disorder (PD), obsessive compulsive disorder (OCD), and major depressive disorder (MDD). It also yields a total anxiety scale (sum of the five anxiety subscales) and a total internalising scale (sum of all six subscales). Only the total anxiety and depression subscales were of relevance to the current study. Items are rated on a four-point Likert-scale from 0 ('never') to 3 ('always'). The scale has demonstrated good test-retest reliability, with intraclass correlation coefficients (ICCs) of .76-.86 over a three-month period (Kösters, Chinapaw, Zwaanswijk, van der Wal, & Koot, 2015) and the subscales have demonstrated good internal consistencies (Cronbach's αs = .78–.88, and were .71-.86 in the current sample) as well as good convergent and discriminant validity against both clinical interview and self-report criteria (Chorpita, Moffitt, & Gray, 2005).

Autism Quotient – 10-item version (AQ-10; Allison, Auyeung, & Baron-Cohen, 2012)

The AQ-10 was included in the study to ensure equivalence between groups, as well as to control for ASD traits when investigating the relationship between social anxiety and MSR. The AQ-10 was developed from the original 50-item version as a screening tool for clinicians. Responses are on a four-point scale ('definitely disagree', 'slightly disagree', 'slightly agree' and 'definitely agree') with responses indicating autistic traits scoring one, and others scoring zero. Higher scores correspond to more autistic traits, with a recommended clinical cut-off of six. Internal consistency was shown to be high in the validation study (α= .85), although for the current sample it was low (α = .22).

Social Competence with Peers Questionnaires (SCPQs; Spence et al, 1995)

The social competence with peers questionnaires assess social competence in relationships with peers in young people aged 8–18. It has different versions of the same items adapted for

different informants. The SCPQs were included in the study to explore relationships between social competence, social anxiety and MSR ability. The pupil version (SCPQ-P) is a 10-item self-report scale with items (e.g. 'I have at least one close friend') rated from 0 ('not true') to 2 ('mostly true'). The scale has been shown to have a single factor, and good internal consistency (Cronbach's α = .75; Spence et al., 1995)**.** Cronbach's α for the SCPQ -P in this study was .84. The teacher version (SCPQ-T) has nine items (e.g. 'has at least one close friend') also rated from 0 ('not true') to 2 ('mostly true') and has also been shown to have a single factor structure and good internal consistency (Cronbach α = .95; Spence et al., 1995). Cronbach's α for the SCPQ -T in this study was .94.

2.5.2 MSR tasks

'Movie for the Assessment of Social Cognition' (MASC; Dziobek et al., 2006)

The MASC task involves watching a 15-minute film of four people interacting at a dinner party. The film is paused at 45 points, and questions concerning the characters' feelings, thoughts and intentions are asked (e.g., 'What is Betty feeling?', 'What is Cliff thinking?', 'Why is Michael doing this?'). The task incorporates several aspects of social cognition (i.e., first- and second order false beliefs, deception, faux pas, persuasion, metaphor, sarcasm and irony). Each correct response is scored as one point and responses are added together to form a total score. In addition, the MASC allows for separate quantification of individuals' error types, including a tendency to make overly complex inferences ('hyper-MSR' errors), overly simplistic inferences ('less-MSR' errors), or a complete lack of inferences (no-MSR). The instructions to the MASC task, as well as an example question can be found in Appendix D. The MASC has high test-retest reliability (*r*=.97) and is sensitive to individual differences in the healthy population (Dziobek et al., 2006; Hezel & McNally, 2014).

Frith-Happé Animations (FH-A; White et al., 2011)

The FH-A task involves watching 12 silent animations featuring two triangles representing three different types of motion: no interaction (NI), physical interaction (PI) and mental interaction (MI). For the MI clips the triangles interact in such a way as to convey intentions toward the other character's mental state (e.g. coaxing, mocking, seducing, surprising). For each clip participants select what type of interaction they think is being displayed. For MI clips, participants are additionally asked to select what emotion they believe to have been felt by the two different triangles. The multiple-choice version of the task has been shown to be sensitive to the ToM deficit in individuals with ASD, who were specifically impaired in identifying the MI

clips, but not the PI or NI clips (White et al., 2011). The instructions provided to participants can be found in Appendix D.

2.5.3 Manipulation checks

Manipulation check measures can be found in Appendix E.

Situational Self-awareness scale (SSAS; Govern & Marsch, 2001)

The Situational Self-Awareness Scale (SSAS; Govern & Marsch, 2001) is a 9-item scale assessing different types of self-focused attention in participants, and was used in the current study to investigate the role of self-focused attention in MSR. It also served as a manipulation check of the social stress induction. Each item is rated on a 7-point scale, ranging from 1 ('strongly disagree') to 7 ('strongly agree'). It yields three subscales, two of which measure self-focus: 'private self-awareness' (e.g., 'Right now, I am conscious of my inner feelings'), which demonstrated a Cronbach's α of .70 in the validation study and was .60 in the current sample, and 'public self-awareness' (e.g., 'Right now, I am concerned about the way I present myself') which demonstrated a Cronbach's α of .82 in the validation study and was .80 in the current sample. The third subscale is a measure of 'situational awareness' (or non-self-focus) (e.g., 'Right now, I am conscious of all objects around me') with a Cronbach's α of .72 in the validation study and .61 in the current sample[4]. The measure has been shown to be sensitive to changes in self-focus over time and across situations, including laboratory manipulations aimed to increase self-awareness (Govern & Marsch, 2001).

Likert-scales

After the completion of the MSR tasks, 10-point Likert-scales were used as a further manipulation check. Participants rated the following questions: 'how anxious/nervous were you feeling during the video tasks?' and 'how much were you focused on your own thoughts and feelings during the video tasks?' on a scale of 0 ('not at all) to 10 ('extremely').

[4] Since the current study's aim was to investigate the effects of self-focused attention on MSR, only the two self-focus subscales were considered to be of relevance. Therefore, the 'situational awareness' subscale, which is not a measure of self-focused attention, was not included in the analyses.

2.5.4 Social Behaviour Assessment

A subset of randomly selected participants from the social stress condition also completed a behavioural assessment that comprises two standardised social tasks: a two-minute conversation with two strangers followed by a two-minute speech to the same two individuals. After each social task, the same two individuals (assessors) complete a 15 and 14 item checklist, for the conversation and speech task respectively, adapted from checklists used by Clark et al. (2006) and Cartwright-Hatton et al. (2003). These checklists covered their perception of how anxious the students appeared, and how well they performed (e.g., whether they appeared embarrassed or boring or whether their voice was quivering). The full checklists are in Appendix F. Items are rated on a five-point Likert scale ('not at all' to 'very much') and positive items (such as whether they appeared confident or socially skilled) are reverse scored. Items are then summed to give a total score, with higher scores indicating poorer social skill. In addition to the two independent adult assessors, an unfamiliar peer observer (from the same age group) completed the same checklists, using video footage of the original social interactions. The inclusion of peer-ratings was aimed to improve the ecological validity of the method, as it has been suggested that same age peers may be more sensitive than adults to small deviations in the behaviours of fellow young people (Miers et al., 2010). This social task was included as a behavioural measure of social competence in order to explore relationships between social competence, social anxiety and MSR ability. In the current study the scales showed good internal consistencies (Cronbach's α's =.78–.95) and excellent inter-rater reliability between all three assessors for both the conversation (ICCs = .88) and speech (ICC = .84) tasks.

2.6 Statistical Analysis

2.6.1 Missing data

Data were entered into and analysed using SPSS for Mac Version 22. Self-report scores for participants with missing data were scale corrected, whereby raw scores were divided by the maximum score that could have been obtained given the missing items, and multiplied by the maximum score of the full scale. No participant had more than 22% missing items and so none were excluded from analyses for any self-report measure[5]. For 17 individuals (14 from the no-stress condition) the FH-A task was not completed due to either technical errors or time

[5] One participant missed 2 items from the SSAS (equating to 22% missing data), one participant missed 2 items from the AQ-10 (equating to 20% missing data) and one participant missed 1 item from the SASA (equating to 4.55% missing data). None of the other study measures had missing data.

constraints, and nine participants (seven from the social stress condition) did not complete the MASC task, due to time constraints. None of these participants differed significantly in terms of demographic information or scores on any of the baseline measures from other participants in their experimental condition. This resulted in a final sample of N=142 (10.6% missing data) who completed the FH-A task and N=150 (5.7% missing data) who completed the MASC task.

2.6.2 Data analysis over-view

Differences between groups in demographic characteristics were assessed using Chi-squared analysis for categorical variables and t-tests for continuous variables. The MSR tasks were then examined for ceiling effects by calculating the percentage of participants scoring at the highest, and second highest marks. Additionally, the association between the two tasks (FH-A and MASC) was examined to assess their construct validity.

Pearson's correlations were used initially to explore relationships between all baseline measures and MSR variables, which then informed which variables to control for in the regression analyses.

To test hypothesis 1, that social anxiety predicts MSR ability, hierarchical multiple linear regressions were conducted, with potentially confounding variables (as shown in the correlation analyses) entered at Step 1 and social anxiety scores (as measured by the SAS-A) entered at Step 2. Analyses were repeated for each MSR dependent variable (on the MASC and FH-A tasks) separately. To take into account the nested structure of the data (that individuals were grouped in classes) cluster analysis around class ID was also conducted. To test hypothesis 2, that social anxiety would interact with the social stress condition to predict further variance in MSR performance, an interaction term (SAS-A * condition) was added at Step 3 in each of the regression analyses.

To test hypothesis 3, that measures of state anxiety and self-focused attention would explain any variance in the relationship between social anxiety and MSR, mediation analyses were conducted using the SPSS bootstrap macro PROCESS (Hayes, 2013). For each mediation analysis, 1000 bootstrap samples were used and the bias corrected accelerated 95% confidence interval reported. Additionally, the Sobel tests (a more conservative test of mediation) were included, as an additional test of the significance of each mediation analysis.

To test hypothesis 4, that social anxiety is related to poorer social competence, linear regressions were conducted, using the SCPQ scores for the whole sample analyses, and assessor-rated outcome measures for analyses involving the subset of students who completed the social behavior task. To test hypothesis 5, that MSR ability will explain any relationship between social anxiety and measures of social competence, mediation analyses using the PROCESS macro were conducted.

2.6.3 Data assumptions

All variables were assessed for normality and homogeneity of variance. Normality was assessed by calculating z-scores for skewness and kurtosis and using a recommended cut-off of 2.58 (Field, 2009). Variables that violated parametric assumptions were transformed (log 10 and square root transformations) and their normality re-assessed. All variables met parametric assumptions after transformation. All reported values are from untransformed data for readability. All regression models were checked for multicollinearity, and all variables reported a tolerance of more than 0.1 and Variance Inflation Factors of less than 10. Assumptions of independent errors for each model were also met (Durbin-Watson=1.88-2.21) and examination of scatter-plots of standardised residuals against standardised predictors revealed no violations of homoscedasticity or linearity, and the histograms and normal P-P plots revealed no concerns over normality of standardised residuals in any analysis.

A two-tailed significance level of $\alpha = .05$ was used throughout the analysis.

3. Results

3.1 Group characteristics

3.1.1 Demographics

Group characteristics can be seen in Table 1. The majority of the sample were female (76%) and 'Black British' (72%). Fisher's exact test revealed no differences between the control and social stress group in terms of gender (p= .189) or ethnicity (p= .871). There was a statistically significant difference between the groups in terms of age, with participants in the no-stress

condition being older (mean difference = 0.13 years) than those in the social stress condition [$t(155) = -2.16$, $p = .025$, $d = .36$].

3.1.2 Baseline measures

Means and standard deviations of scores on the baselines measures, per condition, can be found in Table 1. Scores on the SAS-A demonstrated that 17% of the total sample were scoring above the recommended clinical cut off of 50, which is a similar proportion to that demonstrated in the original scale validation study, where 14% of the community sample scored above cut off (La Greca & Lopez, 1998). There were no significant differences between the conditions in terms of mean SAS-A scores [$t(1, 152) = -.19$, $p = .852$], or the proportion of students scoring above the recommended clinical cut off [$\chi^2(1) = .02$, $p = .895$]. There were also no differences between groups on the AQ-10, SCPQ-P or RCADS subscales (ps > .237). However, there was a significant group difference on the teacher-rated version of the SCPQ [$t(1, 118) = 2.81$, $p = .006$, $d = .51$), which is likely to reflect rater bias due to one teacher rating students from the no-stress condition, and a different teacher rating students from the social stress condition. Therefore, the SCPQ-T was excluded from further analyses.

3.1.3 Measures of self-focus and state anxiety

Measures of self-focus and state anxiety (as measured by the SSAS before completing the MSR tasks, and the SUD scales post-task) also revealed no group differences (all ps > .253), suggesting that the social stress manipulation may not have been effective.

Table 1: Group Characteristics

	No-stress (N = 94)	Social stress (N = 65)
Demographics		
Age	17.23 (SD .39)	17.36 (SD .33)
Gender	67 females (71.3%) 27 male (28.7%)	53 females (81.5%) 12 male (18.5%)
Ethnicity	Black / Black British: 65 (69.1%) Mixed / Multiple ethnicities: 12 (12.8%) White: 7 (7.4%) Asian / Asian British: 3 (3.2%) Chinese: 1 (1.1%) Other: 6 (6.4%)	Black / Black British: 49 (75.4%) Mixed / Multiple ethnicities: 6 (9.2%) White: 4 (6.2%) Asian / Asian British: 3 (4.6%) Chinese: 0 (0%) Other: 3 (4.6%)
Baseline Questionnaires	Mean (SD)	Mean (SD)
SAS-A	42.05 (11.99)	41.70 (10.26)
SAS-A above cut off	16: 17.02%	11: 16.92%
AQ-10	3.17 (1.57)	3.02 (1.41)
SCPQ-T	14.12 (4.17)	11.19 (5.25)
SCPQ-P	14.73 (4.17)	15.18 (3.42)
RCADS – A	30.02 (15.33)	34.26 (16.74)
RCADS – D	10.16 (5.08)	9.68 (5.00)
Manipulation Check Measures	Mean (SD)	Mean (SD)
SSAS – Situ	15.39 (3.25)	14.45 (3.98)
SSAS – Private	15.35 (3.43)	14.85 (3.56)
SSAS-Public	11.95 (5.03)	12.40 (4.84)
Anxiety SUD	1.73 (2.13)	2.10 (2.26)
Self-focus SUD	3.97 (2.49)	3.86 (2.50)

Note. SAS-A = Social Anxiety Scale-Adolescents; AQ-10 = Autism Quotient – 10 item version; SCPQ-P = Social Competence with Peers Questionnaire – Pupil Version; SCPQ – T = Social Competence with Peers Questionnaire – Teacher Version; RCADS – A = The Revised Child Anxiety and Depression Scale – Anxiety Total; RCADS – D = The Revised Child Anxiety and Depression Scale – Depression Total; SSAS – Situ = Situation Self Awareness Scale – Situation Subscale; SSAS – Private = Situation Self Awareness Scale – Private-Self Awareness Subscale; SSAS – Public = Situation Self Awareness Scale – Public-Self Awareness Subscale

3.2 MSR task validity and ceiling effects

Total scores on the two MSR tasks, the MASC and the FH-A, were weakly but significantly correlated (r = .17, p = .044), indicating that to some extent they are tapping into shared underlying constructs. Possible ceiling effects were also investigated, which was defined as when more than 15% of respondents achieved the highest score possible (Terwee et al., 2007). On the MASC, 0% of participants scored within the top two highest scores. On the FH-A task 3% of participants scored the highest possible score, and 10% scored at the second highest score. This indicates that neither task suffered from ceiling effects, and were therefore suitable for assessing MSR ability within this age group.

3.3 Primary analyses

3.3.1 Bivariate correlations between baseline measures and MSR variables

Pearson's correlations between all baseline measures and MSR variables are presented in Table 2, with significant relationships highlighted in bold. Significant positive relationships were demonstrated between SAS-A and the RCADS subscales (anxiety, depression and social anxiety subscales), and negative relationship was demonstrated between SAS-A and the self-report social competence scale (SCPQ–P). No significant relationship was demonstrated between SAS-A scores and social communication scores (AQ-10). Significant relationships were demonstrated between age and a number of the MSR outcome variables, and between AQ-10 and one of the FH-A task subscales. Additionally, social anxiety (as measured by the SAS-A) but not depression or general anxiety symptoms (as measured by the RCADS) demonstrated significant relationships with MSR variables.

Table 2: Bivariate correlations between study measures

	Age	AQ10	SCPQ	SAS-A	RCADS-A	RCADS-Tot	RCADS-SAD	RCADS-MD	FHA-Total	FHA-MI	FHA-PI	FHA-NI	FHA-E	MASC-Total	MASC-HYP	MASC-Less	MASC-No	MASC-Con
Age	1																	
AQ10	-.04	1																
SCPQ	-.14	-.07	1															
SAS-A	-.01	.08	**-.53****	1														
RCADS-A	.19	**.22***	**-.41****	**.60****	1													
RCADS-Total	.19	**.25***	**-.41****	**.60****	**.98****	1												
RCADS-SAD	.08	.16	**-.38****	**.65****	**.80****	**.77****	1											
RCADS-MD	.17	**.26***	**-.27***	**.44****	**.68****	**.80****	**.44****	1										
FHA-Total	**.23****	-.10	-.06	**.19***	.04	.05	.16	.06	1									
FHA-MI	**.20***	-.15	-.04	**.19***	.01	.02	.09	.04	**.83****	1								
FHA-PI	.18*	.15	.02	.07	.12	.12	.18	.07	**.64****	**.34****	1							
FHA-NI	-.12	.18*	.10	-.12	.07	.06	-.04	-.01	**-.60****	**-.32****	-.03	1						
FHA-E	**.30****	**-.24****	.05	.10	-.02	-.03	.02	-.08	**.72****	**.75****	**.31****	**-.39****	1					
MASC-Total	**.21****	-.15	.10	-.14	-.03	-.02	-.08	.02	**.17***	.14	.06	-.16	**.23****	1				
MASC-HYP	**.22****	.12	-.15	.15	.02	.03	.04	.02	-.04	-.03	-.07	-.01	-.13	**-.73****	1			
MASC-Less	-.12	.11	.06	-.01	-.15	-.16	-.06	-.14	-.16	-.13	-.01	**.20***	-.20*	**-.53****	-.06	1		
MASC-No	.06	.00	-.05	.05	.21	.17	.18	.02	-.17	-.12	-.06	.16	-.07	**-.46****	.03	**.18***	1	
MASC-Con	**-.18***	**-.20***	.00	.11	-.01	-.04	-.04	-.10	.10	.14	-.01	-.07	**.20***	.05	.03	-.12	-.01	1

AQ-10 = Autism Quotient – 10 item version; SCPQ-P = Social Competence with Peers Questionnaire – Pupil Version; SAS-A = Social Anxiety Scale for Adolescents ; RCADS –A = Revised Child Anxiety and Depression Scale – Anxiety Total; RCADS – D = The Revised Child Anxiety and Depression Scale – Depression Total ; RCADS – SAD = *The Revised Child Anxiety and Depression Scale – Social Anxiety Disorder Subscale; FHA-Total = Frith-Happe Animations – Total score; FHA-MI = Frith-Happe Animations – Mental interaction items; FHA-PI = Frith-Happe Animations –Physical interaction items; FHA-NI = Frith-Happe Animations –No interaction items; FHA-E = Frith-Happe Animations –Emotion items; MASC-Total = MASC – Total score; MASC-HYP = MASC hyper-MSR errors; MASC-Less = MASC less-MSR errors; MASC-No = MASC No MSR errors; MASC-Con = MASC Control items*

**Significant at the α = .05 level; **Significant at the α = .01 level*

3.3.2 Hypothesis 1: Social anxiety will predict poorer overall MSR performance on the FH-A task and the MASC task. On the MASC task, the reduced total score will be due to a higher number of hyper-MSR errors

The primary hypothesis, that social anxiety will predict MSR ability, was tested using hierarchical multiple linear regressions, with each MSR dependent variable (DV) from the FHA task (FHA total score, FHA-MI items, FHA-PI items, FHA-NI items and FHA-Emotions items) and the MASC task (MASC total Score, MASC hyper-MSR errors, MASC less-MSR errors and MASC no-MSR errors), with SAS-A as the independent variable (IV).

As performance on the MASC correlated with age, this was entered at Step 1 along with the control scores of the MASC and condition to reflect the two experimental groups. SAS-A accounted for a significant portion of MASC total score [R^2 change = .03, F(1, 139) = 4.04, p = .046] and the MASC-hyper-MSR error score [R^2 change = .04, F(1, 139) = 5.60, p = .019] over and above that accounted for by age, control question performance and condition. This indicated that higher SAS-A scores significantly predicted both lower MASC total scores, and higher numbers of hyper-MSR errors. The full regression models, with significant findings highlighted in bold, are displayed in Tables 3 (DV = MASC total score) and 4 (DV = MASC hyper-MSR error score). SAS-A did not predict 'less-MSR' errors [R^2 change < .001, F(1, 139) = .03, p= .871] or 'no-MSR' errors [R^2 change = .002, F(1, 139) = .33, p= .565] on the MASC task.

Table 3: Output for hierarchical regression with MASC total score as the DV

	B	SE B	β	*p*	R^2	ΔR^2
Step 1					**0.08****	
Constant	-22.42	17.28		.197		
Condition	-.01	.75	-.00	.991		
Age	2.94	.97	.26	**.003**		
MASC Control Qs	.77	.36	.18	**.034**		
Step 2					**0.11****	**0.03***
Constant	-20.15	17.13		.242		
Condition	-.00	.74	-.00	.996		
Age	2.94	.96	.26	**.003**		
MASC Control Qs	.82	.36	.19	**.023**		
SAS-A	-.06	.04	-.16	**.046**		

DV: MASC total Score: , F(1, 139) = 4.04, p = .046, R^2 = .11

**Significant at the α = .05 level; **Significant at the α = .01 level*

Table 4: Output for hierarchical regression with MASC hyper-MSR error score as the DV

	B	SE B	β	*p*	R^2	ΔR^2
Step 1					**.08***	
Constant	38.90	13.07		**.003**		
Condition	1.02	.56	.15	.074		
Age	-1.84	.73	-.15	**.014***		
MASC Control Qs	-.204	.27	-.06	.453		
Step 2					**.11***	**.04***
Constant	36.89	12.89		**.005**		
Condition	1.01	.56	.15	.071		
Age	-1.84	.72	-.21	**.012***		
MASC Control Qs	-.25	.27	-.08	.413		
SAS-A	.06	.02	.19	**.019***		

DV: MASC hyper-MSR errors: F(1, 139) = 5.60, p = .019, R^2 = .11
**Significant at the α = .05 level; **Significant at the α = .01 level*

Since both age and AQ-10 were found to correlate with performance on the FHA, these variables were entered at Step 1, along with experimental condition. SAS-A accounted for a significant portion of FHA total score [R^2 change = .03, F(1,132) = 4.98, p= .027] and FHA-MI item score [R^2 change = .03, F(1, 132) = 5.07, p= .026] over and above that accounted for by age, AQ-10 score and condition. Higher SAS-A scores predicted both higher FHA total scores and higher MI item scores. The full regression models, with significant findings highlighted in bold, are shown in Tables 5 (DV = FH-A total score) and 6 (DV = FHA-MI score). SAS-A did not predict scores for the FHA-PI items [R^2 change = .003, F(1, 132) = .41, p= .521], FHA-NI items [R^2 change = .02, F(1, 132) = 2.80, p= .096], or FHA-Emotion items [R^2 change = .01, F(1, 132) = 1.67, p= .198]. As an additional precautionary step, the analyses were re-run without the AQ-10 measure in the models, since its reliability in this sample was questionable (Cronbach's α = .23). This had no effect on the significance of social anxiety predicting MSR in any of the regression models.

Table 5: Output for hierarchical regression with FH-A total score as the DV

	B	SE B	β	*p*	R^2	ΔR^2
Step 1					**.09****	
Constant	-11.46	8.76		.193		
Condition	-.78	.39	-.17	**.046***		
Age	1.16	.50	.20	**.022***		
AQ-10	-.11	.12	-.07	.380		
Step 2					**.12****	**.03***
Constant	-12.60	8.64		.147		
Condition	-.78	.38	-.17	**.043***		
Age	1.14	.50	.19	**.023***		
AQ10	-.12	.12	-.08	.316		
SAS-A	.04	.02	.18	**.027***		

DV: FHA total score: F(1, 132) = 4.98, *p*= .027, R^2 = .12

Table 6 Output for hierarchical regression with FH-A MI score as the DV

	B	SE B	β	*p*	R^2	ΔR^2
Step 1					**.08****	
Constant	-5.88	4.75		.219		
Condition	-.38	.21	-.15	.073		
Age	.52	.27	.16	.058		
AQ-10	-.11	.07	-.13	.114		
Step 2					.**11****	**.03***
Constant	-6.50	4.69		.168		
Condition	-.38	.21	-.15	.068		
Age	.51	.27	.16	.060		
AQ10	-.11	.07	-.14	.086		
SAS-A	.02	.01	.19	.**026***		

DV: FHA-MI: F(1, 132) = 5.07, *p* = .026, R^2 = .11

All regressions were repeated in STATA in order to perform cluster analysis around class group ID. Social anxiety continued to significantly predict FH-A total scores (*p* = .018) and FH-A MI scores (*p* =.047), but fell short of significance in predicting MASC total score (*p*=.083) and MASC hyper-MSR score (*p*=.070).

3.3.3 Hypothesis 2: Social Anxiety will be more strongly related to MSR under conditions of social stress

There were no significant differences found between the conditions in terms of ratings of anxiety or self-focus (see Table 1; all *p*s > .253), which was also true when only including those students scoring above clinical cut off on the SAS-A in the analysis (all *p*s > .333). Therefore it appears that the social stress manipulation may not have been effective in eliciting anxiety and enhanced self-focus in participants, even in those reporting high levels of social anxiety.

As an additional manipulation check, Pearson's correlations were used to investigate whether social anxiety was differentially related to the measures of state anxiety (Anxiety SUD scores) and the two subscales of self-focused attention (SSAS-Private and SSAS-Public), across the two different conditions. Correlations indicated that the relationship between social anxiety and state anxiety, and between social anxiety and SSAS-Public might be different in the two different conditions (see Table 7).

Table 7: Pearson's bivariate correlations between SAS-A and the state measures

	No Stress Condition	Social Stress Condition
State measures:	Correlations with SAS-A	
Anxiety SUD	**.25***	**.30***
SSAS-Private	-.01	.17
SSAS-Public	**.45****	**.29***

**Significant at the α = .05 level; **Significant at the α = .01 level*

In terms of testing hypothesis 2, exploratory Pearson's correlations (see Table 8) indicated that the relationship between social anxiety and MSR (as measured by the MASC) may be stronger in the social stress condition. However, when the interaction term (SAS-A*condition) was added at Step 3 to the regression models conducted previously, this did not improve the prediction of MSR performance in any of the models [MASC total score: R^2 change = .01, $F(1, 138) = 1.91$, p= .169; MASC hyper-MSR errors: R^2 change = .01, $F(1, 138) = 1.57$, p= .212; FHA total score: R^2 change = .001, $F(1, 131) = .09$, p= .767; FHA-MI score: R^2 change = .001 , $F(1, 131) = .13$, p= .716].

Table 8: Pearson's partial correlations between SAS-A and the MSR variables

	No Stress Condition	Social Stress Condition
MASC	Partial[6] correlations with SAS-A	
Total	-.10	**-.30***
Hyper-MSR	.14	**.33***
FHA	Partial[7] correlations with SAS-A	
Total	.20	.16
MI items	.16	.24

**Significant at the α = .05 level; **Significant at the α = .01 level*

3.3.4 Hypothesis 3: Measures of state anxiety and self-focused attention will mediate the relationship between SAS-A and MSR Variables

As Pearson's correlations revealed that social anxiety was significantly associated with ratings of state anxiety (r= .26, p = .001) and public self-awareness (r= .39, p < .001), but not with private self-awareness (r=.05, p= .504), only Anxiety SUD scores, and SSAS-Public ratings were entered as mediators in the analyses, with SAS-A as the IV. Separate analyses were run for each of the MSR DVs (MASC total scores, MASC hyper-MSR scores, FHA total scores and FHA-MI scores), which were included on the basis of significant associations with SAS-A in the previous regression models. The same variables that were entered at Step 1 of the hierarchical regression models were entered as covariates in the mediation analyses: for the MASC DVs these variables were age, condition and MASC control questions, and for the FH-A DVs these were age, condition and AQ-10 scores.

When Anxiety SUD scores were entered as the mediator, with MASC total scores as the DV (see Figure 2), higher SAS-A scores predicted higher Anxiety SUD ratings (B = .01, SE = .002, p= .001) and higher Anxiety SUD ratings predicted lower MASC Total Score (B = -2.98, SE = 1.14, p =.010).

[6] *Partial correlations: controlling for age & MASC control questions*

[7] *Partial correlations: controlling for age & AQ10.*

With bootstrapping confidence interval testing there was a significant indirect effect of SAS-A on MASC total score through state anxiety (95% CI = -.058, -.007), where Anxiety SUD scores accounted for roughly 40% of the total effect. A Sobel test also found the partial mediation to be significant (*p*=.044). However, it should be noted that with bootstrapping, the total effect of SAS-A on MASC total score was only marginal (B = -0.06, SE = .03, *p*=.064).

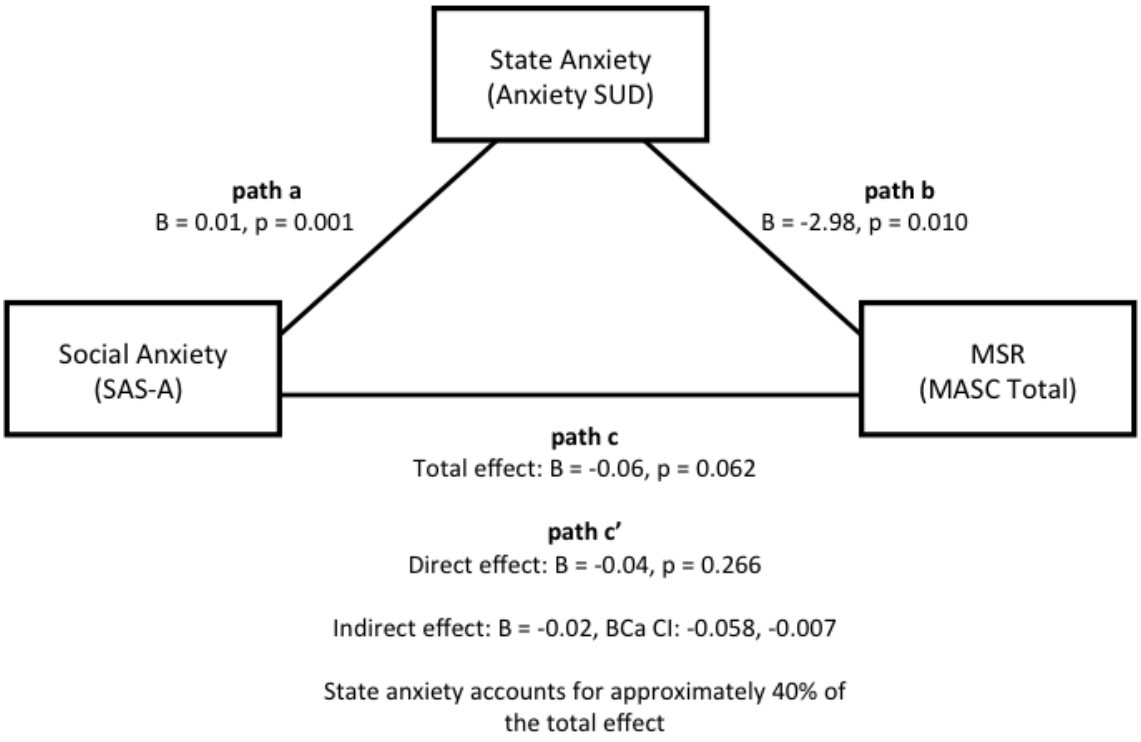


Figure 2: Model of social anxiety as a predictor of MASC total score mediated by state anxiety

When MASC hyper-MSR was entered as the DV (see Figure 3), higher SAS-A predicted higher Anxiety SUD ratings (B = .01, SE = .002, *p*=.001) and higher Anxiety SUD ratings predicted more MASC hyper-MSR errors (B = 2.26, SE = .86, *p*=.009). With bootstrapping confidence interval testing it was shown that the indirect effect of SAS-A on MASC hyper-MSR errors through state anxiety ratings was significant (95% CI = .006, .040), where Anxiety SUD scores accounted for roughly 32% of the total effect. A Sobel test also found partial mediation in the model (*p* = .043).

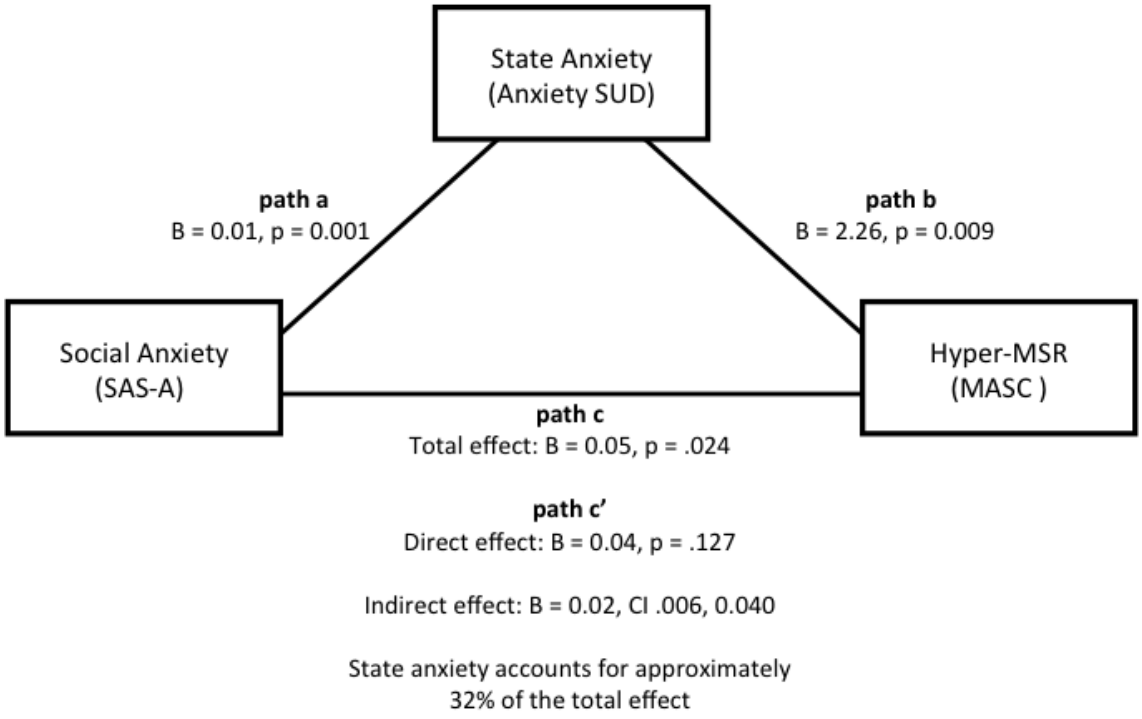


Figure 3: Model of social anxiety as a predictor of MASC hyper-MSR errors mediated by state anxiety

Anxiety SUD ratings did not mediate the relationship between SAS-A and FHA total scores or FHA-MI scores.

When public self-awareness (SSAS-Public) was entered as the mediating variable (see Figure 4), SAS-A predicted higher SSAS-Public ratings (B = .18, SE = .033, $p < .001$) and higher SSAS-Public ratings predicted fewer MASC hyper-MSR errors (B = -.12, SE = .06, p=.039). There was a significant indirect effect of SAS-A on MASC hyper-MSR errors through SSAS-Public (95% CI = -.049, -.002). As there is a positive indirect effect and a negative direct effect in this model, this is 'inconsistent mediation', suggesting a 'suppressor effect' of SSAS-Public on the relationship between SAS-A and MASC hyper-MSR errors. Thus, calculating the percentage of the total effect explained by the mediator variable is not meaningful. Using completely standardized coefficients, the effect size of this mediation was estimated to be medium [ab_{CS} = .08, BCa CI (-.162, -.011)].

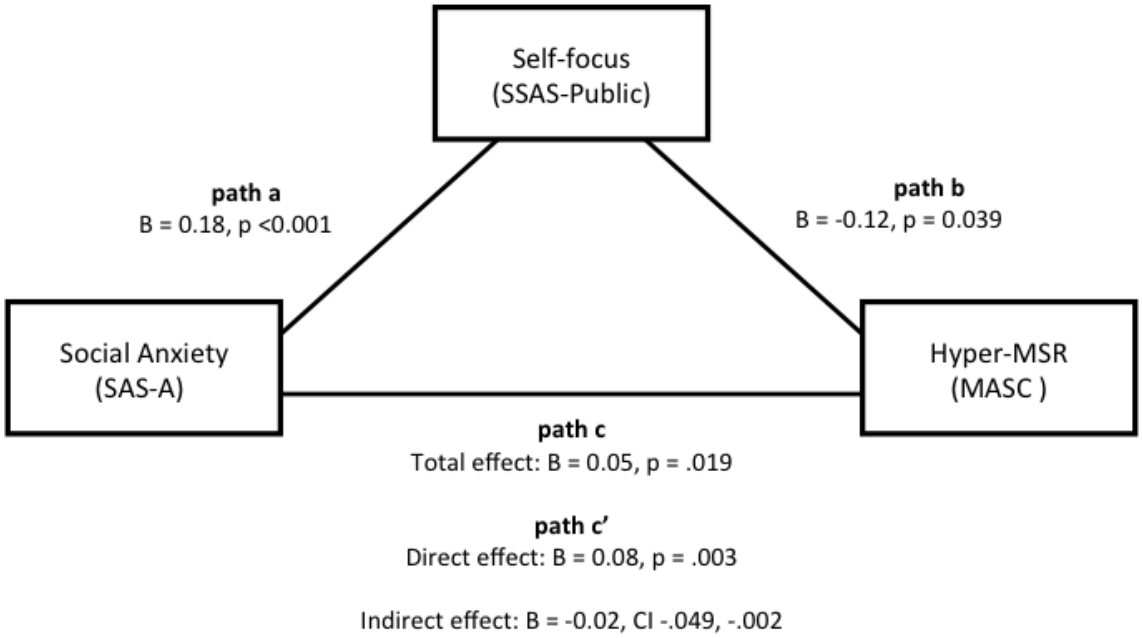


Figure 4: Model of social anxiety as a predictor of MASC hyper-MSR errors mediated public self-awareness

SSAS-Public did not have a significant mediation effect on the relationship between SAS-A and any of the other MSR outcome variables (MASC total, FHA total or FHA-MI).

3.4 Secondary analyses

3.4.1 Hypothesis 4: Social Anxiety will be related to poorer social competence

As predicted, SAS-A significantly predicted SCPQ scores [$F(1, 152) = 59.35$, $p < .001$, $R^2 = .28$], indicating that higher levels of social anxiety was associated with lower levels of self-reported social competence.

In addition to self-report ratings of social competence, independent assessors completed a social behavior assessment checklist for the 23 students who were randomly selected to take part in the conversation and speech task. The demographic and baseline characteristics of the students who completed the behaviour task are presented in Appendix G. There were no significant differences in terms of demographic or baseline characteristics between the group who completed the behaviour task group and the rest of the sample (all *ps* > .588). As there was strong agreement in scoring between the two adult assessors (ICC = .939 for conversation task

ratings and .906 for speech task ratings), the two scores were combined into one. Linear regressions demonstrated that SAS-A was not significantly related to assessor scores on either the conversation task (p=.383) or the speech task (p= .242). This was also true for the peer rated scores, in that SAS-A was not related to either peer-rated conversation scores (p= .615) or speech scores (p= .179).

3.4.2 Hypothesis 5: Measures of MSR performance will explain any relationship between social anxiety and social competence

To investigate whether MSR performance could explain any of the relationship between SAS-A and poorer social competence (as measured by the SCPQ) each of the MSR variables (MASC total, MASC hyper-MSR, FHA total, FHA-MI) were entered as mediating variables into the analyses, with SAS-A as the IV and SCPQ as the DV. None of the indirect effects were found to be significant, suggesting that the MSR variables did not account for a significant portion of the total variance in the relationship between SAS-A and SCPQ ratings.

4. Discussion

The current study sought to investigate whether social anxiety symptoms in young people were associated with dysfunctional MSR ability, as measured by the MASC and the FHA tasks. The study also investigated whether impairments in MSR could be explained by state factors, including anxiety and self-focused attention. Finally, the study investigated whether social anxiety was associated with reduced social competence, and if so, whether MSR deficits may explain this.

In partial support of hypothesis 1, young people reporting more social anxiety symptoms demonstrated dysfunctional MSR, as measured by the MASC task. This was due to making more hyper-MSR errors, which is line with findings from the adult literature (Hazel & McNally, 2014; Washburn et al., 2016). However, on the FHA task, social anxiety predicted *superior* performance overall, related to significantly better performance in identifying the clips depicting a 'mental interaction' between the two animated shapes. Hypothesis 2 was not supported in that social anxiety did not significantly interact with the social stress condition, to further improve the prediction of MSR performance. Hypothesis 3 was partially supported, whereby state anxiety was shown to have a significant mediation effect on the relationship between

social anxiety and dysfunctional MSR, as measured by the MASC. Self-focused attention, in the form of public self-awareness, was also shown to mediate an indirect effect of social anxiety on MASC hyper-MSR errors, where higher public self-awareness predicted fewer hyper-MSR errors. Hypothesis 4 was also partially supported, in that higher social anxiety was related to lower self-report ratings of social competence, although this was not true for the assessor-rated behavioural measures of social skill. Finally, hypothesis 5 was not supported as MSR performance was not shown to mediate the relationship between social anxiety and poorer social competence.

The study's main findings with respect to the relation between social anxiety and MSR are perhaps seemingly contradictory, in that social anxiety was related to impaired MSR ability on the MASC task, but superior ability on the FH-A task. However, given the rather modest correlation between scores on the two tasks, it may be that they are tapping into slightly different kinds of MSR ability. Another tentative explanation for this differential pattern of performance across tasks, is that socially anxious individuals exhibit MSR skills that lie at the upper end of the normal continuum, with a general increase in sensitivity to the thoughts and feelings of others. On some tasks, such as the MASC, this enhanced sensitivity may manifest as hyper-MSR (going 'above and beyond' the information provided) but on the simpler FH-A task, this manifests as improved detection of mental interaction clips. Perhaps given their heightened social concerns, socially anxious individuals engage in more, and put increased effort into, MSR, although if over-interpretation is taking place then the accuracy of their reasoning may be impaired.

In terms of what might be underpinning the relationship between social anxiety and hyper-MSR (on the MASC) or enhanced MSR (on the FH-A), this study explored whether state anxiety and / or self-focused attention has any mediating effect. As expected, social anxiety was related to higher state anxiety, which was shown to partially mediate the relationship between social anxiety and poorer overall performance on the MASC, as well as between social anxiety and increased hyper-MSR errors, accounting for 40% and 32% of the total effects respectively. This implies that state anxiety could have a detrimental effect on performance on more complex MSR tasks, such as the MASC, in the form on increased hyper-MSR errors. This is in keeping with a recent study that found that amongst female participants (as were the majority in the current sample), a higher cortisol stress response after completing the Trier Social Stress Test was associated worse performance on the MASC (Smeets, Dziobek & Wolf, 2008). Such findings

provide support for cognitive models of social anxiety that propose that state anxiety has a detrimental effect in terms of processing social information (e.g. Clark & Wells, 1995).

In addition to state anxiety, the potential influence of self-focus was also investigated. It was predicted that increased self-focus (as measured by the private and public subscales of the situational self-awareness scale) would have a detrimental impact on MSR performance. In the current study, social anxiety was found to be associated with increased public-self-awareness (the form of self-focus that enhances awareness of self-impression) but not private self-awareness (the form of self-focus that enhances awareness of internal thoughts and feelings), which is in line with previous research that has also found this dissociation (Darvill, Johnson, & Danko, 1992; George & Stopa, 2008). Interestingly, unlike the apparent detrimental effects of state anxiety, public self-awareness was found to have a 'suppressor effect' (MacKinnon, Fairchild, & Fritz, 2007) on the relationship between social anxiety and increased hyper-MSR errors. This suggests that socially anxious individuals who display greater self-awareness may in fact be more accurate in their MSR abilities, than their less self-aware counterparts. This is not in keeping with cognitive models that posit a detrimental role for self-focused attention. However, since public self-awareness did not account for the improved performance on the FHA task, this is not likely to be a robust finding.

Although there is some indication from the current study of the potential role of state anxiety and self-focus in MSR, it is surprising then that the experimental condition of social threat did not differentially affect MSR performance in individuals with higher levels of social anxiety. This may be due to the social threat manipulation not being effective in inducing greater anxiety or self-focus, even in those scoring above the clinical cut off on the SAS-A. This is surprising given that the presence of cameras, and the upcoming threat of a speech and conversation task with strangers, has proved an effective manipulation in previous research (e.g. Govern & Marsch, 2001; Hass & Eisenstadt, 1991; George & Stopa, 2008). It is possible that the measures of self-focus and state anxiety were not sensitive enough to detect group differences, particularly the measure of state anxiety, which asked participants how anxious they were feeling retrospectively ("how nervous were you feeling during the video tasks?"). However, the issue of sensitivity is less likely for the SSAS, which has been shown to be sensitive to laboratory manipulations (such as the presence of mirrors and cameras) in previous research. Nonetheless, the role of state factors on MSR ability appears to be an avenue worth exploring, as other researchers have demonstrated that experimentally induced negative emotional reactivity may mediate the relationship between clinical symptoms and poorer social problem solving, in

individuals with borderline personality disorder (BPD; Dixon-Gordon, Chapman, Lovasz, & Walters, 2011). Future studies in social anxiety might therefore consider piloting the social stress manipulation to ensure its relevance, in order to investigate the role of state factors on MSR more effectively.

Relationships between social anxiety, social competence and MSR were also explored and as expected social anxiety predicted lower self-report ratings of social competence. However, social anxiety did not predict poorer performance on a behavioral measure of social skill, as rated by both adult and peer assessors. This is in line with previous research that has also found a dissociation between self-report and assessor rated measures of social competence (Cartwright-Hatton et al., 2005; Erath et al., 2007; Rapee & Lim, 1992). Such findings support interventions that help individuals become aware that other people do not necessarily notice their anxiety-related symptoms and behaviours, nor does appearing anxious inevitably give rise to negative judgement on social performance, even by peers. However, given the small sample of students who completed the behavior task (n = 23), whether social anxiety is related to more objective assessor-ratings of social competence remains uncertain. Indeed, evidence within the social anxiety literature is mixed with regards to how social anxiety is related to social competence. It is likely that whilst for some individuals problematic social skills plays a key role in their social anxiety, for others it may be that distorted negative cognitions in relation to self-evaluation are more relevant (Miers et al., 2011).

Finally, MSR performance was not found to explain any of the relationship between social anxiety symptoms and the self-report ratings of social competence. Indeed, as the measure of social competence was self-report, it is difficult to know whether this is an accurate assessment of individuals' social ability. Additionally, since the effect sizes of the relationships between social anxiety and the MSR variables were small, the real-world impact (e.g. on social functioning) of enhanced or hyper-MSR in social anxiety remains unclear.

Speculatively, hyper-MSR, which can be considered as reading 'too much' into how others are feeling or what they are thinking, might give rise to social misunderstandings. Additionally, increased sensitivity to others' mental states, combined with a tendency to make over-interpretive inferences, could feed into cognitive processes such as negative biases, or increased social rumination. In fact it has been suggested by some authors that enhanced MSR abilities may be related to internalising disorders more generally. Zahn-Waxler et al., (2012) have

proposed that excessive levels of empathy and social sensitivity, may increase liability to disorders such as anxiety and depression, particularly in individuals with high levels of internalising emotions, such as guilt, shame, sadness and embarrassment. Indeed, although the current study did not find any association between MSR ability and depression or general anxiety symptoms (as measured by the RCADS), other studies have demonstrated enhanced or excessive MSR in subclinical and mild depression (e.g. Harkness, Jacobson, Duong, & Sabbagh, 2010; Harkness, Washburn, Theriault, Lee, & Sabbagh, 2011) as well as BPD (Sharp et al., 2013). Thus, enhanced or excessive MSR may serve as a more general vulnerability to (or indeed be a product of) a number of mental health problems.

4.1 Clinical implications

Psychological interventions that target MSR impairments have already been developed in the form of mentalisation-based therapies for BPD (Bateman & Fonagy, 2010) and meta-cognitive training programmes for individuals with schizophrenia (e.g. Moritz, Veckenstedt, Randjbar, Vitzthum, & Woodward, 2011; Penn, Roberts, Combs, & Sterne, 2007). Such treatments aim to encourage 'thinking about thinking' as well as help individuals develop greater awareness of biases in their MSR. Importantly, mentalisation-based therapies for BPD have been able to demonstrate that hyper-MSR is malleable to such treatments (Sharp et al., 2013). Therefore, if hyper-MSR can be shown to have negative psychological or social consequences, some of the therapeutic techniques developed for these other clinical populations to target dysfunctional MSR, could be used for socially anxious individuals. Current CBT interventions may also consider emphasising the importance of behavioural experiments that use stooge feedback, in order to help socially anxious individuals who hyper-mentalise, to gain more accurate perceptions of what others might be thinking or feeling. Finally, MSR errors could also be targeted through computer-based training programmes, such as cognitive bias modification protocols. As the social brain and metacognitive skills are continuing to develop through the teenage years (Blakemore, 2008; Selman, 1980) targeting dysfunctional MSR in this age group may reduce vulnerability to mental health conditions such as social anxiety.

4.2 Limitations

This study had a number of limitations. The cross sectional nature of the study means that the temporal links between MSR and social anxiety are not known. The non-clinical sample used also means that the findings cannot necessarily be generalised to the clinical population. Another limitation is the mass testing format of the study, which although aids study feasibility,

the tradeoff is less control over data collection, leading to some missing data. Another limitation is that no measure of intellectual ability was included, which is a potential confound in terms of task performance. However, IQ has not been shown to be related to MASC task performance in previous studies (Hazel and McNally, 2014; Dziobek et al., 2006) and the multiple-choice version of the FHA task is likely to rely less on language abilities, thus reducing the potential confound of verbal IQ. However, controlling for general intellectual ability would further validate the study's findings.

Another consideration is the reliability and validity of the included measures. The AQ-10 demonstrated very low internal consistency in this sample, which limits how reliably ASD traits were measured in this study. Some aspects of validity of the MSR tasks were also questionable. Although the MASC task was chosen due to its suitability for use in older populations, as well as its ability to quantify MSR error types, it may lack cultural validity in the current sample of inner-city London adolescents. Indeed, the MASC was originally filmed in German over 10 years ago, and so might be of less relevance in terms of the social contexts and characters depicted. Both the MASC and FHA tasks also suffer in terms of ecological validity, since they are unable to tap into online MSR processes that take place in real-life contexts. Although this reflects a general difficulty with the assessment of MSR, it is likely that direct involvement in a 'live' social interaction, would be more sensitive at tapping into any problematic socio-cognitive skills. Future research may therefore wish to take advantage of virtual reality technology, which provides a more realistic and immersive experience, thus increasing the likelihood of tapping into the relevant, self-related cognitive processes (Canty et al., 2015).

4.3 Conclusions

The current study has demonstrated that social anxiety is related to both enhanced and excessive MSR, on two different tasks. This is in keeping with the notion that socially anxious individuals are preoccupied with the impression that they give to others, which is likely to enhance their sensitivity to social stimuli. If the thoughts and feelings of other people are consistently being over-interpreted, this is likely to give rise to social misunderstandings, which may further increase social anxiety. However, the precise psychological and functional consequences of excessive MSR are speculative and warrant further investigation. If future research demonstrates a role for enhanced or excessive MSR in the development or maintenance of SAD, then treatment strategies that consider individual differences in MSR are needed. Fostering the optimal level of social sensitivity may help to improve social functioning and alleviate some of the interpersonal difficulties that characterise this disorder.

5. References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Alfano, C. A., Beidel, D. C., & Turner, S. M. (2006). Cognitive correlates of social phobia among children and adolescents. *Journal of Abnormal Child Psychology*, *34*(2), 189–201.

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "Red Flags" for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in toddlers in 1,000 cases and 3,000 controls [corrected]. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*(2), 202–212.e7.

Banerjee, R. (2008). Social cognition and anxiety in children. In C. Sharp, P. Fonagy & I. Goodyer (Eds.), *Social cognition and developmental psychopathology* (pp. 239-269). Oxford: Oxford University Press.

Banerjee, R., & Henderson, L. (2001). Social-cognitive factors in childhood social anxiety: a preliminary investigation. *Social Development*, *10*(4), 558-572.

Bateman, A., & Fonagy, P. (2010). Mentalization based treatment for borderline personality disorder. *World Psychiatry : Official Journal of the World Psychiatric Association (WPA)*, *9*(1), 11–5.

Beesdo, K., Bittner, A., Pine, D. S., Stein, M. B., Höfler, M., Lieb, R., & Wittchen, H. U. (2007). Incidence of social anxiety disorder and the consistent risk for secondary depression in the first three decades of life.*Archives of General Psychiatry*, *64*(8), 903-912.

Blakemore, S.-J. (2008). The social brain in adolescence. *Nature Reviews. Neuroscience*, *9*(4), 267–77.

Brunello, N., Den Boer, J. A., Judd, L. L., Kasper, S., Kelsey, J. E., Lader, M., ... & Montgomery, S. A. (2000). Social phobia: diagnosis and epidemiology, neurobiology and pharmacology, comorbidity and treatment.*Journal of affective disorders*, *60*(1), 61-74.

Cartwright-Hatton, S., Hodges, L., & Porter, J. (2003). Social anxiety in childhood: The relationship with self and observer rated social skills. *Journal of Child Psychology and Psychiatry*, *44*(5), 737-742.

Cartwright-Hatton, S., Tschernitz, N., & Gomersall, H. (2005). Social anxiety in children: social skills deficit, or cognitive distortion? *Behaviour Research and Therapy*, *43*(1), 131–41.

Chen, J., Furukawa, T. A., Nakano, Y., Ietsugu, T., Ogawa, S., Funayama, T., … Rapee, R. M. (2010). Video feedback with peer ratings in naturalistic anxiety-provoking situations for social anxiety disorder: Preliminary report. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*(1), 6–10.

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: a revised child anxiety and depression scale. *Behaviour Research and Therapy*, *38*(8), 835–855.

Chorpita, B. F., Moffitt, C. E., & Gray, J. (2005). Psychometric properties of the Revised Child Anxiety and Depression Scale in a clinical sample. *Behaviour Research and Therapy*, *43*(3), 309–22.

Clark, D. M., & McManus, F. (2002). Information processing in social phobia. *Biological Psychiatry*, *51*(1), 92–100.

Clark, D. M., Ehlers, A., Hackmann, A., McManus, F., Fennell, M., Grey, N., ... & Wild, J. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: a randomized controlled trial. *Journal of consulting and clinical psychology*, *74*(3), 568.

Clark, D. M., & Wells, A. (1995). A cognitive model of social phobia. In R. G. Heimberg, M. Liebowitz, D. Hope & F. Scheier (Eds.), *Social phobia: Diagnosis, assessment and treatment* (pp. 69-93). New York, NY: Guildford.

Darvill, T. J., Johnson, R. C., & Danko, G. P. (1992). Personality correlates of public and private self consciousness. *Personality and Individual Differences*, *13*(3), 383–384.

Dixon-Gordon, K. L., Chapman, A. L., Lovasz, N., & Walters, K. (n.d.). Too upset to think: The interplay of borderline personality features, negative emotions, and social problem solving in the laboratory.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., … Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*(5), 623–36.

Elkind, D., & Bowen, R. (n.d.). Imaginary audience behavior in children and adolescents.

Erath, S. A., Flanagan, K. S., & Bierman, K. L. (2007). Social anxiety and peer relations in early adolescence: behavioral and cognitive factors. *Journal of Abnormal Child Psychology*, *35*(3), 405–16.

Fedoroff, I. C., & Taylor, S. (2001). Psychological and pharmacological treatments of social phobia: a meta-analysis. *Journal of Clinical Psychopharmacology*, *21*(3), 311–24.

George, L., & Stopa, L. (2008). Private and public self-awareness in social anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*(1), 57–72.

Ginsburg, G. S., Kendall, P. C., Sakolsky, D., Compton, S. N., Piacentini, J., Albano, A. M., … March, J. (2011). Remission after acute treatment in children and adolescents with anxiety disorders: findings from the CAMS. *Journal of Consulting and Clinical Psychology*, *79*(6), 806–13.

Gould, R. A., Buckminster, S., Pollack, M. H., Otto, M. W., & Massachusetts, L. Y. (1997). Cognitive-Behavioral and Pharmacological Treatment for Social Phobia: A Meta-Analysis.

Clinical Psychology: Science and Practice, *4*(4), 291–306.

Govern, J. M., & Marsch, L. A. (2001). Development and validation of the situational self-awareness scale. *Consciousness and Cognition*, *10*(3), 366–78. doi:10.1006/ccog.2001.0506

Greca, A. M. La, & Lopez, N. (1998). Social Anxiety Among Adolescents: Linkages with Peer Relations and Friendships. *Journal of Abnormal Child Psychology*, *26*(2), 83–94.

Haller, S. P. W., Cohen Kadosh, K., Scerif, G., & Lau, J. Y. F. (2015). Social anxiety disorder in adolescence: How developmental cognitive neuroscience findings may shape understanding and interventions for psychopathology. *Developmental Cognitive Neuroscience*, *13*, 11–20.

Harkness, K. L., Jacobson, J. A., Duong, D., & Sabbagh, M. A. (2010). Mental state decoding in past major depression: Effect of sad versus happy mood induction.

Harkness, K. L., Washburn, D., Theriault, J. E., Lee, L., & Sabbagh, M. A. (2011). Maternal history of depression is associated with enhanced theory of mind in depressed and nondepressed adult women. *Psychiatry Research*, *189*(1), 91–6.

Hass, R. G, & Eisenstadt, D (1990) The effects of self-focused attention on perspective-taking and anxiety, Anxiety Research, 2:3, 165-176

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Hezel, D. M., & McNally, R. J. (2014). Theory of Mind Impairments in Social Anxiety Disorder. *Behavior Therapy*, *45*(4), 530–40.

Higa, C. K., Phillips, L. K., Chorpita, B. F., & Daleiden, E. L. (2008). The Structure of Self-Consciousness in Children and Young Adolescents and Relations to Social Anxiety. *Journal of Psychopathology and Behavioral Assessment*, *30*(4), 261–271.

Hope, D. A., Heimberg, R. G., & Klein, J. F. (1989). Social Anxiety and the Recall of Interpersonal Information. *Journal of Cognitive Psychotherapy*, *4*(2), 185–195.

Hudson, J. L., Keers, R., Roberts, S., Coleman, J. R. I., Breen, G., Arendt, K., … Eley, T. C. (2015). Clinical Predictors of Response to Cognitive-Behavioral Therapy in Pediatric Anxiety Disorders: The Genes for Treatment (GxT) Study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *54*(6), 454–63.

Inderbitzen-Nolan, H. M., Anderson, E. R., & Johnson, H. S. (2007). Subjective versus objective behavioral ratings following two analogue tasks: a comparison of socially phobic and non-anxious adolescents. *Journal of Anxiety Disorders*, *21*(1), 76–90.

Kösters, M. P., Chinapaw, M. J. M., Zwaanswijk, M., van der Wal, M. F., & Koot, H. M. (2015). Structure, reliability, and validity of the revised child anxiety and depression scale (RCADS) in a multi-ethnic urban sample of Dutch children. *BMC Psychiatry*, *15*(1), 132.

La Greca, A. M., & Lopez, N. (1998). Social anxiety among adolescents: Linkages with peer relations and friendships. *Journal of abnormal child psychology*, *26*(2), 83-94.

La Greca, A. M., & Harrison, H. M. (2005). Adolescent peer relations, friendships, and romantic relationships: Do they predict social anxiety and depression?. *Journal of Clinical Child and Adolescent Psychology*, *34*(1), 49-61.

Lau, J. Y. F. (2013). Cognitive bias modification of interpretations: a viable treatment for child and adolescent anxiety? *Behaviour Research and Therapy*, *51*(10), 614–22.

Leigh, E., & Clark, D. M. (2016). Cognitive Therapy for Social Anxiety Disorder in Adolescents: A Development Case Series. *Behavioural and Cognitive Psychotherapy*, *44*(1), 1–17.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614.

Mansell, W., & Clark, D. M. (1999). How do I appear to others? Social anxiety and processing of the observable self. *Behaviour Research and Therapy*, *37*(5), 419–434.

Melfsen, S., & Florin, I. (2002). Do Socially Anxious Children Show Deficits in Classifying Facial Expressions of Emotions? *Journal of Nonverbal Behavior*, *26*(2), 109–126.

Miers, A. C., Blöte, A. W., de Rooij, M., Bokhorst, C. L., & Westenberg, P. M. (2013). Trajectories of social anxiety during adolescence and relations with cognition, social competence, and temperament. *Journal of Abnormal Child Psychology*, *41*(1), 97–110.

Miers, A. C., Blöte, A. W., & Westenberg, P. M. (2010). Peer perceptions of social skills in socially anxious and nonanxious adolescents. *Journal of Abnormal Child Psychology*, *38*(1), 33–41.

National Institute for Health and Care Excellence (2013). Social Anxiety Disorder: recognition, assessment and treatment of social anxiety disorder. London: Clinical Guideline, No. 159.

Niels Christensen, P., Stein, M. B., & Means-Christensen, A. (2003). Social anxiety and interpersonal perception: a social relations model analysis. *Behaviour Research and Therapy*, *41*(11), 1355–71.

Petersen, A. C., & Leffert, N. (1995). Developmental Issues Influencing Guidelines for Adolescent Health Research: A Review. *Journal of Adolescent Health*, *17*(5), 298–305.

Pfeifer, J. H., Masten, C. L., Borofsky, L. A., Dapretto, M., Fuligni, A. J., & Lieberman, M. D. Neural correlates of direct and reflected self-appraisals in adolescents and adults: when social perspective-taking informs self-perception. *Child Development*, *80*(4), 1016–38.

Ranta, K., Kaltiala-Heino, R., Fröjd, S., & Marttunen, M. (2013). Peer victimization and social phobia: a follow-up study among adolescents. *Social Psychiatry and Psychiatric Epidemiology*, *48*(4), 533–44.

Ranta, K., Kaltiala-Heino, R., Pelkonen, M., & Marttunen, M. (2009). Associations between peer victimization, self-reported depression and social phobia among adolescents: the role of comorbidity. *Journal of Adolescence*, *32*(1), 77–93.

Rapee, R. M., & Lim, L. (1992). Discrepancy between self- and observer ratings of performance in social phobics. *Journal of Abnormal Psychology*, *101*(4), 728–31.

Runchman, C., Yazdi, H., Burnett Heyes, S., Pile, V., & Lau, J. Y, F. (in preparation). Is social anxiety characterised by poor social understanding?

Scharfstein, L. A., Beidel, D. C., Sims, V. K., & Rendon Finnell, L. (2011). Social skills deficits and vocal characteristics of children with social phobia or Asperger's disorder: a comparative study. *Journal of Abnormal Child Psychology*, *39*(6), 865–75.

Schwartz, J. S., & Maric, M. (2014). Negative Cognitive Errors in Youth: Specificity to Anxious and Depressive Symptoms and Age Differences. *Behavioural and Cognitive Psychotherapy*, 1–12.

Selman, R. L. (1980). An analysis of "pure" perspective taking: Games and the delights of deception. In R. L. Selman (Ed.), The growth of interpersonal understanding. New York: Academic.

Sharp, C., Ha, C., Carbone, C., Kim, S., Perry, K., Williams, L., & Fonagy, P. (2013). Hypermentalizing in Adolescent Inpatients: Treatment Effects and Association With Borderline Traits.

Simonian, S. J., Beidel, D. C., Turner, S. M., Berkes, J. L., & Long, J. H. (2001). Recognition of facial affect by children and adolescents diagnosed with social phobia. *Child Psychiatry and Human Development*, *32*(2), 137–45.

Spence, S. H. (2003). Social Skills Training with Children and Young People: Theory, Evidence and Practice. *Child and Adolescent Mental Health*, *8*(2), 84–96.

Spence, S. H., Donovan, C., & Brechman-Toussaint, M. (n.d.). Social skills, social outcomes, and cognitive features of childhood social phobia.

Spurr, J. M., & Stopa, L. (2003). The observer perspective: effects on social anxiety and performance. *Behaviour Research and Therapy*, *41*(9), 1009–28.

Storch, E. A., Masia-Warner, C., Dent, H. C., Roberti, J. W., & Fisher, P. H. (2004). Psychometric evaluation of the Social Anxiety Scale for Adolescents and the Social Phobia and Anxiety Inventory for Children: construct validity and normative data. *Journal of Anxiety Disorders*, *18*(5), 665–79.

Taylor, S. (1996). Meta-analysis of cognitive-behavioral treatments for social phobia. *Journal of Behavior Therapy and Experimental Psychiatry*, *27*(1), 1–9.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., … de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42.

Tibi-Elhanany, Y., & Shamay-Tsoory, S. (2011). Social cognition in social anxiety: First evidence for increased empathic abilities. *Israel Journal of Psychiatry and Related Sciences, 48*(2),

98-106.

Veale, D. (2003). Treatment of social phobia. *Advances in Psychiatric Treatment*, *9*(4), 258–264.

Vernberg, E. M., Abwender, D. A., Ewell, K. K., & Beery, S. H. (1992). Social Anxiety and Peer Relationships in Early Adolescence: A Prospective Analysis. *Journal of Clinical Child Psychology*, *21*(2), 189–196.

Washburn, D., Wilson, G., Roes, M., Rnic, K., & Harkness, K. L. (2016). Theory of mind in social anxiety disorder, depression, and comorbid conditions. *Journal of anxiety disorders*, *37*, 71-77.

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: a quick and objective test of Theory of Mind for adults with autism. *Autism Research : Official Journal of the International Society for Autism Research*, *4*(2), 149–54.

Woody, S. R. (1996). Effects of focus of attention on anxiety levels and social performance of individuals with social phobia. *Journal of Abnormal Psychology*, *105*(1), 61–9.

Woody, S. R., Chambless, D. L., & Glass, C. R. (1997). Self-focused attention in the treatment of social phobia. *Behaviour Research and Therapy*, *35*(2), 117–29.

Woody, S. R., & Rodriguez, B. F. (n.d.). Self-Focused Attention and Social Anxiety in Social Phobics and Normal Controls. *Cognitive Therapy and Research*, *24*(4), 473–488.

6. Appendices

6.1 Appendix A: Ethical approval letters

Initial letter of approval

Georgina Bremner
Department of Psychology
P078
Institute of Psychiatry, Psychology and Neuroscience
Addiction Sciences Building
4 Windsor Walk
London SE5 8AF

10 December 2014

Dear Georgina,

PNM/14/15-41 Mental state reasoning in adolescent social anxiety

Review Outcome: Full Approval

Thank you for submitting your application for ethical approval. This was reviewed by the PNM RESC on 9th December 2014. As a result, the Committee have granted full ethical approval for your study.
Provisos
Your approval is based on the following provisos being met:
1. Section 1.4: Please ensure that copies of permission letters from gatekeeper organisations are available upon request.
2. Sections 2.2 and 2.3: Please note that ethical approval for PhD studies is normally granted for a period of three years.
3. Section 2.7: Contact the schools, with whom you will be collaborating, to determine whether you should obtain DBS clearance.
4. Section 7.1:
I. The Committee recommends that participants are allowed the opportunity to discreetly withdraw from the study if they so wish.
II. Please ensure the information sheets are distributed to the correct participant groups.
5. Section 7.2 and Information Sheet: Specify an exact date in September 2015 as the deadline for withdrawal of participant data.
6. Information Sheets:
I. The paragraph beginning with 'If this study has harmed you in any way…' should appear before the contact details of your academic supervisor.
II. Provide more specific information about the inclusion criteria for participation.
7. Demographic questionnaire: Please use UK Census 2011 categories of ethnicity.
8. AQ-10: The Committee assumes that the sentence beginning with 'A quick referral guide for adults with suspected autism…' will be removed.
You are not required to provide evidence to the Committee that these provisos have been met, but your ethical approval is only valid if these changes are made. You must not commence your research until these provisos have been met.

Please ensure that you follow all relevant guidance as laid out in the King's College London Guidelines on Good Practice in Academic Research (http://www.kcl.ac.uk/college/policyzone/index.php?id=247).

For your information ethical approval is granted until 10 December 2014. If you need approval beyond this point you will need to apply for an extension to approval at least two weeks prior to this explaining why the extension is needed, (please note however that a full re-application will not be necessary unless the protocol has changed). You should also note that if your approval is for one year, you will not be sent a reminder when it is due to lapse.

Ethical approval is required to cover the duration of the research study, up to the conclusion of the research. The conclusion of the research is defined as the final date or event detailed in the study description section of your approved application form (usually the end of data collection when all work with human participants will have been completed), not the completion of data analysis or publication of the results.
For projects that only involve the further analysis of pre-existing data, approval must cover any period during which the researcher will be accessing or evaluating individual sensitive and/or un-anonymised records.
Note that after the point at which ethical approval for your study is no longer required due to the study being complete (as per the above definitions), you will still need to ensure all research data/records management and storage procedures agreed to as part of your application are adhered to and carried out accordingly.

If you do not start the project within three months of this letter please contact the Research Ethics Office.

Should you wish to make a modification to the project or request an extension to approval you will need approval for this and should follow the guidance relating to modifying approved applications: http://www.kcl.ac.uk/innovation/research/support/ethics/applications/modifications.aspx

Please would you also note that we may, for the purposes of audit, contact you from time to time to ascertain the status of your research.

If you have any query about any aspect of this ethical approval, please contact your panel/committee administrator in the first instance (http://www.kcl.ac.uk/innovation/research/support/ethics/contact.aspx)
We wish you every success with this work.

Yours sincerely,

James Patterson - Senior Research Ethics Officer
For and on behalf of
Professor Gareth Barker, Chairman
Psychiatry, Nursing and Midwifery Research Ethics Subcommittee (PNM RESC)
Cc: Jennifer Lau

Approval for modification to ethical approval

Georgie Bremner
Department of Psychology
P078, Institute of Psychiatry, Psychology and Neuroscience
Addiction Sciences Building
4 Windsor Walk
London SE5 8AF

27 April 2015

Dear Georgie,

PNM/14/15-41 Mental state reasoning in adolescent social anxiety

Thank you for submitting a modification request for the above study. I am writing to confirm approval of this. The approved modification is summarised below:

1. Section 8.2: Entry of participants into prize draw for vouchers worth £30 and £10.

If you have any queries, please let me know.

Yours sincerely,

James Patterson – Senior Research Ethics Officer

Cc Jennifer Lau

6.2 Appendix B: Information sheets and consent forms

Information sheet for 'no-stress' condition

INFORMATION SHEET FOR PARTICIPANTS

REC Reference Number: PNM/14/15-41

YOU WILL BE GIVEN A COPY OF THIS INFORMATION SHEET

This information sheet is for young people

Social worries and how we understand others

We would like you to be part of a study that is investigating how social worries might impact on how we understand other people. We are asking students aged 16-18 to complete some questionnaires and video tasks in their classrooms. Since the age range of your class is within this, we are inviting you to take part in this study. However, before you decide whether you would like to take part or not, it is important that you understand why we are doing this study and what it involves. **Please remember that you do not have to take part in the study.** Please take time to read the following information carefully and discuss it with others if you wish. Ask your teacher or the project psychologist if there is anything that is not clear or if you would like more information.

What are we trying to find out?

Feeling anxious in social situations (known as social anxiety) is a common problem in young people and it can generate a lot of distress and worry. A lot of the fears and worries are to do with the way that they think about social situations. This study aims to find out more about social anxiety in 16-18 year olds in terms of how young people think about and understand others, which may play an important part in why people feel anxious.

Who are we asking to take part?

We will be asking all young people aged 16 to 18 at your school to take part in this study. Only students who have difficulties understanding written English will not be able to take part. This is because all participants need to be able to fully understand this information sheet and the consent form provided.

If you decide that you don't want to take part, you are free to withdraw at any time and without giving a reason. A decision not to take part, or a decision to withdraw at any time, will not affect you at school. If you have any questions about this project and what you are being invited to take part in, please ask the researcher before you decide whether to join in.

What will happen if I take part?

We will ask all participants to complete three questionnaires in the classroom, as well as some more interactive social tasks, including two video tasks. The total testing session will last up to 90

minutes. All your answers provided are completely private and will **not** be shared with the other pupils or with your teachers.

The first questionnaire will be asking about whether you ever experience things such as worry, nervousness or low mood. The second questionnaire will be asking about how you feel in social situations (e.g. I get nervous when I talk to peers I don't know very well) and the third questionnaire will be asking about your social understanding and communication (e.g. I find it easy to 'read between the lines' when someone is talking to me). There are no correct or incorrect answers.

The rest of the study will involve watching two videos of social interactions and completing a multiple choice answer sheet about what you saw. We may also ask you to take part in a more interactive task, where we may invite you to share your views about a chosen topic, with members of our research team.

Are there any benefits?

In taking part in this study you will be entered into a raffle with the rest of your classmates who are taking part. This means that you will have the chance to win either one £30 voucher, or one of two £10 vouchers.

Another benefit is being part of a study that will help us understand the difficulties in social anxiety, which in turn may help better treat young people with social anxiety. We will also be providing you with information about common mental health difficulties and where to seek support if you feel you might benefit from this.

Additionally we will be working with your school to organise a talk or workshop (for example about an area of psychology, or careers advice) that will hopefully be helpful to you.

Are there any risks?

This project does not have any likely risks associated with it. Sometimes, when we ask young people questions about their feelings, this can be upsetting but there will always be someone available to talk to and to help. Additionally it is possible that some people may feel distressed whilst doing some of the tasks. You are free to withdraw from the study at any time without giving a reason. Additionally, we will ensure that there is always someone available to talk to, either someone at school or someone from our research team. The project psychologist's email is at the bottom of this document.

Everything you tell us is private and confidential (i.e. we won't tell anyone, including your teachers). However, if you tell us something that makes us worry that there might be risk to you (for example you want to hurt yourself or that someone is hurting you) then we would need to break confidentiality. We would always try to speak to you about this first before contacting anyone and would be able to help you get support for this.

How will we ensure your data is kept confidential?

To ensure confidentiality, you will be randomly allocated a code. This unique code will be used on questionnaires, measures and data analysis so no personally identifiable information will be

associated directly with your data. The consent form that you sign, if you choose to take part, will be kept separately from the data in a locked filing cabinet which only the main researchers have access to. However, confidentiality will be broken in the unlikely event that you indicate potential harm to yourself or others as it is the researcher's duty to pass on this information. In this case, the researcher would speak with you and possibly your GP if necessary.

It is up to you to decide whether you would like her to take part or not. If you decide to take part, you are free to withdraw at any time and without giving a reason. In addition to withdrawing from the study, you may also withdraw any data/information she has already provided up until 30th September 2015.

The project is funded by King's College London

What will happen to the results of the study?
The results of the study will be used help us understand some difficulties in young people with social anxiety. The results will also be submitted for publication but no personally identifiable information will be included in this. If you wish to receive a copy of the published article, please let the researcher know.

What if I have further questions, or if something goes wrong?
If this study has harmed you in any way or if you wish to make a complaint about the conduct of the study you can contact King's College London using the details below for further advice and information:

The Chair, Psychiatry, Nursing & Midwifery (PNM RESC), rec@kcl.ac.uk
Who should I contact for further information?

If you have any questions or require more information about this study, please contact me using the following contact details:

Georgie Bremner
Trainee Clinical Psychologist
Institute of Psychiatry, Kings College London
Department of psychology, PO78, Institute of Psychiatry, Psychology & Addiction Sciences Building, 4 Windsor Walk, London SE5 8AF Email: georgina.bremner@kcl.ac.uk

Thank you for reading this information sheet and for considering taking part in this research.

Consent form for the 'no-stress' condition

CONSENT FORM FOR PARTICIPANTS IN RESEARCH STUDIES

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

This consent form is for young people

Social worries and how we understand others

King's College Research Ethics Committee Ref:

Thank you for considering taking part in this research. The person organising the research must explain the project to you before you agree to take part. If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

I confirm that I understand that by ticking/initialling each box I am consenting to this element of the study. I understand that it will be assumed that unticked/initialled boxes mean that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element I may be deemed ineligible for the study.	**Please tick or initial**

	Please tick or initial
1. I confirm that I have read and understood the information sheet dated 09/11/14, Version Number 1, for the above study. I have had the chance to consider the information and am happy with the answers to my questions.	
2. I understand that taking part is voluntary and that I am free to stop at any time without giving any reason. I also understand that I will be able to ask to have my data removed until September 2015.	
3. I consent to the use of my personal information for the reasons explained to me. I understand that such information will be handled in line with the terms of the UK Data Protection Act 1998.	
4. I understand that my information may be part of a review by responsible individuals from the College for monitoring and audit purposes.	
5. I understand that my information will be kept private and it will not be possible to identify me in any publications	
6. I agree to being informed of the results of this study If YES: Email address……	

Name of Participant	**Date**	**Signature**

Name of Researcher	**Date**	**Signature**

INFORMATION SHEET FOR PARTICIPANTS

REC Reference Number: PNM/14/15-41

YOU WILL BE GIVEN A COPY OF THIS INFORMATION SHEET

This information sheet is for young people

Social worries and how we understand others

We would like you to be part of a study that is investigating how social worries might impact on how we understand other people. We are asking students aged 16-18 to complete some questionnaires and video tasks in their classrooms. Since the age range of your class is within this, we are inviting you to take part in this study. However, before you decide whether you would like to take part or not, it is important that you understand why we are doing this study and what it involves. **Please remember that you do not have to take part in the study.** Please take time to read the following information carefully and discuss it with others if you wish. Ask your teacher or the project psychologist if there is anything that is not clear or if you would like more information.

What are we trying to find out?

Feeling anxious in social situations (known as social anxiety) is a common problem in young people and it can generate a lot of distress and worry. A lot of the fears and worries are to do with the way that they think about social situations. This study aims to find out more about social anxiety in 16-18 year olds in terms of how young people think about and understand others, which may play an important part in why people feel anxious.

Who are we asking to take part?

We will be asking all young people aged 16 to 18 at your school to take part in this study. Only students who have difficulties understanding written English will not be able to take part. This is because all participants need to be able to fully understand this information sheet and the consent form provided.

If you decide that you don't want to take part, you are free to withdraw at any time and without giving a reason. A decision not to take part, or a decision to withdraw at any time, will not affect you at school. If you have any questions about this project and what you are being invited to take part in, please ask the researcher before you decide whether to join in.

What will happen if I take part?

We will ask all participants to complete three questionnaires in the classroom, as well as some more interactive social tasks, including two video tasks. The total testing session will last up to 90

minutes. All your answers provided are completely private and will **not** be shared with the other pupils or with your teachers.

The first questionnaire will be asking about whether you ever experience things such as worry, nervousness or low mood. The second questionnaire will be asking about how you feel in social situations (e.g. I get nervous when I talk to peers I don't know very well) and the third questionnaire will be asking about your social understanding and communication (e.g. I find it easy to 'read between the lines' when someone is talking to me). There are no correct or incorrect answers.

The rest of the study will involve watching two videos of social interactions and completing a multiple choice answer sheet about what you saw.

Are there any benefits?
In taking part in this study you will be entered into a raffle with the rest of your classmates who are taking part. This means that you will have the chance to win either one £30 voucher, or one of two £10 vouchers.

Another benefit is being part of a study that will help us understand the difficulties in social anxiety, which in turn may help better treat young people with social anxiety. We will also be providing you with information about common mental health difficulties and where to seek support if you feel you might benefit from this.

Additionally we will be working with your school to organise a talk or workshop (for example about an area of psychology, or careers advice) that will hopefully be helpful to you.

Are there any risks?
This project does not have any likely risks associated with it. Sometimes, when we ask young people questions about their feelings, this can be upsetting but there will always be someone available to talk to and to help. Additionally it is possible that some people may feel distressed whilst doing some of the tasks. You are free to withdraw from the study at any time without giving a reason. Additionally, we will ensure that there is always someone available to talk to, either someone at school or someone from our research team. The project psychologist's email is at the bottom of this document.

Everything you tell us is private and confidential (i.e. we won't tell anyone, including your teachers). However, if you tell us something that makes us worry that there might be risk to you (for example you want to hurt yourself or that someone is hurting you) then we would need to break confidentiality. We would always try to speak to you about this first before contacting anyone and would be able to help you get support for this.

How will we ensure your data is kept confidential?
To ensure confidentiality, you will be randomly allocated a code. This unique code will be used on questionnaires, measures and data analysis so no personally identifiable information will be associated directly with your data. The consent form that you sign, if you choose to take part, will be kept separately from the data in a locked filing cabinet which only the main researchers have

access to. However, confidentiality will be broken in the unlikely event that you indicate potential harm to yourself or others as it is the researcher's duty to pass on this information. In this case, the researcher would speak with you and possibly your GP if necessary.

It is up to you to decide whether you would like her to take part or not. If you decide to take part, you are free to withdraw at any time and without giving a reason. In addition to withdrawing from the study, you may also withdraw any data/information she has already provided up until 30th September 2015.

The project is funded by King's College London

What will happen to the results of the study?

The results of the study will be used help us understand some difficulties in young people with social anxiety. The results will also be submitted for publication but no personally identifiable information will be included in this. If you wish to receive a copy of the published article, please let the researcher know.

What if I have further questions, or if something goes wrong?

If this study has harmed you in any way or if you wish to make a complaint about the conduct of the study you can contact King's College London using the details below for further advice and information:

The Chair, Psychiatry, Nursing & Midwifery (PNM RESC), rec@kcl.ac.uk

Who should I contact for further information?

If you have any questions or require more information about this study, please contact me using the following contact details:

Georgie Bremner
Trainee Clinical Psychologist
Institute of Psychiatry, Kings College London
Department of psychology, PO78, Institute of Psychiatry, Psychology & Addiction Sciences Building, 4 Windsor Walk, London SE5 8AF Email: georgina.bremner@kcl.ac.uk

Thank you for reading this information sheet and for considering taking part in this research.

Consent for the social stress condition

Version 1: 09/11/14

CONSENT FORM FOR PARTICIPANTS IN RESEARCH STUDIES

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

KING'S College LONDON

This consent form is for young people

Social worries and how we understand others

King's College Research Ethics Committee Ref:

Thank you for considering taking part in this research. The person organising the research must explain the project to you before you agree to take part. If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

I confirm that I understand that by ticking/initialling each box I am consenting to this element of the study. I understand that it will be assumed that unticked/initialled boxes mean that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element I may be deemed ineligible for the study.

Please tick or initial ☐

Please tick or initial

1. I confirm that I have read and understood the information sheet dated 09/11/14, Version Number 1, for the above study. I have had the chance to consider the information and am happy with the answers to my questions. ☐
2. I understand that taking part is voluntary and that I am free to stop at any time without giving any reason. I also understand that I will be able to ask to have my data removed until September 2015. ☐
3. I consent to the use of my personal information for the reasons explained to me. I understand that such information will be handled in line with the terms of the UK Data Protection Act 1998. ☐
4. I understand that my information may be part of a review by responsible individuals from the College for monitoring and audit purposes. ☐
5. I understand that my information will be kept private and it will not be possible to identify me in any publications ☐
6. I agree that I will be filmed as part of the study ☐
7. I understand that video-footage will only be used for research purposes by responsible individuals within psychology research, and will not be available to the general public. ☐
8. I agree that the research team can use clips of the video for teaching purposes or conferences. (This is an optional additional consent and declining will not prevent you taking part in the project). ☐
9. I would like to be informed of the results of this study ☐
   If YES: Email address.....

Name of Participant	**Date**	**Signature**
Name of Researcher	**Date**	**Signature**

6.3 Appendix C: Study Baseline Measure

SAS-A

This is not a test, there are no right or wrong answers. Please answer each item as honestly as you can.

Use these numbers to show HOW MUCH YOU FEEL something
is true for you: 1 = Not at all
2 = Hardly ever
3 = Sometimes
4 = Most of the time
5 = All the time

1.	I worry about doing something new in front of others……	1 2 3 4 5
2.	I like to do things with my friends……	1 2 3 4 5
3.	I worry about being teased……	1 2 3 4 5
4.	I feel shy around people I don't know……	1 2 3 4 5
5.	I only talk to people I know really well……	1 2 3 4 5
6.	I feel that peers talk about me behind my back……	1 2 3 4 5
7.	I like to read……	1 2 3 4 5
8.	I worry about what others think of me……	1 2 3 4 5
9.	I'm afraid that others will not like me……	1 2 3 4 5
10.	I get nervous when I talk to peers I don't know very well ……	1 2 3 4 5
11.	I like to play sports……	1 2 3 4 5
12.	I worry about what others say about me ……	1 2 3 4 5
13.	I get nervous when I meet new people……	1 2 3 4 5
14.	I worry that others don't like me……	1 2 3 4 5
15.	I'm quiet when I'm with a group of people……	1 2 3 4 5
16.	I like to do things by myself……	1 2 3 4 5
17.	I feel that others make fun of me……	1 2 3 4 5
18.	If I get into an argument, I worry that the other person will not like me……	1 2 3 4 5
19.	I'm afraid to invite others to do things with me because they might say no…….	1 2 3 4 5
20.	I feel nervous when I'm around certain people……	1 2 3 4 5
21.	I feel shy even with peers I know well……	1 2 3 4 5
22.	It's hard for me to ask others to do things with me……	1 2 3 4 5

RCADS

Please put a circle around the word that shows how often each of these things happens to you. There are no right or wrong answers.

1	I worry about things	Never	Sometimes	Often	Always
2	I feel sad or empty	Never	Sometimes	Often	Always
3	When I have a problem, I get a funny feeling in my stomach	Never	Sometimes	Often	Always
4	I worry when I think I have done poorly at something	Never	Sometimes	Often	Always
5	I would feel afraid of being on my own at home	Never	Sometimes	Often	Always
6	Nothing is much fun anymore	Never	Sometimes	Often	Always
7	I feel scared when I have to take a test	Never	Sometimes	Often	Always
8	I feel worried when I think someone is angry with me	Never	Sometimes	Often	Always
9	I worry about being away from my parent	Never	Sometimes	Often	Always
10	I am bothered by bad or silly thoughts or pictures in my mind	Never	Sometimes	Often	Always
11	I have trouble sleeping	Never	Sometimes	Often	Always
12	I worry that I will do badly at my school work	Never	Sometimes	Often	Always
13	I worry that something awful will happen to someone in my family	Never	Sometimes	Often	Always
14	I suddenly feel as if I can't breathe when there is no reason for this	Never	Sometimes	Often	Always
15	I have problems with my appetite	Never	Sometimes	Often	Always
16	I have to keep checking that I have done things right (like the switch is off, or the door is locked)	Never	Sometimes	Often	Always
17	I feel scared if I have to sleep on my own	Never	Sometimes	Often	Always
18	I have trouble going to school in the mornings because I feel nervous or afraid	Never	Sometimes	Often	Always
19	I have no energy for things	Never	Sometimes	Often	Always
20	I worry I might look foolish	Never	Sometimes	Often	Always

21	I am tired a lot	Never	Sometimes	Often	Always
22	I worry that bad things will happen to me	Never	Sometimes	Often	Always
23	I can't seem to get bad or silly thoughts out of my head	Never	Sometimes	Often	Always
24	When I have a problem, my heart beats really fast	Never	Sometimes	Often	Always
25	I cannot think clearly	Never	Sometimes	Often	Always
26	I suddenly start to tremble or shake when there is no reason for this	Never	Sometimes	Often	Always
27	I worry that something bad will happen to me	Never	Sometimes	Often	Always
28	When I have a problem, I feel shaky	Never	Sometimes	Often	Always
29	I feel worthless	Never	Sometimes	Often	Always
30	I worry about making mistakes	Never	Sometimes	Often	Always
31	I have to think of special thoughts (like numbers or words) to stop bad things from happening	Never	Sometimes	Often	Always
32	I worry what other people think of me	Never	Sometimes	Often	Always
33	I am afraid of being in crowded places (like shopping centers, the movies, buses, busy playgrounds)	Never	Sometimes	Often	Always
34	All of a sudden I feel really scared for no reason at all	Never	Sometimes	Often	Always
35	I worry about what is going to happen	Never	Sometimes	Often	Always
36	I suddenly become dizzy or faint when there is no reason for this	Never	Sometimes	Often	Always
37	I think about death	Never	Sometimes	Often	Always
38	I feel afraid if I have to talk in front of my class	Never	Sometimes	Often	Always
39	My heart suddenly starts to beat too quickly for no reason	Never	Sometimes	Often	Always
40	I feel like I don't want to move	Never	Sometimes	Often	Always
41	I worry that I will suddenly get a scared feeling when there is nothing to be afraid of	Never	Sometimes	Often	Always
42	I have to do some things over and over again (like washing my hands, cleaning or putting things in a certain order)	Never	Sometimes	Often	Always
43	I feel afraid that I will make a fool of myself in front of people	Never	Sometimes	Often	Always
44	I have to do some things in just the right way to stop bad things from happening	Never	Sometimes	Often	Always
45	I worry when I go to bed at night	Never	Sometimes	Often	Always
46	I would feel scared if I had to stay away from home overnight	Never	Sometimes	Often	Always
47	I feel restless	Never	Sometimes	Often	Always

AQ-10

Please tick one option per question only:		Definitely Agree	Slightly Agree	Slightly Disagree	Definitely Disagree
1	I often notice small sounds when others do nc				
2	I usually concentrate more on the whol picture, rather than the small details				
3	I find it easy to do more than one thing at onc				
4	If there is an interruption, I can switch back t what I was doing very quickly				
5	I find it easy to 'read between the lines' whe someone is talking to me				
6	I know how to tell if someone listening to me getting bored				
7	When I'm reading a story I find it difficult to wor out the characters' intentions				
8	I like to collect information about categories c things (e.g. types of car, types of bird, types c train, types of plant etc.)				
9	I find it easy to work out what someone thinking or feeling just by looking at their face				
10	I find it difficult to work out people's intentions				

SCPQ-P

Please put a circle around the rating which best describes you over the past 4 weeks. Please answer all questions.

	Not True	Sometimes True	Mostly True
1. Has at least one close friend	0	1	2
2. Has stable friendships with peers	0	1	2
3. Peers like to sit next to him/her in class	0	1	2
4. Finds it easy to make friends	0	1	2
5. Is chosen by peers to be on their team	0	1	2
6. Peers invite him/her to parties or social events	0	1	2
7. Is popular amongst peers	0	1	2
8. Is chosen by peers as a partner to work on a project	0	1	2
9. Has good relationships with classmates	0	1	2

SCPQ-T

Please put a circle around the rating which best describes this pupil over the past 4 weeks. **Please circle the 0 if the item is not true. Circle the number 1 if the item is sometimes true. It the item is mostly true, then circle the number 2.** Please answer all items.

	Not True	Sometimes True	Mostly True
1. Has at least one close friend	0	1	2
2. Has stable friendships with peers	0	1	2
3. Peers like to sit next to him/her in class	0	1	2
4. Finds it easy to make friends	0	1	2
5. Is chosen by peers to be on their team	0	1	2
6. Peers invite him/her to parties or social events	0	1	2
7. Is popular amongst peers	0	1	2
8. Is chosen by peers as a partner to work on a project	0	1	2
9. Has good relationships with classmates	0	1	2

6.4 Appendix D: Instructions and description of the MSR tasks

The Movie for the Assessment of Social Cognition: Instructions and Example Question

Instructions (each bullet presented on a separate slide):

- You will be watching a 15 minute film. Please watch very carefully and try to understand what each character is feeling or thinking.
- Now you will meet each character: Sandra, Michael, Betty, and Cliff (a photo is shown of each character)
- The film shows these four people getting together for a Saturday evening.
- The movie will be stopped at various points and some questions will be asked.
- All of the answers are multiple choice and require one option to be selected from a choice of four.
- If you are not exactly sure of the correct answer, please guess.
- When you answer, try to imagine what the characters are feeling or thinking at the very moment the film is stopped.
- The first scene is about to start. Are you ready? Again, please watch very carefully because each scene will be presented only once.

Question 1:

The movie begins with a doorbell ringing, and a young woman (Sandra) opens the front door. Upon opening the door, a man of a similar age, enters the apartment. Sandra says "Hi" and the man asks her whether she is surprised. Before she can answer, he tells her that she looks "terrific", and asks her whether she did something with her hair. Sandra smiles and touches her hair and starts to say something when the man compliments her again, saying that her hair looks "classy".

The movie then stops and the following question is presented with four options to choose from:

What is Sandra feeling?

1. (1) that her hair does not look nice (no mentalizing)
2. (2) that she is pleased about his compliment (less MSR)
3. (3) that she is exasperated about the man coming on too strong (hyper-MSR)
4. (4) that she is flattered but somewhat taken by surprise (accurate MSR)

Instructions for Frith-Happe Animations Task

Instructions:

You are going to watch 12 short video clips showing two animated triangles. Each clip lasts for less than 1 minute and has no sound.

Please watch each sequence very carefully and make sure that you watch the clip right until it finishes.

The clips are divided into 3 categories:

- No interaction: There is no obvious interaction between the triangles and movement appears random.
- Physical interaction: An interaction between the triangles in which actions are directed toward each other in order to achieve specific goals.
- Mental interaction: An interaction between the triangles involving the manipulation of the emotions and thoughts of one triangle by the other.

After each clip you will be asked to decide which of these 3 categories best describes what was happening.

For some of these clips there will be some more detailed questions regarding what you think was happening, so please pay careful attention to them.

First you will be shown two practice items.

6.5 Appendix E: State Anxiety and Self-focused attention measure

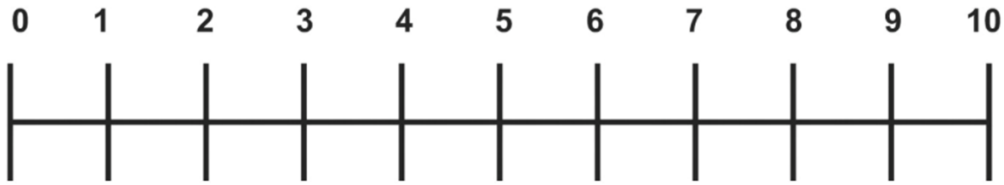
Situational Self-awareness scale

Please respond to each statement based on how you feel **RIGHT NOW, AT THIS INSTANT**—not the number that how you feel in general, or at this point in your life. **Circle the statement that best fits with your answer.** There are no 'right' or 'wrong' answers—just be honest.

1. Right now I'm, keenly aware of everything in my environment	**Strongly disagree**	**Moderately disagree**	**Slightly Disagree**	**Neither agree not disagree**	**Slightly Agree**	**Moderately agree**	**Strongly agree**
2. Right now, I am conscious of my inner feelings.	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree
3. Right now, I am concerned about the way I present myself	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree
4. Right now, I am self-conscious about the way I look.	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree
5. Right now, I am conscious of what is going on around me.	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree
6. Right now, I am reflective about my life.	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree
7. Right now, I am concerned about what other people think of me.	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree
8. Right now, I am aware of my innermost thoughts.	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree
9. Right now, I am conscious of all objects around me.	Strongly disagree	Moderately disagree	Slightly Disagree	Neither agree not disagree	Slightly Agree	Moderately agree	Strongly agree

Anxiety Likert Scale

How anxious/nervous/worried were you feeling during the video tasks?



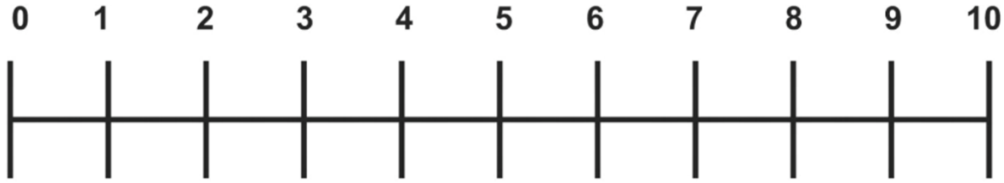
Not at all nervous

Moderately Nervous

Extremely nervous

Self-focus Likert Scale

How much was your attention focused on your own feelings/thoughts during the video tasks?



Not at all focused

Moderately focused

Fully focused

6.6 Appendix F: Behavioural task checklist for assessors

Conversation Task

Please rate how you think the student appeared during the task on the following items:

	Not at all				Very much
Anxious	**1**	**2**	**3**	**4**	**5**
Confident	**1**	**2**	**3**	**4**	**5**
Embarrassed	**1**	**2**	**3**	**4**	**5**
Boring	**1**	**2**	**3**	**4**	**5**
Socially Skilled	**1**	**2**	**3**	**4**	**5**
Voice quivering	**1**	**2**	**3**	**4**	**5**
Left long gaps	**1**	**2**	**3**	**4**	**5**
Interesting	**1**	**2**	**3**	**4**	**5**
Shaky	**1**	**2**	**3**	**4**	**5**
Awkward	**1**	**2**	**3**	**4**	**5**
How loud and clear was their voice?	**1**	**2**	**3**	**4**	**5**
Did they stumble over their words?	**1**	**2**	**3**	**4**	**5**
How much did they smile?	**1**	**2**	**3**	**4**	**5**
How much did you look at the other two people?	**1**	**2**	**3**	**4**	**5**
How much did they speak compared with either of you?	**1**	**2**	**3**	**4**	**5**

Speech Task

Please rate how you think the student appeared during the task on the following items:

	Not at all				Very much
Anxious	**1**	**2**	**3**	**4**	**5**
Confident	**1**	**2**	**3**	**4**	**5**
Embarrassed	**1**	**2**	**3**	**4**	**5**
Boring	**1**	**2**	**3**	**4**	**5**
Voice quivering	**1**	**2**	**3**	**4**	**5**
Left long gaps	**1**	**2**	**3**	**4**	**5**
Interesting	**1**	**2**	**3**	**4**	**5**
Shaky	**1**	**2**	**3**	**4**	**5**
Awkward	**1**	**2**	**3**	**4**	**5**
How loud and clear was their voice?	**1**	**2**	**3**	**4**	**5**
Did they stumble over their words?	**1**	**2**	**3**	**4**	**5**
How much did they smile?	**1**	**2**	**3**	**4**	**5**
How much did they look at the either of you?	**1**	**2**	**3**	**4**	**5**
How good was their speech?	**1**	**2**	**3**	**4**	**5**

6.7 Appendix G: Demographic and baseline characteristics of students who completed the behaviour task

Age	17.31 (0.25); 16.89 – 17.78
Gender	Female: 17 (73.9%) Male: 6 (26.1%)
Ethnicity	Black / Black British: 17 (73.9%) Mixed / Multiple ethnicities: 2 (8.7%) Asian / Asian British: 2 (8.7%) White: 1 (4.3%) Chinese: 0 (0%) Other: 1 (4.3%)
AQ-10	3.15 (1.35); 1-5
SAS-A	40.96 (10.83); 20-71
SCPQ	15.35 (2.56); 11-20
RCADS-A	33.13 (12.26); 8-46
RCADS-D	8.25 (3.24); 2-12

Service Related Project

Facilitating an acceptance and commitment therapy group in an acute inpatient setting: A service evaluation

Supervised by Dr Rumina Taylor

Croydon & Southwark Psychology Acute Inpatient Services

Psychosis Clinical Academic Group

Abstract

There is growing recognition of the need for increased access to psychological therapies for the acute inpatient population. The use of psychology groups is an important means of achieving this, given the high patient to therapist ratio seen on most inpatient wards. Acceptance and commitment therapy (ACT) has an emerging evidence base for use with individuals with psychosis as well as those within acute care. The present study describes the introduction of an ACT group delivered to two acute inpatient wards in Croydon and Southwark, both within the Psychosis Clinical Academic Group (CAG). The effectiveness, feasibility and acceptability of the intervention were evaluated using clinical outcome measures as well as service user and staff feedback questionnaires. The results and limitations of the study are discussed and service recommendations are provided.

Contents

1. Introduction 132

1.1. Need for Evidence-based Psychological Therapies in Acute Inpatient Settings 132

1.2 Acceptance and Commitment Therapy 134

1.3 Evidence Base 136

1.4 Aims and Objectives 138

2. Method 138

2.1 Design 138

2.2 Participants 138

2.2.1. Recruitment 139

2.3 The ACT Group 139

2.3.1 Equipment and Materials 140

2.4 Outcome Measures 141

2.4.1 Effectiveness 141

2.4.2 Feasibility and Acceptability 142

2.5. Data Analysis 143

3. Results 143

3.1 Participant Characteristics 143

3.2 Attendance and Attrition 144

3.3 Outcome Measures 144

3.3.1 Effectiveness 144

3.3.2 Feasibility and Acceptability 147

3.4 Summary of Results 150

4. Discussion 151

4.1 Feasibility, Acceptability and Effectiveness of the ACT Group 151

4.2 Systemic Challenges 153

4.3 Limitations 154

4.4 Recommendations 155

4.5 Dissemination of Findings 155

4.6 Leadership Skills 155

4.7 Conclusions 156

5. References 157

6. Appendices 160

Appendix A: Group Worksheets 160

Appendix B: Feedback Questionnaires 163

List of Tables

Table 1: Six Core Processes Targeted by ACT to Enhance Psychological Flexibility 135

Table 2: Outline of Group and Session Content 140

Table 3: Demographic and Clinical Characteristics of Participants 144

Table 4: Means and Standard Deviations of Scores on Clinical Outcome Measures at Baseline and Post-intervention. 145

Table 5: Medians and Ranges of Scores on the Stress and Symptom Inference Bubbles 146

List of Figures

Figure 1*:* Means of scores on the MHCS, CORE-10, SMQ and VQ clinical outcome measures at baseline and post-intervention..... 145

Figure 2: Medians of scores on the Stress Bubbles before and after each session. 146

Figure 3: Medians of scores on the Symptom Interference Bubbles before and after each session. 147

1. Introduction

1.1. Need for Evidence-based Psychological Therapies in Acute Inpatient Settings

A number of recent policy initiatives have been set out to improve the quality of acute inpatient mental health care (Department of Health (DoH) 2002, 2007; Mind, 2011; Schizophrenia Commission, 2012). The reports have highlighted that the experience of hospitalisation for many service users is aversive, with some even describing acute settings as 'anti-therapeutic' (e.g. the Schizophrenia Commission, 2012). MIND (2004; 2011) found that medication continues to be prioritised over psychological interventions, despite service users wanting access to psychological therapies during crisis. Not only are anti-psychotic medications associated with significant side effects, but also the majority who take them continue to experience residual positive and negative symptoms (Pfammatter et al., 2006). Thus, psychological interventions, aimed to help service users cope with symptoms that medication does not eliminate, are greatly needed.

Another issue that has been raised is the 'intense boredom' that many service users experience on acute wards (MIND, 2004). In a review of social interaction and activity among inpatients on 16 acute psychiatric wards in six hospitals, Radcliffe and Smith (2007) found 84% of patients to be socially disengaged and mainly inactive. Inactivity and boredom have been associated with more challenging behaviour and violent incidents (MIND, 2004; McGeorge & Rae, 2007), which impacts negatively on the mental and physical health of patients (Sommers & Vodanovich, 2000). It is not surprising then, that the Care Quality Commission (CQC; 2011) found that just under half of service users felt that the ward had a negative effect on their mental health.

A negative ward environment represents a false economy for trusts as this results in delayed recovery and thus longer admissions (MIND, 2004). Service users are also less likely to choose to come into hospital voluntarily when needed, or may try to avoid services because they fear being admitted, placing themselves at significant risk and increasing the numbers detained under the Mental Health Act (MHA; DoH, 2007). Inpatient admissions are costly to individuals as well as to society and so it is of critical importance to improve the quality of care provided in order to reduce the number and length of stays.

Many have argued that an important step towards raising the standards of acute inpatient care is to increase access to psychological therapies (e.g. DoH, 2002; Sainsbury Centre for Mental

Health Study, 2005; Schizophrenia Commission, 2012). This is in line with recommendations to reduce the medical emphasis and move towards a more "collaborative and person centred approach" which promotes "healing and recovery" (MIND, 2011; p. 42). The recommendations also state that ''a wide range of effective psychological therapies'' should be accessible for all service users accessing inpatient and crisis services, including brief interventions (MIND, 2011; p. 45). However, despite these recommendations, less than one ward in five psychiatric wards has been found to regularly offer talking therapies (Sainsbury Centre for Mental Health Study, 2005) and less than half of service users who wanted talking therapy had been able to access any (CQC, 2009).

A major challenge for psychologists working within acute inpatient settings is to find ways of working that will benefit the largest number of service users possible. Within the South London and Maudsley NHS Foundation Trust (SLaM), where the current evaluation took place, a total of six full-time equivalent clinical psychologists cover a total of 16 wards (serving approximately 260 patients). Thus, in order to benefit larger numbers of service users, input from clinical psychologists should not be restricted to individual therapy but should also support other staff members with individual and group interventions, in addition to helping the wider development of a therapeutic milieu (DoH, 2002; Royal College of Psychiatrists, 2010). This latter part of the role is thought to be particularly pertinent in inpatient settings, where the psychological impact needs to extend to the whole ethos of the institution. Thus, supporting other staff in delivering group interventions (via co-facilitation) can be a useful training model to disseminate expertise and skills in the practice of psychological therapies, which in turn helps develop a more therapeutic ward environment.

Delivering group interventions offers not only a useful training opportunity, but also offers a cost-effective and efficient means of increasing service user access to psychological therapies. Groups also have the added benefit of reducing isolation, increasing activity and providing structure to the ward environment. However there are a number of challenges in developing groups within the acute inpatient setting. Firstly, the acute inpatient population can be difficult to engage, since many are detained under the MHA and are not necessarily treatment-seeking. Secondly, the high turnover of service users means that the time-scale of interventions needs to be relatively short-term. Finally, the heterogeneity of mental health needs and level of functioning of service users means that providing a group intervention that is interesting, meaningful, and appropriate for all in attendance, is greatly challenging (Radcliffe & Smith, 2007; Newell, Harries, & Ayers, 2012). These challenges highlight the appeal of delivering evidence-

based trans-diagnostic approaches in the inpatient setting, as well as underline the need for monitoring and evaluation to ensure that interventions are relevant to participants (CQC, 2011).

1.2 Acceptance and Commitment Therapy

Acceptance and commitment therapy (ACT) is a 'third wave' therapy within the cognitive behavioural therapy (CBT) tradition. Whereas first and second wave therapies have focused on changing behavior and cognitions, ACT focuses on changing an individual's *relationship* with their thoughts and feelings, thus helping them develop alternative ways of relating to unwanted internal experiences. Many mental health disorders involve experiential avoidance, whereby individuals try to avoid, escape or suppress distressing thoughts, feelings or unusual experiences (Hayes, Luoma, Bond, Masuda, & Lillis, 2006). Since research suggests that attempts to suppress such psychological phenomena can often make it worse (Wegner, 2009), ACT encourages clients to recognise and reduce their struggles and instead develop a more accepting stance to be able to move in a valued direction (Bloy, Oliver, & Morris, 2011).

ACT identifies six core clinical processes that are central to the development of psychological flexibility, and the intervention aims to enhance these processes (see Table 1 for a description of these processes, adapted from Hayes and Strosahl, 2004).

Table 1: Six Core Processes Targeted by ACT to Enhance Psychological Flexibility

Acceptance	ACT aims to foster acceptance and willingness towards experience, while undermining the dominance of emotional control and avoidance as responses.
Defusion	Defusion is separating or distracting oneself from thoughts, and letting them go instead of being caught up with them. Being fused with one's thoughts enables difficult thoughts to dominate behaviour and act as a barrier to life-promoting activities.
Contact with Present Moment	ACT encourages living in the present moment with greater awareness, contacting more fully the ongoing flow of experience as it occurs.
Self-as-Context	ACT aims to enable individuals to make experiential contact with a sense of self that is a safe and consistent perspective from which to observe and accept changing inner experiences.
Values	Values are chosen directions for living, which cannot be fully satisfied or achieved (unlike a goal). ACT aims for individuals to identify their personal values, in order to live a more satisfying and meaningful life.
Committed Action	ACT aims to enable individuals to build larger and larger patterns of committed action that are consistent with values.

The ACT approach of enhancing psychological flexibility is especially appropriate for the acute inpatient population, as many will be suffering with chronic mental health difficulties. For this population, developing a more accepting stance towards persistent and recurring symptoms, and engaging with more valued and meaningful living, is likely to be very beneficial. This is in line with the 'Recovery Approach', which was developed in the 1990s in response to the appreciation that people with mental illness need more than just symptom relief. This is because the consequences of severe mental illness include significant reductions in personal and social functioning, which greatly impedes an individual's chance of true recovery. For this reason there have been a number of policy initiatives in recent years to promote greater use of the recovery approach in mental health settings (e.g. Shepherd, Boardman & Slade, 2008).

Despite growing appreciation of the recovery approach, mainstream attitudes continue to view mental health symptoms (particularly psychotic symptoms) as unacceptable. This has meant that in many mental health settings, recovery continues to be equated to symptom elimination, which can have harmful consequences. In particular, it can mean that individuals learn not to

refer to psychotic or other distressing experiences, so that they can be discharged from hospital. Thus, individuals are more likely to avoid or suppress symptoms, leading to greater preoccupation with their inner world, becoming increasingly dominated by their illness and often retreating from previously valued activities and aspirations (Mitchell & McArthur, 2013). ACT aims to reverse this cycle by enhancing processes such as acceptance, defusion and value based committed action to enable individuals to live a more meaningful and valued life.

1.3 Evidence Base

The ward settings in the current study were part of the Psychosis CAG, which means that the majority of patients had a diagnosis on the psychosis spectrum (ICD-10 codes: F.20 – F.29; WHO, 1993). These are a group of severe mental disorders characterised by delusions and hallucinations, which disrupt a person's perception, emotion and behaviour (NICE, 2014). There is emerging evidence for the effectiveness of ACT as an intervention for psychosis in terms of managing distressing symptoms and reducing relapse rates. A randomised controlled trial comparing a four-session individual ACT intervention with treatment as usual, found that ACT participants had lower rehospitalisation rates over both a four month and one year follow-up period (Bach & Hayes, 2002; Bach, Hayes, & Gallop, 2012). This finding of reduced rehospitalisation rates was replicated by Gaudiano and Herbert (2006) who additionally found that the ACT participants showed advantages in mood, social impairment and overall distress associated with hallucinations. Thus, there is growing evidence that ACT can be developed into an effective intervention for people with serious mental illness, such as psychosis, and that it can be facilitated within inpatient settings (Peterson & Zettle, 2009).

There is less evidence for the effectiveness of delivering ACT interventions in group formats for individuals experiencing psychosis within acute settings. However, evidence from the CBT literature would suggest that group interventions can be as effective as individual interventions for this population (e.g. Wykes, Steel, Everitt, & Tarrier, 2008). Moreover, other mindfulness and acceptance-based psychological interventions, delivered in a group format to those with psychosis, have been reported to be both acceptable and feasible (e.g. Chadwick, Hughes, Russell, Russell, & Dagnan, 2009; Jacobsen, Morris, Johns, & Hodkinson, 2011). Indeed, many aspects of ACT lend themselves to delivery in a group format, such as metaphors and experiential exercises, which can benefit from involvement of more than two individuals (Walser & Pistorello, 2004). Additionally, setting goals and making commitments in a social context may strengthen the likelihood of action, and involvement of peers may help reduce stigma and enhance self-compassion.

There are a small but growing number of clinicians who are using ACT in a group format for individuals with psychosis. McArthur, Mitchell and Johns (2013) outline a six-session 'ACT-for-Psychosis Group Protocol', which they run as a closed programme in both inpatient and outpatient settings. A closed programme allows for the progressive development of the ACT model and approach with the same participants across several sessions. The authors report that they use the protocol flexibly to suit the specific context. For instance, for outpatient settings the protocol is delivered in six (1.5 - 2 hour) weekly sessions, whereas for inpatient settings it is delivered over just two weeks to better suit the shorter duration of acute admissions. Their group has a focus on helping clients to get in touch with their values and identify barriers, by having discussions about "what's important in life?" and "what gets in the way of us doing what's important?". Values are discussed in terms of being 'life directions' (see Table 1 for definition of values) that are chosen by an individual and which guide action. The authors reflect that exploration of values can be distressing for clients, as it can involve recognising how disconnected their lives have become from the things they care about most. However, the authors also note that working with clients' values is an essential component of ACT, as they provide a powerful source of motivation and allow individuals to emotionally connect with the costs of their avoidance. The authors also state that basing the session content around just one metaphor (the 'Passengers on the Bus'[8]; Hayes, Strosahl, & Wilson, 1999) is important for simplicity and makes the sessions more accessible for all participants. Despite these useful recommendations, to our knowledge there has been no published evaluation of this group intervention to date.

Another research group have developed a four-session 'Act for Life' protocol for outpatients with psychosis (O'Donoghue, Oliver, Morris, Johns, Jolley, & Butler, submitted), which outlines four (2 hour) weekly sessions, with two (2 hour) follow-up booster sessions. The sessions are broadly aimed to encourage more flexible (accepting, mindful, defused) responding to symptoms and associated emotions and thoughts, in order to increase values based behaviour. The sessions are also based around the 'Passengers on the Bus' metaphor, to illustrate the concepts of values, mindfulness, willingness, defusion and committed action. The current study

[8] The 'Passengers on the Bus' metaphor is designed to help individuals become aware of their choices and values (by identifying where they want their bus to go), whilst emphasising the control they have over their choices (only they can drive their bus) and also highlighting difficult psychological phenomena such as anxieties, doubts, voices (passengers on their bus) which can serve as barriers to moving in a chosen valued direction. In this way the metaphor encourages individuals to recognise their passengers, and instead of struggling with them, to accept them whilst still being able to make active choices about how they want to live their life.

adapted the Act for Life (O'Donoghue et al., submitted) protocol for the first time, in order to better suit the inpatient context. This was done by making sessions just one hour long and running them bi-weekly, so that the four sessions were completed within two weeks, rather than four. No follow up booster sessions were provided as this was deemed to be less feasible for the acute environment where discharge rates are relatively high.

1.4 Aims and Objectives

The aim of the current study was to adapt the Act for Life (O'Donoghue et al., submitted) protocol and to implement the intervention within two acute inpatient wards; Gresham 2 Ward at the Bethlem Royal Hospital and Aubrey Lewis 3 Ward at the Maudsley Hospital, both within the Psychosis CAG.

The primary objective of the current study was therefore to assess the feasibility, acceptability and potential benefits of this ACT group for inpatients with psychosis, in order to add to the limited evidence base. It was expected that the ACT group would improve access to psychological therapies for those within an acute inpatient setting in line with recent national policy recommendations. Finally, it was hoped that the intervention would also disseminate skills and understanding of a new psychological approach to ward staff, with the aim of creating a more therapeutic ward environment.

2. Method

2.1 Design

This study was conducted using an AB (Baseline-Intervention) single group design. Each ACT group consisted of four bi-weekly one-hour sessions. The four-session group was run a total of five times across the two sites (three times at the Bethlem site and twice at the Maudsley site).

2.2 Participants

Participants were a total of 30 males (aged 19 – 66) from two male acute inpatient wards: Gresham 2 Ward (Bethlem Royal Hospital) and Aubrey Lewis Ward 3 (Maudsley Hospital) both within the Psychosis CAG.

2.2.1. Recruitment

The recruitment strategy was proactive, targeting those patients who were approaching discharge, regardless of diagnosis, since this seemed to fit with the ethos of ACT with its focus on changing relationships with symptoms and working towards one's values for living. However, anyone who was experiencing distress and wanting to work towards their recovery goals and values was invited to attend the group, as long as their attendance would not significantly disrupt or disturb other participants. Information was provided to the multi-disciplinary team (MDT) about ACT and the nature of the group. Invitations to attend the group were extended to ward staff to increase a sense of ownership of the group among non-psychology MDT staff. Posters and flyers were pinned up on notice boards around the ward and distributed among staff and patients.

2.3 The ACT Group

The intervention was based on the 'ACT for Life: Group Intervention for Psychosis Manual' (O'Donoghue et al., submitted) which describes an intervention consisting of four two-hour weekly sessions spanning four weeks. The protocol was adapted to become four one-hour bi-weekly sessions, spanning just two weeks, to better suit the high turnover of patients in the inpatient setting. As set out in the manual, the intervention was based principally around the 'Passengers on the Bus' metaphor (Hayes et al., 1999), which provided a consistent narrative thread throughout the groups. This metaphor allowed the exploration of issues such as personal values, committed action, barriers, thought-fusion and mindfulness. It was also decided to place more of an emphasis on participant values, as research has suggested that during crisis individuals engage in less values based action (Mitchell & McArthur, 2013). Every session therefore included a values exercise, discussions around barriers to values and addressing these through mindfulness practice, and setting values-driven actions. It was also felt that a focus on recovery goals and values for living rather than symptoms would foster greater engagement and motivation of service users to attend the group, in a setting where building a rapport with clients can be difficult.

Each session followed the same format: a values exercise, a short mindfulness exercise, a barriers to values exercise, discussions of committed action from the previous session, value-driven goal setting, group discussion/activity (including role plays) and setting between session committed actions (see Table 2 for an outline of the group with session content, and Appendix

A, p. 46-48 for worksheets used). Groups aimed to have six to eight participants, with two to three facilitators, who remained the same during each block of the intervention.

Table 2: Outline of Group and Session Content

Session Number	Summary of Content
1	Introductions and ground rules. Introduction of key concepts: values, barriers and committed action (using flip chart/white board, card sorting exercise, 'Passengers on the Bus' video and worksheets). Concept of mindfulness concept introduced using 3-minute breathing space. Participants helped to identify a personal value and to set a corresponding committed action plan to complete out of the session.
2	Introductions and reminder of ground rules. Key concepts revisited: values, barriers and committed action (using the 'Compass' metaphor and 'Passengers on the Bus' metaphor). Mindfulness exercise ('leaves on the stream') used to illustrate struggles with unwanted psychological phenomena. Participants facilitated to identify their personal barriers or 'passengers' that prevent values-based action. Values-based action plan revisited before close of group.
3	Introductions and reminder of ground rules. Exercise of 'noticing others' values' before a 5-minute mindfulness exercise. Introduction of new concept 'willingness' (noticing without engaging) using the 'clipboard metaphor' exercise and the 'Passengers on the Bus' role-play, to practice different ways of being with 'passengers'. Values based-action plan revisited before close of group.
4	Introductions and reminder of ground rules. Mindfulness eating exercise. Recap of values, goals, internal and external barriers, committed action before closing thoughts from participants.

2.3.1 Equipment and Materials

The group was delivered using a number of different materials. A computer (or laptop) and projector were used to display a presentation with key points as well as watching the 'Passenger on the Bus' video (https://www.youtube.com/watch?v=u0VPUudQ7kc), which helped make the groups more interactive and provided structure. During group discussions, key points were written up on a whiteboard or flipchart. Individual worksheets were also used in every session, where participants could write down personal values, barriers, goals and committed action plans (see Appendix A, p.46-48 for worksheets). Refreshments were also provided in many of the groups to create a relaxed and welcoming atmosphere.

2.4 Outcome Measures

2.4.1 Effectiveness

Four outcome measures were completed before the first session (baseline) and after the final session (post-intervention). Additionally measures of stress and symptom interference (fusion) were administered before and after each session to measure within session changes.

2.4.1.1 Confidence and Coping with Symptoms. The Mental Health Confidence Scale (MHCS; Carpinello, Knight, Markowitz, & Pease, 2000) was used to measure self-efficacy in relation to mental health. The questionnaire consists of 16 items that measure three factors: confidence in the development of hope, confidence in the ability to manage symptoms and emotions, and confidence in the ability to advocate for one's needs and rights. Reponses are given on a six-point Likert scale (0= very non confident to 6 = very confident) with higher total scores indicative of greater self-efficacy. The questionnaire has high construct validity and low error variance making it a reliable measure (Carpinello et al., 2000).

2.4.1.2 Psychological Distress. Global psychological distress was measured using a 10-item version of the Clinical Outcomes in Routine Evaluation (CORE-10; Barkham et al., 2013). This assesses domains of functioning, problems or symptoms and well-being, with the responses given on a five-point Likert frequency scale (0 = never to 4 = most or all of the time) and total scores are gained by a sum of the items. A higher total score is indicative of greater psychological distress. The CORE-10 has been shown to correlate with the CORE-OM at 0.94 in a clinical sample, and has good internal reliability (Cronbach's alpha = 0.9) (Barkham et al., 2013).

2.4.1.3 Mindfulness Skills. The Southampton Mindfulness Questionnaire (SMQ; Chadwick et al., 2008) was used to assess the extent to which participants respond mindfully to distressing thoughts and images. It consists of 16 items that measure four aspects of mindfulness: full observation, letting something be, not having aversion, and the absence of judgment. Items are scored on a seven-point Likert scale, worded "strongly disagree" to "strongly agree", yielding a range of possible scores from 0 to 96. The authors indicate it has a one-factor structure and exhibits adequate internal and external validity (Chadwick et al., 2008).

2.4.1.4 Pursuing Values. The Valuing Questionnaire (VQ; Smout, Davies, Burns, & Christie, 2014) is an eight-item questionnaire that measures the extent to which an individual is living in line with their values, with higher scores indicating more values-based behavior. The authors indicate it has a two-factor structure (factor one = progress towards valued living and factor two = obstructions to valued living) and that it has good psychometric reliability and validity.

2.4.1.5 Stress and Symptom Interference. In addition to the above four outcome measures, within session changes in stress and symptom interference were assessed using *stress bubbles* (Jacobsen et al., 2011). These are in the form of a visual analogue scales, with six bubbles gradually increasing in size from "no stress" (score=0) to "very stressed" (score=5) for the stress scale, and "no interference" (score=0) to "very interfering" (score=5) for the symptom interference scale.

2.4.2 Feasibility and Acceptability

At the end of the four sessions, participants were asked to complete a service-user feedback questionnaire. Any staff members who were involved with running the group (who were not project leads in the current study) were also asked to complete a qualitative feedback questionnaire. Additionally, participant group attendance and attrition were recorded.

2.4.2.1 Service User Satisfaction. To quantitatively measure service user satisfaction with the group, a satisfaction questionnaire, developed by Gledhill, Lobban and Sellwood (1998) was given to all participants to complete after the final session. The questionnaire consists of nine dichotomous (yes/no) questions relating to the person's experience of being part of a psychology group.

Three extra open-format questions were added to the questionnaire for the purpose of this study, to elicit qualitative feedback from participants that could help inform future development of the service. These included questions about ease of access to the group, what aspects of the group were found to be helpful/unhelpful and how the group could be improved in the future (see Appendix B, p.49 for questionnaire).

2.4.2.2 Staff Feedback. A five-item open format questionnaire was developed to elicit qualitative feedback from staff involved with running the group. Questions focused on practical

and logistical issues of running the group and whether their involvement in the group impacted their own practice (see Appendix B, p.51 for questionnaire).

2.5. Data Analysis

Frequencies and distributions were inspected. Participant group attendance and attrition was recorded. Satisfaction questionnaires were reviewed at the end of treatment and percentages were calculated for each item to give an indication of whether clients enjoyed the group, found it beneficial, and would participate in the intervention again. Paired samples t-tests, where participants had provided both baseline and post-intervention outcome measures, were used to assess any benefits from the intervention on the outcome measures and where data were non-normally distributed non-parametric tests were employed. Sample sizes vary in different analyses due to missing data. Analyses were conducted using SPSS version 21. All tests were two-tailed and the significance level was set at $p<0.05$.

Thematic analysis was conducted on qualitative data in order to identify emerging themes within individual questions. Particular attention was paid to areas of consensus between respondents and to topics pertinent for the group's development.

3. Results

3.1 Participant Characteristics

The mean age of group members was 40.4 years (*SD* = 14.6 years; range = 19-66), with an average contact with mental health services of 6.9 years (*SD* = 4.8 years; range = 0-15 years) and an average of 2.6 previous hospital admissions (*SD* = 2.5; range = 0–10). Just under half (46.7%) of participants were classified as being Black or Black British, 43.3% as White or White British and 10% as Asian or Asian British, which reflects the population served by the trust. The majority of participants (63%) were under a section of the MHA and 76.7% had a diagnosis on the psychosis spectrum (ICD-10: F.20 – F.29; WHO, 1993). A summary of demographic and clinical characteristics of participants is shown in Table 3.

Table 3: Demographic and Clinical Characteristics of Participants

Demographic and Clinical Characteristics	Mean (SD)
Age	40.4 (14.6)
Ethnicity	
White/White British	43.3%
Black/Black British	46.7%
Asian/Asian British	10%
Diagnosis	
Psychosis Spectrum	76.7%
Other[9]	23.3%
Mental Health Act Status	
Informal	37%
Section 2	13%
Section 3	50%
Contact with mental health services (years)	6.9 (4.8)
Previous admissions	2.6 (2.6)

3.2 Attendance and Attrition

Each session had a mean of 5.1 participants (*SD* = 1.5; range: 2-7) with earlier sessions tending to have more participants than later sessions (*M* = 6.0, *SD* = 1.4 for session 1; *M* = 5.4, *SD* = 1.1 for session 2; *M* = 4.4, *SD* = 1.7, for session 3; *M* = 4.4, *SD* = 1.5, for session 4), reflecting an attrition rate of 26.7%. Each participant completed a mean of 3.3 sessions (*SD* = 0.9), which reflected an average attendance rate of 83%.

3.3 Outcome Measures

3.3.1 Effectiveness

Two participants did not provide baseline measures (due to personal choice) and five participants did not complete the post-intervention assessment. Participants who did not complete the post-intervention assessment did not differ significantly from the other participants in terms of their demographic and clinical characteristics, nor their baseline outcome data. The mean scores and standard deviations of the outcome measures administered

[9] Other diagnoses included borderline personality disorder, mental health disorder not otherwise specified and non-psychotic disorder.

are presented in Table 4 and Figure 1. Paired samples t-tests showed statistically significant reductions in global psychological distress between baseline (*M*=16.5, *SD*=9.4) and end of treatment (*M*=12.7, *SD*=9.6), [t (22) = – 2.39, p = 0.03] as assessed by the CORE-10. There was also a statistically significant increase in mindful responding to distressing thoughts and images, as measured by the SMQ, between baseline (*M*=48.3, *SD*=17.5) and post-intervention (*M*=57.7, *SD*=15.3) [t (20) = -2.87, p = 0.009]. No significant changes were observed on reported levels of confidence and coping (as measured by the MHCS), nor of valued living (as measured by the VQ).

Table 4: Means and standard deviations of scores on clinical outcome measures at baseline and post-intervention.

	N	Baseline	Post Intervention
MHCS	23	59.5 (17.6)	62.7 (16.0)
CORE-10	23	16.5 (9.4)	12.7 (9.6)
SMQ	21	48.3 (17.5)	57.7 (15.3)
VQ	22	25.3 (9.7)	28.6 (9.3)

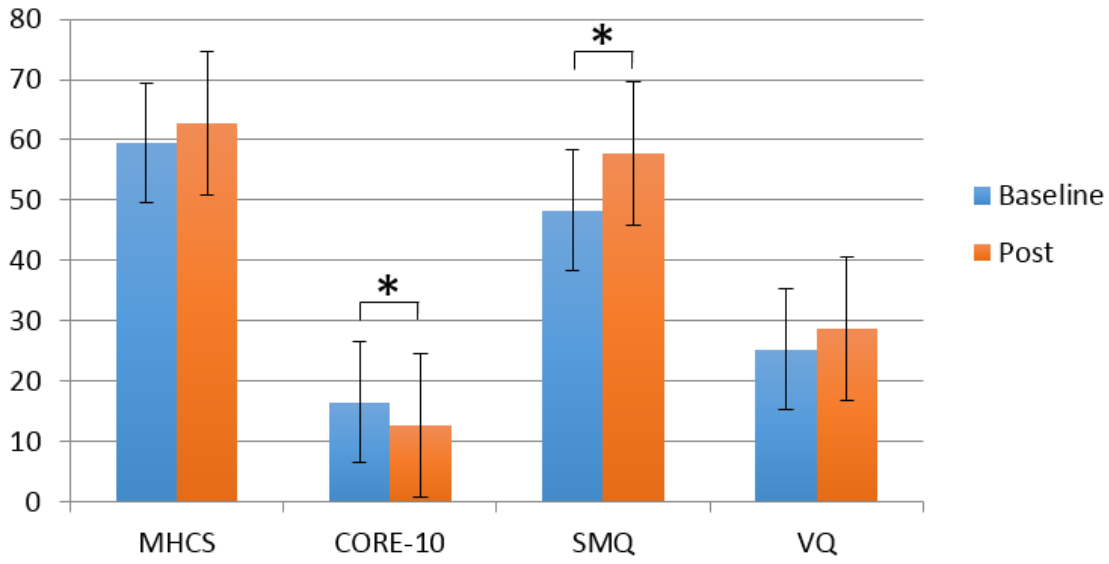


Figure 1: Means of scores on the MHCS, CORE-10, SMQ and VQ clinical outcome measures at baseline and post-intervention. Significant differences between baseline and post-intervention were found on the CORE-10 and on the SMQ (* = p < 0.05).

Scores on the stress and interference measure were non-normally distributed and therefore Wilcoxon signed-rank tests were used. The medians and ranges of scores are presented in Table 5 and Figures 2 and 3. For all sessions of the intervention, there were significant within-session reductions in both stress [session 1: Z = 5.28, p < 0.001; session 2: Z = 3.58, p = 0.002; session 3: Z = 3.43, p = 0.003; session 4: Z = 2.41, p = 0.024] and symptom interference [session 1: Z = 2.17, p = 0.039; session 2: Z = 2.14, p = 0.045; session 3: Z = 3.07, p = 0.011; session 4: Z = -2.77, p = 0.012]. Stress and symptom interference ratings prior to the start of each session did not differ significantly across sessions.

Table 5: Medians and ranges of scores on the stress and symptom inference bubbles

	Stress		**Symptom interference**	
	Before session	**After session**	**Before session**	**After session**
Session 1	2 (0-5)	0 (0-4)	2 (0-5)	1 (0-4)
Session 2	2.5 (0-5)	1 (0-4)	2 (0-5)	1.5 (0-5)
Session 3	2 (0-5)	1 (0-4)	2 (0-4)	0 (0-3)
Session 4	0.5 (0-5)	0 (0-4)	1 (0-5)	0 (0-5)

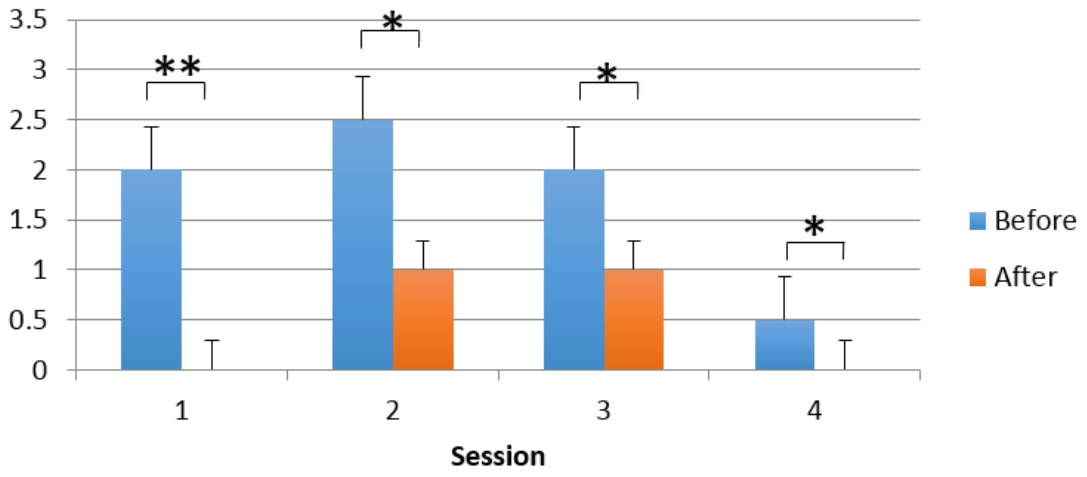


Figure 2: Medians of scores on the Stress Bubbles before and after each session. All sessions produced a significant reduction in stress ratings (* = p < 0.05; ** = p <0.001).

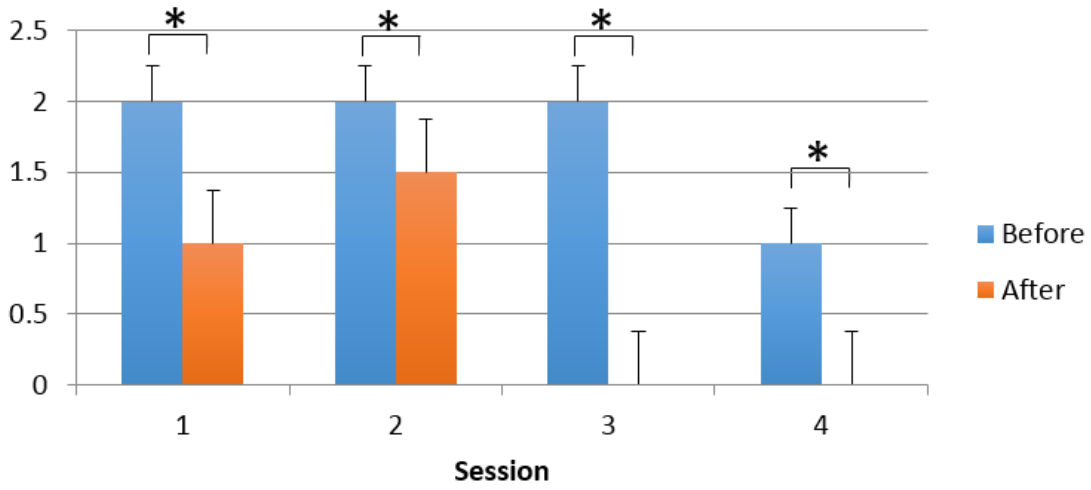


Figure 3: Medians of scores on the Symptom Interference Bubbles before and after each session. All sessions produced a significant reduction in symptom interference ratings (* = $p < 0.05$).

3.3.2 Feasibility and Acceptability

A total of 21 participants independently completed the feedback questionnaire. Participants from the first group (a total of four) were not given feedback questionnaires as this was not yet part of the protocol. A further five participants from later groups did not complete a questionnaire due to non-attendance at the final session.

3.3.2.1 Service User Quantitative Feedback. All 21 participants stated they had enjoyed the group and would recommend the group to other service users. Ninety per-cent felt that they benefitted from the group and would like to be involved in a similar group in the future. Seventy four per-cent stated that they felt better able to cope with their problems after attending the group. Just over half of the participants (53%) fed back that they would have preferred to have engaged in individual work and 30% reported finding it difficult to discuss their problems in the presence of others. However, 79% felt that they had benefited from meeting other group members who had similar difficulties and only 5% indicated there were components of the group they did not like.

3.3.2.1 Service User Qualitative Feedback. Thematic analysis was conducted in order to identify emerging themes within individual questions. Particular attention was paid to areas of consensus between respondents and to topics pertinent for the group's development. The following three themes were identified:

3.3.2.1.1 Feasibility of attendance. Participants identified the group as being easy to access, which is likely to be due to the fact that the group took place on the ward where they were residents.

3.3.2.1.2 Aspects that had been found helpful. Many participants commented on the value of being part of a group. Example quotes included "the group helped me feel a bit more independent - talking to friends and hearing ideas", "I enjoyed the fact it was group based as there was loads of input from different individuals" and "I liked the experience of being in the group". Many participants also valued the mindfulness exercises, for example one participant found that "it was nice to concentrate on the present and learn not to be distracted by the past or let future worries get to you". Another found the exercises to be "important and relaxing". Finally, participants also appeared to find the values-based exercises helpful, for example "I liked learning to focus in and stop blocking off from what's important to you".

3.3.2.1.3 Suggestions for group improvements. A few participants felt that a longer intervention would have been helpful. For example, one participant said that "a longer time frame would have been better" and another said that he "would have liked more sessions".

Another area for improvement suggested by participants was to have more interactive aspects of the group, such as role-plays. For example, one participant suggested greater use of "props and acting parts" as he had "found these fun".

3.3.2.2 Staff Qualitative Feedback. A total of seven staff members were involved with running the five groups across the two sites. Staff comprised two clinical psychologists (CP), two trainee clinical psychologists, one occupational psychologist (OT) and two healthcare assistants (HCA). Feedback was collected from five staff members who were not directly involved in carrying out the current service evaluation. A total of four questionnaires were returned (one HCA left the service before completing a questionnaire) which provided data from each of the different professionals (one CP, one trainee CP, one OT and one HCA). Thematic analysis was conducted in order to identify emerging themes within individual questions. Particular attention was paid to areas of consensus between respondents and to topics pertinent for service development. The following themes were identified:

3.3.2.2.1 Overcoming Practical Difficulties. Staff highlighted a number of practical issues with running the group. For instance, finding a room "*big enough and at the right time*" was difficult, which meant that flexibility both in terms of the group's start time and the space available was required.

Staff members commented on how sudden discharge of participants could be disruptive for group attendance. For example, one staff member noted that service users "*tend to move on from the ward quickly*" meaning that they might be *"discharged during the course of the two week intervention, and therefore miss the final sessions"*. A solution to this difficulty was that participants were "*informed that they were welcome back on the ward to attend the group sessions, even if they were discharged"*. However, in reality very few participants took up this option.

Another area of difficulty reported by staff was how to manage participants that were deemed unsuitable for the group. It was noted that some participants could present very differently from one session to the next. For example, one staff member said that *"the mental states of some patients could be quite changeable"* and indeed *"one participant became unsuitable for the group after the first session, as his temperament became very volatile"* which would have been highly disruptive for the rest of the group members. One strategy that was used to prevent this outcome was to consult ward staff "*in advance of each session to help reduce the likelihood that an unsuitable patient (e.g. particularly volatile in mood) attended"*. However, the difficulty of how to tell a service user that they would not be suitable for the group remained, and was something that had to be done as sensitively as possible.

Another challenge was ensuring that all ward staff on duty understood that the group was 'closed'. For instance, problems were encountered when *"staff could send in patients who had not been selected for the group"* or *"send patients in to join a session a long time after the start"* which was *"quite disruptive*". Feedback from one staff member noted that "*with changing shifts, bank staff etc.*" it was difficult to inform all staff about the group and its particular format. Another staff member who made a similar comment noted that this was probably because ward staff were used to "*psychology groups being more drop in drop out"*. One staff member commented on the effectiveness of *"having more of a presence on the ward"* in terms of *"helping communication between us and the rest of the ward staff*". This also made it easier to "*ensure that all staff understood that the group was closed, and also that they were aware of*

who should be attending the group so that they are able to remind the relevant clients". One staff member found that putting the names of those patients participating in the group up "*in the nursing office as well as in the ward diary, helped prevent situations whereby unsuitable patients would be sent in, but also situations where group members would be inadvertently be taken off the ward to attend other appointments*".

3.3.2.2.1 Perceived Benefits of the Group. All staff members that were involved in the group felt that the group had a beneficial impact on the wellbeing on the participants and all felt that the group should continue to be run. One staff member commented that "*patients seemed to really enjoy the group*" and that *"some objectively appeared more settled and euthymic over the course of the group sessions*". Another commented on how the ACT principles could successfully be understood by patients "*as evidenced by their verbal input and reflections during group discussions"*. One staff member commented on how the video used was "*very accessible*" and another member commented on benefit of the group's emphasis on values "*which everyone can relate to – things like family, friends, independence etc. and these things are often not at the forefront of people's minds when in crisis in hospital*".

3.3.2.2.2 Impact of ACT group on Staff Practice. All staff members felt that their involvement in the ACT group changed their own practice. One staff member said how it encouraged them to "*think about replacing some of the current groups with ACT groups instead*" as well as *"revise some other group material to incorporate some ACT ideas (such as thinking about values)*". This was echoed by another staff member who said that they have started to use more ACT techniques in their individual work. One staff member felt that the staff who were involved in facilitating the ACT groups are perhaps "*more enlightened anyway*" and that teaching ACT principles to staff who are "*less likely to think in an ACT way, would be really beneficial*". Staff beyond those directly involved in the group also fed-back informally how patients had told them of how they had enjoyed the group. One psychologist therefore "*spent some time communicating some of the key metaphors (such as driving your bus, and barriers to doing so) so that staff were able to follow up this work if it came up with patients*".

3.4 Summary of Results

The outcomes of these first five ACT groups provide encouraging indications of the potential benefits of running such groups within acute inpatient settings.

Firstly, participants reported significant within-session reductions in both Stress as well as Symptom Interference, in every session. Secondly, scores across a range of outcome measures, collected before and after the four-session intervention, point to significant reductions in the reporting of general psychological distress and a significant increase in the reporting of mindful responding.

In terms of the participants' experience of the group, feedback from service-users shows that the group was both accessible and acceptable. Feedback from staff also indicates that the group was perceived to be both feasible and beneficial.

4. Discussion

4.1 Feasibility, Acceptability and Effectiveness of the ACT Group

This study aimed to assess the feasibility, acceptability, and effectiveness of running a four-session ACT group within an acute inpatient setting. The positive changes in scores across a range of outcome measures, between baseline and post intervention suggest that the four-session ACT group had a positive impact on participants' psychological wellbeing. Moreover, both the service-user and staff feedback indicated that this group was feasible to run for staff facilitators, and it was found to be accessible, enjoyable and beneficial for service-users to attend.

The significant reductions in both stress and symptom inference within each of the four sessions highlights the benefit of attending even just one session. This is particularly pertinent in the unpredictable ward environment, where many patients will only attend one or two sessions.

Another promising outcome of this evaluation was the significant reduction in scores on the CORE-10, which indicates that service-users were reporting mild levels of distress as the post-intervention time point, compared with moderate levels at baseline. Additionally, the significant increase in scores on the SMQ suggests that the group was successful in developing those ACT skills that were being targeted during the sessions. Indeed the mindfulness or 'noticing' exercises appeared to be accessible, helpful and enjoyable for participants who were objectively and often subjectively calmer after doing these exercises. This significant increase in more mindful responding is impressive given that unlike other groups that have employed mindfulness

techniques (e.g. Chadwick et al., 2009), the current protocol did not give instructions for members to practice these skills outside of the sessions. Considering that participants enjoyed this aspect, perhaps some simple between session mindfulness exercises may have further enhanced and consolidated these skills.

Increased scores on the VQ indicated that, although not statistically significant, participants were living slightly more in line with personal values at the post-intervention time point, than at baseline. However, this lack of a significant increase might reflect the difficulty, from a clinician's perspective, in helping clients within the inpatient setting, to set meaningful and relevant committed actions that they can complete between sessions. For example, for a number of clients their chosen personal value was 'education', but whilst they were in hospital they felt restricted in what meaningful actions they could commit to that would be in line with this. Since the VQ specifically measures the enactment of personal values over the past week, it is likely that difficulties in setting and completing such actions, perhaps due to the inherent restrictions of the inpatient environment, limits the potential for change on this measure. Perhaps working with those values that better lend themselves to the ward environment (e.g. 'friendship' or 'helpfulness') would have helped to over-come this challenge.

Similarly to the VQ, the MHCS also demonstrated a slight, but not significant, increase in scores at the post-intervention time point, suggesting a small shift towards an increased sense of self-efficacy. However, participants were already scoring highly on this measure at baseline, indicating consistently high self-efficacy ratings across the time points and perhaps limiting any further significant increase. This was perhaps unusual for patients within an acute setting where the majority were under section of the MHA and not necessarily treatment seeking. However, the intervention was targeted at those nearing discharge and who were aware that their time on the ward was coming to an end. In this way, the group was also pitched to participants as an opportunity to consider their lives in the community after discharge to facilitate recovery, which may have instilled feelings of hope.

Beyond the positive changes on clinical outcome measures, the group appeared to provide an opportunity for service users to access psychological therapy in an environment that fostered mutual respect between individuals. All participants, many of whom were experiencing chronic, distressing and often treatment-resistant psychosis, were able to tolerate the hour-long group sessions and were able to engage with the group material and reflect on personal experiences.

More importantly, all participants stated that they had enjoyed the group and that they would recommend it to other service users. This is especially promising given heterogeneity of the groups in terms of the different ethnicities, ages, backgrounds and clinical presentations of the participants. It appears that the 'Passengers on the Bus' metaphor and personal values were concepts that most people could relate to.

4.2 Systemic Challenges

There were some systemic challenges in running a closed group within the inpatient setting. This is likely because most other groups run on the wards were structured as 'open' and transdiagnostic, with stand-alone sessions to accommodate the varied nature of acute admissions and high patient turnover. It was therefore a challenge to implement a closed group that required some homogeneity in clinical presentation (i.e. some more disruptive clinical presentations were deemed unsuitable for this group) and to ensure that both ward staff and patients understood this. As highlighted in the staff feedback, it could be quite disruptive when staff inadvertently sent in unsuitable clients, halfway through a session, and it could be difficult to explain to some service users why they were not able to attend. Additionally, potentially suitable service users who were admitted to the ward after the intervention had started were unfortunately unable to access the group. However, a major benefit of a closed group was that it allowed for each session to build upon material from the last, and it provided a safe and supportive space for members to develop a sense of connectedness and share their personal experiences. ACT also being a trans-diagnostic therapeutic approach was especially appropriate and fit with the heterogeneous presentations observed across participants.

MDT staff involvement was crucial in a number of ways, such as helping to identify suitable participants and encouraging attendance by both reminding participants but also ensuring that other appointments did not coincide with the group where possible. Such support from the MDT is likely to have played a role in the high attendance rate of 83%. It was also helpful to have staff be involved in the group as participants, to promote ideas of common humanity and reduce perceptions of 'them and us'. Although invitations were extended to all staff members to attend the groups only one staff member chose to participate. This likely reflects a number of systemic challenges including low staffing numbers and therefore a limited capacity to run a group. A continuing challenge will therefore be to integrate ACT principles to the wider ward milieu. This is especially challenging given that ACT principles, such as non-judgmental awareness of difficult internal experiences, are in contrast to the medical model that often continues to dominate inpatient settings.

4.3 Limitations

There were a number of limitations with the current study. Firstly, there was no control or comparison group, which means that it is difficult to draw strong conclusions that the intervention was responsible for positive changes in clinical outcome measures. Many other factors are likely to have been influencing the well-being of participants, such as medication, natural recovery or simply becoming more settled within the ward environment. Where possible participants completed outcome measures independently but when support was required this came from group therapists, which may have affected responding. Additionally, lack of follow up data precludes any examination of the potential longer-term impact of the groups, although practically this data would have been difficult to collect given that many participants were close to discharge. Moreover, given the brief nature of the intervention its effectiveness in producing highly significant and long-term changes on clinical outcome measures is likely to be compromised. However, this group was designed and implemented to accommodate the fast-paced nature of acute inpatient care and to allow as many service users as possible to access and benefit from a psychological intervention.

Another limitation was that service user and staff feedback was collected in the format of written questionnaires. Not only does this require participants to understand and produce written information but it also meant that some questions were left out, especially those that were open ended and aiming to elicit qualitative feedback. It is likely that use of a semi-structured interview format would enable the collection of richer and more thorough feedback data, which would better inform future service development.

An additional limitation is that service user feedback was only received from those participants who were present at the final session, which may have created a biased sample. Although no data was formally collected on reasons for non-attendance, only seven participants out of the total sample of 30 did not attend a final session. Discharge or leave from the ward seemed to be the main reasons for this, but this should be formally assessed in future evaluations. Lastly, as this group was an all-male sample, these findings are not representative of the female inpatient population.

4.4 Recommendations

Based on the clinical outcomes and feedback from current evaluation, the following recommendations are made for the ACT Group:

- To continue to facilitate the ACT group intervention on both Gresham 2 Ward and Aubrey Lewis 3 wards, and to continue to monitor its effectiveness, feasibility and acceptability.
- Future groups should consider the following:
  - To collect staff and service user feedback in the format of semi-structured interviews.
  - To collect longer-term follow up data.
  - To formally collect data on reasons for non-attendance.
  - To continue to disseminate both the group's outcomes as well as ACT principles to non-psychology ward staff
  - To introduce this intervention to other male or female inpatient settings, within the Psychosis CAG.

4.5 Dissemination of Findings

The results of this evaluation were fed back to the services (Gresham 2 ward and Aubrey Lewis 3 ward) in presentations delivered to the MDTs in one of the weekly meetings. Feedback also involved teaching staff ACT skills to use on the ward. Based on the outcome of the current evaluation the group is planned to be continued to run at both sites but with the recommendations described above. A preliminary report of the evaluation was also published on the Psychosis CAG intranet site and circulated.

4.6 Leadership Skills

A number of leadership skills were developed in the running of this service evaluation, which were required at each stage. Firstly, when initially introducing the ACT group to the wards, the clinical team had to be engaged through presentations at MDT meetings and individual conversations with relevant staff members, such as ward managers. Indeed, close team working was a vital aspect of this evaluation, since constant coordination with the MDT was required both to enable the use of the required facilities (such as rooms, laptops, projectors, refreshments) and to help identify suitable participants and encourage attendance.

The running of the group also involved leadership with respect to the teaching of ACT principles and skills, in a way that influenced and developed the practice of other staff members. This occurred both via co-facilitation but also via teaching ACT skills at MDT meetings (in which the

ACT group was first introduced and when the results were fed-back) in order to reach staff members who were not directly involved with running the groups.

More generally, leadership skills were required for managing difficult situations as and when they arose, including managing unsuitable participants and overcoming practical difficulties such as the availability of rooms and staff.

4.7 Conclusions

This evaluation provides valuable 'practice-based evidence' (Beail, Newman & Kellett, 2004) to support the growing evidence base of the effectiveness of using ACT in a group format, within acute inpatient settings.

The results of this evaluation demonstrate that the ACT group intervention is a workable way to engage inpatient service users in a way that promotes a reduction in psychological distress, reduces stress and fusion, maintains a high sense of self-efficacy, increases mindfulness based skills and increases valued-based action.

The evaluation also represents an important move towards changing staff attitudes towards distress and creating a genuinely therapeutic ward environment. Increased involvement of non-psychology MDT staff in the future running of the group will be needed to ensure that the intervention is sustainable and supported within the broader context.

5. References

Bach, P., & Hayes, S. C. (2002). The use of acceptance and commitment therapy to prevent the rehospitalization of psychotic patients: a randomized controlled trial. *Journal of consulting and clinical psychology*, *70*(5), 1129.

Bach, P., Hayes, S. C., & Gallop, R. (2012). Long-term effects of brief acceptance and commitment therapy for psychosis. *Behavior modification*, *36*(2), 165-181.

Barkham, M., Bewick, B., Mullin, T., Gilbody, S., Connell, J., Cahill, J., ... & Evans, C. (2013). The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. *Counselling and Psychotherapy Research*, *13*(1), 3-13.

Beail, N., Newman, D. W., & Kellett, S. (2004). Routine monitoring: Developing practice based evidence for psychotherapy. *Journal of Intellectual Disability Research*, 48, 300-300.

Bloy, S., Oliver, J. E., & Morris, E. (2011). Using acceptance and commitment therapy with people with psychosis: A case study. *Clinical Case Studies*, 1534650111420863.

Care Quality Commission (2009). *Mental Health Acute Inpatient Services Survey 2009*. London: Care Quality Commission.

Care Quality Commission. (2011). *Monitoring the Mental Health Act in 2012/11.* Newcastle: Care Quality Commission.

Carpinello, S. E., Knight, E. L., Markowitz, F. E., & Pease, E. A. (2000). The development of the Mental Health Confidence Scale: A measure of self-efficacy in individuals diagnosed with mental disorders. *Psychiatric Rehabilitation Journal*, *23*(3), 236.

Chadwick, P., Hughes, S., Russell, D., Russell, I., & Dagnan, D. (2009). Mindfulness groups for distressing voices and paranoia: a replication and randomized feasibility trial. *Behavioural and Cognitive Psychotherapy*, *37*(04), 403-412.

Chadwick, P., et al. (2008). Responding mindfully to unpleasant thoughts and images: reliability and validity of the Southampton mindfulness questionnaire (SMQ).*British Journal of Clinical Psychology* 47(Pt 4): 451-455.

Department of Health. (2002). *Mental Health Policy Implementation Guide. Adult Acute Inpatient Care Provision.* London: Department of Health.

Department of Health. (2007). *The Mental Health Act.* London: The Stationery Office.

Gaudiano, B. A., & Herbert, J. D. (2006). Acute treatment of inpatients with psychotic symptoms using Acceptance and Commitment Therapy: Pilot results. *Behaviour research and therapy*, *44*(3), 415-437.

Gledhill, A., Lobban, F., & Sellwood, W. (1998). Group CBT for people with schizophrenia: a preliminary evaluation. *Behavioural and Cognitive Psychotherapy*, *26*(01), 63-75.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour research and therapy*, *44*(1), 1-25.

Hayes, S. C., & Strosahl, K. D. (Eds.). (2004). *A practical guide to Acceptance and Commitment Therapy*. New York: Springer-Verlag.

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (1999). *Acceptance and commitment therapy: An experiential approach to behavior change*. Guilford Press.

Jacobsen, P., Morris, E., Johns, L., & Hodkinson, K. (2011). Mindfulness groups for psychosis; key issues for implementation on an inpatient unit. *Behavioural and cognitive psychotherapy*, *39*(03), 349-353.

McArthur, A., Mitchell, G., & Johns, L. C. (2013). Developing Acceptance and Commitment Therapy for Psychosis as a Group-based Intervention. *Acceptance and Commitment Therapy and Mindfulness for Psychosis*, 219-239.

McGeorge, M., & Rae, M. (2007). Acute in-patient psychiatry: service improvement – the time is now. *Psychiatric Bulletin, 31,* 259-261.

MIND. (2004). *Ward Watch: Mind's campaign to improve hospital conditions for mental health patients.* London: Mind.

MIND. (2011). *Listening to Experience: An Independent Inquiry into Acute and Crisis Mental Healthcare*. London: Mind.

Mitchell, G. & McArthur, A., (2013). Acceptance and Commitment Therapy for Psychosis in Acute Psychiatric Admission Settings. *Acceptance and Commitment Therapy and Mindfulness for Psychosis*, 206-218.

Newell, S. E., Harries, P., & Ayers, S. (2012). Boredom proneness in a psychiatric inpatient *population*. . *International Journal of Social Psychiatry, 58,* 488-495.

National Institute for Health and Care Excellence (2014). *Psychosis and Schizophrenia in adults: treatment and management* (CG178). London. National Institute for Health and Care Excellence.

O'Donoghue, E., Oliver, J., Morris, E., Johns, L., Jolley, S. & Butler, L. (Submitted). ACT for Recovery: Group intervention for psychosis Manual.

Petersen, C. L., & Zettle, R. D. (2009). Treating Inpatients with Comorbid Depression and Alcohol Use Disorders: A Comparison of Acceptance and Commitment Therapy versus Treatment as Usual. *Psychological Record*, *59*(4).

Pfammatter, M., Junghan, U. M., & Brenner, H. D. (2006). Efficacy of psychological therapy in schizophrenia: conclusions from meta-analyses. *Schizophrenia bulletin*, *32* (supplement 1), S64-S80.

Radcliffe, J., & Smith, R. (2007). Acute in-patient psychiatry: how patients spend their time on acute psychiatric wards. *Psychiatric Bulletin, 31,* 167-170.

Royal College of Psychiatrists (2010). Accreditation for Inpatient Mental Health Services. Standards for Inpatient Wards – Working Age Adults (4th edn).

Shepherd, G., Boardman, J. & Slade, M. (2008) Making Recovery a Reality. London: Sainsbury Centre for Mental Health.

Smout, M., Davies, M., Burns, N., & Christie, A. (2014). Development of the Valuing Questionnaire (VQ). *Journal of Contextual Behavioral Science*, *3*(3), 164-172.

Sommers, J., & Vodanovich, S. J. (2000). Boredom proneness: Its relationship to psychological and physical health symptoms. *Journal of Clinical Psychology, 56,* 149-155.

The Sainsbury Centre for Mental Health. (2005). *Acute Care 2004: A national survey of adult psychiatric wards in England.* London: The Sainsbury Centre for Mental Health.

The Schizophrenia Commission. (2012). *The abandoned illness: a report from the Schizophrenia Commission.* London: Rethink Mental Illness.

Walser, R. D., & Pistorello, J. (2004). ACT in group format. In *A practical guide to acceptance and commitment therapy* (pp. 347-372). Springer US.

World Health Organization. (1993). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Wykes, T., Steel, C., Everitt, B., & Tarrier, N. (2008). Cognitive behavior therapy for schizophrenia: effect sizes, clinical models, and methodological rigor. *Schizophrenia Bulletin*, *34*(3), 523-537.

6. Appendices

Appendix A: Group Worksheets

All worksheets taken from the 'ACT for Recovery: Group intervention for psychosis Manual' (O'Donoghue et al., submitted) with permission from authors.

My Values and Goals

My goals are to:

_____
_____
_____
_____
_____
_____
_____

My actions to move me closer to my goals:

_____
_____
_____
_____
_____
_____
_____

My values guiding these goals and actions:

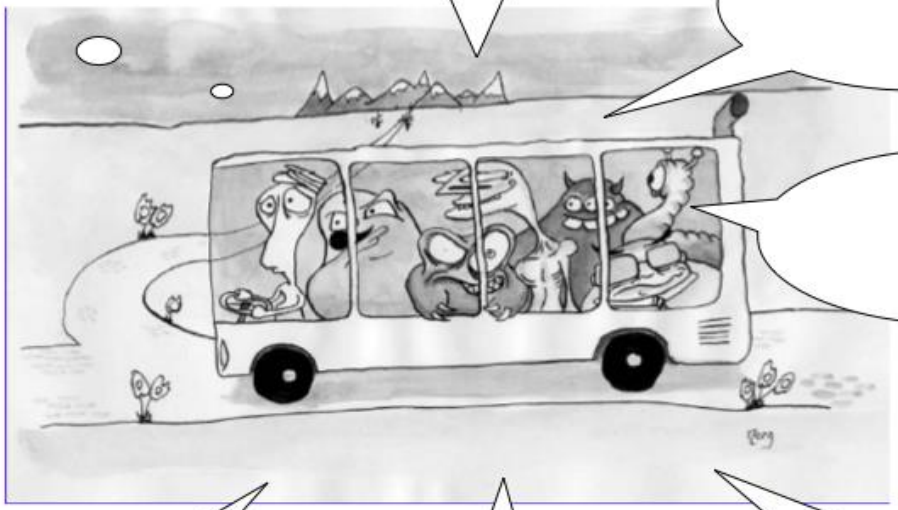
_____
_____
_____
_____
_____
_____

Passengers on the Bus Worksheet

Value:

Action related to your value:

_____
_____
_____
_____
_____
_____
_____



Out of Session planning Worksheet

My goal is to (be specific):

_____

_____

_____

_____

My values guiding this goal:

_____

_____

_____

My action for this week to move me closer to my goals:

_____

_____

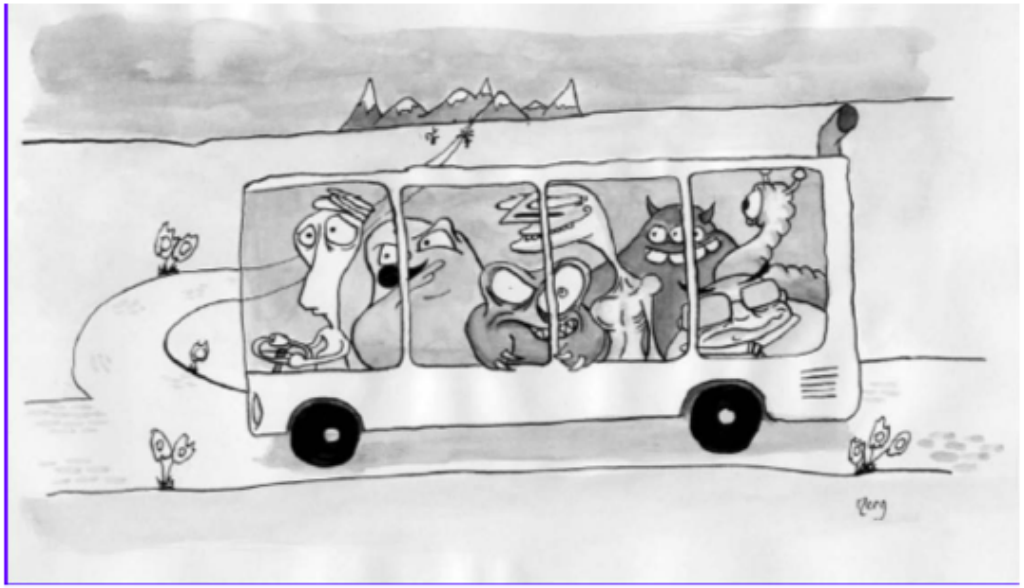
_____

_____

_____

_____

_____



Passengers that might show up as I work towards my goal: _____

_____

_____

_____

_____

_____

Appendix B: Feedback Questionnaires

Satisfaction questionnaire

(Adapted from Gledhill et al., 1998)

Thank you for attending the ACT group.

As this was the first time that this group has been run on this ward, it would be most helpful if we could have some feedback on how you found it. Please complete the following questions by circling the appropriate answer. Please feel free to add my additional comments; they will be most welcome.

1. I found the group enjoyable. Yes/No

2. I feel that I benefited in some way by attending the group. Yes/No

3. There were some things about the way the group was run that I did *not* like. Yes/No

4. I feel that I benefited from meeting people who had similar problems to my own. Yes/No

5. I found it difficult to discuss my problems in the presence of others. Yes/No

6. I would prefer to have been seen by a psychologist on my own. Yes/No

7. I feel more able to cope with my problems since attending the group. Yes/No

8. I would like to be involved in another group like this in the future. Yes/No

9. I would recommend this group to other service users? Yes/No

10. How easy/difficult was it for you to attend this group?

11. Do you have any suggestions about how it might be run in the future?

12. Any additional comments about what you found helpful or unhelpful about the group? (Feel free to use reverse side of this page)

Staff Feedback Questionnaire

Thank you for your involvement with the ACT group.

As this was the first time that this group has been run on this ward, it would be most helpful if we could have some feedback on how you found it.

1. Were there any difficulties that you encountered whilst trying to organise/ run the group in terms of feasibility/practicalities?

2. Were you able to overcome any practical issues? If so, how?

3. Did you think the group was beneficial in any way? If so, how?

4. Did your involvement with the ACT group change your own practice at all? Or the therapeutic approach of the ward?

5. Would you like to see the group continue to be run? Or you think individual work would be better?