
Examination number 4669800

Psychology (Master of Science by Research)

The University of Edinburgh

2009
ABSTRACT

Recent research on cognitive functioning in autistic spectrum disorders (ASDs) has seen a divergence in approach between accounts that emphasise a general cognitive style (Baron-Cohen, 2002; Happé & Frith, 2006) versus those who focus on specific cognitive deficits and difficulties (Plaisted et al, 2006; Minshew et al, 2002). The present study attempted to address this by investigating categorisation abilities in problem-solving; a set of specific processes which also have implications for cognitive style and general information processing. Two tasks were devised based on the Twenty Questions verbal inquiry paradigm (Mosher & Hornsby, 1966) to examine factors affecting category use in ASD problem-solving. A group of high-functioning children with autism (n =14) and a group of typically-developing controls (n = 14) were tested on both tasks. Participants were required to ask questions to identify a series of items selected by the experimenter from a closed set, in a task structure similar to guessing games such as Guess Who?. Effects of task content, cognitive flexibility, memory and language were analysed. ASD participants were significantly impaired in successful category use on trials containing primarily conceptual content, but demonstrated unimpaired levels of performance on trials where perceptual content dictated grouping. However, ASD participants also consistently asked questions of lower quality than controls, indicating some form of persistent local focus on the level of sets. In addition, verbal IQ was found to specifically support performance in ASD participants but not controls. The implications of these findings are discussed with reference to understanding of the autistic cognitive style, methodological issues in the matching of ASD individuals, and the possible neural basis of cognitive abnormalities in autistic spectrum disorders.
Autism and associated disorders on the autistic spectrum (ASDs) are characterised by difficulties in social functioning and communication alongside the presence of restricted interests and repetitive behaviours (Wing & Gould, 1979; APA, 1993). This profile is often associated with cognitive impairment at a level of learning disability, with up to 74% of individuals on the spectrum estimated as falling within the “low-functioning” range, i.e. with a full-scale IQ estimate of 70 or below (Chakrabarti & Fombonne, 2005). Accordingly, much research has focussed on the specific cognitive profile of individuals with autism (Dawson, 1996; Joseph, Tager-Flusberg, Lord, 2002; Mayes & Calhoun, 2003), with the aim of pinpointing core cognitive deficits which may underlie the range of functioning and behaviour seen across the spectrum.

Most contemporary theories of autism have conceptualised this profile in terms of a generalised deficit in core cognitive processing (Baron-Cohen, 1995; Frith & Happé, 1994; Russell, 1997). For some theorists this deficit has been primarily social; Baron-Cohen (1995) and Leslie (1987) for example have argued that a deficit in theory of mind, or the ability to represent the mental states of others, is at the centre of difficulties seen in ASD. This, they suggested, underpins the typical problems with understanding social situations, learning language and engaging in pretend or imaginary play that many children with autism have (Baron-Cohen, 1995). This approach is supported by evidence of impaired ASD performance on a range of tasks testing false belief understanding (Baron-Cohen et al, 1985; Baron-Cohen, 1989) and interpretation of mental states (Baron-Cohen et al, 2001). Another account which centred on the social dimension of ASD was supplied by Hobson (1993), who argued that early deficits in intersubjective communication between parents and children were the driving factor in the development of ASD pathology. Many children who go on to acquire a diagnosis exhibit reduced eye-contact since birth (Zwaigenbaum et al, 2005); for Hobson, this delays or interrupts the usual processes of shared attention which are required for successful parent-child communication and subsequent socio-emotional development (Hobson, 1993).

Other theories have sought to explain the symptomatology of autism in terms of irregular cognitive processing in both social and non-social domains. For example, Frith & Happé’s weak central coherence hypothesis (Frith, 1989; Frith & Happé, 1994) posited a general deficit in the processing of global, coherent wholes from local features. A number of individuals with autism have enhanced abilities in specific aspects of perceptual processing; for instance, ASD participants are typically faster at completing visual tasks such as the Embedded Figures Test (Shah & Frith, 1983; Joliffe & Baron-Cohen, 1997), have been reported to be less susceptible to visual illusions (Happé, 1996) and often demonstrate pitch-perfect
sensitivity to changing auditory tones (Frith, 1989). Frith & Happé argued that such skills result from a lack of interference from the global or holistic whole, leaving the ASD individual to focus on local details instead (for instance, they may recognise a familiar face by looking at specific features rather than their overall configuration) (Frith, 1989; Langdell, 1978). In basic perceptual terms this may be considered an advantage (indeed, some theorists have questioned why this implies a deficit at all; Mottron et al, 2006), but in other domains the impairment of global processing is proposed to create considerable difficulties, in both social and non-social domains. For example, a strong attention to detail may lead to problems accepting small changes in patterns and routines, while a focus on local features of a sentence may cause errors in the recognition of appropriate contextual cues (Happé, 1997).

Lastly, the executive functioning theory of autism (Ozonoff et al, 1991; Russell, 1997) has proposed that the majority of symptoms seen in ASD can be explained through deficits, or combined cognitive impairments, in aspects of planning, working memory, response inhibition and flexibility (Pennington & Ozonoff, 1996). Using techniques originally developed for assessing cognitive functioning in patients with frontal damage, researchers have reported impaired performance by ASD individuals on problem solving tasks such as the Towers of Hanoi (Ozonoff et al, 1991, Ozonoff & McEvoy, 1994), cognitive flexibility tasks such as the Intradimensional/Extradimensional (ID/ED) task (Hughes et al, 1994, Ozonoff et al, 2004), and various measures of spatial working memory (Goldberg et al, 2005; Verté et al, 2006). The central hypothesis of executive functioning theory posits that problems within these core cognitive processes lead to the range of social and non-social difficulties in ASD; for example, Russell (1997), has argued that apparent difficulties in theory of mind, such as a failure to anticipate another’s knowledge state, could result from a deficit in the ability to inhibit the state of one’s own knowledge (Russell, 1997). As with central coherence research, this approach has been developed based on findings from a battery of neuropsychological measures (Ozonoff et al, 2004), providing a picture of generalised executive dysfunction in autism.

However, while general deficit accounts offer an important guide to some key elements of ASD cognition, their utility is often limited by the sheer range of symptoms seen in individuals on the autistic spectrum. The mainly social-cognitive theories have been criticised for their failure to account for the non-social aspects of autism, especially concerning characteristics like repetitive behaviours or sensory sensitivities (Frith & Happé, 1994; Russell, 1997). For example, it is not clear why a child with theory-of-mind difficulties would also demonstrate enhanced visual discrimination. At the same time, the hypotheses of weak central coherence and executive dysfunction have not always been successful in fully accounting for the specific social-cognitive difficulties seen in ASD (Baron-Cohen, 2002); many ASD
children seem to have a genuine lack of interest in the social realm which is not easily or directly accounted for by differences in executive skills or coherent processing.

This spectrum of symptoms is mirrored by an equally heterogeneous spectrum of severity in ASD. Firstly, while difficulties in the three main “triad” areas are required for a diagnosis (social, communicative and restricted/repetitive behaviours; APA, 1993), the severity of those difficulties and the extent to which they persist with age may vary considerably (Beglinger & Smith, 2001). Furthermore, IQ undoubtedly plays an important but as yet unclear role in the expression of symptoms in autism (Joseph, Tager-Flusberg & Lord, 2002). As outlined above, a large proportion of diagnosed individuals will score below the normal range on standardised IQ tests (Chakrabati & Fombonne, 2005), but a sizeable high-functioning minority does exist, and some individuals (particularly those with Asperger Syndrome) may record extremely high scores on IQ tests (i.e. Full-scale IQ >135; Mayes & Calhoun, 2003). Symptom severity should not by any means be taken as equivalent to IQ score, but it will often vary roughly in accordance with this IQ range, with higher-functioning individuals tending to demonstrate much less profound symptomatology (e.g. subtle executive differences, first- and second-order reasoning about mental states), while at the same time still demonstrating some solidly ASD behaviours (e.g. attention to detail, obsessional interests) (Joseph et al, 2002)\(^1\). This variation in functioning and severity across the spectrum poses just as much of a difficulty for generalised theories as the variation in social and non-social difficulties: not only is there a large range of symptoms to account for, there is also a large variance in their expression across ASD individuals.

Even within the generalised accounts, however, there has been further difficulty in maintaining the coherence of a singular hypothesis (Happé, Ronald & Plomin, 2006). For example, in central coherence research a recent study (Lopez et al, 2008) has indicated that difficulties in global processing on perceptual tasks may be unrelated or even inversely related to other types of weak central coherence in autism. Specifically, the study in question reported that some ASD individuals struggled on tests requiring understanding of coherent semantic concepts, while others had difficulty on purely visual central coherence tasks. Very few ASD participants exhibited difficulties in both domains. The authors interpreted this as evidence of a dissociation in central coherence difficulties between what might be termed conceptual coherence on the one hand, and perceptual coherence on the other (Lopez et al, 2008).

\(^1\) Another way of interpreting this point is that varied symptom severity will often determine to what extent an individual with autism can comprehend and complete standard IQ measures. It is a matter of current debate as to whether measures such as the standardised Wechsler tests (Wechsler, 1991) provide the best index of general cognitive functioning in autism, given the irregular nature of the autistic cognitive profile (Dawson et al, 2007).
In this sense, the idea of a general deficit in central coherence would itself seem to lack coherence, and if there are enduring problems with coherent processing in ASD, they would seem to exist as specific difficulties in selective domains.

Similarly, in executive functioning research subtle deficits in ASD planning and spatial working memory have been demonstrated alongside relatively intact inhibition abilities (Russell, Hala & Hill, 2003; Hill, 2004; Goldberg et al, 2005), suggesting the presence of selective strengths and weaknesses even within the executive domain. The umbrella of executive processes no longer provides an overall “executive dysfunction” hypothesis, but a varied collection of cognitive peaks and troughs (as is also seen in ASD performance on standardised IQ tasks; Mayes & Calhoun, 2008). The consequence of this is that the general theories have become more specific in their outline of the ASD cognitive profile, but as a result they have also become fractionated from an original whole.

For researchers this has led to a refocus on particular processes in ASD which, though specific, still play core roles in cognitive functioning and symptomatology. Refocussing on specific processes, in the tradition of some of the earliest cognitive research on ASD (e.g. Hermelin & O’Connor, 1970) can be seen in recent work done on visual discrimination (Plaisted et al, 1998), free recall (Bowler et al, 2004) and sequencing (McGonigle-Chalmers et al, 2008). Plaisted & colleagues have conducted a series of experiments on the nature of low-level visual processes in autism (Plaisted et al, 1998, 1999). By examining discrimination and search abilities for very basic perceptual stimuli over multiple training and testing conditions, their studies have demonstrated specific roles for attentional processes and selective biases in ASD visual cognition (O’Riordan & Plaisted, 2001; Plaisted et al 2003). Measurement of reaction times (O’Riordan & Plaisted, 2001) and matching abilities at a range of very short presentation times (Plaisted et al, 2006) has also provided a picture of the time-course of specific processing abilities in this area. The thorough nature of this programme of research has done much to highlight the continuities and discontinuities in basic sensory processing for ASD individuals; differences which cannot be ignored in any attempt at a generalised account of higher-order cognitive processes in autism.

The work of Bowler and colleagues (Bowler et al, 1997; 2004; Gaigg et al, 2008) has investigated a selection of relatively higher-order cognitive processes than Plaisted et al, but with a similarly focussed and detailed approach. Primarily studying factors supporting recall in ASD, Bowler et al have used multiple variations on the classic free recall paradigm to demonstrate important effects of contextual information (Bowler et al, 2004; Bowler, Gaigg & Gardiner, 2008a), emotional arousal (Gaigg & Bowler, 2008), idiosyncratic ordering principles (Bowler, Gaigg & Gardiner, 2008b) and irregular semantic encoding (Gaigg et al, 2008) on memory functioning for individuals with autism. In the typical free recall
paradigm participants must recall a list of words or items to an experimenter following a learning or familiarisation phase, and often semantic relations between items play an important role in allowing individuals to group their response into chunks of related information (Tulving, 1962). By systematically varying the content of what must be recalled and the level of contextual cues available to participants, Bowler and colleagues have demonstrated subtle irregularities in the organisation and utilisation of semantic concepts by ASD individuals, leading to impaired performance on memory tasks (Bowler et al., 1997; Bowler et al., 2008a). In this sense, while similar methods have certainly been used before in ASD memory research (Boucher & Warrington, 1976, Tager-Flusberg, 1991), the work of Bowler et al. has been especially notable for hinting at what might underlie differing memory performance in autism.

Processes related to memory have also been investigated by McGonigle-Chalmers and colleagues in a series of studies on sequencing (McGonigle & Chalmers, 2002; McGonigle-Chalmers et al, 2008; McGonigle-Chalmers & Alderson-Day, forthcoming). Utilising techniques developed originally for primate research (McGonigle, Chalmers & Dickinson, 2003), McGonigle-Chalmers et al have examined the respective roles of working memory and planning in ASD sequencing via the measurement of performance on simple touch-screen computer games. The standard task requires participants to produce an unbroken, unrepeated sequence of touches when faced with an array of basic items. Variation in the specific conditions of this format has demonstrated specific effects of cognitive load, size ordering (McGonigle-Chalmers et al, 2008) and item grouping (McGonigle-Chalmers & Alderson-Day, forthcoming) in the sequencing of individuals with autism. As with the work of Bowler et al., this series of studies has worked towards showing why certain processes fail in ASD sequencing; in this case, this has been achieved through the simultaneous measure of multiple outcomes relating to sequence length, strategy, speed and error (McGonigle & Chalmers, 2002).

Typically what these approaches share is the use of a single experimental task, with multiple conditions and measures of performance, to examine a specific ability and the associated processes which support it. This broadly stands in contrast to much of the work that has supported the generalised theories, in which batteries of tasks with singular outcome measures have been used to provide general performance scores (e.g. Ozonoff et al, 2004). The advantage of a more focussed approach is that it allows for greater

---

2 The Wisconsin Card Sorting Task (Heaton, 1981), which has been widely used in executive functioning research (Hill, 2004; Kenworthy et al, 2008), might be considered the exception to this rule in that it also provides multiple measures of a singular process (namely, card sorting). However, arguably the general use of the WCST has been to include it in multiple test batteries and only extract one or two of the possible outcomes (usually measures of perseverative error).
accuracy in establishing the conditions under which specific cognitive processes default (McGonigle-Chalmers et al, 2008); rather than simply knowing whether a group tended to pass or fail a group of tasks, one can point to where errors occurred, where strategies diverged, and, to a certain extent, why this happened.

What is gained by the accuracy of such approaches, however, is arguably matched by the loss of an overall picture of ASD. The undoubted advantage of ambitious, general theories of autism is that they attempt to make sense of the disorder as a whole, aiming at an explanation of why such a strange collection of symptoms and severity exists (Klin, 2009). In the face of fractionation, some of the generalised theories have attempted to retain this “big picture” by beginning to move towards a more flexible account of processing styles and biases in ASD, rather than talking about a central deficit in processing (Happé, 1999). For example, the weak central coherence account has been adapted into an account about bias away from global processing and towards local processing (rather than a deficit in global processing itself; Happé & Frith, 2006). Baron-Cohen’s theories, meanwhile, have been expanded to encompass “systemising”, a hypothesised drive to group and understand information in terms of rigid, ordered systems (Baron-Cohen, 2002; 2006). Talking about an autistic cognitive style or bias retains generality and potentially covers a wider range of symptoms than theories about modular deficits. Furthermore, it arguably gets much closer to providing an account of day-to-day ASD functioning, in that it also covers tendencies and preferences in everyday behaviour (rather than simply explaining how ASD individuals perform on arbitrary and artificial neuropsychological tasks; Kenworthy et al, 2008).

The challenge, though, is to somehow measure styles or biases with the same level of accuracy as one can for specific cognitive processes; to combine the bigger picture with empirical data relating to actual cognitive functioning. One way of doing this is to try to operationalise the workings of a cognitive style, but this can be very difficult to achieve, especially if the definition of that style itself is possibly too general to explicitly measure (see Falter et al, 2007; 2008; and Knickmeyer et al, 2008, for an interesting exchange of views on how best to measure “systemising”). Alternatively, one can attempt to measure specific cognitive skills which could in some way provide an example of the autistic style at work; a snapshot of cognitive functioning in which one could reasonably expect a cognitive style or bias to be expressed. The aim of the present study is to do just that: an examination of the autistic cognitive style via the measurement of a specific process; namely the process of categorisation, and associated factors which support its functioning in ASD.
Categorisation: a window on the autistic cognitive style?

Categorisation is one example of a specific cognitive process with real relevance to examining potential style or biases in information processing. The ability to categorise exemplars into sets, based on perceptual or conceptual similarities with other set members or possible prototypes, is a key skill essential to the efficient organisation of new information. The intuitive and automatic use of sets to organise exemplars saves cognitive effort (allowing an individual to focus on truly novel stimuli or situations) and can be used to guide future behaviour (Rosch, 1978; Quinn & Eimas, 1987). If the stimulus is familiar, or similar to previously seen stimuli, it can be processed quicker and responded to with previously used behaviours. If a stimulus is new it must be processed for longer and in detail, and appropriate responses must be planned and executed. In addition, consistent and automatic use of sets is essential to other core components of cognitive functioning; it allows for effective abstract reasoning (via set inclusion, exclusion, conjunct membership), the deciphering of communicative symbols, and organised search of a complex set (McGonigle-Chalmers & Alderson-Day, forthcoming).

The suggestion that some form of abstract processing or concept manipulation is impaired in autism is not new, with various theorists suggesting a failure of “abstract attitude” (Goldstein, 1939), failure to extract meaning from new information (Hermelin & O’Connor, 1970) or a fall-off in abstract reasoning with level of intellectual disability (Rutter, 1978). However, outside of accounts of language in autism, a specific emphasis on categorisation abilities is relatively new, with a range of recent studies proposing specific abnormalities in the category processing of individuals with ASD (Bowler et al, 1997; Klinger & Dawson, 2001; Minshew et al, 2002). One reason for this interest is the anecdotal evidence that many individuals with ASD have difficulty generalising across specific objects or events (Klinger & Dawson, 1995), suggesting a possible problem with grouping similar things together. This can also be seen to a certain extent in the over-literal understanding of language that many ASD individuals demonstrate (Mitchell et al, 1997), and the desire for routiniz ed sameness in everyday objects and situations (Kanner & Eisenberg, 1956). In this sense, it could be that problems with categorisation relate to and mediate a range of characteristic tendencies and abnormalities in ASD. In addition, further examination of categorisation not only provides a snapshot of a specific set of abilities in ASD, but an opportunity to look more generally at how individuals with autism go about organising information; how the autistic cognitive style, if one exists, is expressed through the grouping of particulars into sets.

The “new look” at categorisation in autism has examined this issue at various levels of processing, from basic category learning, to performance on explicit categorisation tasks, to spontaneous use of categories in goal-directed cognition. The following review will consider the evidence of ASD processing
abnormalities of each of these domains in turn, with a view in particular to their potential implications for an autistic cognitive style. Firstly, category learning in neurotypical infants and ASD individuals will be discussed; that is, the process by which new groupings are learned based on exposure to set exemplars. It will be suggested that basic category learning processes are largely intact in autism. Secondly, findings from sorting and matching-to-sample paradigms will be evaluated for evidence of impairments in categorisation performance in ASD; here it will be argued that subtle irregularities in processing arise in complex categorisation, possibly indicating qualitative or stylistic differences in ASD category organisation. Lastly, the implications of such differences will be examined for goal-directed cognitive tasks. It will be argued that spontaneous category use, that is, the ability to generate one’s own categories to organise cognitive operations, may be differentially impaired in ASD under specific conditions, leading to the expression of abnormal styles or biases in information processing. To further investigate this, paradigms from reasoning research will be put forward and developed as an effective tool in examining real-world utilisation of categories and the potential expression of a cognitive style in individuals with ASD.

Category learning in neurotypical infants

Studies on typically developing infants have indicated the early presence of a range of abilities related to category learning. Although tests of category recognition are difficult to conduct with pre-verbal children, a number of studies have attempted to indirectly measure categorisation in infants by recording habituation responses to different types of visual stimuli (Bombar & Siqueland, 1983; Younger & Cohen, 1983; Younger & Gottlieb, 1988; Younger & Fearing, 1999) or, in some cases, sequential touching of groups of objects (Starkey, 1981). Most notably, Younger and colleagues have claimed to demonstrate categorical habituation to different geometric shapes in 3-7 month-olds (Younger & Gottlieb, 1988); habituation to artificial animal images with correlating features in 7 month-olds (Younger & Cohen, 1986); and, in 10 month-olds, habituated responses to previously unseen category members (Younger & Cohen, 1983) and categorical discrimination of dog, cat and horse natural kinds (Younger & Fearing, 1999). The extent to which this constitutes explicit and consistent category learning in young infants has been questioned by some theorists, most notably Sugarman (1981)\(^3\). Nevertheless, the general picture

---

\(^3\) Sugarman (1981), for instance, has suggested that the ability to discriminate two separate categories simultaneously does not develop until 24 months, while studies of explicit sorting of objects have only shown this to be consistently apparent in 18-24 month olds (Ricciuti, 1965; Sugarman, 1982). Moreover, studies which have utilised object examination methods instead (in which the amount of time an infant is either handling or looking at
from infant research does suggest early tendencies to implicitly recognise similar individual exemplars and, by the end of the first year at least, a tendency to recognise new exemplars in a categorical manner.

In the following years the development of category learning is thought to extend beyond basic categories to further classification of specific sub-classes (*subordinate* categories), and understanding of abstract groupings of basic categories themselves (*superordinate* categories). Cohen & Younger (1983) have reported response to superordinate labels such as toys in 18-24 month olds, although the ability to sort according to such sub- or superordinate categories would not appear to develop until later (Mervis & Crisafi, 1982). It is generally proposed that with the development of such distinctions comes an understanding of simple hierarchy in categorisation, underpinning the move from predominantly perceptual groupings of objects to more abstract, conceptual classification (Quinn & Eimas, 1987). Such a process is presumably assisted and facilitated by the development of language in the second year of life.

At the age of 3 new category learning may still be based primarily on correlations in perceptual characteristics (see Younger & Mekos, 1992, for a discussion of differing strategies in new category learning), but between the ages of 4 and 7 years a taxonomic understanding of flexible and context-independent groupings of objects would seem to begin to develop (Lucariello et al, 1992).

**Category learning in autism**

The types of studies which have been conducted with young neurotypical infants are, for various reasons, very difficult to conduct with children with autism. Diagnosis for core autism, as opposed to other pervasive developmental disorders, is difficult to reliably carry out prior to 2-3 years of age (Lord, 1995; Baird et al, 2001). Moreover paradigms which rely on measures of visual attention in early infancy or require verbal responses in toddler and pre-school children are highly impractical when dealing with young children with autism, many of whom will not engage in shared attention and may not even speak.

As such, a direct comparison of ASD category learning in early infancy is nigh on impossible. However, studies on older children and adults with autism have managed to examine perceptual-based category

---

an object is measured) have highlighted limiting factors on early category development. For example, 10 month-olds would still appear to strongly depend on perceptual similarities in their category recognition (Oakes et al, 1997; Horst et al, 2005), while the extent to which different types of images are learnt as being part of the same category would appear to also depend on previous exposure to other exemplars in the environment around them (Oakes & Spalding, 1997; Kovack-Lesh et al, 2008), suggesting that the developmental rate and course of categorisation processes may vary for some young children.
learning when dealing with novel sets of exemplars. At a very basic level it has previously been suggested that individuals with autism exhibit reduced perceptual similarity (Plaisted et al, 1998), that is, a tendency to notice the dissimilarities between two objects over their similarities. Evidence for this is derived from clinical descriptions of strong attention to detail in autism (Frith, 2003) and enhanced performance on visual discrimination and embedded figures tasks (Shah & Frith, 1983, O’Riordan & Plaisted, 2001). If this was the case it would have clear implications for the ability to form categories, as perceptual-based grouping would be made considerably harder. Furthermore, abnormalities at such a basic level of processing would plausibly have considerable knock-on effects for higher-order cognitive processes, potentially leading to styles or biases across a whole range of operations.

An alternative hypothesis, put forward by Klinger & Dawson (2001), is that individuals with autism have no problem noticing perceptual similarities, but nevertheless encode individual exemplars in a degraded form. Typically individuals will recognise and respond faster to prototype examples of a previously learnt set even if they had never actually seen the prototype item (a so-called “prototype effect”). It is suggested that on some level exposure to set exemplars leads to the implicit and automatic derivation of a prototype representation, which is then used to guide future classification and recognition (Posner & Keele, 1968; 1970; although see Sherman, 1985). Klinger and Dawson (2001) propose that individuals with autism fail to extract a prototype representation of a category member following exposure to set exemplars, and must instead rely on recognition of explicit, rule-based similarities in grouping.

This claim is based on their own findings of irregular prototype recognition in ASD individuals (Klinger & Dawson, 1995; 2001). Using a task where participants had to identify new members of a set following a phase of familiarisation, the authors measured categorisation abilities for a range of artificial animal stimuli. The stimuli consisted of five different types of imaginary cartoon animals (“mips”, “daks”, “sops”, “pevs” and “tuzs”) that differed in specific perceptual features (e.g. legs, arms, wings, head-shape etc.). Artificial stimuli have been used before in studies of neurotypical infants (e.g. Younger, 1985) and are useful because they allow for the precise variation of perceptual similarities and dissimilarities within a set. In addition, because they are novel stimuli, effects of previous exposure and experience in participants can be controlled for. (Unless the individual has taken part in a lot of categorisation research previously they should hopefully have never before seen a “mip”). In Klinger & Dawson’s study, the features of these animals were varied so that they either formed a prototype animal (i.e. the mean set of features for that animal, such as the average “dak”), or instead could be recognised because they all conformed to a specific rule (e.g. all “sops” have triangle heads; Klinger & Dawson, 2001).
Using this method, the authors found that ASD children exhibited reduced recognition of new prototype cases in comparison to controls; that is, they failed to exhibit a prototype effect when faced with a prototype item. Alongside this, ASD participants exhibited unimpaired identification of animals which could be recognised by a specific rule, suggesting intact categorisation where a prototype wasn’t required. This was still the case even when participants were not told to look for a specific rule. For Klinger & Dawson (2001) these results implied that rule-based category learning may be relatively automatic and unimpaired in ASD, but the representation of that category’s central concept, the prototype, is not created.

However, recent studies of basic category learning in autism do not support either of the above hypotheses put forward by Plaisted et al (1998) or Klinger & Dawson (2001). For example, Bott et al (2006) examined the ability of adults with high-functioning autism to learn novel categories of rectangles varying in length and width. Although ASD adults took longer than ability-matched controls to learn the initial category structure (which in itself was arguably highly arbitrary), their ability to recognise new shapes as category exemplars was no more or less accurate than controls. It should be stressed here that the stimuli used were very basic shapes, much more so than those used in Klinger & Dawson (2001), for example. While it might be suggested that this doesn’t add much to understanding of higher-level categorisation, it does suggest a significant degree of intact processing in low-level perception. Moreover, the participants in question would seem to have had no difficulty in assessing perceptual similarity, contrary to the predictions of Plaisted et al.

Similarly, a study by Soulieres et al (2007) of adolescents and adults with autism found normal sensitivity to categorisation and discrimination of a set of basic ellipse shapes. When asked to categorise different ellipses following a learning phase, ASD individuals demonstrated comparable levels of categorisation ability to controls, exhibiting a similar “classification curve” (Soulieres et al, p. 481) in their discrimination of progressively dissimilar ellipses. Where there was a difference was at the boundary of two categories, where effects of explicit knowledge are usually called upon to assist correct discrimination (Goldstone, 1998). The authors interpreted this as evidence of ASD individuals having a degree of low-level cognitive autonomy in the face of top-down influences like explicit knowledge (Soulieres et al, 2007). Nevertheless, as in the study by Bott et al., ASD individuals would seem in this case to have successfully learnt new categories in a relatively regular manner, implying an intact basis for perceptual learning at the most basic level.

The findings of Bott et al. and Soulieres et al. are consistent with other studies looking at prototype formation, which have also pointed towards fundamentally intact category learning processes. Earlier categorisation studies by Tager-Flusberg (1985a, 1985b) found no differences in ASD and control
participants when identifying prototypes in pictorial and verbal matching-to-sample tasks. Furthermore, subsequent studies by Molesworth and colleagues have reported evidence of intact prototype effects in the recognition memory of children with Asperger’s syndrome (Molesworth et al, 2005) and high-functioning autism (Molesworth et al, 2008). The participants in these studies were mostly older than those studied by Klinger & Dawson, but if their hypothesis was correct, ASD participants of any age should exhibit impaired prototype recognition of some sort. The evidence from the above studies would appear to suggest that prototypes facilitate recognition of learned categories just as well as they would for neurotypical individuals in a range of different ASD individuals.

In this sense, on a basic level ASD category learning would appear to be relatively unimpaired. Despite hypotheses of reduced recognition of similarities, or degraded representation of category examples, individuals on the autistic spectrum learn categories to a similar level of accuracy and sensitivity as controls. Although the methods used are not directly comparable to those used in young infants, they have largely used similar stimuli to the original Younger studies. ASD individuals would appear to respond to new categories (whether they are artificial shapes or made-up animals) in similar ways to controls, suggesting intact learning processes and a relatively regular style or approach to category learning at the simplest levels of cognitive processing. If there are differences in the categorisation process which may reflect an autistic cognitive style, they would not appear to be at the learning stage, and may instead be happening further downstream. To investigate this we need to look at categorisation performance, particularly via the way that explicit category knowledge is demonstrated in experimental tasks.

**Categorisation performance in ASD**

- **Basic and complex sorting**

Card or object sorting tasks provide one of the most basic indexes of categorisation performance. It also provides a prime outlet for the expression of styles or biases in categorisation, as typically participants will be asked to sort a collection of items according to their own preferred criterion, or based on a rule which they need to infer. The Wisconsin Card Sorting Task (WCST; Heaton et al, 1981) is an example of the latter method, and has been used extensively in research with the ASD population. The task requires participants to sort piles of cards (displaying icons varying in colour, form and number) according to a hidden criterion, which must be deduced from experimenter feedback. After 10 trials the criterion changes without warning (a set change) and participants must adapt their sorting principle to continue correctly. Although primarily used as a test of flexibility within executive functioning research (Ozonoff et al, 1991;
Bennetto et al, 1996), it also provides a measure of categorisation abilities between set changes. Success and consistency in categorisation can be assessed through the number of categories completed by a participant, and the number of times a participant successfully sorts cards for at least 3 consecutive trials (referred to by the task authors as the *conceptual level response*, in that it indicates “insight into the correct sorting strategy”; Heaton et al, 1993, p12). As such, difficulties with categorisation on a conceptual level or preferences away from grouping items in a conceptual manner can be directly analysed for indications of an overriding cognitive style.

A number of studies have reported impaired performance by ASD individuals on the WCST (Rumsey, 1985; Prior & Hoffman, 1990; Ozonoff et al, 1991; Szatmari et al, 1990; Bennetto et al, 1996; Pascalvuaca et al, 1998; Ozonoff & Jensen, 1999; Shu et al, 2001; Geurts et al, 2004; Verté et al, 2005; South et al, 2007), although reviews by Hill (2004) and Kenworthy et al (2008) have indicated that this is mostly based on increased numbers of perseverative errors following changes in set. Perseverative errors are instances in which a participant persists with a sorting strategy following a set change even when it is no longer correct. In studies by Rumsey (1985), Szatmari et al (1991), Bennetto et al (1996), Pascalvuaca et al (1998), Shu et al (2001), Lopez et al (2005) and Tsuchiya et al (2005) ASD participants completed significantly fewer categories than controls, possibly indicating a bias away from conceptual categorisation in sorting. However, if a participant demonstrates increased perseveration, they will also inevitably complete fewer categories on the WCST (due to the structure of the task). Importantly, in none of these studies was reduced category completion observed *without* increased perseveration, and six studies have reported no significant deficits in category completion (Ozonoff et al, 1991; Nyden et al, 1999; Liss et al, 2001; Salmond et al, 2001; Ambery et al, 2006; Kaland et al, 2008). Therefore, any interpretation of a possible problem or bias in category completion in ASD individuals cannot be demonstrated consistently and independently of problems with perseveration; it could be that differences in category completion, where present, are simply by-products of perseverative error.

The *conceptual level response* measure on WCST arguably offers a better index of categorisation, in that it only measures within-set runs of correct responses. Thus it is less likely to be affected by perseveration proximal to the hidden criterion changes. Of those that have reported using this variable, only two studies (Pascalvuaca et al, 1998; Shu et al, 2001) have found significant impairments in conceptual level response, while Rumsey (1985) and Kaland et al (2008) reported no impairments in such responses. Combined with the evidence from category completion this would suggest that specific abnormalities in categorisation, such as a bias away from conceptual grouping, have not been demonstrated. Rather,
studies using the WCST can only reliably relate differences in ASD performance to problems with flexibility.

Earlier sorting studies which did not utilise the WCST also demonstrated largely normal categorisation performance in ASD individuals. In studies using object (Ungerer & Sigman, 1987) and semi-cued sorting (Tager-Flusberg, 1985b) similar levels of performance were reported for children with autism compared to mentally-aged matched controls. This pattern of regular and unimpaired sorting abilities was true for categories based on basic perceptual criteria (such as colour or form), but also for sorting of a small group of basic functional groupings (animals, fruits, vehicles and furniture), suggesting the presence of categorisation capability in both concrete and abstract domains (Ungerer & Sigman, 1987).

However, it has been suggested that sorting abilities in autism only appear intact in cases where sorting criteria are simple, or the cognitive load is low (Klinger & Dawson, 2001; Gastgeb et al, 2006). Evidence for this claim is mainly supplied by ASD individuals’ performance on complex categorical sorting tasks. For example, Shulman et al (1995) compared categorisation abilities in a group of children with autism with a group of children with learning disabilities (LD) and a group of typically developing (TD) children. On a free-sorting task based on perceptual similarities, participants with autism performed as well as participants in either control group. However, when free-sorting of more conceptual objects was required (this time according to six possible taxonomic groupings: trees, beds, humans, animals, tools & vehicles), ASD participants performed poorer than either LD or TD children. Similarly, where sorting was deliberately made more complex (via the requirement to sort according to two simultaneous criteria) ASD participants performed worse than typically-developing controls when sorting based on perceptual criteria, and worse than either group again when sorting according to conceptual criteria (ibid.).

Evidence of a possible problem with more abstract and complex categorical sorting is also provided by tasks in which participants are given a choice of criterion by which to sort exemplars. Recently, Ropar & Peebles (2007) have replicated a finding from Kovattana & Kraemer (1974) that ASD participants, when given a choice, prefer to sort objects into categories based on concrete, perceptual similarities rather than conceptual/ semantic similarities. Using a novel sorting task, participants were asked to sort a selection of books into two separate piles. The set was organised so that the books could be sorted into a pile on games and a pile on sports, but in contrast to age and IQ-matched controls, ASD participants tended to sort the books according to concrete similarities in the book covers (for example, the number of balls on the cover illustration), rather than conceptual criteria (such as games vs. sports). This does not necessarily demonstrate an impairment in sorting or problem with categorising according to concepts (indeed, ASD participants could recognise that some books were on sports while others were on games), but it does
indicate a possible bias away from more abstract forms of category processing; a clear example of an autistic style in processing. Taken together with the findings from Tager-Flusberg (1985a), Ungerer & Sigman (1987), and Shulman et al (1995) this would appear to indicate that sorting abilities in autism may be relatively intact, but can be notably affected by (possibly interacting) constraints of complexity and abstraction, biasing participants towards alternative methods of grouping. Specifically, this appears to somehow hinge on the distinction between perceptual and conceptual categories.

- Identification and discrimination tasks

A number of studies have examined categorisation via the classification of exemplars in forced choice or matching-to-sample formats. Typically, following a learning phase, participants are presented with two alternatives from which a category member must be identified. As with the evidence from sorting tasks, the use of such identification paradigms has demonstrated largely intact categorisation abilities in ASD. For example, Tager-Flusberg (1985a) found no significant differences between children with autism and mental age-matched controls on a matching-to-sample task for items grouped according to basic (cars, chairs and dogs) and superordinate (vegetables, fruits, animals, clothing, furniture & vehicles) categories, in contrast to the findings of Shulman et al (1995) and consistent with the patterns of sorting seen in Ungerer & Sigman (1987). More recent studies by Geurts et al (2004) and Verté et al (2005) have also supported this finding; using the categories subtest of the Snijders-Oomen Non-verbal Intelligence Test (SON-R; Snijders et al, 1989), both studies reported reduced scores on ASD participants’ identification of shared categories between groups of cards, but these differences were found to disappear once age and full-scale IQ effects were controlled for.

However, other matching paradigms have been used to demonstrate subtle differences in the ASD profile amidst relatively intact categorisation performance. In contrast to the use of the SON-R, Gastgeb et al (2006) used a reaction-time based paradigm to assess category identification in children, adolescents and adults with autism. In the study, participants were asked to verify category membership of visual stimuli when following auditory prompts. While no differences in verification accuracy were observed, speed of categorisation revealed differences between the two groups. In contrast to typically-developing controls, children with autism exhibited significantly slower reaction times in categorisation, especially for more non-typical class exemplars (an effect also seen to a lesser extent in adolescents and adults with autism).

---

4 This procedure was also used by Klinger & Dawson (2001), but their findings were discussed earlier because of the hypothesised implications for basic category learning in ASD. The following studies, in contrast, have largely
The pattern of results seen in Gastgeb et al (2006) is important, as it suggests a more nuanced picture than any general categorisation impairment. The main finding here is the sensitivity to atypicality: ASD participants consistently demonstrated slower reaction times for more atypical exemplars, relative to controls. This could be explained simply as a speed-of-processing effect (indeed, baseline reaction times in categorisation were also slower for ASD participants), but such an explanation would not account for ASD performance worsening with increasing atypicality. When one considers that overall accuracy was not impaired for ASD participants, this suggests that there is a qualitative difference in the processing of such categories, rather than a general deficit.

In this sense, it could be that the apparent impairments observed by Plaisted et al (1998) and Klinger & Dawson (2001) in category learning can instead be explained by subtle irregularities in the organisation of different category exemplars. Whether or not they actually automatically form prototypes (and the evidence discussed above suggests that they probably do), the ASD participants studied were slower to categorise, slower to recognise atypical exemplars, and benefitted most when with explicit rules were available by which to categorise. This suggests that while ASD participants can accurately identify category exemplars, the process by which they achieve that may be in some way different to neurotypical individuals.

In summary, the evidence from classic categorisation tasks presents a mixed profile of category use in ASD individuals. Primary classification abilities would appear intact: when sorting according to a single criterion, and when sorting based on concrete or perceptual similarities, individuals with autism generally perform as well as controls. In comparison, sorting according to more than one criterion and sorting according to conceptual criteria appear to pose more problems, and evidence from sorting preferences suggests a possible bias away from the use of more abstract categories. Although claims of a prototype formation deficit are not clearly supported, the evidence from studies examining identification of typical and atypical exemplars indicates potentially irregular organisation of categories, or the utilisation of differing processes in achieving categorisation, which plausibly may underlie the problems seen when complex or abstract sorting is required.

---

utilised this method to examine categorisation performance, as opposed to basic learning. (This is demonstrated by their use of familiar categories and taxonomies rather than artificial stimuli).
Categorisation in action: implications of subtle processing deficits for the use of categories in goal-directed behaviour

If irregular category organisation is present in ASD, what would the implications be for cognitive tasks where categorisation must be utilized to achieve a certain goal? Classical categorisation tasks as already discussed offer a straightforward index of primary category processing in ASD, but arguably such paradigms only provide a partial picture of categorisation abilities, in that participants are always prompted in some way to address particular sets when ordering or recognizing new exemplars. In real world categorisation a vast range of different types of information must be categorised automatically and without prompting. A more naturalistic approach is to examine category use in the performance of other cognitive processes towards a particular goal; in essence, categorisation as a strategy in organised information processing. If there are subtle irregularities in how ASD individuals come to process and apply categories and if this is supposed to have direct consequences for an autistic cognitive style, then this should become apparent in the application of categorisation strategies in cognitive tasks.

Paradigms from memory research provide one such window on category use. A wide literature exists on various aspects of memory functioning in autism, concentrating mainly on the role of possible working memory deficits within executive dysfunction (Bennetto et al, 1996; Ozonoff et al, 2004; Goldberg et al, 2005). However, categorisation can also be examined via the use of concepts and categories to organise memory processes.

In ASD various studies have pointed to either deficits or differences in spontaneous grouping and facilitation effects in free-recall tasks. As outlined above, the classic paradigm consists of participants being presented with a list of words or objects and then being and asked to reproduce them from memory after a delay. Studies on normal participants have demonstrated a spontaneous tendency to group recalled objects according to ordering principles (usually semantic in nature) and facilitation effects when participants are asked to recall lists of semantically-related words (Tulving, 1962). Early studies on ASD individuals reported overall impaired performance on free recall tasks as opposed to cued recall (Boucher & Warrington, 1976; Boucher, 1981), while other studies have indicated reduced or idiosyncratic use of semantic clustering in ASD responses (Hermelin & O’Connor, 1970; Tager-Flusberg, 1991; Bowler et al, 2008a). Although research has not consistently demonstrated the former pattern of impaired recall (Prior & Chen, 1976; Minshew & Goldstein, 1993; Mottron et al, 2001; Toichi & Kamio, 2003), studies by Bowler and colleagues have supported the latter finding by the observation of reduced semantic facilitation effects in adults with autism (Bowler et al, 2008b; Gaigg et al, 2008) and Asperger Syndrome (Bowler et al, 1997). Such effects appear to persist even with explicit rehearsal and training (Smith et al,
2007). In each of these studies explicit cued recall has been reported as intact and facilitated by semantic relations alongside free recall abnormalities, suggesting a problem with the organised and unprompted retrieval of relevant relations, rather than their encoding (Bowler et al, 2004).

Drawing a parallel with the evidence from sorting research, this would suggest primarily intact grouping abilities (as evidenced by unimpaired cued recall), but degraded use of grouping to work towards a goal other than categorisation itself (Gaigg et al, 2008). This may also possibly explain why abstract categorisation appeared intact in matching-to-sample tasks (Tager-Flusberg, 1985a) but not in complex free sorting tasks (Shulman et al, 1995), in that matching-to-sample provides a much stronger cue towards the goal of categorisation itself than a request to sort a set of items according to a self-generated rule.

A similar picture has arisen from research examining categorisation in verbal reasoning processes. Just as categorisation can be used to group together items in a string to be recalled, or a visual array to be searched, it may be utilised to methodically search through a set of possible solutions in a problem. Measures tapping these abilities, such as the Twenty Questions Test (Mosher & Hornsby, 1966), or Verbal Absurdities Test (Thorndike et al, 1986) require an understanding of categorical groupings and an ability to use conventional categories in order to reach a solution. For example, in Twenty Questions, an experimenter chooses an item from a fixed array of everyday forms (such as animals, tools, vehicles etc.) and participants must establish which item is chosen via a series of deductive inferences. In order to do this within the minimum number of questions, they must ask questions which efficiently sub-divide the set of possible options into smaller and smaller categories: if category use is in some way impaired or irregularly organised, this will greatly hinder an individual’s ability to find the target.

Using the Twenty Questions task, Minshew et al (1994) examined verbal reasoning abilities in a sample of children and adults with autism compared to typically developing controls. While controls successfully used questions which eliminated categories of at least two items at a time (known as constraint-seeking questions), participants with autism were significantly less likely to use such questions, and tended to rely on serial guessing (or hypothesis-testing questions) as their main strategy. That is, rather than use category groupings to strategically sub-divide the set, ASD participants tended to simply eliminate individual items one by one, despite it being highly inefficient to do so (e.g. “Is it the banana? Is it the sheep? Is it the knife?”…..). Comparing the results with evidence from a previous study of intact ASD performance on basic categorisation tasks such as the WCST (Minshew et al, 1992), the authors interpreted the result as evidence of a “deficiency in the forming of verbal concepts as an essential strategy to solve the problem.” (Minshew et al, 1994, p38). As with the more recent evidence from free recall tasks, the suggested problem here is one of spontaneous category use; ASD individuals would seem
to learn categories and apply them in unimpaired way, but not use them when they must be self-generated to solve a problem.

Further evidence for this potential abnormality in category and concept use was provided by a much larger multi-task study by Minshew, Meyer & Goldstein (2002). In a factorial design comparing ASD and control performance on a range of categorisation and reasoning tasks, Minshew and colleagues replicated the hypothesised dissociation between intact recognition of categories on certain tasks and impaired spontaneous formation of categories on others. Forming one underlying performance factor (referred to as concept formation by the authors), ASD participants asked significantly fewer constraint-seeking questions than controls on the Twenty Question Test, and also performed poorly on the Verbal and Pictorial Absurdities subtests of the Stanford-Binet scale (Thorndike et al, 1986). Forming a second factor (referred to as concept identification), tests of sorting such as the Halstead Category Test (Reitan & Wolfson, 1993) did not demonstrate group differences to the same degree. Moreover, discriminant analysis of the task battery revealed that the Twenty Questions and Verbal Absurdities Tests alone accurately classified 78.4% of participants with autism, suggesting that they were relatively powerful indicators of the main abnormalities in ASD category processing (Minshew et al, 2002).

There was an attempt by the group to replicate their findings in younger children with ASD (Williams, Goldstein & Minshew, 2006) which actually found a lower discriminant value for a larger battery of reasoning tasks. However, as the authors themselves note, group differences in this area may have been obscured by the tasks being too hard for both ASD and control participants (indeed, scores from the WCST and Halstead tests could not be used because too few participants were able to complete them). Importantly, on the Twenty Questions Task ASD participants still asked a noticeably lower percentage of constraint-seeking questions than controls (45.3 vs. 57.5%, significance tests not reported; Williams et al, 2006).

Considering their apparent utility it is perhaps surprising then that measures such as the Twenty Questions Task have not been applied more extensively to ASD populations since Minshew et al’s (2002) study. An abnormality in the use of categories to organise information would certainly have important implications, not only in terms of ASD problem-solving but also in terms of the wider understanding of an autistic cognitive style. The lack of further work is unfortunate, as although the proposed irregularity in concept processing raises interesting questions for the categorisation debate, it also requires further examination. Specifically, while ASD performance on concept formation tasks was more notably impaired, this profile, particularly on the Twenty Questions Task, represents something of a black box in terms of clear interpretation. To unpack this we must consider the various demands that are posed by the task structure.
The Twenty Questions Task: a simple structure with multiple demands

It can be shown that a number of relevant processes are open for examination via the Twenty Questions Task. The standard paradigm consists of an experimenter selecting an item from an array of 42 everyday objects and animals (presented as black and white line drawings in a grid formation), which participants must then identify by asking a series of “yes/no” questions to eliminate possibilities (Mosher & Hornsby, 1966). Participant performance can be assessed via the percentage of trials successfully completed (i.e. with the target found within 20 questions) and the number of questions needed on average to find the target, providing a gross measure of competence and efficiency in participant reasoning. In addition, quality of search strategy can also be assessed by examining the value of individual questions. Effective questions establish constraints which eliminate the largest number of items irrespective of whether the answer is yes or no (e.g. is it an animal?); less effective questions only pick out a small number of items, or test a particular hypothesis (e.g. is it a goldfish?).

As discussed above, this is essentially a process of categorisation; to succeed the participant must recognise the groupings which are most effectively going to subdivide the set each time, moving from the general through to the particular. But the range of applications for the test stretch beyond that; though it was first used by Hornsby and colleagues to simply assess inquiry strategies in school-age children (Mosher & Hornsby, 1966), the test was later utilised in a range of studies to examine effects of categorisation strategy (Laughlin et al, 1969), modelling (Denney, 1975), structural analogy (Siegler, 1977) and cognitive style (Howie, 1975) on reasoning in children and adolescents. Latterly such tests have been used to investigate executive dysfunction in patients with frontal damage (Klouda & Cooper, 1990; Upton & Thompson, 1999) and chronic alcoholism (Laine & Butters, 1982), and today form part of standardised executive functioning test batteries such as the Delis Kaplan EF System (DKEFS; Delis, Kaplan & Kramer, 2001).

As such there are a range of factors to consider when interpreting irregular performance on the Twenty Question Task. Firstly, as with the evidence from sorting tasks, it is important to ask what kind of categorisation is being talked about. Recent research on working memory in ASD has indicated intact spontaneous grouping of icons in supraspan sequencing by ASD children: McGonigle-Chalmers & Alderson-Day (forthcoming) have reported adoption of colour- and shape-based grouping to support working memory in the formation of error-free sequences. It may be that perceptual-based groupings, if available, are easier to utilise in goal-directed cognitive operations for ASD individuals than more abstract, or hierarchically-related groupings (this would also be consistent with the sorting preferences observed by Ropar & Peebles, 2007). Furthermore, many individuals on the spectrum would seem to
actively categorise the information around them in their day-to-day behaviour; whether it is naming different types of cars, labelling animals or even examining washing machines (Baron-Cohen, 2002; Frith, 2003). Such behaviours are often strongly linked to highly specialised interests held by ASD individuals, but the point remains - if there is a problem with spontaneous, goal-directed category use, this cannot be the case across the board in ASD. Rather, the ability to utilise categories to work towards a goal must be mediated in some way by the nature of the information available; is it perceptually salient? Does it require understanding of abstract hierarchies?

Secondly, the task in its standard form places considerable memory demands on participants. Faced with an unchanging array of items, participants must ask a sequence of questions, mentally eliminating possibilities as they go while remembering what has already been asked. Considering that working memory is a central function which may be differentially impaired in ASD (Bennetto et al, 1996; Kenworthy et al, 2008), and that prospective working memory especially may be a problem for ASD individuals (McGonigle-Chalmers et al, 2008), it may be that apparent impairments in generating categories stem more from a difficulty with remembering what has been eliminated and what can be asked next, rather than a problem or abnormality within concept formation per se.

Finally, the task is highly verbal; verbal instructions are given, participants are expected to ask questions to gain information, and they must integrate verbal responses into a plan of action and elimination (Siegler, 1977). Given that delayed language development and poor verbal IQ are diagnostic factors and typical characteristics respectively of autistic disorder (APA, 1994; WHO, 1993), any task that relies so much on verbal mediation would be expected to pose a significant challenge for individuals with autism (in contrast to directed card sorting or paradigms where experimenters ask the questions).

In this sense, tests such as Twenty Questions would seem to offer a powerful paradigm for assessing category use in ASD, but need to be adapted in order to answer a range of questions. Specifically, effects of stimulus content, working memory and language ability must be controlled for, in order to a) properly elucidate the nature of abnormal category use in ASD, and b) explain why this occurs when basic category learning and performance would seem to be relatively spared. The remainder of this project will seek to outline and test a method which combines these possible factors in a modification of the Twenty Questions paradigm.
“Guess Who?” and “Guess What?”: Adapting Twenty Questions to examine categorisation processes.

There are a range of different simple reasoning games that mimic the essential structure of Twenty Questions (Mosher & Hornsby, 1966; Siegler, 1977). One example is the popular children’s board game Guess Who?, an inquiry-based game with a fixed set of possibilities in which participants must search for a selected individual rather than a particular item. Like the classic Twenty Questions paradigm, participants have to try to find the target within as few questions as possible (usually from a set of 24 possibilities) and must rely on simple grouping strategies to sort and subdivide the set (for example participants may ask “Do they have a moustache?” or “Do they have white hair?”). However, unlike Twenty Questions, set items in Guess Who are presented on a special board of hinged frames which can be placed in an upright or down position (see figure 1). This means that participants can physically reject eliminated possibilities following opponent feedback, rather having to hold in memory the list of previous questions and answers.

Figure 1: The Guess Who Board with remaining items (in yellow) and eliminated (in red). Once an item is eliminated it can be placed in a horizontal position.
In addition, the items used in a Guess Who set can be easily replaced and customised, allowing for comparison of task performance across different stimulus sets with varying content. This opens up the opportunity to utilise sets of items which can be designed to control for varying levels of similarity and grouping between individuals, in a similar way to the use of artificial animal sets in previous categorisation tasks (Younger, 1985, Klinger & Dawson, 2001). Lastly, the game is fun; although Twenty Questions is not a typical executive functioning task, recently a number of tasks in the field have been criticised for lacking ecological validity and being overly artificial in their measurement of participant abilities (Kenworthy et al, 2008). The game structure of Guess Who, in contrast, provides immediate feedback for players and is designed to entertain. This makes it ideal for examining the different factors affecting category and concept use in the reasoning of ASD children.

- **Varying task content: the role of conceptual vs. perceptual stimuli**

The first question to address concerns content. Although Minshew et al (2002) found impaired performance in ASD individuals on the Twenty Questions task, the evidence from categorisation performance tasks suggests that the level of abstractness and complexity in content may play an important role (Shulman et al, 1995; Ropar & Peebles, 2007). Specifically, it may be that the stimuli used by Minshew et al (42 line drawings of common animals, plants and objects) require an understanding of abstract, taxonomic hierarchies, and that it is this which is causing impairments in performance rather than a problem with category use. Such hierarchies are often conventionally assumed by inquiry games, but their structure and the concepts that they consist of are shared meanings (Rosch, 1978). Such concepts are negotiated, often via communication with others, and often based on family resemblances, rather than essential definitions (Wittgenstein, 1956).

In this sense, it is not at all clear that a child with autism (who may have engaged in very little shared meaning with others) will immediately recognise an array of items as a collection of living vs. non-living things, animals vs. plants, mammals vs. non-mammals in a structurally ordered way. Such a shared “concept space” might still be something very alien to them (Hobson, 2009). Instead, it may be that ASD individuals can and will solve the Twenty Questions task effectively if faced with a different set, one with no conventional hierarchy but available properties for categorisation. A set, essentially, which can be searched on purely concrete or perceptual criteria, rather than demanding a certain level of conventional, conceptual understanding about the items involved.

To explore this, two Twenty Questions tests were devised to be used on the Guess Who board. Firstly, in order to replicate Minshew et al’s previous test, a typical set was devised of 24 common items including
animals, plants, vehicles and kitchen implements. Available categories were roughly similar to the original Mosher & Hornsby (1966) 42-item set. The stimuli of the task, while varying in a range of perceptual and more abstract criteria, could most effectively be searched based on conceptual groupings, such as plants, animals, mammals, etc. In essence, this provided a standard from of the Twenty Questions task; if Minshew et al. were correct in their proposal of a difficulty in spontaneous category use, it should definitely be expressed in this setting.

Secondly, a novel set was devised, to examine category use where no understanding of abstract properties or their relations was required. The set could be categorised, but in a way entirely up to the participant, based on the information in the items presented. The set devised consisted of 24 unique robot characters varying in name, colour, shape and individual features. Like the artificial animal stimuli used by Younger et al., this was a set which no participant had seen before. In addition, almost all the variation between items was perceptual in nature; participants could ask about robots with 4 legs, or 3 eyes, or square heads, for example. Although Younger’s animals or similar stimuli could have been used, robots were chosen in this case to provide a template which was readily modifiable (i.e. robots can come in all shapes and sizes, with any number of features) and potentially appealing to participants with autism, given the mechanical nature of the stimuli (Baron-Cohen, 2002). If ASD participants had a general problem with spontaneous category use, this would be expected to be present not only in the standard form of the task, but also in the novel, robot-based form of the set. However, if their problem was not with category use, but with the use of conceptual items and their groupings, then performance on this task would be relatively unimpaired; the ASD performance would dissociate across the standard and novel forms of the task.

Based then on Minshew et al.’s previous findings, it was hypothesised that ASD participants would perform worse than controls on the standard Twenty Questions set. However, it was also predicted that ASD participants would not demonstrate impaired performance on the novel Robot set, or if they did, this would be to a lesser degree than on the standard Twenty Questions set.

- Varying task demands: the role of memory

The second question to address concerns memory load. As explained above, the classic Twenty Questions task is not only a test of categorisation and reasoning, but also a test of working memory. Based on the evidence from executive functioning research, a plausible alternative hypothesis is that the verbal working memory demands of Twenty Questions pose the bigger problem for ASD individuals, rather than a general problem with category use. The use of the Guess Who board structure allows for physical elimination of possibilities, so potential memory problems in participants should be eliminated as a
possible confounding factor in this case. To test this possibility, “bonus rounds” were built into the game where participants were asked to search the set without pushing any of the cards down following their elimination. It was predicted that ASD participants would be relatively more impaired than controls where they were not allowed to physically eliminate possibilities from the set.

- Varying content within-tasks: the role of flexibility

Thirdly, a within-set change was built into the design, to examine flexibility. As shown by studies using the WCST (Pascalvuaca et al, 1998; Shu et al, 2001; Lopez et al, 2005) ASD individuals often demonstrate perseverative errors when the category set is switched, suggesting a level of inflexibility in their category use. Although not directly related to their use of the Twenty Questions task, Minshew et al (2002) have argued that only flexible application of categories represents genuine concept formation, as it shows the ability to separate the abstract concept from the concrete properties of the available set. To explore this issue, both the standard and novel Twenty Questions sets were adjusted to include slightly different items following the first three trials. The additional items were chosen because they notably changed the effectiveness of particular category questions, giving participants a reason to change their questioning strategy (for example, whereas only 6 out of 24 items were animals in the first standard set, this was then raised to 12 out of 24, making “Is it an animal?” a more effective opening question). Based on the previous evidence of cognitive inflexibility in individuals with autism, it was hypothesised that ASD participants would be less sensitive than controls to the set change, and would persist with previously used questions despite their change in utility.

- Supporting performance: the influence of verbal IQ and age

Lastly, it was predicted that language ability and age would significantly influence task performance. As a verbal inquiry game, the Twenty Questions task probably uses more language than most standard experimental tasks for children with autism. Furthermore, success on the task requires the generation of effective verbal labels for sets. While the Method below outlines various procedures to try and minimise the amount of language required, linguistic factors could not be eliminated without constraining the types of questions participants could ask and drastically altering the nature of the task. As such it was hypothesised that verbal IQ would be positively related to task success. Based on the original Twenty Questions studies in typically-developing children (Mosher & Hornsby, 1966; Olver & Hornsby, 1966), it was also predicted that older children would perform better on all forms of the task. In previous research, 6- and 8-year old participants have been found to take longer and ask questions of poorer quality than 11-year old participants (Mosher & Horsby, 1966) while older adolescents have also been reported to
ask more efficient questions than younger adolescents on the standard Twenty Questions Task (Drumm & Jackson, 1996).

Finding the answers: the use of multiple measures to assess performance

A final note concerns the measurement of performance on the Twenty Questions task. As discussed above, an important development in autism research over the past 15-20 years has been the use of multiple outcome measures on a single or small range of cognitive tasks (Plaisted et al, 1998; McGonigle-Chalmers et al, 2008). The structure of the Twenty Questions task allows for a similar level of detail in its analysis. Primarily, task success has been assessed via the percentage of trials completed on the task, and the mean number of questions used per trial (Mosher & Hornsby, 1966; Minshew et al, 1994). To this can be added question quality, an index of the proportion of items eliminated per question which has also been used under previous versions of the task (Delis et al, 2001). If ASD task performance was generally impaired, or categories were being used abnormally used, then it would be primarily evident on this collection of measures.

In addition, the more recent applications of the task have come to focus on the different question types used by participants (Minshew et al, 2002; Williams et al, 2006). Some of these have already been outlined; constraint-seeking questions refer to questions which eliminate at least two or more items within the set, while hypothesis-testing questions eliminate singular items via a direct reference. A third main question type; pseudo-constraint questions, also exists: these are questions which do not directly refer to an item, and have a similar grammatical structure to constraint-seeking questions, but nevertheless only eliminate one item (Mosher & Hornsby, 1966). (For example, rather than ask “Is it the umbrella?” one might ask “Do you put it up when it rains?”). Analysis of question types offers a window on the act of categorisation itself on the Twenty Questions task; if performance is impaired due to abnormalities within category use, it will be evident in the use of differing question types.

Finally, the content of participant questions can also be assessed. Although not often recorded in previous Twenty Questions research, the target of questions can be very informative in illustrating qualitative differences in participant preference and strategy (in a similar way to the analysis of sorting preferences in previous categorisation research; Ropar & Peebles, 2007). The two-task structure of the present study, comparing largely conceptual content on the one hand with perceptual content on the other, will most likely dictate question content in each task (See the Method section below for an exact outline of the differing category codings possible). Nevertheless, analysis of question content within each task offers
important information about the categories participants choose to utilise, and it is here that aspects of
cognitive style would be expected to be most evident. If there was an aversion to more abstract questions,
or an over-reliance on perceptual characteristics in ASD participants, then it would plausibly be present in
their questioning strategies on the two tasks.

In summary, the performance of ASD performance across the two tasks, under varying task demands, will
be assessed in three main areas: overall task success, the use of different question types, and variations
in question content.
METHOD

Participants
A mixed clinical group of 14 ASD participants (12 male, 2 female) were recruited from schools in the Lothian area based on previous study participation. All had a diagnosis of either autism (n = 13) or Asperger syndrome (n = 1) in accordance with ICD-10 research diagnostic criteria (WHO, 1993). All participants were diagnosed by professionals in local child & adolescent mental health services using the Autism Diagnostic Interview – Revised (ADI-R; Lord, Rutter & Le Couteur, 1994) and Autism Diagnostic Observation Schedule (ADOS; Lord et al, 2000). 10 participants had also had their diagnosis reconfirmed using the ADI-R by a trained university researcher within the last two years. Exclusion criteria for the clinical group included any presence of conditions known to be comorbid and similar in symptomatology to ASD (such as Fragile X or Rett’s Syndrome) or other neurological conditions (such as Tourette’s Syndrome) which may be expected to affect cognitive functioning. In addition children with specific and severe impairments in language processing over and above their ASD symptoms (such as verbal dyspraxia) were not recruited. Over 75% of the group currently attend a special education school. Participants were aged between 8 years 4 months and 17 years 2 months, with a mean age of 13 years 4 months (SD = 2y 5m).

- Cognitive ability profile

Study participants completed the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) with a trained researcher as part of testing. The WASI consists of four subtests and can be administered in less than 30 minutes. It includes two verbal measures (vocabulary and similarities) and two non-verbal measures (matrix reasoning and block design), and can be used to provide estimates for full-scale, verbal and non-verbal or performance IQ (Wechsler, 1999). 12 out of 14 participants completed all 4 subtests of the WASI, but 1 participant (DS) did not complete the similarities subtest while another did not complete similarities or block design (JB). Their full-scale IQ scores were estimated based on the minimal 2-subtest criterion (vocabulary and matrix reasoning) defined in the WASI manual (Wechsler, 1999). Verbal and non-verbal IQ estimates were also derived manually for these participants based on the T-scores of their completed subtests, but these scores were only used in covariate analysis and are not included in the VIQ/NVIQ summaries and T-tests described below. Based on their test scores all ASD participants were estimated to fall within the normal IQ range of 70 and above, (mean full-scale IQ = 103.29, SD = 15.72, range = 74 – 126), indicating the high-functioning nature of the group. Table 1 displays a summary of full-scale, verbal and performance IQ estimates. (See appendix A for specific subtest scores).
- Controls

14 typically-developing children (11 male, 3 female) were recruited for an age-matched control group. Participants were matched to within 12 months of their clinical counterpart (mean age = 13y 4m, SD = 2y 2m, range = 8y 9m - 16y 5m). All of the control participants currently attend mainstream school. Where possible, controls were also matched for full-scale IQ to within 10 points. All controls completed the full version of the WASI. Control FS-IQ scores ranged from 97 to 125 (mean FSIQ = 113.71, S.D. = 7.56). Ten participant pairs in total could be matched for full-scale IQ in addition to age. Despite the verbal nature of the task, participants were not explicitly matched on verbal IQ. This was done deliberately for two reasons. Firstly, language delay forms part of the diagnostic criteria for autism (APA, 1994; WHO, 1993) and so may be expected, as a function of diagnosis, to result in a lower verbal IQ estimate for participants with autism. Secondly, the similarities subtest of the WASI verbal subscale explicitly measures the ability to apply category names to pairs of common items (for example, a participant may be asked to name a common category for “cow” and “bear”; Wechsler, 1999). Therefore it may in part be measuring similar processes to the target tasks outlined below. If ASD and control participants were to be matched on this criterion, this could in effect be deselecting possible group differences in categorisation ability.

- Recruitment and consent

Participants were initially contacted via an information letter to families. The letter invited them to take part in a new study on reasoning and problem-solving. On return of a reply slip, families were contacted by phone to further explain the study and provide the opportunity to ask questions about the research.

Table 1: Age, Full-scale and subscale IQ scores for ASD and Control groups

<table>
<thead>
<tr>
<th></th>
<th>ASD</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (months)</td>
<td>Mean 159.50</td>
<td>Mean 160.29</td>
</tr>
<tr>
<td></td>
<td>SD 28.66</td>
<td>SD 26.89</td>
</tr>
<tr>
<td>Full-Scale IQ</td>
<td>Mean 102.29</td>
<td>Mean 113.71*</td>
</tr>
<tr>
<td></td>
<td>SD 15.73</td>
<td>SD 7.56</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>Mean 104.42</td>
<td>Mean 117.64^</td>
</tr>
<tr>
<td></td>
<td>SD 20.51</td>
<td>SD 13.19</td>
</tr>
<tr>
<td>Non-verbal IQ</td>
<td>Mean 102.67</td>
<td>Mean 107.14</td>
</tr>
<tr>
<td></td>
<td>SD 10.84</td>
<td>SD 9.92</td>
</tr>
</tbody>
</table>

* Significant difference at p < .05
^ Approaching significance at p < .1
after which a testing session was arranged. All ASD participants and all but two of the control participants were accompanied by a parent to the testing session. Before testing parents of participants were asked to provide full written consent for study participation and participants were reminded that they could withdraw at any time. Each participant received a £10 gift voucher as a “thank you” following their participation. Study procedures were approved by the University of Edinburgh School of Philosophy, Psychology and Language Sciences (PPLS) Ethics Committee prior to testing.

**Design**

A 2 x 2 x 3 (group x task x trial block) factorial design was used to compare a population of clinical-control matched pairs on standard and novel forms of the Twenty Questions Task. Task order was counterbalanced across each group so that half of the participants completed the standard then novel forms while the other half completed novel then standard forms.

Each task is divided into 3 trial blocks. Block 1 consists of 3 trials (the *initial* condition), which are then followed by a within-set change. Block 2 consists of 3 trials with the changed set (the *flexibility* condition). Block 3 consists of 2 further bonus trials with the original set, but with physical elimination of cards prohibited (the *memory* condition). See Figure 2 for a diagram of the testing order.

**Figure 2: Testing order for task conditions with sets used.**

<table>
<thead>
<tr>
<th>Block</th>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Initial Condition</td>
<td>Flexibility Condition</td>
<td>Memory Condition (bonus round)</td>
</tr>
<tr>
<td>Trial</td>
<td>1 2 3</td>
<td>4 5 6</td>
<td>7 8</td>
</tr>
<tr>
<td>Set used</td>
<td>Set 1</td>
<td>Set 2</td>
<td>Set 1</td>
</tr>
</tbody>
</table>
Materials

- The Guess Who board

Four 30cm x 30cm plastic boards from the popular children’s game *Guess Who?* (© Hasbro) were used for testing. Each board contains 24 hinged frames which may be placed in an upright or closed position (closed indicating that the item has been eliminated as a possibility). Frames hold cards indicating possible items chosen by an experimenter. Two boards were made up of items from the standard Twenty Questions Task and two were made up of the novel Robot stimuli. Coloured plastic counters were used to denote the number of questions used and successful identification of target items. Participants were given 10 question counters at the start of each round, denoting the maximum number of questions they could ask.  

- Twenty Questions stimuli: standard items

Two sets of 24 items were assembled based on the Twenty Questions subtest of the Delis-Kaplan Executive Functioning System (DKEFS; Delis et al, 2001). This set was selected as it is one of the most recent Twenty Questions sets to have been developed and is accompanied by high-quality, coloured images of items. (In the original form of the task (Mosher & Hornsby, 1966) and subsequent replication by Minshew and colleagues (Minshew et al, 1994; 2002) black and white line drawings were used). In addition the DKEFS set is standardised for familiarity of items and is designed to provide participants with a conventional structural hierarchy, which may be used for organised elimination of possibilities (DKEFS manual; Delis et al, 2001). Appendix B contains a diagram of the full DKEFS set structure.

The initial set (set 1) consisted of four main item groupings: animals, plants, vehicles and household goods or objects. Although the DKEFS set consists of 30 objects, the 24 items which were selected for the Guess Who board were chosen in such a way that would preserve the basic hierarchical structure of the 30 item set. The full set could be equally divided into living vs. non-living things, and then further subdivided into plants or animals, objects or vehicles. Each of these basic groupings could also be equally

---

5 Although, as the name suggests, there are usually a maximum of twenty questions in the Twenty Questions task, this is usually based on searching through a set of 30-40 items. As the Guess Who board is limited at 24 possible items, and many of the participants had played both Twenty Questions and Guess Who before, a limit of ten questions was chosen to provide more of a challenge.
divided in at least one way (for example, the vehicles could be divided into land vs. non-land transportation.). See figures 3a and 3b for the full array for set 1 (as seen by the participant) and a diagram of its hierarchical structure.

Figure 3: The Twenty Questions task (set 1).

a) Set items

![Image of set items]

b) Set structure

<table>
<thead>
<tr>
<th>12 living</th>
<th>6 animals</th>
<th>2 air</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2 water</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 land</td>
</tr>
<tr>
<td>6 plants</td>
<td>2 plant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 veg</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 fruit</td>
<td></td>
</tr>
<tr>
<td>12 non-living</td>
<td>6 transport</td>
<td>2 air</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 land</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 water</td>
</tr>
<tr>
<td>6 kitchen</td>
<td>3 cutlery</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 crockery</td>
</tr>
</tbody>
</table>
Set 2 was devised for the within-set change. It also consisted of classic Twenty Questions items but the frequency of various category members was adjusted. In set 2 the number of living things was increased to 18, with 6 new animal items (CAT, COW, DUCKS, HORSE, SHEEP, and SNAKE). All the vehicles from set 1 were removed, as were the CUPS and PLATE items. These were replaced by FRIDGE and OVEN items. The adjustments for set 2 were again made while preserving a hierarchical structure, although the main full set division was now animals vs. animals (12/12). In order to maintain an equal structure some additional items were included which were not drawn from the DKEFS 30-item set (specifically the snake, sheep, cat and horse). These items were selected to be of equal familiarity and visual quality to the DKEFS items. See appendix C for a diagram of the set 2 array.

As in Guess Who, but unlike Twenty Questions, the name of each item was printed on the card below the image. This was done so that participants did not have to generate their own labels for different items, and also to ensure that participant and experimenter were referring to the same items in questioning.

- Twenty Questions stimuli: novel (robot) items

Two sets of 24 novel items were assembled based on a set of custom-drawn robot characters. Robot stimuli were chosen in order to appeal to ASD participants, many of whom had a keen interest in mechanical objects. Each character was designed by the author to hold a unique combination of perceptual features, varying in colour, head-shape, number of eyes and number of wheels and feet. Robot characters could have one of four colours (blue, red, green or black), one of four head-shapes (square, circle, triangle or x-shaped), between 1 to 4 eyes, and have wheels (2 or 4) or feet (1 or 2). Individual robot characters were arbitrarily assigned common names, which were printed below the image. Names were of specific genders but sets were not divided equally according to gender (and thus the abstract label was not necessary for effective categorisation of the set).

Set 1 of the robot items was designed to allow for structured search but without a necessary hierarchy (See figures 4a and 4b). Head-shapes were distributed so that an initial question about square-headed robots would be most effective (12 square, 4 each of circle, triangle, x-shaped). Colours were equally distributed (6 each of blue, red, green and black), but numbers of eyes, feet and wheels were unevenly apportioned so that some questions would be more effective than others if selected (eyes: 9 x 1, 6 x 2, 4 x 3, 5 x 4; feet: 7 x 1, 5 x 2; wheels: 5 x 2, 7 x 4).
Figure 4: The Robots task (set 1)

a) Set items

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cath</td>
<td>Jonny</td>
<td>Sarah</td>
<td>Will</td>
<td>Anna</td>
</tr>
<tr>
<td>Doug</td>
<td>Tony</td>
<td>Colin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chris</td>
<td>Gillian</td>
<td>Jeff</td>
<td>Paula</td>
<td>Lee</td>
</tr>
<tr>
<td>Dave</td>
<td>Sean</td>
<td>Sophie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cilla</td>
<td>Annabelle</td>
<td>Wendy</td>
<td>Barry</td>
<td>Norman</td>
</tr>
<tr>
<td>Karl</td>
<td>Billy</td>
<td>Jack</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) Set structure

<table>
<thead>
<tr>
<th></th>
<th>3 blue</th>
<th>2 feet</th>
<th>2x 1 eye</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 wheels</td>
<td>1x 3 eye</td>
<td></td>
</tr>
<tr>
<td>12 square-heads</td>
<td></td>
<td>1 feet</td>
<td>1x 1 eye</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 wheels</td>
<td>1x 3 eye, 1x 2 eye</td>
<td></td>
</tr>
<tr>
<td>3 green</td>
<td></td>
<td>2 feet</td>
<td>1x 1 eye, 1x 3 eye</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 wheels</td>
<td>1x 2 eye</td>
<td></td>
</tr>
<tr>
<td>3 red</td>
<td></td>
<td>1 feet</td>
<td>1x 4 eye</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 wheels</td>
<td>2x 4 eye</td>
<td></td>
</tr>
<tr>
<td>3 black</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                | 3 blue | 1 feet | 1x 1 eye |        |
|                |        | 2 wheels | 1x 1 eye, 1x 4 eye |        |
| 12 non-square  |        | 1 feet |          |        |
| (4 circle, 4 triangle, 4 x-shape) |        | 2 wheels | 1x 2 eye, 1x 4 eye |        |
|                |        | 3 red | 1x 1 eye, 1x 3 eye |        |
|                |        | 1 wheels | 1x 2 eye |        |
|                |        | 3 black | 2 feet | 2x 1 eye |
|                |        |        | 1 wheels | 1x 2 eye |
The second set of robots (following the within-set change) was adjusted so that colour, rather than shape, was the more effective criterion on which to begin searching. With set 2 a question about red robots would offer the most productive split as an initial question (12 red, 4 blue, 4 green, 4 black). Head-shapes distributions were adjusted so that all circle-headed robots were removed (8 square, 8 triangle, 8 x-shaped). The numbers of eyes for each robot were again distributed to make some questions more profitable than others in eliminating items (eyes: 7 x 1, 5 x 2, 8 x 3, 4 x 4) although this time the distributions of feet and wheels provided an even split (feet: 6 x 1, 6 x 2; wheels: 6 x 3, 6 x 4).

For bonus (i.e. memory) rounds on the standard and novel stimuli tasks, the first sets on each task were used. New sets were not used as this would, in effect, create a second flexibility condition.

- Feedback to participants: the role of the experimenter

In the classic version of the task an experimenter will choose an item from the set array and hold it in their mind while the participant asks them questions. Although this is simple to implement, it is arguably not ideal in ASD research for a combination of reasons. Firstly, the concept of having to find out “what’s in someone’s head” is not necessarily interesting or easily achievable for a child with autism (Hobson, 1993). Secondly, the feedback given by an experimenter will be almost entirely verbal, and does not necessarily reinforce any sense of achievement or reward in completing a trial.

To combat this, computer-based feedback was used on both the standard and novel forms of the task. The experimenter could not be fully eliminated from the task, as a fully-computer based task would require a programme that was receptive and responsive to almost any type of question. Instead, a procedure (outlined below) was devised to create the illusion that the experimenter was “using the computer” to choose each item prior to the start of the trial. The advantages of this were a) that the participant could trust that an item had indeed been selected (as opposed to the experimenter just stringing them along), and b) that the computer in effect became a “mystery box” with the answer in. In this way, the answer was externalised, hidden but mutually accessible by both experimenter and participant. The method of finding the answer still relied on asking the experimenter questions (as they already knew what was selected), so the primary form of feedback, the yes/no answer after each question, was still provided by the experimenter. But the answer itself, “revealed” by the computer feedback at the end of the trial, could be seen as external to the experimenter, rather than “in the mind”. The use of a computer also allowed for a range of visual and auditory cues to be presented as game feedback, instead of having to rely on the entertainment skills of the experimenter.

A 17” Dell Inspiron Laptop was used to provide the visual and auditory feedback, with Microsoft
Powerpoint being used to display a slideshow accompanying the task rounds. Computer animations were designed in order to make the task as game-like as possible for participants. Prior to the start of the trial, item selection was accompanied by 2-3 second “random generator” animation, in which the experimenter would “use the computer to choose the first object/robot”. The animation consisted of a black question mark symbol centred in the screen, followed by a rapid presentation of item silhouettes (each present for 0.1s), after which the question mark symbol returned. Following this the experimenter would invite the participant to start asking him questions about the target item. During the trial, the question mark symbol remained in the centre of the screen until either the target was identified or the participant had used up their 10 questions (see below for a full explanation of the procedure during each trial). On completion, successful trials were accompanied by a 3 second fanfare, applause or game-show organ noise clip and the replacement of a question-mark symbol (150-size Arial font) by the target item (see figure 5 for an example). Unsuccessful trials were followed by the revelation of the item, but without sound.

Figure 5: Example of feedback presentation

1. Full array displayed
2. Question mark displayed as mask
3. “Random generator” animation (experimenter chooses the item)
4. Mask reappears (item selected)
5. Item revealed at end of trial
Procedure

Testing sessions took place at the Laboratory for Cognitive Development in the Psychology department or in some cases in participants’ homes (if this was more convenient for families). Participants were seated at a desk alongside the experimenter, approximately 20-30 cm away from the game-board and 60 cm away from the computer. 10 coloured question counters were placed between the experimenter and participant, with another 16 orange and yellow trial counters to the left of the experimenter. At the start of testing, all frames on the set 1 game-board (standard or robots) were placed in the upright position, facing the participant (see figure 6).

Figure 6: Diagram of testing setup.

Prior to testing participants were asked if they had ever played Twenty Questions or Guess Who before. If they had they were asked to explain the rules as they understood them, and instructions were tailored to their level of understanding accordingly. To begin the experimenter explained the rules as follows:

“Today we are going to play two games: one involves asking questions about different types of objects, and the other involves asking questions about robots. As you can see, all of these cards have a different object or robot on them. In a moment I am going to use the computer to choose one of them, and your job is to try and find out which one it is. You can do this by asking me questions about them, but I can only
answer yes or no to your questions. Every time I answer a question, you can push down all the ones you
don’t need on the board, like this. (Experimenter demonstrates the hinged frame system on the game-
board). The aim of the game is to find the object/robot in as few questions as possible, and you are
allowed ten questions on each trial to try and find it. These counters mark out how many questions you
have asked (experimenter points to question counters), while these other counters are for how many times
you manage to find the right one (experimenter points to trial counters). Every time you find one, you will
get one of these counters. Do you have any questions? (If necessary:) Would you like a practice go before
we start?”

If participants were unsure or it seemed that they had not understood the instructions, a practice trial of 8
items was played. The experimenter would choose one item and a round would be played, with the
experimenter demonstrating the elimination of items on the board if needed. If the participant found the
target in 8 questions or less (i.e. at least demonstrating a serial hypothesis-testing strategy) then the full
task was begun. If participants could not do this, or failed to ask any questions, the roles were reversed:
the participant was asked to choose one and the experimenter modelled an average questioning strategy
(i.e. asking questions of medium utility). As early Twenty Questions research indicated considerable
modelling effects in neurotypical children when exposed to another player’s game strategy (Denney,
1975), participants who required this step were denoted in later analysis and their responses examined for
any apparent benefit due to modelling. In addition, if parents indicated any concerns about their child’s
ability to read the different item labels, participants were asked to name each of the items in the set prior
to playing the game.

Before the start of each trial, the experimenter would use the following cue:

“I’m going to use the computer to choose the first / next one. (Experimenter starts and stops “random
generator” on laptop presentation). OK, I’ve chosen one. What’s your first question?”

Following block 1 of each task, the following prompt was provided by the experimenter:

“Now we are going to use a different set of things. Some of the objects/robots in this set have changed.
Before we play again, take a moment to have a look at the new set. (Participant given approximately 30
seconds to examine the set). OK, I am going to choose the next one…. “.

Finally, prior to the bonus rounds, the following instructions were given:

“Now we are on to the bonus rounds. In the bonus rounds, there is a new challenge. I am going to choose
another one, but this time you need to try and find it without pushing any of them down. Do you think you
could have a go at it?".

On each trial participants were given a maximum of ten questions to identify the target item. After each question the experimenter removed one of the coloured question tokens. Specific responses were used in some cases to replicate the procedures of the DKEFS Twenty Questions Task. If a participant asked a question which could not answered dichotomously (for example, “Is the robot red or is it green?”), the experimenter would remind the participant:

“I can only answer your question yes or no. Would you like to ask your question again?”.

If the participant asked an ambiguous question (e.g. “Is it big?” Or “Can you hold it?”), the experimenter would ask the participant to clarify:

“I'm not sure how to answer. Could you be more specific?”

If the participant asked a question where the answer is sometimes or mostly true (e.g. “Would you find it in a zoo?”, majority answers were used, i.e.:

“I think most people would say yes/no”.

Participants could ask questions to clarify categories (e.g. “Do you count fruit as a living thing?”). Such “questions about questions” were not deducted from their ten possible questions. On the bonus memory rounds participants could also ask if they had previously asked a specific question (e.g. “Did I ask if it was black?”), as this is also allowed in the DKEFS form of the task (DKEFS manual; Delis et al, 2001). Additionally, specific verbal prompts were used at certain trial points. For instance, if after three questions participants had only asked questions which eliminated individual items, the experimenter would say:

“Remember, the aim of the game is to find it in as few questions as possible”.

After six questions and eight questions, the experimenter would also remind the participant how many questions they had left. If after ten questions only one possible item remained in the set (but had not been directly referred to), this was counted as a completed trial. At the end of each trial participants received their ten question tokens back and the correct item was revealed using the computer presentation, which was accompanied by sound effects if the participant was successful or muted if not. Participants received an orange and yellow trial token for each target successfully identified within the ten questions.

For each task the order of selected items was always the same. In the standard form, the targets chosen
were TREE, DOG, and BUS in block 1, SHARK, BANANA and FORK in block 2, and ELEPHANT and BOAT in block 3. In the novel form, robots COLIN, LEE and ANNA were chosen in block 1, GREG, ZACK and KENNY in block 2, and WENDY and JACK in block 3.6

**Trial scoring and coding of responses**

- **Task success**

A number of outcome variables can be drawn from the Twenty Questions task which measure aspects of categorisation efficiency, strategy preference and memory functioning. However, the primary outcomes measured here were selected to be as close as possible to previous versions of the Twenty Questions task while also addressing the key questions and factors affecting task performance. As in Mosher & Hornsby (1966) and Minshew et al (1994), overall task success was assessed by the overall percentage of trials completed and the mean number of questions taken on each trial by participants. Question quality was assessed continuously using a version of the “initial abstraction” score from the DKEFS Twenty Question Task (Delis et al, 2001). This measure assesses question quality in terms of the minimum proportion of items eliminated by each question. For example, a question which refers to a property possessed by 3 of 8 items will receive a score of 3/8 (0.375), irrespective of whether the target item has that property (leaving 3 possibilities) or does not (leaving 5 possibilities). As each question must divide the set in some way, the only way to guarantee a maximal number of items eliminated is to split the set in half in each time. Because of this the maximum score on this measure is 0.5.

---

6 It is significant to note that some original versions of the Twenty Questions task did not previously use a set target, instead providing answers which eliminated the least amount of possibilities for each question (i.e. if three animals and two vegetables remained in the set and a participant asked “Is it a vegetable?” the experimenter would say no, thus only eliminating two items rather than three). This method, which guarantees that participants cannot “luck out” early on with an inspired guess, has been used previously to establish a baseline number of questions which can then be assessed for quality (Siegler, 1977). However using an experimental strategy such as this arguably detracts from the game-like nature of the task, which is essential in ensuring participant compliance and motivation. Furthermore, it is not the method used by either of Minshew et al’s (1994; 2002) studies, or the standardised DKEFS test, and there are methods of analysis which can be deployed to minimised the effect of lucky guesses by participants.
- Question types

The assessment of question types is another measure of performance used in many of the previous Twenty Questions studies. Participant questions were coded following Mosher & Hornsby’s original definitions (Mosher & Hornsby, 1966). Questions which referred to two or more items in the set and eliminated at least one item were defined as constraint-seeking questions (for example, “Is it a living thing?”). Questions which only referred to and eliminated one item in the set were scored as hypothesis-testing questions (e.g. “Is it the bowl?”). Questions which did not explicitly name an item, but nevertheless only referred to one item were marked as pseudo-constraint questions (e.g. “Do you use it to spread butter?”). Based on these definitions an efficient player of the game would ask a series of constraint-seeking questions to sub-divide the set until only two or three items remain. Then a hypothesis-seeking or pseudo-constraint question must be used to finally refer to the target item.

In addition to the coding labels used previously, redundant and repeated questions were also noted. Redundant questions were defined as questions which did not eliminate any of the remaining set members (for example, asking “Can you keep it as a pet?” when only DOG and CAT items remain). Repeat questions denoted questions which had already been asked within that trial. Both redundant and repeated questions are specifically important when considering the role of memory and flexibility in task performance. If, for example, verbal working memory difficulties were significantly affecting ASD performance, then one would plausibly expect an increase in repeated or redundant questions during the memory-based bonus rounds.

- Question content

As already alluded to, the content of the two tasks would be expected to largely dictate the content of participants’ questions. Specifically, the Twenty Questions task might be expected to prompt more concept-based questions, referring to abstract groupings, or functions of items, while the Robots form of the task would prompt mostly perceptual questions, referring to colour, form or specific features. However, in theory the two sets do share properties which allow for common methods of sorting. For example, though the standard form of the task largely consists of conceptual groupings, there is nothing stopping a participant searching the set based on purely perceptual similarities and dissimilarities between the set items. Similarly, on the novel form of the task the individual names of robots do allow for grouping based on gender, an abstract label which is independent of the perceptual features in the given set. (Participants who have played Guess Who before might even be expected to utilise exactly this sorting criterion).
Nevertheless, the contents of each participant question were coded for themes or prevalent strategies that could be grouped into roughly two areas: *conceptual* categories and *perceptual* categories. For each question, one of the following labels was assigned:

**Conceptual Categories**

*Abstract questions* – a question about an abstract label or taxonomic grouping, such as “Is it a mammal?” or “Is it male?”. Such questions require both a degree of conceptual understanding and knowledge of an abstract label which denotes the conceptual grouping. In this sense, this form of questioning represents the least perceptually-bound and most conceptually-based search strategy.

*Function questions* – a question about the use, purpose or function of an item, e.g. “Is it for DIY?” or “Do you use it to get to school?”. Functional questions represent a level of search which is more abstract than questions referring to immediate perceptual properties, but nevertheless still display a more concrete form of search than a genuinely abstract questioning strategy. Overly function-based questions might also be interpreted as evidence of a preference for systemic or mechanical concepts.

*Specific feature questions (conceptual)* – a question about a specific feature or characteristic of an item which is neither functional nor perceptually bound, e.g. “Does it live in the sea?” or “Can it carry more than 4 people?”. This criterion is relatively broad due to the many different levels of sub-categories which could be specified without any one being prominent. However, it is useful to record as it displays a level of conceptual grouping which is independent of perceptual properties, despite often being narrower in focus than more abstract forms of grouping. In this sense, a strong use of this strategy might suggest a local bias in conceptual processing (Happé, 1999).

**Perceptual Categories**

*Colour questions* – a question that refers to the colour of an item or part of that item, e.g. “Is it red?” or

---

7 Ropar & Peebles’ (1997) comparison between searching based on the use of sporting equipment vs. searching based on sports or games demonstrates this distinction; the former being functional and more concrete, the latter representing a strategy based on two abstract groupings.

8 It could be argued here that specific feature questions might be more conceptual than certain functional questions, and that function questions are simply one sub-class of feature questions. (Compare, for instance, “Does it lay eggs?” with “Can you use it to cut wood?”; one question refers to all oviparous animals, while the other refers to saws). However, the teleological nature of function arguably also adds a level of abstractness to the search criterion, making it an awkward sub-class of feature questions. For this reason, and its specific interest in terms of cognitive style, it is recorded separately.
“Does it have white bits on it?” On the novel form of the task, and for most items on the standard form of the task, this will be a relatively global perceptual property, in that it often refers to the predominant colour of the whole object.

*Form questions* – a question about the shape or form of the item, e.g. “Is it round?” or “Does it have a square head?”. As with colour questions, this is usually a more global perceptual criterion on which to search, in that it will often refer to the whole object’s shape or size.

*Name questions* - a question about an item’s name or the letters it is made up of e.g. “Does it begin with a G?”. Although an item’s relation to its name is often abstract, the name is included in the set as an intrinsic feature of the perceptual array. In this sense asking questions about it is just as perceptually bound as asking about the number of eyes or ears an item possesses.

*Specific feature questions (perceptual)* - a question about a perceptual feature which is not any of the above. Typically this will refer to local details of an item, such as the number of feet or wheels an item has. Questions about specific perceptual features are notable because they would be a plausible target for participants with any sort of local bias or preference in perceptual processing (Mottron et al, 2006). In contrast, a perceptually-bound strategy which is not pulled towards local features would be expected to be expressed in use of colour, form and feature-based questions more equally.

In summary, the various categories of question content can all be used to examine aspects of participant strategy which indicate certain types of cognitive style. The main distinction here, as with the two tasks, is the overall contrast between conceptual vs. perceptual strategies (with abstract questions being the most conceptual, and specific perceptual feature questions being the least conceptual). However, within these groupings there are further areas of interest which have specific implications for a cognitive style: in particular, differing levels of abstractness and potential functional preferences in conceptual grouping, and evidence of local feature biases in conceptual and perceptual domains.

*Data recording and analysis*

Testing sessions were recorded using a portable table microphone and Audacity sound editing software. (Parents were explicitly informed of this prior to consenting). All analyses were conducted using SPSS 16.0 statistical software. Prior to analysis all variables were checked for departures from assumptions of parametricity using Kolmogorov-Smirnov and Shapiro-Wilk tests of normality. (Shapiro-Wilk tests were used alongside the more standard test of normality in cases where the Kolmogorov-Smirnov test could only provide rounded estimates). Where non-normal variables were identified they were marked for later
analysis using non-parametric statistics. Where analyses of variance were conducted, variables were still included if they only violated one of two assumptions (either distribution normality or homogeneity of variance), and they were supplemented with an appropriate non-parametric test (one of the Mann-Whitney U-Test, Wilcoxon Signed Ranks Test, Kruskal-Wallis or Friedmann’s ANOVA). If the results of parametric and non-parametric analysis on the same variable differed, the non-parametric result was reported. If any variables violated both of the above assumptions they were excluded from parametric analyses and only analysed using non-parametric methods.

To assess the influence of covariates on task performance in the two groups, age and full-scale, verbal and non-verbal IQ were examined for contributions to the main outcome variables using correlational analysis. Predictors with significant associations were retained for use in between-groups analysis as covariates. Non-significant variables were excluded from covariate analysis, although any non-significant but moderate correlations with outcome variables were noted. This approach was chosen specifically to examine the influence of different IQ components. Most between-groups designs in this area of research will seek to primarily match clinical and control participants on chronological and mental age, or utilise age and full-scale IQ scores as primary predictors in covariate analysis (Jarrold & Brock, 2004). As it was only possible to FSIQ-match a subset of participants in the present study (10/14 pairs), full-scale IQ must be controlled for in some way. However, within the influence of full-scale IQ it is arguably verbal IQ, rather than non-verbal IQ, which will be the relatively greater contributor to task performance, given the linguistic demands of the Twenty Questions task. For the reasons outlined above, the groups have deliberately not been matched on this factor. As such, verbal IQ is likely to be the main predictor (along with age) which must be controlled for when comparing the two groups, although some aspects of non-verbal IQ may also make a contribution. Preliminary analysis of predictor-outcome associations allows for a thorough examination of the relative contributions of each of these scores, so that they can be controlled for in the most effective way during between-groups analysis.

Finally, multiple comparisons were controlled for by analysing groups of potentially related dependent variables together in multivariate analyses of covariance, followed by running planned univariate ANCOVAs. Homogeneity of regression slopes and covariance matrices were checked for multivariate and covariate analyses. Unless otherwise specified, other post-hoc comparisons and multiple comparisons with non-parametric tests were conducted with appropriately adjusted alpha-levels using the Bonferroni method.
RESULTS

1. Preliminary analysis

i) Age and IQ differences

Independent samples T-tests were used to assess age and IQ differences between the ASD group and typically-developing control group (TD henceforth). ASD participants scored significantly lower than TD participants on full-scale IQ (ASD m = 102.29, TD m = 113.71; t = -2.51, df = 26, p = .024) and there was also a trend towards a significant difference on verbal IQ (ASD m = 104.42, TD m = 117.64; t = -1.984, df = 24, p = .059). No significant differences in non-verbal IQ were observed (ASD m = 102.67, TD m = 107.14; p = .283, n.s.). Given the overall difference in full-scale IQ, it is likely that the observed difference in verbal IQ would have been significant had all of the ASD participants completed the full WASI battery.

ii) Selection of covariates

Age, full-scale and subscale IQ scores were assessed for possible associations with the main outcome variables. Pearson’s Product Correlation Co-efficient (r$_p$) analysis was used to assess all parametric predictor-outcome associations and Spearman’s Rho (r$_s$) was used for all non-parametric correlations. Outcome variables included in the correlation matrix were as follows:

**Mean Question Quality:** the overall mean percentage of items eliminated by each question.

**Mean Questions to Target:** the overall mean number of questions used by participants to find the target on each trial.

**Trials completed (%)**: the overall percentage of trials where the target was successfully identified in 10 questions or less.

**Constraint-seeking questions (%)**: the percentage of questions asked which referred to and eliminated at least 2 items.

---

9 The estimates for VIQ and NVIQ added to the group for covariate analysis confirm this: with the addition of VIQ/NVIQ scores for the two missing ASD participants, verbal IQ was found to significantly differ between the two groups (t = -2.427, df = 26, p = .024) but any differences in non-verbal IQ were still far from significant (p = .392).
Hypothesis-testing questions (%): the percentage of questions asked which directly referred to and eliminated only one item.

All but two predictors (trials completed and percentage of hypothesis-testing questions) were assessed using parametric methods. Due to considerable skew in the data for these outcomes (positive skew in the case of the former, negative for the latter), non-parametric correlation coefficients were used instead.

Correlational analysis revealed a significant association between age and the mean number of questions asked ($r_p = -.354$, $n = 28$, $p = .065$), indicating that older participants tended to take fewer questions overall to find the target. A trend towards an association with the percentage of hypothesis-seeking questions asked was also observed although not at a significant level ($r_s = -.354$, $n = 28$, $p = .065$). Interestingly, this indicated that older participants tended to ask more, rather than less, hypothesis-seeking questions across the two tasks.

Full-scale IQ scores were associated with a range of outcome measures. FSIQ scores positively correlated with mean question quality ($r_p = .614$, $n = 28$, $p = .001$), the percentage of constraint-seeking questions asked ($r_p = .731$, $n = 28$, $p = .001$) and the percentage of trials completed ($r_s = .389$, $n = 28$, $p = .041$), indicating that participants with higher full-scale scores tended to also ask better quality questions (i.e. eliminating more items per question on average) and successfully identified more target items in general.

Examination of the IQ subscale scores pointed towards the likely source of observed associations between full-scale IQ and task success. While non-verbal IQ scores did not significantly correlate with any outcome variables (all $r$-values < .3, all $p$-values > .13), verbal IQ scores were positively associated with mean question quality ($r_p = .611$, $n = 26$, $p = .001$), mean number of questions asked ($r_p = -.446$, $n = 26$, $p = .022$), percentage of constraint-seeking questions asked ($r_p = .745$, $n = 26$, $p = .0001$) and percentage of trials completed ($r_s = .470$, $n = 26$, $p = .015$).

This range of associations suggests that task success in general was mainly related to the verbal components of participants’ IQ scores, reflecting the highly verbal demands of the two questioning tasks. For this reason verbal IQ was chosen as the main covariate in all subsequent between-groups analyses of task outcomes. Although the groups were demonstrated to be adequately matched on age, the correlations observed suggested the potential for age effects. To monitor these effects age was also included as a covariate in the between-groups analysis of task performance.
### iii) Order effects

Half of the participants completed the standard Twenty Questions task first and the Robots task second (order 1), while the other half completed the tasks in reverse order (order 2). Mean differences on the main outcome variables were compared for those completing orders 1 and 2, to assess possible order effects. No significant differences based on order were observed, although there was a slight tendency towards higher rates of hypothesis-testing questions being asked by those who completed the Twenty Questions Task first (order 1 m = 17.9%, order 2 m = 11.7%; Mann-Whitney U = 56.00, p = .056). In the absence of any other noticeable order effects, the data from the two testing order groups were combined and treated as one sample for the ASD vs. TD analysis.
2. Task success

The main measures of task success were the mean question quality, the mean number of questions used to identify the target item\(^\text{10}\), and the percentage of trials completed. Mean question quality and mean questions-to-target were compared for the two groups using analyses of covariance, which included the selected covariates of verbal IQ and age. The percentage of trials completed was assessed using a Mann-Whitney U-test due to non-normal data in both groups. Table 2 displays group means for all outcome measures on both tasks.

i) The Twenty Questions Task (standard items)

All participants demonstrated an understanding of how to play the Twenty Questions task, with all but one participant in both groups managing to complete at least 5 out of 8 trials. Though scores in both groups were high, ASD participants completed significantly fewer trials on average than TD controls (ASD m = 86.6% (18.6%), TD m = 98.25% (4.5%); U = 59.00, p = .019). Examination of the scores for each trial block indicated that this difference was mainly due to each group’s reaction on block 2 (that is, following the within-set change). Trials completed on blocks 1 and 3 did not significantly differ between the two groups, but on block 2 ASD participants only completed 85.7% (21.5% SD) of the trials, while all the control participants completed 100% (U = 63.00, p = .020).

Mean scores for the number of questions asked and mean quality per question indicated similar differences between the two groups. ASD participants tended to ask more questions than controls across the 3 blocks (ASD block 1 = 6.02 (1.32), block 2 = 5.78 (1.88), block 3 = 6.29 (1.93); TD block 1 = 5.83 (1.38), block 2 = 5.17 (0.72), block 3 = 4.93 (1.22)), and tended to ask questions of lower quality (ASD block 1 = 0.24 (0.10), block 2 = 0.25 (0.10), block 3 = 0.22 (0.09); TD block 1 = 0.34 (0.07), block 2 = 0.34 (0.06), block 3 = 0.32 (0.08)). When controlling for verbal IQ and age no significant effects of group or block were found in the analyses of covariance. However, a group-related interaction effect was

\(^{10}\) The structure of both tasks allowed for the possibility of lucky guesses being correct, and in a small number of cases participants did indeed identify the target item on 2 questions or less. THE DKEFS Twenty Questions Task provides a “weighted achievement” score to address this, which it does by awarding maximum points for identification of an item in the logically minimum amount of questions, i.e. 4 or 5 (Delis et al, 2001). Departures from this are given proportionately smaller scores, so a target identified in 3 questions will achieve 2 points, a target achieved in 6-7 questions will achieve 4 points, etc. In the present study weighted achievement scores were also analysed, but provided no major departures from the pattern of results observed using the mean number of questions asked per trial. Furthermore, the distribution of the weighted achievement score arguably minimises the effect of lucky early guesses at the cost of also minimising differences at the other end, i.e. those taking many more questions on average. As such the adjusted scores are not reported here.
Table 2: Main group outcomes for task success, question types and question content

<table>
<thead>
<tr>
<th>Task success</th>
<th>Twenty Questions Task (standard items)</th>
<th>Robots Task (novel items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASD Mean  SD</td>
<td>TD Mean  SD</td>
</tr>
<tr>
<td>Trials completed (%)</td>
<td>86.61  18.65</td>
<td>98.21*  4.54</td>
</tr>
<tr>
<td>Questions asked (Mean)</td>
<td>6.05  1.29</td>
<td>5.36  0.77</td>
</tr>
<tr>
<td>Question quality (Mean)</td>
<td>0.24  0.10</td>
<td>0.33*  0.15</td>
</tr>
<tr>
<td>Question types (%)</td>
<td>Mean  SD</td>
<td>Mean  SD</td>
</tr>
<tr>
<td>Constraint-seeking questions</td>
<td>51.00  23.90</td>
<td>68.40*  7.90</td>
</tr>
<tr>
<td>Hypothesis-testing questions</td>
<td>28.30  31.70</td>
<td>12.00*  10.10</td>
</tr>
<tr>
<td>Pseudo-constraint questions</td>
<td>18.00  16.10</td>
<td>16.20*  11.80</td>
</tr>
<tr>
<td>Repeated questions</td>
<td>0.70  1.30</td>
<td>0.00*  0.00</td>
</tr>
<tr>
<td>Redundant questions</td>
<td>4.00  4.30</td>
<td>3.90  4.30</td>
</tr>
<tr>
<td>Question content (%)</td>
<td>Mean  SD</td>
<td>Mean  SD</td>
</tr>
<tr>
<td>Colour-based questions</td>
<td>10.90  18.10</td>
<td>10.96  13.60</td>
</tr>
<tr>
<td>Form-based questions</td>
<td>0.21  0.55</td>
<td>0.44  0.89</td>
</tr>
<tr>
<td>Name questions</td>
<td>0.00  0.00</td>
<td>4.21  12.68</td>
</tr>
<tr>
<td>Specific features (perceptual)</td>
<td>6.82  10.41</td>
<td>8.53  6.33</td>
</tr>
<tr>
<td>Abstract questions</td>
<td>30.29  22.77</td>
<td>39.70*  16.46</td>
</tr>
<tr>
<td>Function-based questions</td>
<td>16.50  16.12</td>
<td>6.40  8.77</td>
</tr>
<tr>
<td>Specific features (conceptual)</td>
<td>24.81  17.74</td>
<td>27.49  11.68</td>
</tr>
</tbody>
</table>

*a* Significant difference at $p < .05$

*b* Approaching significance at $p < .1$

*c* Significant main effect ($p<.05$) across both tasks

*d* Approaching significance at $p < .01$ (Bonferroni corrected)
observed for the number of questions asked in the memory condition: while control participants asked fewer questions on average (TD m change = 5.17 to 4.93), ASD participants tended to ask more questions, particularly in response to having to memorise eliminated items (ASD m1 = 5.78, m2 = 6.29; group* block effect F(1, 22) = 5.593, p = .027, η² = .203). This suggests that in general ASD and TD participants performed at similar levels on the Twenty Questions Task, but the variation in conditions within the task exposed specific group differences. In particular, ASD participants were less successful than controls in identifying the targets following the within-set change (block 2), and tended to ask more questions when they could not physically remove eliminated items from the set (block 3).

**Figure 8. The change in number of questions asked across within-trial variations (standard task)**
ii) The Robots Task (novel items)

As with the Twenty Questions Task, every participant was able to play the Robots Task to an analysable standard and all participants completed at least 5 out of 8 trials. Again, completion rates were high but this time no significant group differences were observed across the 3 trial blocks, possible because participants in the control group successfully completed 100% of the robot trials (ASD T% block 1 = 95.2% (12.1%), block 2 = 95.2% (12.1%), block 3 = 89.2% (28.9%). The lack of variance in the control group on this measure suggests that the Robots form of the task was fairly easy for control participants, although the presence of this apparent ceiling effect does not necessarily imply that significant group differences were obscured: ASD and control participants did not significantly differ on the mean number of questions asked either (ASD QT m = 5.64 (0.97), TD QT m = 5.19 (0.55), p = .525), suggesting similar levels of performance. This was also true across the different trial blocks (ASD QT block 1 = 5.47 (1.02), block 2 = 5.38 (0.87), block 3 = 6.25 (1.87); TD QT block 1 = 5.47 (0.71), block 2 = 4.60 (0.57), block 3 = 5.65 (1.09)).

Despite this, analysis of the mean question quality for each group indicated the presence of some group differences even on the novel form of the task. Overall mean question quality was significantly lower in the ASD group (ASD m = 0.30 (0.08), TD m = 0.38 (0.04), F (1, 24) = 3.555, p = 0.036, $\eta^2_p = .129$), indicating that ASD participants tended to eliminate a lower proportion of items than controls on each question. Control participants scored a consistently higher QQ average across the three blocks, but an interaction effect was also observed in the change between blocks 1 and 2 (group*block F (1, 24) = 4.126, p = 0.053, $\eta^2_p = .147$). Both groups tended to increase their quality of questions on the move to the second set (possibly indicating an effect of practice), and then decrease during the memory condition, but control participants increased their question quality by a significantly greater degree in the shift to block 2 (ASD QQ change = 0.30 to 0.32, TD QQ change = 0.32 to 0.41). This suggests that the within-set shift did not markedly impair either group in their questioning, but interrupted a possible improvement in questioning in ASD group which is otherwise seen in the controls.

iii) Comparing the tasks: group interaction effects

Comparison of scores on the two tasks confirmed the suggestion that the novel form of the task with the robot items was generally easier, with completion rates in both groups being significantly higher for robot trials than for trials with the standard items (Twenty Questions T%: ASD m = 86.6%, TD m = 98.25%; Robots T%: ASD m = 93.8%, TD m = 100%; z = -2.11, p = .031). Differences in the number of questions
used were specifically evident on the standard task during the memory condition, and this was demonstrated by a specific group-by-block effect ($F(2, 44) = 2.526, p = .046, \eta^2_p = .103$) for block 3. Importantly, for both variables, where significant group differences were observed on the standard task, they were not observed on the novel task.

The mean question quality across the two tasks consistently differed for the two groups, with a strong effect of group ($F(1, 22) = 4.370, p = .024, \eta^2_p = .166$) even after effects of verbal IQ and age were accounted for ($VIQ: F(1, 22) = 6.368, p = .019, \eta^2_p = .224; Age: F(1, 22) = 8.511, p = .008, \eta^2_p = .279$). No group-by-task interactions were observed, suggesting that controls were consistently asking better questions, irrespective of content. In addition, however, a significant interaction effect between verbal IQ and group was observed ($group*VIQ: F(1,22) = 4.487, p = .046, \eta^2_p = .169$) indicating that verbal IQ may have influenced performance differently depending on which group participants were in.

**Figure 9:** Mean question quality for each group on a) standard and b) novel forms of the task.
**iv) Summary**

As was predicted, the two groups demonstrated significant differences in task success but only in specific conditions. Significant differences on the standard form of the task, in trial completion and the number of questions used, were not apparent on the novel form of the task. Furthermore, it was the within-trial variations in set items and memory demands which best highlighted the group differences on the Twenty Questions task. However, contrary to the original hypothesis, differences in mean question quality were present and significant across both tasks. This supports the idea of a general difficulty with category utilisation in ASD, rather than a content-specific problem: whether the items were standard Twenty Questions items, or the novel robot forms, the difference in question quality between ASD and control participants appears to have been relatively robust. The one qualification here is that verbal IQ has been controlled for as a consistent covariate, but in some cases (as in the analysis of overall mean question quality) it may be that verbal IQ has an inconsistent effect, depending on group. For further exploration of this and other group differences, the various question types must be examined.

**3. Question types**

Participant questions were coded into one of 5 categories: constraint-seeking questions, hypothesis-testing questions, pseudo-constraint questions, repeated questions, and redundant questions. Constraint-seeking questions primarily provide an index of how many questions actually form a category criterion; they are the key to sub-dividing the set. Other question types provide information about sub-optimal searching, whether through minimal elimination of items (hypothesis-testing and pseudo-constraint questions) or errors in search (repeated or redundant questions).

**i) The Twenty Questions Task**

Figure 10 displays the relative proportions of question types used by participants in each group. All but one participant (RH, a member of the ASD group) used at least one constraint-seeking question at some point to search for target items. However, ASD participants asked a significantly smaller proportion of constraint-seeking questions than control participants (ASD m = 51.0 %, TD m = 68.4%; F (1, 22) = 3.351, p = 0.041, \( \eta^2_p = .132 \)) even after VIQ and age effects had been controlled for (VIQ: F (1, 22) = 4.589, p = .043, \( \eta^2_p = .173 \); age: F (1,22) = 5.215, p = .032, \( \eta^2_p = .192 \)). In addition, a group-by-VIQ effect
was indicated at a level approaching significance (F (1,22) = 3.301, p = .083, $\eta^2_p = .130$) suggesting a differing contribution of VIQ in the two groups.

Group scores for the “guess-like” hypothesis-testing questions painted a similar picture. ASD participants tended to ask a higher proportion of hypothesis-testing questions than controls (ASD m = 28.3%, TD m = 12.0%; F (1,22) = 3.800, p = 0.032, $\eta^2_p = .147$), and once again there seemed to be some form of interaction effect between group and verbal IQ (F (1, 22) = 4.119, p = .055, $\eta^2_p = .158$). For pseudo-constraint questions the same group difference was observed (ASD m = 18.0%, TD m = 16.2%; F (1,22) = 3.202, p = .044, $\eta^2_p = .127$), indicating that ASD participants also tended to ask more questions which sounded like they established categories without actually eliminating more than one item. No specific covariate interaction effects were observed on this variable.

Rates for repeated or redundant questions were low in both groups, and skewed towards zero. No control participants repeated any questions on the Twenty Questions task, whereas 0.7% of ASD questions were repetitions of previous enquiries (U = 70.00, p = .049). The proportion of redundant questions in each group was almost exactly equal (ASD m = 4.0% (4.3%), TD m = 3.9% (4.3%)).

**Figure 10. Question types asked on the Twenty Questions task**
Comparison of the within–task variations indicated no major changes in the use of constraint-seeking questions; both groups tended to drop slightly in their proportion of constraint-seeking questions following the within-set change (ASD change = 59% to 44%, TD change = 75% to 64%) although not to a significant degree. The memory condition only made a small difference to ASD scores, while the control scores remained constant (ASD m = 52 %, TD m = 64%). The significant group difference in use of constraint-seeking questions was consistent across the 3 conditions in the task. Block-by-block comparisons for the other question types also indicated no significant block effects, or any specific group-by-block interactions (all p-values > .1). This is especially notable for those questions which reflect error: repeated or redundant questions were not noticeably affected following the change in set or during the bonus rounds, suggesting that both groups could handle the addition of new task demands.

ii) The Robots Task

On the novel form of the task ASD participants recorded lower rates of constraint-seeking questions than controls (ASD m = 57.9% (15.7%), TD m = 71.6% (6.8%) although not at a level reaching significance (F(1, 22) = 2.229, p = .075, $\eta^2_p = .092$). This was also the case for hypothesis-testing questions (ASD m = 17.5% (23.0%), TD m = 6.4% (7.5%); F(1, 22) = 2.330, p = .141, $\eta^2_p = .096$), but not pseudo-constraint questions, where there was a clear difference between the groups (ASD m = 19.6%, TD m = 20.7%; F(1, 22) = 7.388, p = .013, $\eta^2_p = .251$). Interestingly, the direction of this difference indicated that control participants were not less but in fact more likely to adopt the language of pseudo-constraint questions when searching the robot set. For repeated questions no differences were observed (ASD m = 0.7 %, TD m = 0.3%, n.s.) but ASD participants did record significantly more redundant questions than TD controls on this form of the task (ASD m = 5.8%, TD m = 1.2%; U = 53.50, p = .013).

Examination of the within-task variations pointed towards notable changes in the relative rates of constraint-seeking questions. For the ASD group, the percentage of constraint-seeking questions consistently dropped for the within-set change and the memory rounds (ASD m1 = 60.4%, m2 = 58.1%, m3 = 53.3%), while the controls improved between blocks 1 and 2 before dropping on the memory rounds (TD m1 = 73.2%, m2 = 74.4%, m3 = 65.1%; block main effect: F(2, 44) = 4.598, p = .015, $\eta^2_p = .173$; group*block interaction effect: F(2, 44) = 2.347, p = .05, $\eta^2_p = .096$). No other significant block effects or group-by-block interaction effects were found, although again a group-by-verbal IQ interaction was indicated (constraint-seeking questions $\eta^2_p = .135$, hypothesis-seeking $\eta^2_p = .176$, pseudo-constraints...
\[\eta_p^2 = .297, \text{ redundant } \eta_p^2 = .111\], suggesting an uneven effect of verbal IQ on question types.

**iii) Comparing tasks**

A repeated measures MANOVA was used to compare question type rates on the two tasks. No significant effect of task or task-related interaction effects were observed, suggesting that task content overall did not make a big difference to the type of questions asked. The group differences seen in the standard and novel forms of the task were sufficient to produce a strongly significant multivariate effect of group (Pillai’s trace \(F = 5.299, p = .002, \eta_p^2 = .595\)), with univariate ANOVAs indicating this to be true of all question types apart from redundant questions (\(p = .508, \text{ all other p-values < .039}\)). It is worth noting however that group differences in question types were mostly reduced on the novel task as compared to the standard task. Across the two tasks no overall block-by-group interaction effects were observed, indicating that the superior rates of questioning seen in controls were relatively consistent. Any slight effects of within-task variations like those reported above were probably task-specific.

**Figure 11: Question types asked on the Robots task.**
In addition, group-by-VIQ interaction effects were observed for constraint-seeking ($F(1, 22) = 7.472, p = .006, \eta^2_p = .254$), hypothesis-testing ($F(1, 22) = 8.286, p = .005, \eta^2_p = .274$) and pseudo-constraint questions ($F(1, 22) = 4.676, p = .021, \eta^2_p = .175$) across the two tasks. As with the main measures of task success, this indicates a non-homogenous influence of verbal IQ in the two groups.

iv) Summary

The analysis of question types suggests that ASD participants consistently asked a lower percentage of questions which actually grouped and categorised the set, as shown in the less frequent use of constraint-seeking queries. Instead they were more likely to ask “guessing” questions (i.e. hypothesis-seeking questions), and, in some cases, were more likely to repeat previously asked questions, or use redundant questions. The consistency of these differences across both tasks again suggests a robust problem with asking effective questions, as would be predicted by a general difficulty with category and concept use.

However, as seen with the measures of task success, group differences were still to some degree lessened by the use of novel rather than standard stimuli, in that differences tended to be less significant overall on the Robots task. In addition, the influence of verbal IQ on how participants used their questions seems to be very important. Although consistent group differences were observed even once verbal IQ effects were controlled for, the presence of VIQ-by-age interaction effects suggests that the recruitment of verbal IQ is much more important for one group than it is for the other.

One way to examine this is to look at the content of questions used by participants, as it may be that verbally-driven questioning strategies were being used by one group but not another. In addition, the presence of such interaction effects suggests that further modelling of the VIQ influence is needed; not just of the general effect of verbal IQ, but group-specific contributions of verbal IQ and other covariates. The remaining analysis will firstly examine the content of questions used by each group, to explore the different questioning strategies used by participants. Following this, exploratory regression analyses will be used to probe further the influence of verbal IQ on task performance.

4. Question content

Participant questions were coded into the seven content categories of abstract, function, specific features (conceptual), colour, name, form and specific features (perceptual). (See p.41 for definitions of each category). Scores for each variable were heavily skewed and non-normally distributed, with strong
tendencies in each group towards particular question categories on each task. Because of this, non-parametric tests were used to compare the groups across the tasks and block variations. To address the inflation of error due to multiple-testing, the alpha rate was reduced to 0.01 for tests of significance.

\(i)\) Twenty Questions Task

The two groups asked questions of fairly similar content on the standard form of the task, but differed in the proportion of function and abstract questions asked. Neither group utilised perceptually-based criteria to any great degree when searching the Twenty Questions items (colour = 10.9%, form = 0.3%, name = 2.0%, specific perceptual features = 7.65%; see table 2 for question content rates for each group). The majority of questions asked referred to specific conceptual features of items (26.2%) or abstract groupings of items (34.8%). Comparisons of the two groups demonstrated differences approaching significance in the percentages of questions referring to abstract labels (ASD m = 30.3%, TD m = 39.7%; U = 51.50, \(p = .016\)), and the relative rates of function-based questions (ASD m = 16.5%, TD m = 6.4%, U = 55.50, \(p = .024\)). This indicated that ASD participants in general tended to ask fewer questions with abstract labels (i.e. referring to taxonomic categories or groupings), instead utilising more concrete, function-based questions in their search. Given the local bias in feature processing hypothesised to exist in ASD individuals, it is also interesting to note the lack of a significant difference in the percentage of conceptual or perceptual feature-questions asked; indeed, control participants actually tended to ask slightly more questions about specific features than ASD participants overall.

Examining question content across the trial blocks, it was clear that the within-trial variations had an effect on participants’ questioning strategies. For perceptual feature-based queries, ASD participants tended to ask a small but consistent percentage of questions across the 3 blocks (ASD m1 = 5.9%, m2 = 7.9%, m3 = 7.5%). Controls, in contrast, actually demonstrated a trend towards utilising significantly more questions initially (TD m1 = 13.1%, U = 55.50, \(p = 0.039\)), before dropping to similar levels as the ASD participants (TD m2 = 6.1%, m3 = 6.7%). Rates were higher in both group for questions relating to specific conceptual features, but the groups did not consistently differ to any significant degree across the trial variations (ASD m1 = 32.5%, m2 = 22.3%, m3 = 25.0%; TD m1 = 26.4%, m2 = 29.6%, m3 = 30.6%). This suggests that ASD participants were no more likely than controls to utilise a feature-based approach during the various stages of the standard task.

For abstract- and function based questions, the group differences appear to have been particularly exacerbated by the addition of memory demands. Slight differences in the rate of functional questions were present across the trial blocks but significantly so in the bonus rounds, where ASD participants
tended to use more function questions than controls (ASD m = 13.1%, TD m = 1.8%; U = 63.0, p = .023). At the same time, ASD participants dropped in their rate of abstract questions used (ASD m = 30.0%, TD m = 41.8.0%, U = 59.50, p = .039) on this trial block, after improving between blocks 1 and 2 (ASD m1 = 33.4%, m2 = 42.0%). (Controls, in contrast, scored between 41% and 43% across all three trial blocks). This suggests a possible shift towards function-based categorisation on the final bonus rounds in ASD individuals, although it is worth noting that abstract questions still formed the majority of queries for both groups.

ii) Robots Task

On the novel form of the task inquiries in both groups primarily referred to colour, form or specific perceptual features such as eyes or wheels. Most questions concerned colour (42.0%), followed by feature- (27.0%) and form-based (23.1%) questions. Name-based questions were asked with a similar frequency to the rate observed in the standard task (3.5%), but function- and abstract-based questions
were predictably lower. As the names denoted male and female robots, gender was available as an abstract criterion by which to sort and was used in a minority of cases (2.5%). The robots had no clear function, and this was reflected by the fact that no participants used function-based questions (0%). ASD participants asked slightly more colour-based questions than controls (ASD m = 47.6%, TD m = 37.3%) and tended to ask fewer feature-based questions (ASD m = 23.3%, TD m = 30.8%), but no significant differences were observed between the groups on any of the content categories.

Across the trial blocks some slight changes in question use were observed. Controls tended to drop their number of colour questions between blocks 1 and 2, while ASD participants tended to ask a roughly similar proportion of questions (ASD m1 = 51%, m2 = 52%; TD m1 = 38% , m2 = 35%; block 2 diff.: U = 53.00. p = .038). Given that the main change between blocks 1 and 2 on the robots task makes colour a primary sorting criterion (due to the variation in set structure), this may suggest that ASD participants were actually more sensitive to the set change than controls, in that they maintained a useful strategy rather than switching away from it. Rates for questions about form were similar for both groups across the trial variations (ASD m1 = 28.3%, m2 = 24.2%, m3 = 24.5%; TD m1 = 23.1%, m2 = 24.0%, m3 = 22.1%), and though controls generally asked more questions about specific perceptual features (ASD m1 = 23.1%, m2 = 24.7%, m3 = 23.6%; TD m1 = 30.3%, m2 = 32.2%, m3 = 29.4%) none of these differences were statistically significant. If anything, the variation in question content between individuals on the robots task was much greater than the variation between groups. (For example, on feature-based questions there was a mean difference of 7.49% between the two groups, but a standard deviation of 14.15%).

**iii) Comparing the tasks**

As has been already noted, the content of each task largely dictated the questions used by participants. While participants tended to search on predominantly perceptual criteria in the novel form of the task (colour, form and specific perceptual features), they used abstract labels, function, and specific conceptual features to sort the Twenty Questions items. This was reflected by a strong task effect on question content (F (1, 23) = 8.568, p = .008, \( \eta^2_p = .271 \)) that was observed in a repeated measures ANOVA on the two tasks. Elsewhere it is important to note that group differences a) were, for the most part, not significantly apparent on the Robots task, but were on the Twenty Questions task, and b) only significantly affected by within-trial variations on the standard form of the task.
iv) Summary

The content of questions used by participants varied considerably across individuals and with the type of stimuli to be searched through. On the novel form of the task, this was manifested by a general lack of group differences, and an attention to colour-, form- and specific feature-based questions. On the classic task, ASD participants tended to use more function-based questions than controls, and referred to fewer abstract labels – particularly during the bonus rounds. This would seem to indicate a tendency towards more concrete verbal strategies in the ASD group.

Notable also was the lack of any major differences for questions relating to specific features, with both groups using conceptual or perceptual details to a similar degree. Far from ASD individuals demonstrating a perceptually-driven bias, or focus on local details, controls were the group that tended to use slightly more questions of this sort.

The pattern of results suggests that group preferences or biases in strategy largely depend on the demands of the task. It also indicates a sensitivity and consistency in both groups in their response to the set in front of them. Largely different strategies were used across the two tasks, which makes it hard to compare true preference in sorting, but it is important to note that the stimuli content did not necessitate these strategies. Participants could have sorted by colour, shape or name on the Twenty Questions task, but they didn’t; and they could have sorted the robots based on an abstract label (gender), but they didn’t. This was true of both groups, but the degree to which they adopted and maintained suitable search criteria varied with the difficulty of the task.

5. Post-hoc regression analyses

Backwards-method linear regression modelling was used to explore the contribution of verbal IQ to task performance. The outcome variables of mean question quality, percentage of constraint-seeking questions used and percentage of abstract questions used (standard task) were selected for modelling. These outcomes in particular were selected for their demonstration of clear differences and, in the former two cases at least, apparently complex contributions of verbal IQ. Predictors included for analysis were verbal IQ, non-verbal IQ and age. (Non-verbal IQ was included to test the adequacy of the model; if the original correlational analyses were correct, non-verbal IQ should be eliminated in the regression model).
i) Whole group modelling

To test the modelling method the two groups were initially analysed as a whole. Regression analysis on mean question quality returned a significant model based on the variables of verbal IQ and age ($R^2 = .706$, $F (2, 25) = 12.416$, $p = .001$; VIQ: standardised $\beta = .674$, $t = 4.706$, $p = .001$; age: stan.$\beta = .331$, $t = 2.311$, $p = .029$). As expected, non-verbal IQ was excluded from the model for a non-significant contribution to the model (stan.$\beta = .142$, $t = .994$, $p = .330$), suggesting no effect of non-verbal IQ on this outcome. Very similar models were returned for the percentage of constraint-seeking questions used ($R^2 = .800$, $F (3, 24) = 14.244$, $p = .0001$) and abstract questions used ($R^2 = .756$, $F (2, 25) = 16.688$, $p = .0001$). These findings in general support the inclusion of age and VIQ as primary covariates in the between-groups analyses.

ii) ASD model

Scores for ASD participants were modelled first. On mean question quality a similar model including VIQ and age was again returned ($R^2 = .735$, $F (2, 11) = 6.446$, $p = .014$), with NVIQ dropping out (stan.$\beta = -.016$, $t = -.072$, $p = .944$). It is notable that the relative contribution of VIQ (stan.$\beta = .791$, $t = 3.541$, $p = .005$) was almost double that of age; (stan. $\beta = .438$, $t = 1.963$, $p = .075$), indeed, the individual contribution of age to the model did not actually reach significance despite its retention. This indicates that verbal IQ played a particularly important role on this outcome for ASD participants.

For constraint-seeking questions the specific importance of verbal IQ was confirmed: the regression analysis returned a model which only retained VIQ (model: $R^2 = .708$, $F (1, 12) = 12.084$, $p = .005$; VIQ: stan. $\beta = .708$, $t = 3.476$, $p = .005$), while rejecting age (stan. $\beta = .337$, $t = 1.613$, $p = .135$) and NVIQ factors (stan. $\beta = .197$, $t = 0.995$, $p = .343$). For abstract-questions a more even model was returned (model: $R^2 = .625$, $F (2, 11) = 3.522$, $p = .066$), with VIQ and age making similar contributions to scores (VIQ: stan. $\beta = .595$, $t = 2.311$, $p = .041$; age: stan. $\beta = .545$, $t = 2.122$, $p = .057$).

iii) Typically-developing model

The modelling of scores for control participants yielded a notably contrasting set of predictor effects. On mean question quality a model was retained which only retained age (model: $R^2 = .567$, $F (1, 12) = 5.676$, $p = .035$; age: stan. $\beta = .567$, $t = 2.382$, $p = .035$). Verbal IQ in fact made a smaller contribution to control questions ($R^2 = .406$, $F (1, 12) = 4.614$, $p = .056$; VIQ: stan. $\beta = .406$, $t = 2.011$, $p = .056$).

---

3 Non-verbal IQ was actually retained as a predictor in the model for constraint-seeking questions, but the individual contribution of NVIQ was still low (stan. $\beta = .248$) and did not reach significance ($t = 2.011$, $p = .056$).
Figure 13 Group-specific predictors of question quality for a) VIQ and b) Age.

a) Verbal IQ vs. Mean Question Quality

b) Age (months) vs. Mean Question Quality
scores (stan. β = .163, t = 0.598, p = .563) than non-verbal IQ (stan. β = .207, t = 0.775, p = .456), with neither contributing to the model in a significant manner. Very similar models were returned for constraint-seeking questions ($R^2 = .694$, $F(1, 12) = 11.131$, $p = .006$; age: stan. β = .694, t = 3.336, p = .006) and abstract questions ($R^2 = .812$, $F(1, 12) = 23.166$, $p = .0001$; age: stan. β = .812, t = 4.813, p = .001), with no significant influences of verbal IQ found.

Comparing the two groups, this suggests that verbal IQ is only making a big difference to scores in the ASD group, and not the control group. In contrast, the controls demonstrate a predictable influence of age; older participants will tend to ask better questions, establishing more constraints and using more abstract questions. While age and VIQ effects would be expected to influence scores on a task with considerable language demands, the uneven nature of these effects in the two groups is considerable and has important implications for how the results should be interpreted.
DISCUSSION

The aim of the project was to explore, in as comprehensive a manner as possible, the various factors affecting category use in ASD during verbal inquiry tasks. This was done in order to provide a window on potential expressions of an autistic cognitive style, and to examine to what extent such expressions may be related to subtle and underlying impairments in category processing. Building on the classic Twenty Questions format, two tasks were developed which varied the conceptual and perceptual content of various sets to be searched through. In addition, conditions were added to each task to vary the level of flexibility and memory resources required for success, and the contributory effects of verbal IQ and age were measured.

The main question concerned the role of stimulus content in spontaneous category use, and this will be considered first. Following this, the roles of verbal IQ and age will be examined because of their implications for comparing performance in the two groups. The effects of trial variations relating to memory and flexibility, and the implications of question preference for an ASD cognitive style will then be explored in turn, before a discussion of the main limitations of the project and its implications for future research. Lastly, an attempt will be made to outline how the findings of the present study sit within current understanding of the underlying neuroscience of autistic spectrum disorders.

Twenty Questions: an example of a content-specific problem in category use?

The primary prediction of the project was that previously observed deficiencies in category use in ASD individuals would effectively disappear when faced with a novel set of items; specifically, a set of items which could be categorised and searched through without requiring an understanding of abstract and taxonomic hierarchies. To a large extent this was true, in that ASD participants performed better on most outcome measures when searching through the Robots task, in comparison to the standard Twenty Questions items. But the results showed that this was also true for the controls, suggesting that the Robots task was generally easier for both groups. Furthermore, analysis of the quality and kind of questions used suggested more pervasive differences between the two groups.

Examination of the specific outcome variables further elucidates the nature of these differences. Group differences for the percentage of trials completed and the number of questions taken on each trial tended to be larger in the Twenty Questions task than the Robots task, with control participants tending to identify more targets while using fewer questions. As was predicted, where there were differences in
success between the two groups, they were only significant for the standard Twenty Questions task. This implies that ASD performance was either in some way facilitated on the robots task, or selectively impaired on the standard items. What it at least demonstrates is a level of sensitivity in category use by ASD individuals: where predominantly perceptual categorisation was required, ASD performance was high and similar to controls, but for conceptual categorisation, overall success significantly dropped. This sensitivity to task content is consistent with the findings of Shulman et al (1995) and Ropar & Peebles (2007), and stands in contrast to accounts that propose a more general problem with spontaneous category use or formation in ASD (Minshew et al, 2002).

However, the analysis of mean question quality revealed a much more consistent difference between the two groups: ASD participants asked questions on both tasks which tended to eliminate fewer items at a time than controls. This difference was present at each stage of either task and was significant even after age and verbal IQ effects had been controlled for. If each question is taken to be establishing a sub-category or constraint by which the set can be sub-divided, this suggests that ASD participants were actually forming smaller categories than controls on average. In addition, it indicates a problem with inquiry-based games which is seemingly not specific to particular types of content, affecting both simple and complex forms of category use for ASD individuals.

This picture was supported by the analysis of question types used. ASD participants, across the two tasks, tended to ask fewer constraint-seeking questions than control participants. As constraint-seeking questions are required to establish even the most minimal (i.e. two item) category, this indicates a lower rate of category-based questioning in the clinical group. Instead, ASD participants tended to ask more hypothesis-testing questions, thus only eliminating one item at a time. These differences, like the outcomes for success, did appear to be worse on the standard form of the task, but nevertheless their presence across both tasks suggests an underlying contrast in the kinds of questions each group were using. Results for pseudo-constraint seeking questions were mixed, in that ASD participants tended to ask more on the Twenty Questions task, while controls asked more on the Robots task.

The combination of results from the measures of task completion and question types suggest a similar level of competence between the two groups, but a contrasting level of efficiency in questioning strategy. The fact that ASD participants managed to find the target item on the large majority of trials and only took more questions under specific conditions suggests that these differences were not due to a lack of understanding. Rather, the use of questions of poorer quality suggests a failure to utilise categories effectively in searching the set. Although task success was better in the present study, and seemingly sensitive to task content, the findings on question types are broadly similar to the differing rates observed

In this sense, what the present results are broadly replicating is the manner in which ASD individuals are completing the Twenty Questions task, if not the overall outcome. The apparent sensitivity in success, particularly concerning conceptual content, refines the picture originally proposed by Minshew et al., in that category use, like category performance before it, would seem to depend on the nature of the stimuli to be sorted (Shulman et al, 1995; Ropar & Peebles, 2007). But the irregular focus on individual cases and smaller categories matches the abnormal strategies recorded by Minshew et al; here problems with spontaneous category use or concept formation would seem to be expressed in terms of a more narrow focus in grouping. The sensitivity to conceptual content affects the outcome of abnormal category use for ASD individuals, while the efficiency and scope of that categorisation seems more generally abnormal.

The roles of verbal IQ and age

The scores for ASD participants, in terms of the differing question types used, also generally resembled the strategies that would be deployed by younger children (Mosher & Hornsby, 1966; Siegler 1977). While the mean age in the present study was 13 years, ASD participants’ use of a relatively smaller number of constraint-seeking questions alongside an increased number of hypothesis-testing questions most closely resembles the strategies observed when typically-developing 8 year-old children attempt the task (Mosher & Hornsby, 1966).

It would be a mistake, though, to simply interpret these results purely as evidence of an immature or underdeveloped searching strategy. The evidence from the analyses of covariance and subsequent regression modelling indicated significant roles for both age and verbal IQ, but importantly the influence of these factors on performance appears to have been different for each group. Specifically, verbal IQ appears to have supported performance in ASD individuals but not control participants, a finding only partly in line with the hypothesised role of VIQ. In contrast, only controls demonstrated the expected age effects on task performance.

How should this apparent difference between the groups be interpreted? Firstly, given the post-hoc nature of the regression analysis and the relatively small sample size, it is worth proceeding with caution. The group-specific influence of verbal IQ might be taken to suggest that the tasks were too easy for the controls, but not for ASD participants. Given that verbal IQ was significantly higher in the control group, and that overall performance in this group was high, this is certainly a plausible possibility – it might be
that variable performance and variable VIQ were sufficient for a relationship to be observed in the ASD group, but not the control group. The specific covariance of age with control performance suggests though that participants in this group were not performing at ceiling, because if they were, one would expect the variance in scores to be insufficient to covary with age. In addition, the variance in VIQ scores for controls was comparable to the ASD group (roughly 2:3). Arguably, this indicates that their scores were sufficiently variable and free to vary to the extent that they could have correlated with verbal IQ, but nevertheless this did not happen. Rather, it would seem that the respective influences of verbal IQ and age for each group were genuinely contrasting, supporting the style and success of participants in relatively dissociable ways.

If this difference is taken to be indicating a genuine contrast, it could have important implications in a range of ways. Technically it could be seen to violate one the assumptions of an analysis of covariance model, as the covariates in this case would not seem to be influencing each group in a homogenous way. For the present study this is not immediately problematic; the covariates were identified in a methodologically systematic manner, and the group-specific effect of verbal IQ only emerged once a range of different variables had been analysed. (It is important to note that this effect was not present on certain variables, such as the percentage of trials completed and the mean number of questions used). In more pronounced cases, though, this would have serious implications for the reliability of a study seeking to make comparisons between groups, and in a wider sense, it also questions the viability of classic between-groups designs used in autism research. Almost all studies of specific neuropsychological processes in autism seek to match clinical and control participants on IQ capabilities and age, or, if this is not possible, will include an IQ score in a model of covariance (Mottron, 2004). The assumption behind this is that these factors will influence performance consistently in both groups, so if they are controlled for, “genuine” group differences, specific to diagnosis, will be observed. However, controlling for such factors will be affecting groups unevenly, if they are supporting performance in one group but not the other. In the present case, verbal IQ is supporting ASD participants in their task performance, but is not important to the controls, who seem to exhibit an age-dependent profile or performance. This may also be the case for other tasks – for example, it could be that two groups are matched on non-verbal IQ for a working memory task, and the influence of NVIQ is not further examined, but inspection of this factor would reveal group-specific effects on performance. (Jarrold & Brock, 2004, have previously argued for the use of covariate modelling along with matching to fully inspect the influence of IQ on group performance).

If this was the case it would support the idea of a diverging developmental trajectory for autistic spectrum
disorders. The lack of a consistent effect of age in the ASD group suggests that it didn’t matter how old a participant was to their level of performance; instead, verbal IQ resources could be seen to be “bootstrapping” performance. Controls, on the other hand, performed closer to what would be expected within a model of progressive developmental stages, utilising more constraint-seeking questions and more abstract questions the older they were. Bootstrapping roles of verbal IQ, and specifically language, have been seen in examples from other areas; for example, while younger children with ASD often fail first- and second-order false belief tasks (Baron-Cohen et al, 1985), there is evidence to suggest that their language skills predict to what extent they can pass false-belief tasks at later time points (Steele et al, 2003; Lind & Bowler, 2009). There it is suggested that language abilities are being used to “hack” false-belief understanding by ASD individuals, in the absence of an intuitive understanding of mental states (Lind & Bowler, 2009); it might be that something similar is happening with tasks used here.

In terms of the present study then, the complex influence of verbal IQ would seem to imply that the differences observed between the two groups are not simply down to differences in IQ (because if this was the case, it would also predict control performance). Rather, there appears to be some persistent difference in the scope of category grouping, which ASD individuals may be able to counteract if they have the requisite verbal IQ resources.

A wider point can also be made. Given the differences in language and socio-emotional development in autism, and given the potentially complex and group-specific influence of factors like verbal IQ, arguably a differing developmental path should at least be looked for in experimental designs, if not expected by experimenters. Controlling and matching for IQ and age factors is meaningless if the developmental paths of individuals in each group are different, and as such experimental designs, as a minimum, should be examining closely the respective effects of age and IQ for both groups.

The roles of flexibility and memory

Aside from overall group differences, the effects of within-set changes in set structure and memory demands were examined on both tasks. The significance of the within-set change was to test conceptual flexibility; participants would have to adapt the focus of some of their questions in order to search the set efficiently following the introduction of new items. Although perseveration in questioning strategies was hypothesised for the ASD group, largely this was not evident. In fact, sometimes their change in questioning strategy was more sensitive to the set change than controls, as seen in the successful use colour questions on the Robots task. However on the harder of the two tasks, the Twenty Questions task,
the quality and success of their questioning seemed to be disrupted by the within-set change (while controls significantly improved). In this sense, the set-change on the two tasks used in the present study failed to replicate the previous findings of perseveration problems which have been observed on the WCST (Hill, 2004; Kenworthy et al, 2008), but still demonstrated elements of conceptual inflexibility proposed by Minshew and colleagues (Minshew et al, 2002). The only other point to add here is that this possible inflexibility is specifically evident on the Twenty Questions items. This suggests again that more abstract groupings will tend to tease out difficulties in ASD category use which might not otherwise be seen for more concrete or perceptually based groupings.

The findings on the memory variations of the tasks indicate a similar situation. By prohibiting the physical elimination of items, participants were required to hold previous questions in mind while simultaneously looking ahead to the next useful question in the search. Previous research on sequencing (McGonigle-Chalmers et al, 2008) and working memory (Goldberg et al, 2005) suggested that this could have been selectively impaired in ASD participants. Although it tended to reduce question quality, largely the addition of memory demands did not have a major effect on task performance for either group. This rules out the potential explanation that poor performance by ASD individuals on previous versions of the Twenty Questions task were due to the various memory demands involved. Rather, the addition of such demands, where they had an effect, would seem to have exacerbated group differences that were present anyway. For example, mean question quality was generally higher for control participants throughout each task, but group differences were among their largest in the case of memory condition. If anything this arguably suggests a contributory influence of working memory difficulties in category use, rather than a more primary explanatory role.

The addition of memory demands had a slightly more interesting implication for the differing question strategies used by participants, specifically on the standard form of the task. For controls, their response to the addition of memory demands was to maintain a relatively high proportion of abstract questions in the final trial block (42%). ASD participants, on the other hand, dropped their rate of abstract questioning (-11%), while maintaining a significantly higher proportion of function-based questions compared to controls. This suggests that the challenge of the bonus rounds prompted ASD participants to revert to a preferred alternative strategy or, in some way made the maintenance of an abstract questioning strategy too difficult. These possibilities are not strictly exclusive – indeed, the most likely explanation concerns a combination of both. But they will be considered in turn because they provide, in a nutshell, the key contrast in ASD cognitive research: should we interpret autistic behaviours as merely expressions of a different style in processing, or can we better explain them in terms of specific impairment.
An aversion to the abstract: the autistic cognitive style at work?

One interpretation of the response to memory demands is that ASD participants reverted to a preferred cognitive style in their questioning. It may simply be that they preferred not to think in the abstract, and that the bonus rounds provided the best outlet for how they would ordinarily go about acquiring, retaining and using information to solve the problem. The scores for question content at other points of the task provide some support for this: throughout the Twenty Questions task ASD participants asked a larger proportion of function-based questions, and on average they tended to ask fewer abstract questions than controls. This observed tendency in questioning is consistent with Ropar & Mitchell’s (2007) findings concerning sorting preferences, where ASD individuals tended to favour concrete over abstract questioning strategies. In addition, the tendency towards functional questions in particular fits with Baron-Cohen’s concept of the “systemising” cognitive style (Baron-Cohen, 2002), in that the mechanics and purpose of individual items seem to have been selected as a key searching strategy. This proposal is not inconsistent with the idea that using abstract terms might be in some way more challenging for individuals with autism, but it does provide a different emphasis when explaining the results in the memory condition. Namely, it suggests that ASD participants were in a sense “reverting to type” under the additional pressure of the bonus rounds; positively focussing on the concrete, rather than simply having problems with the abstract.

However, the sense of a pervasive, all-encompassing bias towards the non-abstract in ASD participants is not necessarily held up in the present results. Firstly, at their highest point the use of functional questions was never above 18% in ASD participants, while the rate of abstract questions never dropped below 30%. The largest proportion of their questions tended to be abstract questions, suggesting a recognition and adoption among ASD participants of effectiveness of such a strategy. The ASD group may have used more functional questions than controls, but this couldn’t be said to characterise their overall approach to the task.

But what about other strategies? Are there other examples in the data which demonstrate a more consistent bias towards the concrete, or the specific? The “systemising” cognitive style is proposed to include a focus on local details (Baron-Cohen, 2006), as are other accounts of biases towards local processing in perception and attention (Mottron et al, 2006; Happé & Frith, 2006). Conceivably, a move from the abstract could have resulted from a greater tendency to focus in on local features or specific

---

12 Researchers such as Laurent Mottron and Michelle Dawson (Mottron et al, 2006; Dawson et al, 2008) have done much to advance this sort of view in the explanation of autistic cognitive differences.
perceptual properties, as is often alluded to in theories of autism (Frith, 2003). But the results from both tasks did not provide evidence for this. On the Robots task, the content of questions asked was largely similar for both groups; using colour, form and specific perceptual features to an equal degree. A local feature bias could have been expressed here by ASD participants, but wasn’t; indeed, at one stage controls actually tended to ask more questions about specific features. On the Twenty Questions task, a perceptually-driven strategy could have been selected, but importantly ASD participants could and did utilise effective questions relating to the conceptual properties of the items at hand. Within this, it would also have been possible for ASD individuals to exhibit some form of local bias via a focus on questions about specific conceptual features, but again no group differences were observed. As stated above, the group differences were seen in the relative rates of abstract and function-based questions.

So where there were group differences in strategy, they were not expressed as a bias towards local features or perceptually-based categorisation. Nor could the overall ASD strategy be conceptualised as consistently functional or concrete, given the proportion of abstract questions used. In fact, the only way in which ASD participants could be said to exhibit a bias or preference towards the local is in the quality and type of their questions, where they tended to ask about smaller categories and often referred to individual exemplars. If there is a local preference it would seem be on the level of sets rather than their exemplars; a focus on the individual or small number of individuals within a set, rather than the specific features of an individual. It may be that this is an instance of a cognitive style at work, although it should be noted that it is sub-optimal to utilise such a style on the present tasks. Also, as previously discussed, the evidence of a specific role for verbal IQ suggests that outside factors will contribute to and ameliorate this tendency. In this sense, the expression of an autistic cognitive style would appear to be relatively flexible. ASD participants may have asked questions of lower quality, but this varied depending on the individual. Similarly, they may have tended to ask slightly different types of questions, but this was sensitive to the task and stimuli at hand.

The role of language in that expression also provides a speculative explanation as to why abstract groupings may become more difficult to maintain when working memory demands are increased; an explanation concerning subtle impairment, rather than simply preference or overall style. Firstly, it is instructive to consider what an abstract questioning strategy requires. To group items according to abstract categories like “mammals” or “cutlery” requires an ability to form sets based on exemplar characteristics which are not necessarily evident in the perceptual array: they depend on knowledge of a range of semantic qualities that the items in question possess, and an understanding of the relations between such items. But more specifically, they also generally require knowledge of the label itself, the
abstract term which exclusively defines that set of items. If that label is not a readily accessible part of an individual’s vocabulary, then the abstract set cannot easily be named. Instead, features, functions or general characteristics must be specified (for example, one may ask “does it lay eggs” rather than “is it a reptile?”), and once this occurs the strategy inevitably becomes less holistic. Without the actual label, one can’t easily be as abstract.

As previously discussed, the general concepts that such abstract terms refer to are proposed to occupy a socially negotiated and conventional “concept space”, and it is not clear to what extent individuals with autism actually access such a “space” (Hobson, 2009). More concretely, we know that ASD individuals struggle with social interaction and can exhibit considerable language delay (Wing & Gould, 1979). The ordinary process by which an individual comes to understand and use such abstract terms, with confidence of their extension, must in some way be affected individuals with autism. As such, it is at least plausible to suggest that simply using abstract labels poses more of a challenge for ASD individuals. With a greater verbal IQ this can be countered, as seen in the way that VIQ supported the use of abstract question strategies in the ASD group. But in the face of additional verbal demands (like having to remember one’s previous questions while preparing new ones), it might be that using abstract labels is the first thing to fall by the wayside.

To summarise, the evidence from performance across the two tasks supports a varied but informative picture of category use and questioning strategy in ASD individuals. As predicted, some differences were only observed for the trickier, standard form of the two tasks, suggesting content-specific difficulties with category use. This was also the case for changes in success, efficiency and questioning strategy as a result of within-set alterations and the addition of memory demands; where group differences existed, the existed on the classic form of the task only. However, the overall question quality for ASD individuals was consistently low, suggesting some enduring abnormality in the way in which participants formed categories to use for search. This, coupled with the seemingly complex role of verbal IQ, supports a subtly irregular cognitive style, although one whose expression is sensitive to the stimuli and task demands at hand. ASD participants in this case were largely achieving the same goals, but were reaching them via a subtly different and irregular route.

**Study limitations and directions for future research**

This pattern of findings is consistent with a range of recent studies on the neuro-anatomical bases of autistic spectrum disorders. The future development of neuropsychological studies of autism must be
conducted with one eye on this area, but before these aspects are discussed, there are a number of limitations to consider concerning the present study.

Firstly, as had already been referred to, it might be suggested that the Robots task was simply too easy to draw out any consistent group differences. Certainly, the task as it is, and the original Guess Who? game, tend to be aimed at younger children than those studied here, and the reduction of items from over 40 to 24 (necessitated by the structure of the Guess Who board) made achieving the target simpler. However, these criticisms apply to the structure of both tasks equally, and as such the variation in task performance must be down to differences in the task content. And this wouldn’t necessarily be a bad thing; if the Robots task was easier because of the robots, then this in part was what was aimed for; the difference in performance must have been due to differences in the nature of the items and the relations between them, i.e. the set structure.

So which of these possibilities explain the difference in performance? The items themselves differed in their depiction of either common or novel objects. In addition, they differed in the extent to which they could be grouped according to conventional and abstract hierarchies; while the Twenty Questions items could be categorised according to conceptually standard taxonomy, the Robots could only be grouped based on the properties they possessed as presented. Both of these differences were intentional, as the manipulation was designed to measure to what extent such factors affect ASD performance. One consequence of this though is that the set structure could not be made exactly analogous across the two tasks. The abstract groupings that exist for the Twenty Questions items also creates a hierarchy of sets within sets, where a participant can move from the general (living vs. non-living) through to the particular (e.g. cutlery). This creates a natural order of search. In contrast, though the Robot items were assembled to create a set with structure that could be searched through in a systematic manner, the properties of the robots (colours, shapes, eyes, feet) did not create a logical structure of sets within sets in a fully analogous way. Therefore, it is conceivable that ASD participants were selectively impaired on the Twenty Questions items not because of the nature of the items themselves, but the logical hierarchy that can be created with them. This is not necessarily a problem in terms of adding to understanding of the theoretical debate\textsuperscript{13}, but it does reveal a limitation in the design: it doesn’t tell us whether it is the abstract,

\textsuperscript{13} As an aside, it is not entirely clear how a logical, abstract hierarchy could be created from completely novel set of items. Klinger & Dawson’s novel animal stimuli (e.g. the “mips” and “sops”; Klinger & Dawson, 2001) could be denoted with category labels, but without a taxonomic fiction about their origins or external features, they are still only grouped on their perceptual qualities. They could be grouped according to particular forms of inferred biological motion (i.e. some fly, some swim), but these groupings would merely resemble pre-existing groupings of natural kinds (flying animals, swimming animals).
representational nature of the items themselves, or the logical set structure they create, which is affecting ASD performance.

A second point to note about set structure in this regard is that the novel items could be seen as creating a subtly different form of category entity. The Twenty Questions items each represent a prototype of a relatively basic category; the item CAT represents the set of particulars that are counted as “cats”, and questions are generally referring to conventional properties that cats possess. In essence, they are a token that represents a set. For the Robots, on the other hand, each item is part of an overall basic category (“robots”) and each possesses identity; all the Robots have proper names, seemingly denoting individual status. As such, when participants are searching through the sets on the two tasks, they are being asked to perform subtly different forms of categorisation. In the standard form of the task they are faced with basic kinds, and being asked to subdivide them based on superordinate groupings (e.g. mammals, vehicles) – the groupings themselves are inherently more abstract than the items themselves. For the robots, on the other hand, participants are being asked to categorise based on subordinate groupings – blue robots, one-eyed robots, square 2-legged robots. Once again, it may be this difference which is prompting the abnormalities in ASD performance – another form of abstractness which ASD individuals are having difficulty dealing with.

The present study arguably establishes the principle that category use in ASD will depend on the nature of the stimuli, but future work will have to focus on separating out these different factors in the processing of set relations. One method of teasing apart the different impact of abstract factors would be to systematically vary the strength of conceptual relations between items: assessing the impact of “relatedness” between items has recently been used successfully in free recall research (Bowler et al, 2008a) and could easily be applied to verbal enquiry tasks. For the novel stimuli, other alternatives could be to create an artificial taxonomy via the addition of other kinds of items, such as aliens, or the levels or perceptual dissimilarity between items could be varied to a greater extent. These options would allow for a comparison of basic- and super-ordinate class categorisation across item groupings while also making the task of grouping more challenging for participants.

Another potential limitation concerns the assessment of verbal IQ and the interpretation of task results in the light of VIQ differences. As noted, the groups differed significantly on VIQ, and a number of outcome measures covaried with VIQ in participant performance. The use of VIQ as a covariate controls for its effects to a certain extent, and the regression modelling indicated that it probably wasn’t the case that VIQ could explain all the group differences (as it was not related to success for controls). However, the sample size of the study is relatively small, meaning that these methods can’t guarantee that all relevant effects
have been identified. To improve on this, a firmer way of demonstrating “pure” group differences would be to have both groups at a more similar level of verbal IQ. Full matching is not necessarily required, given that a) verbal IQ will be strongly related to ASD diagnostic factors and b) VIQ is not necessarily exerting the same influence on each group anyway, but by minimising VIQ differences one can examine more closely specific trends in ASD performance that exist even when they have the VIQ resources to complete the task (an example here being the persistent difference in mean question quality). Note that this is not the same as saying that these differences are the only true abnormalities in ASD performance on such tasks, as eliminating VIQ differences is to eliminate a key aspect of ASD functioning. (See Jarrold & Brock, 2004, and Mottron, 2004, for a discussion of the best ways to match ASD participants with controls in between-groups designs).

A further way of exploring the role of verbal IQ would be to compare ASD performance with individuals from other developmental disability groups, in particular those with language impairments and developmental dyspraxias. If these tasks are really just measures of linguistic functioning or verbalisable knowledge, then one would expect similar patterns of performance. However, if there really is a problem with spontaneous category use in ASD, or a cognitive style which can create such difficulties, then one would expect diverging profiles among different clinical groups. Speculatively, it seems plausible to suggest that these groups may perform similarly in general measures of task success but may differ in the focus and strategy of their category questions.

A final question about the design concerns the within-trial variations. A balance had to be found between including as many variations as possible, while at the same time not making the tasks too long and over-testing the participants. In an attempt to achieve this, the respective trial blocks consisted of a maximum of three trials, and, in the memory condition, two trials. Given that the standard form of the task typically consists of four trials only (Mosher & Hornsby, 1966; Delis et al, 2001; Minshew et al, 2002), it might be suggested that the trial blocks in the present study were too short to clearly assess participant performance in a reliable way. (Although this criticism could not be made of the tasks as a whole, given that each task contained eight trials in total).

In addition, the order of the trial blocks may have made a difference to participant performance, specifically in the sense that participants had already had the practice of six trials before attempting the memory condition. The memory condition was placed as a “bonus round” at the end of the task because it was expected to be harder for some of the weaker participants and might have deterred participants from playing if it was included at the start of the task, so to a certain extent the practice of doing other trials first was always acknowledged. Nevertheless, the potential effect of the memory condition may have been
masked by practice effects. Because of this, and the relatively short number of trials in each block, a possible replication in future research would be to vary these conditions in separate sessions and with proper randomisation of condition order. This would allow a more in-depth examination of the effect of such variations, building on the indications of possible content-specific effects seen in the present study.

However, despite these limitations, the overall structure of the task has arguably yielded a rich and informative set of results concerning ASD category use. The use of multiple measures to assess success, question types and question strategy has highlighted varying levels of categorisation success alongside more persistent differences in question quality, with indications of style-based irregularities. This level of analysis has not been applied before to the Twenty Questions task, and if the main outcomes of previous studies had simply been used (for example, Minshew et al, 2002, only used the percentage of constraint-seeking questions) then the more detailed picture of category processing in ASD would have been missed.

**Categorisation in the brain: irregular utilisation of neural networks?**

The findings in the present study would seem to fit with a current and prevailing trend in the understanding of the brain functioning in autism. Previously what might be termed the “deficit model” led to an approach which effectively targeted specific areas of the brain for signs of functional or structural abnormality. For example, Baron-Cohen’s findings concerning ASD difficulties on false-belief and mental-state reading tasks (Baron-Cohen et al, 1985, 1989; Joliffe & Baron-Cohen, 1997) contributed to a focus on the amygdala and its surrounding circuitry (e.g. Ashwin et al, 2007; Pierce et al, 2001). However, module-specific explanations have always been limited in the same way that more general cognitive theories in autism have been, in that the observation of focal problems in brain functioning do not easily explain the range of difficulties seen in autism.

In more recent years, findings from neuropsychology and neuroscience have begun to focus more on the functioning of networks of brain areas in autism, rather than isolated brain centres. Advancements in the measurement of signal dependencies in functional MRI and MEG have allowed for the examination of how brain areas “talk” to each other during activation (David et al, 2004), while the development of fractional anisotropy techniques in DTI have shed light on the specific physical structure of connecting pathways within the brain (Barnea-Goraly et al, 2005; Keller et al, 2007). The use of these techniques has allowed an increasing number of studies to explore how brain connectivity, in a general sense, may be abnormal in autism, and that it is this, rather than a specific modular problem, which may draw together the difficulties seen across the spectrums of severity and symptomatology (Belmonte et al, 2004; Rippon
et al, 2007).

Where the findings of the present study come in is in highlighting the differing route towards a specific goal that ASD individuals seem to take. In this case, ASD individuals were generally successful in their search for targets, but demonstrated a sensitivity to stimuli that required more abstract, complex forms of categorisation. This implies that such stimuli place a differential or disproportionate cost on ASD processing, possibly because of something about the style or bias they exhibit in creating categories and searching the set.

In the brain, this conceivably would be expressed firstly by abnormal activation of areas outside of the typical networks for that operation. Secondly, weaker associations between active brain regions during a task might also be observed, as the maintenance of synchrony between different areas is thought to be required for holding in mind holistic representations and sets (Tallon-Baudry & Bertrand, 1999). Lastly, evidence of strong hyper-activation in specific brain areas might also be expected, driving a local focus in sorting through the over-excitation of local neuronal assemblies (Courchesne & Pierce, 2005).

Little neuroimaging research has been done specifically on visual categorisation tasks in autism (see Gaffrey et al, 2007, for an example of lexical categorisation being studied), but studies have demonstrated all three of the above characteristics during related tasks. During visual categorisation of faces and non-face objects, ASD individuals have demonstrated irregular activation of brain networks outside of the fusiform face area (Pierce et al, 2001) and utilisation of differing networks to controls during visual imagery tasks (Kana et al, 2006). Within networks, reduced functional connectivity (usually measured by assessing the correlation in activation between two brain areas) has also been observed during face-based working memory tasks (Koshino et al, 2008) sentence comprehension tasks (Just et al, 2004), visual illusion tasks (Brown et al, 2005) and Tower-of-Hanoi style tasks (Just et al, 2007). Demonstration of increased local connectivity or low-level over-excitation has not been easy to demonstrate in functional studies (mainly due to the insufficient spatial resolution of current imaging techniques; Logothetis, 2008) but structural studies have indicated evidence of abnormally dense packets of neurons on a local level (Courchesne & Pierce, 2005; Casanova et al, 2002), supporting the idea of an overly local focus in some aspects of processing.
It is important to emphasise that in a number of the above studies, abnormal use of networks by ASD individuals had been observed alongside relatively minor behavioural differences on the cognitive tasks used. This supports the overall idea of ASD individuals achieving the same goals, but having to utilise alternative routes towards success. To a certain extent this is also analogous to the way in which the autistic cortex is thought to have developed; Matthew Belmonte has referred to the symptoms of autism as “the developmental reaction of a normal human mind to abnormal neural hardware” (Belmonte et al, 2004, p.649). The combination of the above findings, and the findings of the present study, suggests that ASD individuals\textsuperscript{15} are attempting tasks in a rational way, are mostly completing cognitive tasks to a normal degree, but are always battling with abnormal constraints. In this case, that irregularity in processing seems to be some form of local focus at the level of sets.

\textsuperscript{14} Belmonte et al (2004) have actually proposed that such a local focus would also cause the more global problems in brain areas communicating, as dysfunction and over-excitation of local networks would lead to a poor signal-to-noise ratio in synchronisation with other neuronal assemblies.

\textsuperscript{15} This does of course only refer to high-functioning individuals with autism or Asperger Syndrome. While the vast majority of neuropsychological research in ASD has been conducted only with high-functioning individuals (thought to be approximately 75%; Mottron, 2004), functional neuroimaging research has been conducted almost exclusively with this group, and those structural studies which have included lower-functioning children have usually had to do so under sedation. In the case of language-heavy tasks, such as those used in the present study, it is not clear how such an HFA-bias can be addressed, but there is a definite need and, to a certain extent, an opportunity for more non-verbal research to be conducted with lower-functioning individuals in order to proffer a more representative picture of ASD functioning.
REFERENCES


91


ACKNOWLEDGEMENTS

Participant vouchers were made available with the financial assistance of a PPLS Research Support Grant (2009).

Thanks go to Justin, Andrea, Gillian, Mhairi, Jane and, of course, Helen for assistance with piloting.
**APPENDIX A**: Full, subscale and subtest scores for performance on the WASI battery.

<table>
<thead>
<tr>
<th>Scale</th>
<th>ASD Mean</th>
<th>ASD SD</th>
<th>Control Mean</th>
<th>Control SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-Scale IQ</td>
<td>102.29</td>
<td>15.73</td>
<td>113.71*</td>
<td>7.56</td>
</tr>
<tr>
<td>(n = 28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>104.42</td>
<td>20.51</td>
<td>117.64</td>
<td>13.19</td>
</tr>
<tr>
<td>(n = 26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-verbal IQ</td>
<td>102.67</td>
<td>10.84</td>
<td>107.14</td>
<td>9.92</td>
</tr>
<tr>
<td>(n = 26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>47.71</td>
<td>14.80</td>
<td>62.43**</td>
<td>8.17</td>
</tr>
<tr>
<td>(n = 28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarities</td>
<td>54.58</td>
<td>13.03</td>
<td>58.21</td>
<td>7.95</td>
</tr>
<tr>
<td>(n = 26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix Reasoning</td>
<td>51.64</td>
<td>7.60</td>
<td>50.57</td>
<td>9.07</td>
</tr>
<tr>
<td>(n = 28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block Design</td>
<td>53.31</td>
<td>6.43</td>
<td>59.21*</td>
<td>6.54</td>
</tr>
<tr>
<td>(n = 27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant difference at p < .05

** Significant difference at p < .01
APPENDIX B: DKEFS example set (i) and set structure (ii)

i)

ii)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 living</td>
<td>8 animals</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>15 non-living</td>
<td>7 transport</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 kitchen/household</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX C: Twenty Questions Set 2 items (i) and set structure (ii)

i)

<table>
<thead>
<tr>
<th>Item</th>
<th>Animal / Non-Animal</th>
<th>Mammal / Non-Mammal</th>
<th>Farm / Non-Farm</th>
<th>Bird / Non-Bird</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eagle</td>
<td>Animal</td>
<td>Mammal</td>
<td>Farm</td>
<td>Bird</td>
</tr>
<tr>
<td>Bowl</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Rose</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Carrot</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Spoon</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Elephant</td>
<td>Animal</td>
<td>Mammal</td>
<td>Farm</td>
<td>Bird</td>
</tr>
<tr>
<td>Shark</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Oranges</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Fridge</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Ducks</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Apple</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Owl</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Fork</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Snake</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Banana</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Sheep</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Cow</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Knife</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Oven</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Dog</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Cat</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Tree</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Horse</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
<tr>
<td>Fish</td>
<td>Non-Animal</td>
<td>Non-Mammal</td>
<td>Non-Farm</td>
<td>Non-Bird</td>
</tr>
</tbody>
</table>

ii)

<table>
<thead>
<tr>
<th>Category</th>
<th>Quantity</th>
<th>Subcategory</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 animals</td>
<td>6 mammals</td>
<td>3 farm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 non-mammals</td>
<td>3 non-farm</td>
<td></td>
</tr>
<tr>
<td>12 non-animals</td>
<td>6 plant</td>
<td>3 fruit</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 kitchen</td>
<td>3 cutlery</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 non-fruit</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 non-cutlery</td>
<td></td>
</tr>
</tbody>
</table>


APPENDIX D: Robots Set 2 items (i) and set structure (ii)

i)

![Image of robots with labels]

N.B. It should be noted that although the two robot sets were arranged so that questions about relatively global properties would be fairly effective if asked as the first question (shape questions in the first instance, colour questions in the second), this was by no means the only way to effectively subdivide the...
set. For instance, questions about local features such as whether the robot has feet or wheels, or more than two eyes, would be just as effective in dividing the set into two equal halves. Just as the standard Twenty Questions items typically require a process of elimination which moves from the general (living things, man-made etc) to the specific (use to eat food, keep as a pet), it would have been possible to create a robot set with such a global-to-local structure. However, as discussed in the introduction, evidence from central coherence research suggests this could cause separate problems for ASD participants (Happé & Frith, 2006). Instead the robot sets were arranged so that they could be searched through based on local properties (eyes, feet) or global properties (shape, colour) depending on participant preference and strategy. Similarly, the set could be searched based on abstract criteria (i.e. gender) if a participant was inclined to do so, but could if anything be more effectively searched based on concrete and perceptual criteria.
APPENDIX E: Within-trial scores for (i) overall success, ii) question types and (iii) question content.

i)

<table>
<thead>
<tr>
<th>Trials completed (%)</th>
<th>Twenty Questions Task (standard items)</th>
<th>Robots Task (novel items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASD</td>
<td>TD</td>
</tr>
<tr>
<td>Block 1 (initial)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block 2 (flexibility)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block 3 (memory)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean questions to target</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td>6.02</td>
<td>1.32</td>
<td>5.83</td>
<td>1.38</td>
<td>5.47</td>
<td>1.02</td>
<td>5.47</td>
<td>0.71</td>
</tr>
<tr>
<td>Block 2</td>
<td>5.78</td>
<td>1.88</td>
<td>5.17</td>
<td>0.72</td>
<td>5.38</td>
<td>0.87</td>
<td>4.6</td>
<td>0.57</td>
</tr>
<tr>
<td>Block 3</td>
<td>6.29</td>
<td>1.93</td>
<td>4.93</td>
<td>1.22</td>
<td>6.25</td>
<td>1.87</td>
<td>5.65</td>
<td>1.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean question quality</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td>0.24</td>
<td>0.10</td>
<td>0.34</td>
<td>0.07</td>
<td>0.30</td>
<td>0.06</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td>Block 2</td>
<td>0.25</td>
<td>0.10</td>
<td>0.34</td>
<td>0.06</td>
<td>0.32</td>
<td>0.09</td>
<td>0.41</td>
<td>0.04</td>
</tr>
<tr>
<td>Block 3</td>
<td>0.22</td>
<td>0.09</td>
<td>0.32</td>
<td>0.08</td>
<td>0.28</td>
<td>0.11</td>
<td>0.36</td>
<td>0.05</td>
</tr>
</tbody>
</table>

* Significant difference at p < .05
<table>
<thead>
<tr>
<th>Question types (%)</th>
<th>Twenty Questions Task (standard items)</th>
<th>Robots Task (novel items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASD</td>
<td>TD</td>
</tr>
<tr>
<td>Constraint-seeking</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Block 1</td>
<td>59.17</td>
<td>25.65</td>
</tr>
<tr>
<td>Block 2</td>
<td>44.09</td>
<td>25.16</td>
</tr>
<tr>
<td>Block 3</td>
<td>51.66</td>
<td>26.62</td>
</tr>
</tbody>
</table>

| Hypothesis-testing       | Mean        | SD | Mean   | SD | Mean   | SD | Mean | SD |
| Block 1                  | 22.31       | 27.36 | 9.35   | 9.71 | 17.14 | 22.29 | 5.45  | 6.02  |
| Block 2                  | 33.68       | 36.35 | 14.01  | 14.74 | 16.93 | 22.28 | 5.46  | 9.83  |
| Block 3                  | 28.34       | 36.83 | 12.63  | 9.96  | 18.26 | 29.07 | 8.82  | 12.85 |

| Pseudo-constraint        | Mean        | SD | Mean   | SD | Mean   | SD | Mean | SD |
| Block 1                  | 19.40       | 16.20 | 14.76  | 11.31 | 18.02 | 12.59 | 20.08 | 12.76 |
| Block 2                  | 17.49       | 21.99 | 20.06  | 17.31 | 23.48 | 14.28 | 20.09 | 11.27 |
| Block 3                  | 10.10       | 10.57 | 14.51  | 12.87 | 18.28 | 13.14 | 22.31 | 14.47 |

$^a$ Significant group*block interaction effect (p < .05)
ii) (contd.)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Repeated</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block 1</td>
<td>0.42</td>
<td>1.55</td>
<td>0.00</td>
<td>0.00</td>
<td>1.26</td>
<td>3.41</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Block 2</td>
<td>0.32</td>
<td>1.20</td>
<td>0.00</td>
<td>0.00</td>
<td>0.42</td>
<td>1.55</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Block 3</td>
<td>1.31</td>
<td>3.71</td>
<td>0.00</td>
<td>0.00</td>
<td>0.42</td>
<td>1.57</td>
<td>1.02</td>
<td>3.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Redundant</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block 1</td>
<td>0.00</td>
<td>0.00</td>
<td>1.25</td>
<td>2.60</td>
<td>4.90</td>
<td>6.76</td>
<td>1.20</td>
<td>3.16</td>
</tr>
<tr>
<td>Block 2</td>
<td>3.67</td>
<td>6.75</td>
<td>1.63</td>
<td>3.54</td>
<td>3.42</td>
<td>6.87</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Block 3</td>
<td>8.19</td>
<td>8.74</td>
<td>8.81</td>
<td>13.33</td>
<td>9.03</td>
<td>11.81</td>
<td>2.52</td>
<td>5.20</td>
</tr>
</tbody>
</table>
### iii)

<table>
<thead>
<tr>
<th>Question content (%)</th>
<th>Twenty Questions Task (standard items)</th>
<th>Robots Task (novel items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASD</td>
<td>SD</td>
</tr>
<tr>
<td><strong>Colour</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block 1</td>
<td>13.66</td>
<td>27.29</td>
</tr>
<tr>
<td>Block 2</td>
<td>7.96</td>
<td>10.35</td>
</tr>
<tr>
<td>Block 3</td>
<td>9.89</td>
<td>19.99</td>
</tr>
<tr>
<td><strong>Form</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block 1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Block 2</td>
<td>0.03</td>
<td>1.10</td>
</tr>
<tr>
<td>Block 3</td>
<td>0.05</td>
<td>1.91</td>
</tr>
<tr>
<td><strong>Name</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block 1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Block 2</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Block 3</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
iii) (contd.)

<table>
<thead>
<tr>
<th>Specific features (perceptual)</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td>5.92</td>
<td>12.27</td>
<td>13.14*</td>
<td>11.76</td>
<td>23.09</td>
<td>19.49</td>
<td>30.34</td>
<td>12.78</td>
</tr>
<tr>
<td>Block 2</td>
<td>7.94</td>
<td>11.98</td>
<td>6.05</td>
<td>5.31</td>
<td>24.67</td>
<td>18.90</td>
<td>32.23</td>
<td>14.06</td>
</tr>
<tr>
<td>Block 3</td>
<td>7.50</td>
<td>12.15</td>
<td>6.66</td>
<td>9.90</td>
<td>23.58</td>
<td>17.97</td>
<td>29.44</td>
<td>16.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abstract</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td>33.46</td>
<td>27.76</td>
<td>42.44</td>
<td>22.11</td>
<td>0.48</td>
<td>1.78</td>
<td>0.41</td>
<td>9.16</td>
</tr>
<tr>
<td>Block 2</td>
<td>41.86</td>
<td>29.35</td>
<td>41.80</td>
<td>15.58</td>
<td>0.55</td>
<td>2.05</td>
<td>4.47</td>
<td>9.19</td>
</tr>
<tr>
<td>Block 3</td>
<td>30.01</td>
<td>27.71</td>
<td>41.84*</td>
<td>21.34</td>
<td>0.71</td>
<td>2.67</td>
<td>4.80</td>
<td>9.59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td>21.02</td>
<td>24.73</td>
<td>6.40</td>
<td>11.33</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Block 2</td>
<td>17.85</td>
<td>18.03</td>
<td>10.72</td>
<td>12.43</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Block 3</td>
<td>13.08</td>
<td>19.74</td>
<td>1.79*</td>
<td>6.68</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specific features (Conceptual)</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td>32.52</td>
<td>29.77</td>
<td>26.46</td>
<td>17.74</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Block 2</td>
<td>22.30</td>
<td>18.97</td>
<td>29.62</td>
<td>15.91</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Block 3</td>
<td>24.97</td>
<td>24.35</td>
<td>30.60</td>
<td>19.96</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Significant difference at p < .05
Effect sizes: a note

The effect sizes reported in the results section reflect partial-eta squared values ($\eta^2_p$). As a rule of thumb, small effects are roughly equal to 0.01, medium effects 0.06, and large effects $= 0.14$ (Kittler et al, 2007).