PRAGMATIC UNDERSTANDING: Comprehension of Conversation in Individuals with and without Communication Challenges



Alexander C. Wilson

St John's College

University of Oxford

Doctor of Philosophy in Experimental Psychology

April 2020

Acknowledgements

I am grateful to my supervisor, Prof. Dorothy Bishop, for her expert guidance throughout the work reported in this thesis. I have been inspired by her meticulous approach to the research process and her insight into all things relating to language and communication. As a mentor, she has been kind and attentive, and a brilliant role model.

I am aware how lucky I have been to be part of the OSCCI research group during my DPhil. Thank you to all the members of OSCCI for their encouragement, friendliness and wealth of knowledge, and thank you to our friends across the Department of Experimental Psychology in Oxford. In particular, I am grateful to Prof. Gaia Scerif for helping me clarify my thinking during conversations we shared.

My special thanks must also go to the participants who made this research possible. I have been so heartened by the enthusiasm and engagement of all the individuals, families and schools taking part. There are many things that I have learned from autistic individuals and their families that have not explicitly made their way into this thesis but have and will continue to exert huge influence on my work. I have been moved by the challenges people have described, and these conversations have stressed for me the importance that research is a collaborative exercise with and for autistic people.

Lastly, thank you as ever to my family for their support.

Abstract

Pragmatic processing of language involves inferring a speaker's intended meaning by integrating their utterance in context. Pragmatics is thought to be dissociable from "core" aspects of language processing, such as vocabulary and grammar, and to be a specific area of difficulty for some individuals, such as autistic people. As there are limitations in current pragmatic assessments, a new Implicature Comprehension Test was developed to specifically measure a person's ability to process implied meaning in context. Results supported the hypotheses, showing that processing implied meaning was separable from core language skills, and that the Implicature Comprehension Test was sensitive to cognitive differences between autistic and non-autistic individuals, especially in adults. The situation became more complex when looking at implied meaning in the context of other socio-cognitive-linguistic skills measured through a cognitive test battery, as domain-level differences between autistic and non-autistic people on the battery were subtle at most, even though autistic people were observed and reported themselves as having considerable day-to-day communication difficulties. It is suggested that these subtle differences reflected domain-level information processing differences in judging the meaning of novel communicative stimuli in the context of uncertainty and a preference for explicit information. However, the subtle scale of observed differences supports the view that communication difficulties in autism vary according to the context, with an important contextual factor being the level of structure and uncertainty inherent in the situation. In exploratory analysis, this thesis questions whether difficulties with "theory of mind" account for communication differences in autism. Suggestive evidence is also provided for gender differences in the communication profile in autism and for a selective link between self-reported, but not observed, communication difficulties and mental health challenges.

Contents

Acknowledgements **I**

Abstract **ii**

Contents **iii**

1. Chapter One. Pragmatics: Theory and Methodology **1**
1.1. Theoretical Issues: What do We Mean by "Communicative Use and Understanding of Language"? 3
1.2. Methodological Issues: How do We Measure Pragmatics? 12
1.3. Rationale for Empirical Work: Implicature as an Indicator of Pragmatic Ability 22

2. Two. Relationships between Implicature Comprehension and Other Communication Skills in Non-Autistic Adults **27**
2.1. Study One: A First Exploration of the Implicature Comprehension Test 28
2.1.1. Methods 28
2.1.2. Results 32
2.1.3. Discussion 34
2.2. Study Two: Testing a Two-Factor Model of Language Processing in the General Population 35
2.2.1. Methods 37
2.2.2. Results 47
2.2.3. Discussion 52

3. Chapter Three. Cognitive Differences between Autistic and Non-Autistic Adults in Processing Implicature **58**
3.1. Methods 59
3.2. Results 65
3.3. Discussion 67
3.3.1. Response Tendency on the Implicature Comprehension Test in Autistic Adults 68
3.3.2. Potential Clinical Applications and Implications 72

4. Chapter Four. Receptive Communication Skills in Autistic and Non-Autistic Adults: Group Comparisons and Heterogeneity **74**
4.1. Methods 76
4.2. Results 86
4.3. Discussion 95
4.3.1. Comparing Autistic and Non-Autistic Adults: A Domain-Level Difference in Information-Processing 95
4.3.2. No Evidence for an Isolated "Theory of Mind" Difficulty 98

4.3.3. Multiple Influences on the Communication Phenotype in Autism 100
4.3.4. Gender and Communication Skills in Autistic Adults 104
4.3.5. Conclusion 105

5. Chapter Five. Communication and Mental Health Challenges in Adults on the Autism Spectrum 106
5.1. Methods 107
5.2. Results 111
5.3. Discussion 115

6. Chapter Six. An Attempt to Tease Apart Core Language and Pragmatics in Children Using a Novel Online Assessment 120
6.1. Implied Meaning and Core Language in Children 120
6.2. Implicature as part of Broader Domains of Language Development 122
6.3. Methods 124
6.4. Results 134
6.5. Discussion 146
6.5.1. Theoretical Implications 147
6.5.2. Practical Implications 148

7. Chapter Seven. Testing for Pragmatic Differences between Autistic and Non-Autistic Children 152
7.1. Methods 152
7.2. Results 161
7.3. Discussion 168
7.3.1. Reflecting on Inconsistencies between Child and Adult Performance on the Test Batteries 171
7.3.2. Conclusion 174

8. Chapter Eight. Concluding Reflections 176

9. References 184

10. Appendices 213
10.1. Appendix A. Existing Tests with a Pragmatic Component (Supplementary Material for Chapter One) 213
10.2. Appendix B. Instructions for Adult Assessments and Examples of Test Items (Supplementary Material for Chapter Two) 221
10.3. Appendix C. Meta-Analysis of Performance of Autistic and Non-Autistic Adults on the Frith-Happé Animations (Supplementary Material for Chapter Four) 235
10.4. Appendix D. Instructions for Child Assessments and Examples of Test Items (Supplementary Material for Chapter Six) 240

Chapter One. Pragmatics: Theory and Methodology

According to clinical intuition, some people have specific difficulties with using and understanding language in context. People on the autism spectrum, for instance, may experience difficulties with conversation in the relative absence of problems with core aspects of language, such as vocabulary and grammar (Baird & Norbury, 2016; Lam & Yeung, 2012). While this pattern of developmental strengths and difficulties may be more common in autistic individuals, a range of others may share a similar communication profile characterised by specific challenges with conversation. This profile has been conceptualised as "pragmatic language impairment" (Bishop, 1998) and has recently been formalised as the diagnosis "social (pragmatic) communication disorder" in DSM-5 (American Psychiatric Association, 2013). These labels all suggest that an impairment in using and understanding language in context may co-exist with typical, or indeed advanced, core language development (i.e. vocabulary and grammar skills). This has the broader implication that communication in context and core language are two separate domains. However, it has been surprisingly hard to provide empirical evidence for this using objective tests. The present thesis aims to investigate the possibility of a dissociation between these two domains in non-autistic individuals using novel tests of pragmatics, before exploring the pattern of language and social communication abilities in groups on the autism spectrum. In line with current preferences of autistic people and their families, I use "autistic" and "on the autism spectrum" to describe the groups involved in this project, while appreciating that people differ in their language preferences (Kenny et al., 2016).

Since the 1980s, researchers have proposed that a subset of individuals with developmental language impairments have disproportionate difficulty with the use of language in

social contexts: they are said to have “pragmatic language impairment” or PLI (Bishop, 1998; Botting & Conti-Ramsden, 1999). Rapin (1996) described the conversation of children with pragmatic difficulties as characterised by: atypical word choices and word finding difficulties, problems comprehending and formulating connected discourse, heavily scripted speech, poor topic maintenance, answering beside the point of a question, and speaking aloud to no-one. The original claim was that these features present in the absence of difficulties with core language skills (such as speech articulation and syntax), although it is possible that difficulties with core language get overlooked when social aspects of language use are noticeably unusual. In line with this view, several studies have found that PLI is often associated with impaired core language skills (e.g. Botting & Conti-Ramsden, 2003; Norbury et al., 2004). In addition, Gibson et al. (2013) attempted to use logistic regression to classify individuals as having a particular communication difficulty based on their presentation, and while greater social difficulties and higher expressive language ability characterised children with PLI compared to those with specific language impairment (SLI), the differences were slight, and indeed they were non-existent for receptive language. Together, these studies indicate that it is not reasonable to expect a strong dissociation between core language skills and the communicative use and understanding of language in context. Two independent review papers support this view, indicating a moderate to large correlation between these two domains (Andrés-Roqueta & Katsos, 2017; Matthews et al., 2018). However, it is difficult to be entirely confident in these conclusions for two reasons: one theoretical and one methodological. There are theoretical issues surrounding quite what we mean by “communicative use and understanding of language”, and there is the methodological problem that we lack validated tests of competence in this domain.

Theoretical Issues: What do We Mean by "Communicative Use and Understanding of Language"?

Two terms are often used in research: social communication and pragmatics. Social communication is the more general term, and refers to the combination of "social interaction, social cognition, pragmatics (verbal and nonverbal) and receptive and expressive language processing" that supports conversation (Adams, 2005, p. 128). The broad, multidimensional nature of this construct makes it difficult to study with precision. Pragmatics is a more specific term, and in theoretical linguistics refers to the role of context in communicative meaning (Ariel, 2010). In this thesis, I am interested in whether problems with social communication can be accounted for by more specific difficulties with adjusting our use and understanding of language to the communicative context. As such, it makes sense to focus on pragmatics. As an example of pragmatics in action, consider what someone might mean by saying "it's quite cold". The speaker might just be describing the weather. But they might also intend the listener to read further meaning into their words. They might be suggesting that someone investigate a broken heater or warning a child not to go in the water, or indeed complaining about a decision to sit outside… Pragmatics refers to this variability in meaning depending on context. While this working definition of pragmatics is consistent with most linguistic definitions of the term, it is notoriously hard to delineate the boundaries between pragmatics and other language domains, as demonstrated in a book-length treatment of the problem (Ariel, 2010). As Levinson (1983) said in his earlier review of different definitions of pragmatics, "The most promising are the definitions that equate pragmatics with 'meaning minus semantics', or with a theory of language understanding that takes context into account" (p. 32). Ariel (2010) largely concurred with this view but placed more emphasis on the cognitive process – i.e. inference – that enables context to

play its role in meaning. She concluded that "the code/inference distinction constitutes the best, in fact the only, candidate for drawing a cognitively viable grammar/pragmatics divide. […] Pragmatics is […] responsible for whatever is inferred based on linguistic inputs" (pp. 271-2). Whereas semantics and grammar (i.e. core language skills) involve decoding conventional meaning, pragmatics involves context-dependent inference. In crude but workable everyday words, this is a distinction between accessing "dictionary" meaning on the one hand and "reading between the lines" on the other.

This contrast between inferencing and decoding offers a useful point of entry for practitioners interested in a cognitive approach to assessing pragmatics, and we will return to this distinction. However, one thing to note about the place of "pragmatics" in applied fields is that the term has suffered from being interpreted in rather different ways. Let's consider three definitions from recent review papers. The first parallels the definition I will use: pragmatic language involves "go[ing] beyond what is explicitly stated to reach the intended meaning" (Baird & Norbury, 2016, p. 746). As their definition, Matthews et al. (2018) described a family of pragmatic skills including "the ability to initiate conversation, to respond with contingent, relevant, and new information, to produce and understand utterances by drawing on context (including the perspectives of interlocutors and what is in their common ground), to use an appropriate register (respecting social status), to recount cohesive and coherent narratives and to understand non-literal language including irony" (p. 186). While few would disagree that these skills involve pragmatics, this listing approach exemplifies the problem researchers have in specifying what is common to these skills – i.e. what pragmatic processing is. Without core criteria for what constitutes the pragmatic aspect of communication, we are at risk of blurring the boundaries between pragmatics and more general aspects of communication. A review paper of

pragmatic interventions gave the following definition: “Early definitions of pragmatic language refer to the use of language in context, encompassing the verbal, paralinguistic and non-verbal aspects of language. Contemporary definitions have expanded beyond just communicative functions to include behaviour that includes social, emotional, and communicative aspects of language” (Parsons et al., 2017). This definition demonstrates the “exclude nothing approach to the assessment of language pragmatics” that has met with heavy criticism from linguists (Cummings, 2007, p. 416). On the basis of this definition, it is difficult to have any clear conception of what assessing pragmatics would look like. Cummings (2007) advises practitioners and applied researchers instead to focus on pragmatics as “a process of reasoning based on features of context” by which we infer what a speaker means (2007, pp. 425–6). To return to the example above, the focus ought to be how we understand what “it’s quite cold” means in the particular context.

In thinking about pragmatics alongside other language skills, a key question is: how do we view the interaction between these skills? Linguistic study is typically broken down into different components based on distinct units – so phonology involves the study of speech sounds, morphology the formation of words, syntax the arrangement of words into grammatical entities, and so on. As Scott-Phillips (2017) states, pragmatics is typically “presented as one further level of linguistic analysis, in addition to semantics, syntax, and the rest” (p. 187). In a visualisation of the model, pragmatics is the outermost ring with all other levels of analysis nested inside. The problem with this model is that it views pragmatics as an afterthought, a final stage of processing, which likely misrepresents the true role of pragmatics in language processing. Communication is inherently social and context-bound, so we should probably be applying a more central role to pragmatics. For one thing, social context seems to play a role in the neural processing of

language from the earliest stages (e.g. van Berkum et al., 2008; Bornkessel-Schlesewsky et al., 2013; Egorova et al., 2013), suggesting that any model of language processing should account for the pervasive role of context, and not see pragmatics as a final process of disambiguation based on context. Scott-Phillips contrasts this approach with what he calls "strong pragmatics". He described "a capacity of mind, to communicate in a way that is fundamentally a matter of social cognition. More precisely, it is a capacity to communicate by expressing and recognizing intentions" (2017, p. 187). This "strong pragmatics" approach gives pragmatics primacy in communication. It recognises that language development is fully embedded in social context, both for the individual child and over evolutionary time. For instance, children who are cued into the intentional, context-bound, social nature of communication are likely to find it easier to acquire new linguistic forms (e.g. Tomasello, 2000). If we are to adopt a model of language processing that conforms with the "strong pragmatics" position, we need to be clear on two things. Firstly, any model needs to make clear predictions about how pragmatic processing might separate from other skills. Secondly, a model needs to help us identify linguistic phenomena that might place a particular demand on pragmatics, as the immediate aim of this project is to devise psychometric tests targeting pragmatics.

Relevance Theory (Sperber & Wilson, 1986) qualifies on both these grounds, as described below. Under this framework, language processing is an interactive combination of (1) decoding the "literal meaning" of an utterance using our vocabulary knowledge and grammatical competence, and – because the linguistic code is always incomplete or ambiguous – (2) inferring from context the full extent of the interlocutor's intended meaning. This inferential aspect would equate with pragmatics. In Relevance Theory, inferring communicative meaning is assumed to be a subdomain of our ability to attribute mental states to others (our "theory of mind", or capacity

for “mind-reading”). As part of our more general “theory of mind” skills, we are thought to make specialised use of a domain-specific heuristic – communicative relevance – which, as the name suggests, is specifically activated by communicative behaviour. This heuristic involves the automatic assumption that interlocutors will communicate in a way that is optimally relevant to the context, and we make inferences about what people mean on the basis of this (see Carston, 2002). See Figure 1.1 for a visualisation of the model of language processing adopted in this thesis.

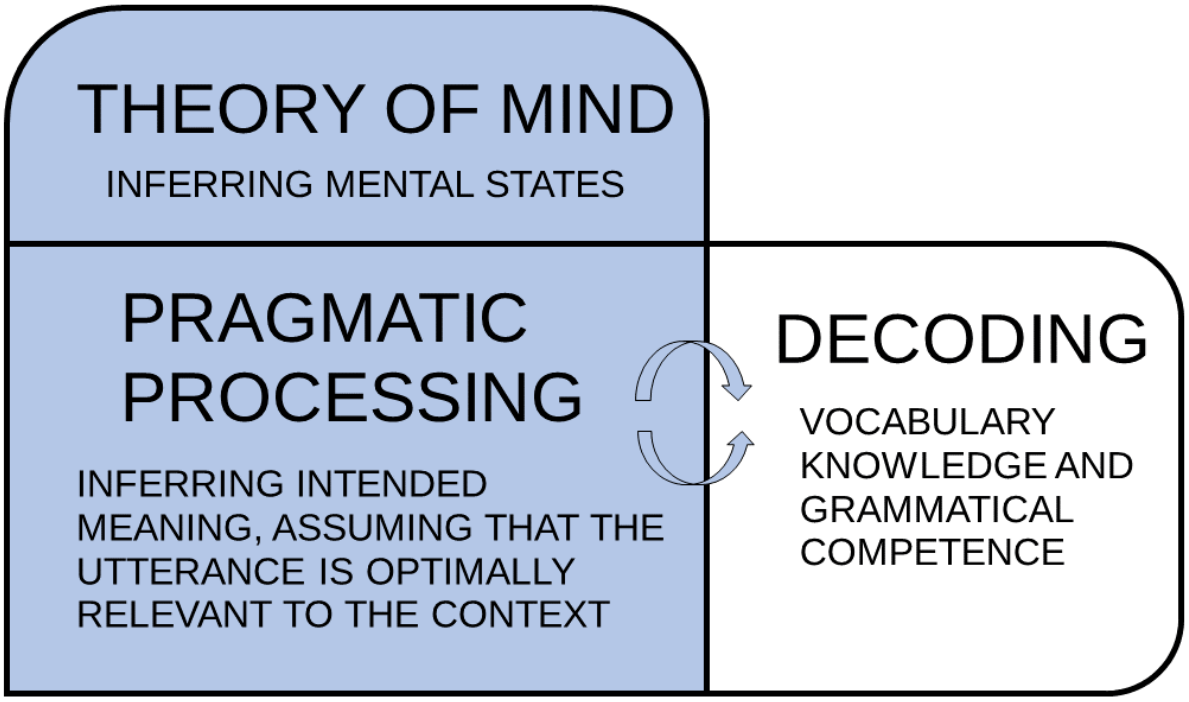


Figure 1.1. My model of language processing, as borrowed from Relevance Theory. Language comprehension is an interaction between decoding and inference. Pragmatic processing is a subdomain of our more general “theory of mind”.

It should be noted that questions have been raised about the necessity of “theory of mind” for pragmatics (e.g. Geurts et al., 2019) – in some circumstances, pragmatic processing may depend on domain-general reasoning without the need for mental state inferences, for instance. On the other hand, plenty of evidence has been provided for a relationship between pragmatics and

"theory of mind" (see Cummings, 2014 for a review), so it makes sense to assume a connection but be open to the possibility that the model of language processing may need revision. I will return to this issue in later chapters.

The focus on inference in Relevance Theory is useful to the psychometrician in devising a test of pragmatics, as this can be operationalised in a specific way and is likely to have face validity to the theoretical linguist (which is not the case with current instruments, as described below). Relevance Theory proposes that we make inferences at two levels. We infer the full extent of the speaker's "explicit" meaning through enrichment of the linguistic code into *explicature*. We also infer further "implicit" meaning, or *implicature*, from a global understanding of the utterance in context. For illustration, consider the following dialogue:

> SPEAKER ONE. Can you pick Sally up from the station?
> SPEAKER TWO. I'm working all day.

The explicature of "I'm working all day" might take the form of "Today I will be doing work tasks at my workplace during my working hours". Here we are making local-level inferences about what the speaker's individual words mean in context. These sorts of "gap filling" inference occur with all language, meaning that all language use involves pragmatics. In addition, we might derive implicature through a global understanding of the utterance in its communicative context; for instance, "I'm working all day" implicitly turns down SPEAKER TWO's request to collect Sally. The implicature of "I'm working all day" is understood as "Sorry, I can't pick Sally up". In other contexts, "I'm working all day" would not communicate anything about picking a person up from the station: the implicated meaning is only expressed here because we assume that an utterance in this context must relevantly address SPEAKER ONE's question and therefore infer what is likely to mean here. The relevance heuristic underpinning communication therefore sets in motion a reasoning process that takes "I'm working all day" as input and produces "Sorry, I

can't pick Sally up" as output. As we can see here, explicature hinges very much on an interaction between decoding and inference, whereas implicature is a purer case of pragmatic inference based on context. In the case of implicature, it does not matter how well a person can decode what they are hearing; they are unlikely to arrive at a satisfactory interpretation of the intended meaning without engaging in contextually sensitive inference (i.e. in pragmatic processing). Implicature would therefore seem to be the quintessential pragmatic phenomenon, and is likely to be a good focus to take in assessing pragmatics.

But to what extent is pragmatic processing a single construct, and relatedly, is it measurable as such? The first issue to raise is that no test offers a pure measure of a single cognitive function, but is "contaminated" by other cognitive factors, such as attention, motivation, conceptual understanding, and so on. Furthermore, any test that aims to home in on a particular linguistic function will encounter the issue that language does not occur in the world in a neatly disentangled form. For instance, if we aimed to measure syntactic processing distinctly from vocabulary knowledge, we would need to devise a language test that involves linguistic structure but no words. Needless to say, that would be quite a challenge, and if it were possible, such a task would likely involve dubious task demands due to the artificial, unfamiliar nature of it, and therefore would not be tapping the construct of interest (syntax) very effectively. The most effective way of measuring a cognitive process is likely to involve, first, devising diverse tests that incorporate items which "press" on the process of interest, while attempting to control other demands as much as possible; and, second, including those tests as part of a battery that aims to measure other cognitive/linguistic processes as well, so that we can (a) extract what is common across tests tapping the construct of interest, and (b) partial out those "confounding" processes as part of the analysis. We might see this as representing an ideal approach to measurement.

However, let's reframe the issue as whether a single test might be sufficient (not perfect, just "sufficient") as a measure of a particular linguistic function. In the case of receptive vocabulary and grammar skills, this perhaps could be argued. For example, combining the British Picture Vocabulary Scales (BPVS; Dunn et al., 1982) and the Test for Reception of Grammar (TROG-2; Bishop, 2003b) would allow us to gain a rough picture of where a child is in terms of their receptive language skills. In both tests, there is a spoken stimulus and the test-taker chooses which of several pictures matches it; in the BVPS, the test-taker hears a sequence of increasingly complex words, whereas the TROG-2 comprises sentences with various grammatical structures. The multiple choice response format of both tests might, for instance, introduce certain task demands that might vary if the tests were given in a different format (e.g. Frizelle et al., 2019), but it would be difficult to argue that these assessments do not have face validity as tests of their respective domains. Moving on to pragmatics: the question is whether it is plausible for a single test to adequately measure pragmatics in a similar way? This comes down to the dimensionality of pragmatic processing: are we viewing it as unidimensional or a set of skills? Empirical evidence regarding the unidimensionality of pragmatics is unsatisfactory for answering this question for two reasons. Firstly, the small amount of evidence is conflicting, with one study indicating unidimensionality for a practitioner-completed pragmatic rating scale (Cordier et al., 2019) and another study showing no clear factor solution for a parent-completed checklist (Geurts et al., 2009). However, the second and more important issue is that rating scales such as these are likely to be measuring more than just pragmatics (as described below) and so do not provide grounds for examining the nature of pragmatic processing. Turning to theoretical linguistics instead, the typical approach would view pragmatic processing as unidimensional.

Theorists often view pragmatics as a single modular system – i.e. distinct from other aspects of language and cognition. However, the degree of modularity claimed for pragmatics has varied. Indeed, a shift has taken place within Relevance Theory itself; the original claim was that pragmatics involves the application of general "mind-reading" (Sperber & Wilson, 1986), whereas more recently pragmatics has been described as depending on a more specific pragmatic comprehension submodule (Sperber & Wilson, 2002). While these different approaches offer different models of the mind, they both make the same prediction regarding the structure of pragmatics: that it is a unidimensional process. As Carston (2002) asserts, Relevance Theory sets out "a single pragmatic comprehension system, informed by a single overarching principle" (p. 142), and all meaning is processed by this one system, including that which is more explicit (*explicature*) and more implicit (*implicature*). This view that pragmatics involves a single system has been challenged (Geurts et al., 2019), on the basis of possible evidence that autistic people have an uneven profile of skills across the pragmatic domain. For instance, there is some evidence to suggest that autistic individuals may have greater difficulty understanding irony than indirect requests (Deliens et al., 2018). This is taken to suggest that pragmatic processing does not rely on one overall system. It should be noted, though, that the evidence does not really support (or refute) this idea. Firstly, studies would need to investigate correlations between tasks, so that we can be confident whether the tasks are tapping similar or different constructs, and secondly, studies would need to make group comparisons on tasks with good psychometric properties (including sufficient within-group variance to detect group differences); neither of these requirements is the case with studies reviewed by Geurts et al. (2019). Also, it seems rather backwards logic to make a case for "fragmented pragmatics" on the basis of communication abilities in autism rather than on the basis of a factor analysis of skills across the general population. While Geurts et al. (2019) should be commended for underscoring the importance of

not assuming that all aspects of communication are a challenge for autistic people, the evidence they recruit does not argue persuasively for a dissociation between aspects of pragmatics. While we have no basis for rejecting the notion that pragmatics is multidimensional, let us take it as a single system, as proposed by Relevance Theory, for the purpose of this working definition of pragmatic processing.

One final point to note in our theoretical discussion of pragmatics is the notable absence of how pragmatically appropriate language is produced. In reading the research literature in linguistics, the reader will be struck by the number of fragments of conversation analysed on the basis of how they might be interpreted. Theories rarely tackle the question of "productive pragmatics". Indeed, the conventional alignment of pragmatics with inference (i.e. a comprehension process) emphasises that fact. Pragmatic comprehension (inferring what someone means in context) will be the focus of this thesis, but we should bear in mind that this is only part of the story.

Methodological Issues: How do We Measure Pragmatics?

There are three forms of pragmatic language assessment: observation of semi-naturalistic conversation, formal language tests, and questionnaires completed by informants or the individual themselves. In determining the scope of a pragmatic *language* test, we might expect the first principle to be the exclusion of non-linguistic features. Of course, if verbal and nonverbal aspects of social communication are inextricably linked, then we might not expect this, but the conceptualisation of communication difficulties such as pragmatic language impairment is that language pragmatics might be usefully isolated. Most nominal tests of pragmatics do not exclude nonverbal aspects. For instance, over half the behaviours rated in the Pragmatic Protocol (a

widely used coding scheme employed in direct observation of conversation; Prutting & Kirchner, 1987) are non-verbal in nature. The authors acknowledge this fact, noting that the coding scheme covers context-dependent aspects of language use but also incorporates rules governing "appropriate" language use, including the expectation that speech is integrated with eye contact, facial expression and gesture. The heterogeneous nature of the coding scheme is due to its purpose: to record as many communication behaviours that may be implicated in day-to-day communication glitches. The Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012) takes a similar approach to the Pragmatic Protocol. Individuals take part in several structured communication activities and free conversation, and the researcher/practitioner codes the individual's performance against a full range of communication behaviours that are known to reliably differentiate autistic and non-autistic people. In comparing the Pragmatic Protocol and ADOS-2, we see two measures designed for clinical reasons to provide an effective summary of the level of communication impairment an individual may be experiencing, and particularly in the case of ADOS-2 to support a possible diagnosis. However, these tools do not allow us to measure pragmatic processing as distinct from other features of social communication. This is also the case with questionnaire measures, of which the most widely used is the Children's Communication Checklist (Bishop, 2003a). Apart from the obvious point that families and school staff (and individuals themselves in the case of adult self-report) are unlikely to have the specialist experience needed to report on pragmatics as understood by linguists, it should be noted that the CCC-2 was designed to measure a full range of communication characteristics that had been prominent in clinical descriptions of individuals with communication challenges but were difficult to assess by formal testing. The CCC-2 is therefore likely to be a heterogeneous measure rather than a specific measure of pragmatics. While structured clinical observation and questionnaire measures are critical in clinical settings for profiling individuals' strengths and

difficulties, these methods do not allow us to identify aspects of cognitive processing that may account for communication difficulties.

It is important to make the distinction between social communication behaviours and the cognitive functions that underpin them. Pragmatic processing may or may not be responsible for these behaviours. Let's consider an example: how might we interpret a difficulty in producing contingent turns in conversation? This was a problem for autistic children tested using the Yale in vivo Pragmatic Protocol, a semi-naturalistic conversational assessment including "specific pragmatic probes" for eliciting target behaviours (Simmons et al., 2014). A lack of contingent turns may arise from insensitivity to what is relevant in the communicative context – i.e. a pragmatic problem. Alternatively, the problem may be a lack of social interest, reduced joint attention, performance anxiety, reduced ideational fluency, word-finding difficulties, perseveration on one's own topic of interest, or other issues besides. These are not pragmatic problems. An observational measure such as the Yale in vivo Pragmatic Protocol or a questionnaire such as the CCC-2 cannot differentiate between these possibilities.

To establish whether pragmatic processing underpins social communication difficulties, psychometric assessment using formal well-controlled tests is necessary. While the range of standardised language tests is extensive, traditional measures do not pick up pragmatic difficulties (Conti-Ramsden et al., 1997), and tests designed for this purpose are possibly not as informative as we might hope. The Test of Pragmatic Language (TOPL-2, Phelps-Gunn & Phelps-Terasaki, 2007) has for a long time (since its first edition in 1992) been the only formal language test focusing wholly on pragmatic language, although other widely-used tests such as the Comprehensive Assessment of Spoken Language (CASL-2, Carrow-Woolfolk, 2017) include pragmatic subtests. These tests require the individual to produce situationally appropriate speech

acts, evaluate the appropriateness of a statement and interpret non-literal/ambiguous language. The TOPL-2 *can* detect pragmatic impairment in groups on the autism spectrum (Young et al., 2005). However, it is not consistently sensitive (Reichow et al., 2008; Volden & Phillips, 2010), and nor is the CASL-2 pragmatic subtest (Klusek et al., 2014). Of course, this might mean that the pragmatic processing of individuals on the spectrum may not be as impaired as assumed, and the social communication difficulties might arise through other cognitive differences. However, there are reasons to question the tests. Firstly, their content does not always square comfortably with what linguists mean by pragmatics, since the tests emphasise codified communication; e.g. what to say when you pick up the phone or how you would greet someone politely. As noted by Cummings (2007), pragmatics is not so much a case of knowing the social rules but of using those rules alongside features of the context to reason about what someone means or what to say in a given situation. To take the example of answering the phone: while it is necessary to know codified formulae for greeting another person, the real pragmatic demand is when you adapt your contributions to the conversation based on what the other person is saying. This is not something that current tests focus on. Pragmatic demands are also likely to be highest in online conversational situations where two or more people are co-constructing a context. Both the TOPL-2 and CASL-2 use static short story prompts that are read to the test-taker slowly and as many times as necessary, and therefore do not assess online pragmatic processing of conversation.

In addition to these questions of content validity, we might also challenge the tests on the basis of their discriminant validity. If our view is that different language skills are potentially separable, we might expect formal pragmatic tests to specifically home in on pragmatics and minimalise other linguistic demands. The TOPL-2 and CASL-2 test manuals do not indicate that

this was their aim in assessing pragmatics. It is perhaps unsurprising, therefore, that there are large correlations between the pragmatic tests and standardised language tests without a pragmatic demand (for example, .73 between TOPL-2 pragmatic index and TOAL-4 spoken language index, which is a composite of two semantic and a word morphology test; Phelps-Gunn & Phelps-Terasaki, 2007). In the case of TOPL-2, Adams (2002) describes this as "a test of high-level language competence, incorporating elements of complex vocabulary, semantics and verbal reasoning", and the authors of the CASL-2 see the large correlations between subtests (including pragmatic and non-pragmatic ones) as desirable, since this supports the combination of subtests into one composite score. However, if tests have not been closely controlled for vocabulary and grammar demands, we cannot be confident that scores truly reflect pragmatics. Of course, any test presented in a linguistic form is going to call on core language skills, and the aim is not to remove this but instead to minimise the amount of variance attributable to vocabulary and grammar skills. If tests used very simple language within the command of all test-takers, we could be more confident that variation in scores reflects differences in pragmatic ability.

There have been several reviews of pragmatic tests available to practitioners (e.g. Adams, 2002; Elleseff, 2015; Norbury, 2014; Russell & Grizzle, 2008; Sobhani Rad, 2014), and these consistently underline the importance of assessing pragmatics but the limitations of existing measures. See the Appendix A for a full list of pragmatic measures collated from these reviews. As the focus of this discussion is on formal tests of pragmatics, I present a subset of that list, comprising just the norm-referenced language tests that apparently tap pragmatics, in Table 1.1. While the measures included in Table 1.1 are certainly useful to the practitioner, none overcomes the problems noted above in relation to the TOPL-2 and CASL-2, and some have additional limitations, the most notable being lack of specificity. While the tests are all marketed or

identified by practitioners/researchers as involving pragmatics, some (e.g. the SLDT and TOPS) incorporate general problem solving and social reasoning skills, which can only be seen as loosely related to pragmatic language. The widespread perception that current measures are limited has led to research interest in new tests, so I carried out a literature search to identify any recent developments in formal tests of pragmatics. The search was conducted in the databases CINAHL, Embase, PsychInfo, and Medline, with the final search on 29/10/2019. Search terms were pragmatic language AND (test# OR assess# OR measur# OR psychometric), and the search period was set as 2014 to the current day, to capture any tests developed since the more recent reviews. As the purpose of the present work is to isolate underlying pragmatic processing, rating scales or questionnaires are not reviewed here, because such measures are arguably more valid in describing overall social communication impairment rather than in tapping pragmatics. Two new formal pragmatic tests with norms were identified, and these are described in detail below. The two pragmatic tests with norms were the Assessment of Pragmatic Abilities and Cognitive Substrates (APACS; Arcara & Bambini, 2016) and Clinical Assessment of Pragmatics (CAPs; Lavi, 2019).

The APACS involves six subtests, four assessing pragmatic comprehension and two assessing pragmatically appropriate production of language. The comprehension subtests include a narrative comprehension task with questions about global meaning, details and non-literal language; a humour comprehension task in which the test-taker needs to decide which is the

Table 1.1

Formal norm-referenced language tests, with a focus on pragmatics or described as having a pragmatic aspect/subcomponent.

Test of Pragmatic Language, Second Edition (TOPL-2)	Phelps-Terasaki & Phelps-Gunn, 2007	6 – 18	Individual is asked a comprehension question or to produce socially appropriate responses across 43-items involving situational and narrative prompts. Organised around six aspects of communication: physical setting, audience, topic, purpose (speech acts), visual-gestural cues, and abstraction.
Comprehensive Assessment of Spoken Language, Second Edition (CASL-2)	Carrow-Woodfolk, 2017	3 – 21	Includes a pragmatic domain, comprising one subtest: individual is presented with social scenarios, and must judge the pragmatic appropriacy of a statement or identify an appropriate thing to say (from age 3). Also includes a supralinguistic domain relevant to pragmatics. Four subtests: (1) nonliteral language (from age 7), (2) double meanings (from age 9), (3) use of context to infer meaning of an unfamiliar word (from age 7), and (4) use of world knowledge to infer full meaning (from age 3).
Diagnostic Evaluation of Language Variation – Norm-Referenced Edition (DELV-NR)	Seymour et al., 2005	4 – 9	Includes a pragmatic domain, assessed by three subtests: (1) producing a contextually appropriate response to a situational probe, (2) narrative generation, and (3) formulating questions to seek specific information.
Clinical Evaluation of Language Fundamentals Fifth Edition (CELF-5) Metalinguistics	Wiig & Secord, 2014	9 – 18	Four subtests: (1) making inferences, (2) manipulating language given a context, (3) understanding multiple meanings; and understanding figurative language. (Superseding Test of Language Competence, TLC)
Assessment of Comprehension and Expression 6 – 11 (ACE 6 – 11)	Adams et al., 2001	6 – 11	General language battery involving five core subtests: (1) sentence comprehension, (2) inferential comprehension, (3) naming, (4) syntactic formulation and (5) semantic decision. In addition, there are two supplementary subtests: (1) nonliteral language, and (2) narrative retelling. Inferential comprehension and the two supplementary subtests are relevant to pragmatics.

Social Language Development Test–Elementary: Normative Update (SLDT – E:NU)	Bowers et al., 2016	6 – 11	Four subtests: (1) drawing mental-state inferences using visual cues, (2) suggesting solutions to peer conflicts, (3) forming multiple interpretations of a picture, and (4) supporting peers with "white lies".
(SLDT-A: NU) Social Language Development Test–Adolescent: Normative Update	Bowers et al., 2017	12 – 17	Five subtests: (1) drawing mental-state inferences using visual cues, (2) interpreting social language such as idioms and nonverbal communication, (3) offering solutions to social problems, (4) supporting others in social interactions, and (5) interpreting ironic statements.
Test of Problem Solving – 3:Elementary (TOPS-3:E)	Bowers et al., 2005	6 – 12	Six subtests: (1) inferences, (2) sequencing, (3) explaining why certain events wouldn't/shouldn't occur, (4) problem solving, (5) predicting, and (6) determining causes.
Test of Problem Solving – 2:Adolescent (TOPS-2:A)	Bowers et al., 2007	12 – 17	Five subtests: (1) inferences, (2) determining solutions, (3) problem solving, (4) interpreting perspectives, and (5) transferring insights.
Social Emotional Evaluation, Second Edition (SEE)	Wiig, 2008	6 – 12	Five subtests: (1) recalling facial expressions, (2) identifying common emotions, (3) recognising emotional reactions, (4) understanding social gaffes, and (5) understanding conflicting messages.
Test of Narrative Language, Second Edition	Gillam & Pearson, 2017	4 – 15	Six subtests, comprising three comprehension tasks and three narration tasks. The comprehension tasks involve literal and inferential questions based on narratives told by the examiner. For the production tasks, the examinee retells a story told by the examiner, tells the story shown by a set of pictures, and makes up a story from a visual prompt.
Bus story	Renfrew, 1995	3 – 8	Individual retells a story told by the examiner using a picture book. Scores can be extracted for story content, sentence length, complexity and level of cueing required.
Expression, Reception and Recall of Narrative Instrument (ERRNI)	Bishop, 2003	4 – Adult	Individual relates a story based on a picture-book, recalls it after a delay and responds to comprehension questions.

funny ending to several short stories; and two subtests focusing on figurative language, specifically metaphors, idioms and proverbs, with one subtest requiring the test-taker to select the correct meaning, and the second requiring them to define the meaning. The two productive tasks include an interview about autobiographical topics, with ratings assigned to several communication parameters, and a description task in which test-takers need to describe the key elements of a series of pictures. The APACS test battery was designed for use with adults with neurologically-based communication difficulties, e.g. following stroke or traumatic brain injury, and showed substantial ceiling effects in the normative sample of 120 healthy adults. While the ceiling effects allow clear identification of individuals with problems, they do mean that the battery would not be appropriate for assessing individual differences in pragmatics across the general population. Other tests, such as the Assessment Battery for Communication (ABaCo, Sacco et al., 2008), have been devised and normed for adults, but as with the APACS, the focus is on clinical groups defined by neurological damage rather than on measuring pragmatic ability as it varies in the general population. Indeed, looking through Table 1.1 showing "pragmatic" tests mentioned in the review papers, there is only one normed test suitable for use in the general adult population – the ERNNI – and this was not even intended as a pragmatic test, as it focuses purely on narrative. Finally, it should be noted that the APACS lacks a focus on flexible conversational contexts where online pragmatic processing may be especially important in shaping communication.

The other normed test battery, the CAPs, is designed for children aged 7 to 18, and includes six subtests, with each based on a sequence of ten videos showing naturalistic social interactions. Three subtests are comprehension-based, with individuals being asked to identify and justify whether anything went wrong in the interaction, and three subtests are production-based, with test-takers being asked what they would say next. The comprehension subtests focus on basic social routines (e.g. how to politely make a request), context-

dependent communication (including sarcasm, colloquial expressions, indirect language etc.) and facial expressions/tone of voice. The production subtests require the individual to appropriately follow social routines, express emotions such as empathy or remorse, and to use appropriate nonverbal communication. The test battery has very good reliability, with all subtests showing test-retest reliability above .9 and inter-rater reliability above .8. The test also seems to have good ecological validity, as it presents videoed conversations that the test-taker responds to in the here-and-now. However, since all subtests require the test-taker to formulate free extended responses, it is questionable whether the assessment isolates pragmatics. In fact, the CAPs looks more like a multidimensional test of complex language function involving pragmatics but also skills in explicit verbal reasoning, expressive vocabulary and grammar, and discourse management. I presented evidence above that the TOPL-2 and CASL-2 conflate a range of language skills, and seeing as there are high correlations between CAPs subtests and the TOPL-2 and CASL-2 (over .85), it seems that the same is likely the case for the CAPs. It should also be noted that the CAPs comprehension subtests are "meta-pragmatic" (i.e. we are asked to judge whether something was inappropriate) and this might involve quite different abilities to online pragmatic comprehension of conversation. During the literature search, two further (non-normed) tests were identified which are worth mentioning here, as they exemplify important distinctions to make between pragmatics and other abilities: the Pragma test (Loukusa et al., 2018) and the Assessment of Metapragmatics (AMP, Collins et al., 2014). The Pragma test assesses several aspects of communication, including contextually-guided inference, the role of relevance in language use, "theory of mind", and emotion recognition. This is clearly not a pure test of pragmatics; indeed, it was not devised as such, but rather to differentiate autistic and non-autistic groups as far as possible. If we were reformulating the test to measure pragmatics more purely, we would need to exclude "theory of mind" reasoning and emotion recognition.

The AMP is similar in format to the comprehension subtests of the CAPs: the test-taker watches videos and explains conversational glitches that occur in the course of interactions. As noted above, this requires "meta-pragmatic" reasoning that might be rather distinct from the comprehension processes that allow us to understand conversation. It is these latter processes that I am especially interested in. One positive point to note about both CAPs and AMP is that they use ecologically-valid videos showing people having live conversations, and this is a format that future tests should aspire to match.

Rationale for Empirical Work: Implicature as an Indicator of Pragmatic Ability

Given the limitations of current formal tests of pragmatics, I set out to design a new test that would tap pragmatic processing. Pragmatics is of course essential in both language production and comprehension. However, I focused on comprehension in devising a test, as it is likely to be easier to isolate pragmatic processing in a receptive rather than productive language test, and scoring can be more objective. As pragmatic comprehension is an open process of inference sensitive to context (Ariel, 2010) or "reading between the lines", it makes sense to home in on uses of language where the meaning is wholly dependent on context, as is the case in particularised conversational implicature. Let us revisit the example discussed above:

> SPEAKER ONE. Can you pick Jane up from the station?
> SPEAKER TWO. I'm working all day.

Here SPEAKER TWO replies indirectly, telling their interlocutor that that they cannot collect Jane. This meaning depends wholly on the communication context; in other contexts, the utterance "I'm working all day" would say nothing about Jane or a station. We only infer this meaning because it follows a question that constrains SPEAKER TWO to respond in a relevant way.

In designing a theoretically motivated test of pragmatics, implicature is therefore a good focus, as implicated meaning can only be interpreted through context-dependent inference (i.e. pragmatic processing). Most existing research into the role of inference in language disorders has been based on the "mental models" approach to narrative processing (e.g. Garnham & Oakhill, 1992). The mental models approach assumes that we build up a mental representation of a narrative (or spoken/written discourse more generally) based on its explicit content, while making inferences to fill any gaps using text-based cues and world knowledge. Some inferences may involve connecting ideas across sentences and others may be more heavily dependent on world-knowledge; terms, such as "bridging inference" and "elaborative inference", are used for these two types, although I will use the term "narrative-based inferencing" to refer to all narrative inferences. Partly, this is to make a clear contrast between narrative-based inferencing and conversational implicature, but also because the two types cannot be reliably measured as separate skills (Language and Reading Research Consortium & Muijselaar, 2018). Narrative-based inferencing when reading a story probably makes different cognitive demands compared to understanding implicature in the course of a conversation. For one thing, implicature involves tracking how interlocutors are collectively creating a communicative context.

In terms of the major research question of this project – whether pragmatics and core language skills can be dissociated – it seems that narrative-based inferencing is highly dependent on core language skills, though other cognitive processes may be implicated too. Individuals with developmental language impairments (including people whose problems are more with social aspects of communication) tend to score lower on inferencing tests, but at a level commensurate with their overall comprehension of literal meaning, suggesting there is unlikely to be a specific issue with inferencing (Adams et al., 2009; Bishop & Adams, 1992; Botting & Adams, 2005). Research has also shown that narrative-based inferencing poses a

challenge for individuals on the autism spectrum, though whether core language skills are responsible for this remains uncertain. Some studies find that autistic individuals underperform compared to those with language disorders (Norbury & Bishop, 2002) and to those matched for language skills (Jolliffe & Baron-Cohen, 1999, 2000), suggesting that forming inferences is a problem in autism, even accounting for other language skills. This would be consistent with the "weak central coherence" account of cognitive differences in autism (Frith, 1989), since this theory stipulates that integrating information into a meaningful global whole might not always come naturally to individuals on the spectrum. However, other studies have found that differences in language ability account for such inferencing deficits (Lucas & Norbury, 2014; Norbury & Nation, 2011). The relationship between autism, language and inference therefore remains an open question.

Where the term implicature has been used in the research literature on language disorders, it has almost exclusively been applied to *generalised conversational implicature*, and usually a particular kind known as scalar implicature. Generalised conversational implicature is a rather different type of meaning compared to the sort of context-dependent implicature discussed above. Grice (1975) used the term for certain expressions that conventionally carry an implied meaning. Consider the use of "some" in this utterance: "There are some well-behaved children in this class." As the speaker could have chosen a stronger word (i.e. "all") from the scale of quantifiers, but deliberately did not, the listener derives a so-called *scalar implicature*: not all of the children are well-behaved. Scalar implicature is a type of generalised conversational implicature because the inferred meaning (i.e. "some" = "not all") is considered typical of this use of language, and not strongly dependent on the context. As a side point, Relevance Theorists view scalar implicature as not implicature at all, but as pragmatically guided development of the logical meaning of particular words/phrases, i.e. *explicature* (Carston, 2002). Therefore, we should be mindful

that implicature is not a single phenomenon, and it is quite possible that some implicature (perhaps particularised conversational implicature, especially) may make greater demands on pragmatic processing, and perhaps cause greater challenges for people with pragmatic difficulties.

As for scalar implicature, individuals on the autism spectrum do not seem to be impaired, and any differences in understanding scalar implicature are best explained by language ability (Chevallier et al., 2010; Hochstein et al., 2018; Pijnacker et al., 2009). Children with SLI do have difficulties with scalar implicature, but only as expected on the basis of their general comprehension difficulties (Katsos & Bishop, 2011). This supports the view that processing scalar implicature depends on structural language skills rather than more complex pragmatic skills. However, individuals on the autism spectrum have been found to have difficulties with processing other types of generalised conversational implicature. For instance, Pastor-Cerezuela et al. (2018) found that autistic individuals were less accurate on a comprehension test in which participants needed to derive implicature; an example item is "Pedro caused Maria's dog to die" (we infer that Pedro did not mean to kill the dog, or else the simpler "Pedro killed Maria's dog" would have been used). This task required application of Gricean conversational maxims; Gricean maxims are rules we are expected to follow when making contributions to a conversation. In the example item above, the specific maxim is that of manner: this is the assumption that the speaker will communicate in the simplest way possible, and if the speaker flouts this maxim, they are attempting to convey some atypical meaning. Surian et al. (1996) found that autistic children also struggled to process utterances that violated such maxims. On the whole, the studies reviewed here indicate that individuals on the autism spectrum have no problem with making rule-based inferences based on the meaning of particular words (e.g. inferring that "some" means "not all") although there is

some reason to believe that difficulties are experienced when making inferences that are guided by conventions of communication (i.e. Gricean maxims).

It is notable that research has rarely engaged with implicature as conceptualised by Relevance Theory and as explored above. A set of papers by Leinonen, Loukusa and others is in some ways an exception (Leinonen et al., 2003; Loukusa, Leinonen, Jussila, et al., 2007; Loukusa, Leinonen, Kuusikko-Gauffin, et al., 2007; Ryder & Leinonen, 2014; Ryder et al., 2008). These researchers report reduced ability in language-impaired and autistic children to make inferences on language tests designed based on concepts from Relevance Theory. Unfortunately, these studies do not focus on implicature in conversational contexts; they focus on making elaborative inferences to fill in semantic gaps in short story material. The overlap between implicature and these types of inference is of theoretical interest – and I explore relationships throughout this thesis – however, issues with implicature may be especially relevant to day-to-day challenges with conversation. As noted above, conversational implicature is likely to make more demand on pragmatic processing, as we need to understand how interlocutors are negotiating communicative contexts.

In the next chapter, I describe development of the Implicature Comprehension Test. The test was designed to measure understanding of this complex pragmatic phenomenon while minimising other linguistic demands, such as vocabulary and syntactic skills. The aim was to develop a test that would be sensitive to individual differences across the general population, in clinical groups and across development. First, I planned to establish whether test scores showed variability in adults, before exploring implicature comprehension in children and in autistic groups, at each step assessing the relationship with other communication skills.

Chapter Two. Relationships between Implicature Comprehension and Other Communication Skills in Non-Autistic Adults[1]

As noted in the previous chapter, there is the clinical intuition that some individuals may have difficulty with using and understanding language in context (i.e. a pragmatic difficulty) but typical grammar and vocabulary development. The upshot of this would be that pragmatic ability and grammar/vocabulary skills might be expected to show some level of dissociation across the general population – though empirical attempts to demonstrate this have largely failed. Researchers have concluded that pragmatics and core language skills are strongly related (Andrés-Roqueta & Katsos, 2017; Matthews et al., 2018). However, this conclusion may be premature if existing tests do not act as pure indicators of particular language skills. My goal for the work presented in this chapter was to devise novel language tests to give a more convincing answer to the question of whether core language and pragmatics are separable skills in the general adult population.

In devising a test of pragmatic processing, I focused on comprehension of particularised conversational implicature, since linguists are unanimous in viewing implicature as a complex pragmatic phenomenon (e.g. Ariel, 2010). As we saw in the previous chapter, implicature is implied meaning that is not readily available in the words themselves but must be inferred from context; for example, I might infer that my friend does not want to sit outside if they say "it's quite cold" in response to my suggestion that we sit in the garden. Since I assume that my friend is communicating relevantly in the context, I interpret their indirect comment about the weather to mean something about where we ought to sit. This sort of meaning is clearly very dependent on context (i.e. on pragmatic

[1] This chapter has been published as Wilson, A. C., & Bishop, D. V. M. (2019). "If you catch my drift…": Ability to infer implied meaning is distinct from vocabulary and grammar skills. *Wellcome Open Research, 4*(68). https://doi.org/10.12688/wellcomeopenres.15210.3

processing) as “it’s quite cold” could mean rather different things in other situations. In my new Implicature Comprehension Test (ICT), I aimed to place a high demand on pragmatics, while also aiming to remove as much as possible the effect of grammar/vocabulary skills on scores by (a) using simple language and (b) incorporating control items (see below). As well as assessing the extent to which pragmatic skill is separable from other language abilities, the test was also designed to be clinically applicable in identifying impairments characteristic of autism and social (pragmatic) communication disorder (SPCD) (e.g. Loukusa & Moilanen, 2009), and to be sensitive to developmental and individual differences. Based on these aims, I was mindful of making the Implicature Comprehension Test (ICT) sufficiently child-friendly, while avoiding ceiling effects in adults.

Study One: A First Exploration of the Implicature Comprehension Test

In this study, I set out to establish the reliability of the Implicature Comprehension Test and the distribution of scores expected in the general adult population. I also planned to assess how implicature processing relates to other language skills and to the broad autism phenotype.

Methods

For the preregistered protocol, please see *https://osf.io/t54hm/*. The study was granted ethics clearance by the Medical Science Interdivisional Research Ethics Committee at Oxford University in April 2018 (Ref: R57087/RE001).

Participants. 120 adults were recruited online through the participant recruitment platform, Prolific (*https://www.prolific.ac/*). All participants reported speaking English as a first language and living in the UK. Individuals were also required to have no significant uncorrected hearing or visual impairment.

Procedure. This was an online study supported by Gorilla (*https://gorilla.sc/*), a cloud-based tool for collecting data in the behavioural sciences, and all data were collected via Prolific in May 2018. Participants completed the study at a time and place of their convenience. The study took around 30 minutes to complete and individuals were remunerated £5 for their time.

Measures. Participants completed four language tests in the order set out below. Following the tests, individuals were asked to complete the Autism-Spectrum Quotient (AQ; Baron-Cohen et al., 2001), a self-report questionnaire measuring autistic traits.

Implicature Comprehension Test (ICT). This task involves watching a series of 55 short cartoon videos (each is approximately 8 s in length). Each video consists of a conversational adjacency pair between two characters: typically, this is a closed question from Character 1 (eliciting a "yes" or "no" response) followed by a response from Character 2. Each utterance is between 6 and 8 words in length, and age of acquisition of the words does not exceed approximately 7 years of age (based on ratings of Kuperman et al., 2012). Following the adjacency pair, the participant hears a comprehension question; typically, this echoes the question posed by Character 1 during the video. The participant then hears a bleep, and is asked to give a "yes", "no", or "don't know" response to the question, recorded by keyboard presses. Example items can be seen here: *https://osf.io/46tyz/*.

In 27 of the 55 trials, Character 2's response provides an answer to Character 1's question that must be inferred via implicature. Half of these represent a "yes" response; half a "no" response.

Example:	
Character 1	"Could you hear what the police said?"
Character 2	"There were lots of trains going past."
Comprehension Question	"Do you think she heard what the police said?"
Correct Answer	"No."

In a further ten trials, Character 2's response provides a more explicit response to Character 1's question. The structure of these items and the overall language level was similar to the implicature items, and therefore they represent positive control items, designed to check that basic language comprehension and task structure did not cause any problems. In this study, mean accuracy on these items was 95.2% (SD = 10.05%), indicating that these items functioned well.

Example:

Character 1	"Did you see the policemen earlier on?"
Character 2	"I saw them standing on the platform."
Comprehension Question	"Do you think he saw the policemen?"
Correct Answer	"Yes."

In a further ten trials, Character 2 provides a "don't know" response to the question, and for these trials, the correct response to the follow-up comprehension question is "don't know". These trials are designed to legitimise "don't know" responding so that participants don't feel that providing a "don't know" response always represents an incorrect answer. The aim of this is that participants who are less sensitive to implicature are likely to provide "don't know" responses to implicature items too. In this study, the average participant gave a "don't know" response to 81.2% of these items (SD = 20.3%).

Example:

Character 1	"Did the police speak to anyone else?"
Character 2	"I wasn't watching them much."
Comprehension Question	"Do you think the police spoke to anyone else?"
Correct Answer	"Don't know."

Finally, the task involves a further ten "open context" implicature items. In these items, Character 1 produces a statement rather than a question. Character 2's response implicitly addresses Character 1's statement. The follow-up comprehension question assesses whether participants have appreciated the implicature; half of the comprehension questions are correctly answered by "yes", half by "no".

Example:

Character 1	“Normally the station doesn’t get busy.”
Character 2:	“Lots of people were coming on holiday today.”
Comprehension Question	“Do you think the station was busy?”
Correct Answer	“Yes.”

Utterance length and psycholinguistic variables (word frequency, word age-of-acquisition and word concreteness) are controlled for across the different item types. There were two measured variables: the sum of implicature items correctly answered (out of 37) and the sum of control items correctly answered (out of 10).

Synonyms Test (Receptive Vocabulary). This 25-item receptive vocabulary task was devised for this study. In each trial, participants choose which of four words is a synonym for the target word. This is a timed task (up to 12 s per item). There was one measured variable: the sum of items correctly answered (out of 25).

Idiom Decision Test. In this test of idiom recognition devised for this study, participants are presented with 40 three-word phrases, each including a transitive verb and its direct object. Half the phrases are idioms (i.e. conventionalised phrases with a meaning that extends beyond the meaning of the component words; e.g. “make your day”) and half are incorrect adaptations of common idioms (e.g. “bite the dirt”, instead of “bite the dust”). Idioms are presented via audio, and participants need to decide as quickly as they can if the phrase is an idiom or not. Responses are given by keyboard presses. The variable of interest was total accuracy (1 point per correct answer).

Test of Complex Syntax-Electronic (TESC-E; Frizelle et al, 2017; see https://osf.io/5ntvc/ for full details). In this test of receptive syntactic processing, participants watch a series of videos and for each video decide whether an auditorily-presented sentence correctly describes activity in it. The task was originally designed for use with children, and so I incorporated some adaptations to make it more appropriate for adults. In the original version, accuracy was the variable of interest, though a ceiling effect was

expected for accuracy in adults, so instead I focused on mean reaction time (RT). In making this change, it was necessary to present the sentence at the end of the video rather than during it, which was the original format; this meant that an individual's response could be timed from the onset of the sentence. Furthermore, I only included the relative clause items, and not the adverbial clause or sentential complement ones. The adverbial clause items make too high a demand on remembering the sequence of events when the video and audio are not presented simultaneously, and the sentential complement items (e.g. "She thinks [that] …", "He wishes [that] ...") are likely to demand "theory of mind" processing that ideally should be removed from this task to ensure that it is a relatively pure measure of structural language skills. In this adapted form of the task, participants watch a series of 20 videos, each around 4 seconds (shorter than in the original version to increase the speed of the task for adults). After each video, participants hear a biclausal sentence incorporating a main clause and a relative clause. There are five types of relative clauses: transitive subject relatives, intransitive subject relatives, direct object relatives, indirect object relatives, and oblique relatives. Participants indicate via keyboard presses whether the sentence correctly describes the video.

Results

See Table 2.1 for descriptive statistics and the reliability of the measures. As shown in the table, all the measures had good internal consistency, indicating that they were subject to relatively little measurement error. The sample was relatively high in autistic traits, as shown by AQ scores. These were significantly higher than the population norm of 16.94 (as identified by Ruzich et al., 2015 in a meta-analysis of 6900 nonclinical individuals), $t(118) = 5.93$, $p < .001$. Eight individuals had an AQ score of at least 32, which is taken to be a cut-off for possible autism.

Table 2.1

Descriptive statistics and reliability coefficients (Cronbach's alpha and standard error of measurement).

	N	Mean	SD	Min	Max	Skew	Kurtosis	Alpha	SEm
Implicature, control items	118	9.62	0.63	8	10	-1.38	0.72		
Implicature	118	27.19	4.08	14	35	-0.79	0.37	.80	1.82
Vocabulary	120	13.43	5.34	3	25	0.12	-1.03	.85	2.07
Idioms	113	35.76	3.19	26	40	-1.31	1.23	.83	1.32
Accuracy on syntax	114	19.31	0.80	17	20	-1.01	0.46		
Mean RT on syntax (ms)	114	3320	447	2188	4721	0.77	0.81	.84	179
AQ score (out of 50)	118	20.65	6.92	5	39	0.33	-0.23	.79	3.17

Table 2.2

Correlations between language tests and AQ. There has been list-wise deletion of cases with outlying scores (N = 108).

	Vocabulary	Idioms	Mean RT on syntax	Total score on AQ
Implicature	-.04	-.10	.01	-.09
Vocabulary		.49	.25	.04
Idioms			.24	.12
Mean RT on syntax				.02

For each variable, I excluded scores according to the criteria of Hoaglin & Inglewicz (1987) for outlier exclusion: 2.2 times the interquartile range below the lower quartile (and above the upper quartile in the case of syntax RT). In analysing the data, I used listwise deletion of any cases with an outlying score. See Table 2.2 for Spearman's correlations between the five test variables.

The correlation matrix suggests that the vocabulary and idiom tests measured closely related semantic skills, and that the syntax test was loosely related to these tests. It is worth noting that the syntax task seemed to have been overly simple for typical adults (as indicated by the strong ceiling effect on accuracy), and so the main demand is unlikely to have been language processing – this perhaps accounts for the low correlations observed between it and the other language tests.

Discussion

In this pilot study, I found that the ICT was internally consistent, and that implicature scores did not correlate significantly with semantic knowledge (measured by tests of vocabulary and recognition of conversational phrases/idioms). We cannot attribute this lack of a relationship between ICT scores and the other language measures to noisy tests, as reliability was consistently good across the measures, as indicated by the Cronbach's alpha values. Similarly, there was no issue with restricted variance on the ICT; while scores were a bit skewed towards the maximum mark, there was nonetheless substantial variability. Clearly the skills assessed by the ICT were not closely related to other language skills measured by the test battery. These results therefore provide some evidence that implicature comprehension dissociates from basic linguistic decoding.

However, the lack of correlations between scores on the ICT and any of the other measures raises the question of precisely what the test is measuring. It was designed to put a high demand on pragmatic processing, and from a linguistic perspective, the test does indeed show face validity as a test of pragmatics, since it focuses on inferring meaning from context, i.e. the linguistic definition of pragmatics. However, it is possible that the ICT measured skills that are very task-specific and not necessarily applicable to understanding day-to-day conversation in context. In other words, it is unclear how well the test targeted pragmatics as a domain. To address this issue, we need to investigate whether the ICT relates to other tests involving conversation and social understanding. This was the aim of Study Two, described below.

One final observation is that the AQ, as a measure of autistic traits, was not closely related to the other measures. This suggests that the broad autism phenotype was not associated with language skills tested here. This was unexpected, as I thought AQ scores

would be negatively related to ICT scores, since autism has been linked to difficulties processing implied and indirect meanings. It should be noted that the sample had a full range of AQ scores, and indeed scores were somewhat higher than the population norm, suggesting that there ought to have been sufficient variance in scores to detect an effect. Similarly, low power is not a plausible explanation, as the sample was powered at 92% to detect a correlation of .3 with the alpha level set to .05 and .3 taken to be the smallest clinically relevant effect size. In summary, the relationship between implicature comprehension and autistic traits as they vary throughout the general population was confidently null. On the other hand, it is possible that a relationship with autistic traits only becomes apparent in clinical groups, and we will explore that question in later chapters.

Study Two: Testing a Two-Factor Model of Language Processing in the General Population

In this next study, I collected some more normative data on the ICT in typical adults. I set out to replicate the finding in Study One that scores on the ICT were distinct from semantics, and also to test for a dissociation between implicature comprehension and grammatical ability too. This was my key question: is pragmatic processing (as measured by the ICT) separable from core language skills (grammar and vocabulary)? My secondary question was whether performance on the ICT was related to performance on other tests measuring social understanding in conversational and other contexts. As noted in the previous chapter, and shown again here in Figure 2.1, the capacity to infer intended meanings from context - pragmatic comprehension - is thought to be a subdomain of our more general ability to recognise other people's intentions, our so-called "theory of mind". This is the theoretical position of Relevance Theory (Sperber & Wilson, 1986) and has found empirical support (Happé, 1993). Conceptually, then, we would expect the test of pragmatic

comprehension - the ICT - to cluster with other tests requiring social inferences in conversational and other contexts, as they involve "theory of mind". As such, I developed a test battery including two tasks that I expected to require sensitivity to conversational contexts as well as a test requiring the individual to infer mental states in different scenarios.

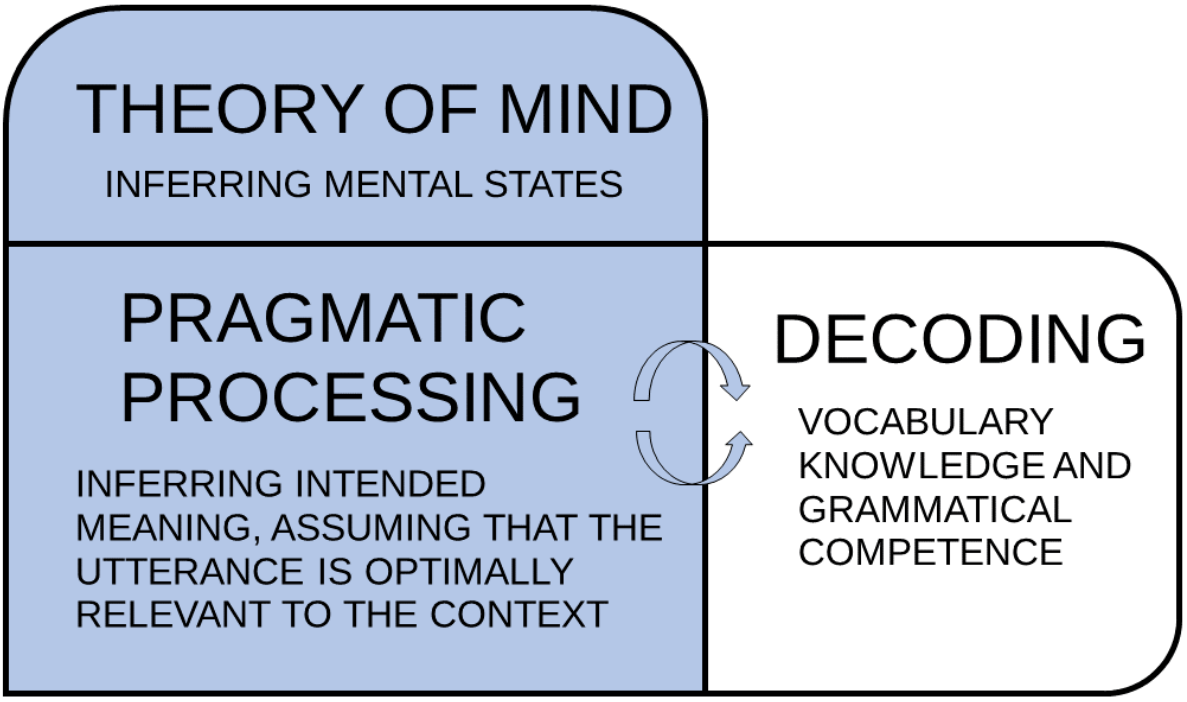


Figure 2.1. My model of language processing, as borrowed from Relevance Theory. Language comprehension is an interaction between decoding and inference. Pragmatic processing is a subdomain of our more general "theory of mind".

I hypothesised that a two-factor model ("core language" and "social understanding") would account for individuals' performance across the tests. Reviews of existing research indicates that "pragmatic" measures tend to correlate with tests of core language skills (Andrés-Roqueta & Katsos, 2017; Matthews et al., 2018), although we should bear in mind that these "pragmatic" measures suffer from problems of poor specificity as noted in the previous chapter, which may inflate the correlations. However, the possibility that individuals can have relatively specific difficulties with pragmatics (Lam & Yeung, 2012) led me to expect that a two-factor structure would fit the battery better than a one-factor structure where all the tests measured some general ability.

Methods

This study was preregistered on the Open Science Framework (OSF) with the title "The relationship between comprehension of conversational implicature, core language skills and social cognition": *https://osf.io/g4bvm*. The study was granted ethical clearance in July 2018 by the Medical Sciences Interdivisional Research Ethics Committee at the University of Oxford: Ref. R57087/RE002. I report below how I determined the sample size, all data exclusions, all manipulations, and all measures in the study.

Power Calculation. The sample size was based on a power calculation using simulations implemented by the R packages simsem (Jorgensen et al., 2018) and lavaan (Rosseel, 2012). I used a p-value of .05 to indicate statistical significance: this reflects a single preregistered statistical test for the confirmatory factor analysis (CFA) used in this study. In determining power for the CFA, I simulated data conforming to a two-factor correlated traits model with a "core language" factor relating to three tests and a "social understanding" factor relating to four tests. Factor loadings were set at .7 and the factors were correlated at .2 (a low correlation was used, as the semantic tests did not correlate with the ICT in Study One). In 10,000 simulations of samples of 120 individuals, 9989 datasets showed a significantly better fit to a two-factor compared to a one-factor model, when compared using a chi-square test. This gives power of over 99% to detect the improved fit of a two-factor over a one-factor model.

Participants. I recruited 120 adults through the online platform, Prolific. Eligibility criteria included: (i) age over 18, (ii) no autism diagnosis, (iii) private access to a computer with a good internet connection, (iv) no significant visual or hearing impairment, (v) native-level fluency in English, and (vi) no participation in Study One.

Mean age of the participants was 30 years;11 months (*SD* = 11 years;3 months, range = 18 - 64 years). 65 identified as women, 54 as men, and one person did not declare their gender. The majority of the sample described their race/ethnicity as White (103 out of 120); four people described themselves as Mixed Race, four as Black, and eight as Asian. 34 of the participants reported that they were students, of whom 28 were completing undergraduate studies and six were postgraduates. Of the 86 individuals who did not describe themselves as students, highest level of education was high school/secondary school for 18 individuals, vocational training/college courses for 13, bachelor's degree for 53, and a higher degree for nine.

Procedure. This was an online study supported by Gorilla, the online tool for presenting experiments and surveys. Participants completed the study at a time and place of their convenience. The study took around 30-45 minutes to complete and individuals were remunerated £5 for their time.

Measures. Seven short language tasks were presented in the order described below. Further details of the tasks are presented in Appendix B and information on accessing the full assessments can be found here (*https://osf.io/xn48e/*).

Implicature Comprehension Test (ICT). This was as described above, although one slight alteration was made. Originally, there was no time limit for responding, but in order to more closely simulate the need for quick processing of implied meaning in actual conversation, a 5-second time-limit for responding was imposed. In this study, there were few instances of time-outs. Among the 117 participants whose data were included for this task, responses were missing for only 0.5% of trials. The most trials missed by any one participant was four, and 94% of participants missed one or fewer trials (83% did not miss any).

Synonyms Test (Receptive Vocabulary). As described above.

Grammaticality Decision Test. In this test of receptive grammatical processing devised for the study, participants listen to sentences and decide if they are well-formed and grammatical. There are 50 items and half are grammatical. Grammatical violations represent mistakes that native speakers would not tend to make, such as using incorrect auxiliary verbs (e.g. “If I will see Ann today, I’ll ask her opinion”) or atypical placing of adverbs (e.g. “If you can’t find it, I can send again the letter”). Participants have up to six seconds to listen to each sentence and indicate by a button press whether or not it is grammatical. There was one measured variable: the sum of items correctly answered (out of 50).

Awkward Dialogues. This task assessing sensitivity to social awkwardness is based loosely on the Faux Pas Recognition Test (Baron-Cohen et al., 1999). Individuals detect discomfort or offence implicitly conveyed in short dialogues. The test was designed not so much to measure a single skill but rather to tap general conversational competence, including pragmatics, mental state attribution, and understanding of paralinguistic cues such as intonation, as well as core language skills.

In the Awkward Dialogues, participants listen to eight short dialogues of around 80-90 words in which two characters each take five conversational turns. In five of the dialogues, one of the characters says something that is socially awkward. Three of the dialogues are control stimuli, in which nothing awkward is said. Participants need to indicate whether something awkward was said, and if they indicate that it was, they are asked to produce a written response to explain what was awkward, why the interlocutors spoke as they did and how they might have felt.

Participant responses on the five awkward dialogues were marked out of two. Two marks were given if the response expressed the main point that made the dialogue awkward, one if the response gave a glimmer of a correct answer, and zero if the participant missed the

awkwardness in the dialogue. Responses from 50 participants were checked by a second independent marker; Cohen's kappa was .78 indicating good inter-rater reliability. Scores for the five awkward dialogues were entered into an item response model, so that factor scores could be extracted as the measure of participant ability on this test. Factor scores were used rather than sum totals, since it was not clear that the intervals in the mark-scheme ought to be seen as equal. Using a polytomous item response model allowed a solution to this issue, as such models treat data as ordinal and model participant ability as a function of the pattern of responses across items.

In addition to being asked whether dialogues are socially awkward, participants are asked a factual recall question to check basic comprehension of each dialogue (one mark for each of five questions). Participants were excluded if they incorrectly answered more than one factual recall question, as these individuals were outliers (according to the outlier definition below). Where participants incorrectly answered one factual recall question and did not score two marks for that dialogue, their mark for that item was deleted and re-imputed by the item response model based on their scores for the other four awkward dialogues.

Test of Fillers and Backchannels. I expected this newly devised test to require sensitivity to the role of conversational fillers in turn-taking. Fillers are thought to have functions in conversation; they are not merely "rubbish" produced by inefficient language production systems. For instance, "um" and "uh" are used to claim or hold the floor (Clark & Fox Tree, 2002). Meanwhile, backchannel continuers, such as "mhmm" and "uhuh", cede the floor to the interlocutor (Jurafsky et al., 1998). As these examples indicate, many fillers are non-lexical speech sounds which do not have substantive meaning but which nonetheless have a communicative function in negotiating who speaks in an interchange. I expected comprehension of fillers/backchannels to allow quite a "pure" test of sensitivity to the turn-

taking mechanics of conversation, as this should make minimal demand on core language skills (i.e. vocabulary and grammar), especially given the non-lexical nature of many fillers/continuers.

Participants watch short videos in which Character 1 makes an utterance of between 5 and 9 words. Next, Character 2 produces a word or non-lexical speech sound. The video then cuts off before anything else can be heard. Participants need to indicate who they think would speak immediately after what is observed in the video. They provide their answer by clicking a button showing the face of the character. A third button shows a question mark that participants may click if it is "very difficult to say". The task includes 40 items. Character 1 says 20 different utterances in the course of the task, with each one produced twice. Character 2 follows up the utterance with a backchannel continuer (or the repair initiator "huh?") one time and a filler claiming the floor the other time. Where a backchannel continuer or "huh?" is used, participants need to select Character 1 as the person likely to speak, and Character 2 where a filler claiming the floor is used. There are four backchannel continuers (mm-hmm, uh-huh, really, right) and five fillers claiming the floor (um, uh, yeah, oh, well). The backchannel continuers are much more often used in this role than to signal an incipient speaker (e.g. in corpus analysis by Jurafsky et al., 1998). With the five fillers, "um" and "uh" are taken to mark a slight pause in which Character 2 formulates what they want to say, and "yeah", "oh" and "well" are fillers claiming the floor. "Yeah" can be used as a backchannel in conversation, though it is much more commonly used to claim the floor than other backchannels (Drummond & Hopper, 1993) and its function is often signaled by the intensity with which it is spoken (Trouvain & Truong, 2012). I therefore use prosodic information to disambiguate "yeah" (and do the same for "oh", since this this is likely to function similarly to "yeah"). There was one measured variable: the sum of items correctly answered (out of 40).

Test of Local Textual Inference. Also devised for the present study, this task was used to assess narrative-based inferencing so that the relationship between this kind of inferencing and other language skills could be evaluated. In processing narrative (or spoken/written discourse more generally), it is assumed that we construct a coherent mental representation of a narrative based on its explicit content, while making inferences to fill any gaps using text-based cues and world knowledge (e.g. Garnham & Oakhill, 1992). This type of inferencing depends heavily on core language skills, such as semantics (Adams et al., 2009; Bishop & Adams, 1992; Botting & Adams, 2005; Lucas & Norbury, 2014) and so we expected it to reflect core language competence rather than pragmatics/social understanding in this study.

Participants read two 100-word sections of a short story and after each section they respond to ten questions with a word or short phrase. Participants have as long as they like to read the text, and then up to 25 seconds to type their response to each individually-presented question. The text remains on the screen when the questions are asked. The questions assess whether participants can make coherence inferences to build up a comprehensive representation of a text. Participants are informed that they may respond that they don't know the response to a question (if there's no relevant information, for example). Four questions are correctly answered by "don't know". Two marks were awarded for a correct response and one mark for a partially correct response, making a maximum total of 40 points.

Frith-Happé Animations (Abell et al., 2000). In this test of mental state attribution, there are a sequence of animations showing two moving triangles that sometimes interact. In this study, individuals were presented with a shorter version of the task, which included all the animations showing a "theory of mind" scenario or goal-directed behaviour; the animations depicting random movement were not presented. The "theory of mind"

animations show the triangles interacting as if they are trying to influence the thoughts and feelings of each other. The "goal-directed" animations show the triangles physically interacting (e.g. fighting). In this version of the task, participants watched animations around 20 s in length (original clips were shortened), before providing a typed answer describing what happened in each animation. As in the original instructions, participants were told that the triangles would interact, and sometimes they would interact as if they were aware of each other's thoughts and feelings. There were five "theory of mind" trials (the four in the original task, plus one of the items originally designated as practice) and four goal-directed trials. The original goal-directed practice item was given as a single practice item. Each of the participant's written descriptions were scored on their appropriateness (i.e. how accurately they inferred the scenario the cartoon represents) out of three according to the mark scheme used by Castelli et al. (2000). A second independent marker also scored the responses given by 50 participants; Cohen's kappa was .79 indicating good inter-rater reliability. Scores for the "theory of mind" animations were entered into an item response model, for the same reason as the scores for the Awkward Dialogues, and factor scores were used as the measured variable for this task.

Data Analysis. Data were analysed using R software (R Core Team, 2017). The data and analysis scripts are accessible on OSF: *https://osf.io/8vjyq*.

I assessed item functioning and the reliability of the measures using classical test theory (CTT) and item response theory (IRT; Embretson & Reise, 2000); R packages psych (Revelle, 2017) and mirt (Chalmers, 2012) were used for this analysis. For each item in each test, I inspected accuracy and the correlation between item-accuracy and total-accuracy on the test with that item excluded. Items were identified as poor if they had low accuracy and a low item-total correlation. I also inspected item characteristic curves (ICCs) produced using

IRT analysis. This involves determining difficulty and discrimination parameters for each item based on patterns of responses across the test, and then using these parameters to plot the probability of a correct response to a specific item across the ability spectrum. The best ICCS will show a steep curve, as individuals of greater ability are more likely to score correctly on an item; on the other hand, low flat curves indicate items that are ambiguous, with no consensus answer at any point along the ability spectrum. Any items showing this pattern were excluded. I then computed the reliability of the measures using CTT coefficients: Cronbach's alpha (with 95% CI), standard error of measurement, and Revelle's beta. These are reported below, alongside a summary of the item-total correlations and item-level accuracy for each test. Revelle's beta is the worst split-half reliability – i.e. the correlation between the halves of a test when it is split in such a way that the correlation is minimised (Zinbarg et al., 2005). Beta was used to give an estimation of general factor saturation that removes the influence of "test-lets", i.e. clusters of items within a test that are more highly correlated than other items because the stimuli may be more similar, etc. Where there is a large discrepancy between alpha and beta, the measure is heterogenous in its content. I also carried out IRT analysis, modelling test performance as a function of item characteristics and participant ability. IRT modelling assumes that the test is unidimensional, and to indicate whether this was indeed the case, I report root mean square error of estimation (RMSEA) for a unidimensional model for each test. ICCs were inspected, as described above, when assessing item functioning, and generated estimates of reliability by ability level. IRT reliability plots were created using R package ggplot2 (Wickham, 2009). IRT modelling is also useful in testing how well all items on a test represent a single latent ability (i.e. whether the test is unidimensional), and report the root mean square error of estimation (RMSEA) for each model.

Any individual with at least one outlying score on any test was excluded from the dataset. Outliers were defined according to Hoaglin and Iglewicz (1987) as 2.2 times the interquartile range below the lower quartile. I also excluded any individuals who had outlying scores on the explicit-response items of the ICT (this threshold was 8/10) or on the factual questions of the Awkward Dialogues (4/5), as these were taken to be control variables identifying individuals who did not engage well with the tasks.

The data were inspected for normality and multivariate outliers using R package MVN (Korkmaz et al., 2014). Multivariate outliers were defined as individuals whose adjusted Mahalanobis' distance was above the 97.5th percentile of the chi-distribution. Maximum likelihood estimation was used for the factor analysis. As this is based on the multivariate normal distribution, it was important to check whether the data departed from this distribution (Flora et al., 2012), although given that robust estimation was used in the factor analysis, assumptions of normality were relaxed (Finney & DiStefano, 2013). Nonetheless, in the case of multivariate outliers being present, I planned to transform variables using the Tukey ladder of power transformations using R package rcompanion (Mangiafico, 2018) to reduce skew, and then test for multivariate outliers again. Remaining outliers would be excluded. The sensitivity of the analysis to any data transformation or outlier exclusion was evaluated by comparing results based on transformed and non-transformed data with and without outliers included.

I had one preregistered analysis for this study: a confirmatory factor analysis (CFA). I specified a two-factor correlated-traits model with a "core language" factor and a "social understanding" factor; see Table 2.3 for which variables were set to load on which factor. I used maximum likelihood estimation with robust standard errors and a Satorra-Bentler corrected chi-square test to evaluate whether a two-factor model fitted the data significantly

better than a one-factor model on which all seven variables were set to load on a single factor. The conventional alpha level of .05 was used to indicate statistical significance, as there was only preregistered one statistical test. This analysis was carried out using R package lavaan (Rosseel, 2012). I report confirmatory fit indices (CFIs) and root mean square error of estimation (RMSEA) with 90% confidence intervals. R package SemPlot (Epskamp & Stuber, 2017) was used to make a visualisation of the factor model.

Table 2.3

Measured variables for each factor.

Test	Study Variable	Details
"Social understanding" factor		
Implicature Comprehension Test	Implicature Comprehension	Total score (36 items, 1 point each)
Test of Fillers/Backchannels	Comprehension of Fillers/Backchannels	Total score (40 items, 1 point each)
Awkward Dialogues	Sensitivity to Social Awkwardness	Factor score derived from IRT model
Frith-Happé Animations	Mental State Attribution	Factor score derived from IRT model
"Core Language" factor		
Test of Local-level Inferencing	Narrative-based Inferencing	Total score (20 items, 2 points each)
Grammaticality Decision Test	Receptive Grammar	Total score (44 items, 1 point each)
Synonyms Test	Receptive Vocabulary	Total score (25 items, 1 point each)

Finally, one exploratory analysis was carried out. As I was particularly interested in whether implicature comprehension was distinct from core language skills, I quantified the proportion of variance shared between implicature comprehension and core language skills (receptive vocabulary and receptive grammar). I also quantified the proportion of variance in implicature scores predicted by the tests tapping pragmatic processing/inferential skills over and above the effect of core language skills. For this analysis, a hierarchical multiple regression was used. Receptive vocabulary and receptive grammar were entered as predictors

of implicature comprehension in the first stage. In the second stage, the conversation comprehension variables (comprehension of fillers/backchannels and sensitivity to social awkwardness), narrative-based inferencing, and mental state attribution were added to the model. I report *F*-statistics, *p*-values, and adjusted *R*-square values for the stages.

Results

Table 2.4 shows descriptive statistics for each language measure. The scores for the Frith-Happé Animations and the Awkward Dialogues are IRT factor scores; for context, the mean raw total for the former (just the "theory of mind" animations) was 10.04 out of 15 (SD = 2.59) and for the Awkward Dialogues was 6.65 out of 10 (SD = 2.49).

Reliability Analysis. When assessing item functioning, there were a few weak items in the Grammaticality Decision Test: six showed chance-level accuracy and low item-total correlations, and their ICCs were low and flat. This suggests that the six items were ambiguous, with judgements of grammaticality being essentially random across the ability range, so they were dropped from analysis; maximum score on the final version of the test was therefore 44. In the ICT, there was one similarly poor item, and it was removed from analysis. Maximum score on this test became 36. No issues with item functioning were detected in any other test.

Table 2.4

Descriptive Statistics.

	N	Mean	SD	Min	Max	Skew	Kurtosis
Implicature	117	28.90	4.50	8	36	-1.20	2.79
Fillers/Backchannels	120	26.80	5.93	7	37	-0.83	0.79
Awkward Dialogues	118	0.01	0.75	-1.83	0.94	-0.57	-0.57
Frith-Happé Animations	120	0.00	0.74	-2.23	1.47	-0.38	-0.18
Inferencing	115	34.40	3.24	24	40	-0.73	0.30
Grammar	118	35.20	4.21	25	44	-0.52	-0.37
Vocabulary	120	12.10	4.45	3	24	0.31	-0.26

The final versions of the tests were reliable, as indicated by the CTT reliability coefficients shown in Table 2.5. It is notable that item-level accuracy was quite variable on the tests. For some items, there was very high agreement between participants; these items were clearly very easy to interpret or involved highly salient communicative cues. Other items were more difficult. These more difficult items correlated well with total scores excluding that item, indicating that they reliably tapped a particular skill.

Table 2.5

Reliability Analysis, including Cronbach's alpha and 95% confidence intervals, standard error of measurement (SEm), Revelle's beta, and IRT RMSEA for a unidimensional model for each test. I also present item-level statistics: corrected item-total correlations (totals excluding the item) and item-level accuracy.

						Item-total correlations			Item-level Accuracy		
Test	Alpha	Alpha 95% CI	SEm	Beta	IRT rmsea	Q1	Q2	Q3	Q1	Q2	Q3
Implicature	.82	.77, .86	1.91	.68	.01	.23	.30	.38	0.68	0.86	0.91
Fillers	.81	.76, .86	2.58	.55	.08	.20	.27	.35	0.51	0.67	0.85
Inferencing	.83	.78, .87	1.34	.68	.04	.34	.42	.52	1.60	1.78	1.89
Grammar	.81	.76, .86	1.84	.65	.05	.15	.27	.36	0.76	0.89	0.93
Vocabulary	.76	.70, .82	2.18	.58	.04	.22	.31	.35	0.28	0.49	0.58

See Figure 2.2 for reliability plots of the tests that were analysed using IRT modeling. IRT models compute standard error of measurement as a function of participant ability, and so this can be used this to estimate reliability [SEm = SD*sqrt(1-reliability)]. I present reliability curves for the Awkward Dialogues and Frith-Happé Animations, and also include the IRT curve for the ICT, as I was particularly interested in how this task functioned across the ability spectrum. The ICT seems particularly reliable in the lower end of the ability spectrum. With respect to the Awkward Dialogues and Frith-Happé Animations, the low number of items (only five) seems to have limited reliability. This IRT analysis is only valid insofar as a unidimensional model fits the data, so I report the RMSEA values for these models, which support unidimensionality: Awkward Dialogues, RMSEA = .06; Frith-Happé

Animations, RMSEA = .00; ICT RMSEA = .01.

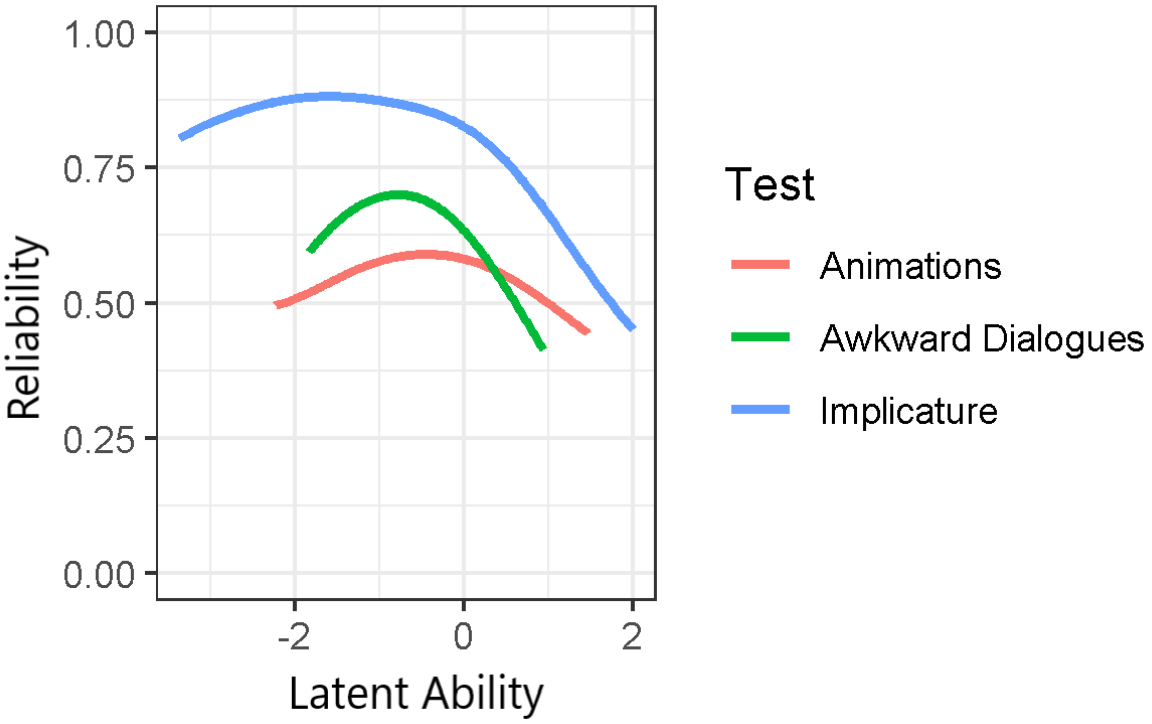


Figure 2.2. A plot showing reliability as a function of latent ability, modelled for three of the social communication tests. For each test, I only model reliability for the range of scores achieved by the participants in the sample.

Confirmatory Factor Analysis (CFA). A few scores were identified as univariate outliers. I performed listwise deletion of cases including any outliers, meaning that data from 112 individuals were used in the CFAs reported below. Accuracy on six of the language tests was not normally distributed; the Shapiro-Wilk test was only non-significant for vocabulary. Also, there were five multivariate outliers in the non-transformed data. Therefore, the six variables were transformed. Following transformation, there were no multivariate outliers. Note that the CFAs reported below were also run using the non-transformed data, and results were very similar. Table 2.6 presents the correlation matrix for the transformed language variables.

Table 2.6

Correlations between language tests. Variables have been transformed and there has been list-wise deletion of cases to show only the data to be modelled in the CFA (N = 112).

	Fillers/ Backchannels	Awkward Dialogues	Frith-Happé Animations	Inferencing	Grammar	Vocabulary
Implicature	.34	.22	.29	.23	.30	.19
Fillers/ Backchannels		.12	.20	.17	.23	.15
Awkward Dialogues			.12	.29	.20	.11
Frith-Happé Animations				.22	.06	.16
Narrative-based Inferencing					.09	.11
Grammar						.45

I ran two CFAs comparing one and two factor models. A two-factor model fitted the data well, CFI = .95, RMSEA = .05, 90% CI [.00, .11], whereas a one-factor model did not, CFI = .87, RMSEA = .08, 90% CI [.00, .13]. The Satorra-Bentler scaled chi-square difference test showed that the two-factor model gave significantly better fit, $\chi 2(1) = 5.48$, $p = .020$. Inspection of the residuals of the two-factor model indicated that there was some mis-specification in this preregistered model. The four highest residuals were all for correlations between narrative-based inferencing and the tests clustering as the "social understanding" factor, indicating that these relationships were not well accounted for by the model. Therefore, I re-specified the two-factor model with vocabulary and grammar loading on one factor and all the other tests loading on a second factor. This new two-factor model showed good fit, CFI = 1, RMSEA = .00, 90% CI [.00, .08]. The Satorra-Bentler scaled chi-square difference test indicated that this two-factor model was significantly better than the one-factor model, $\chi 2(1) = 12.47$, $p < 0.001$. See Figure 2.3 for a visualisation of this two-factor model.

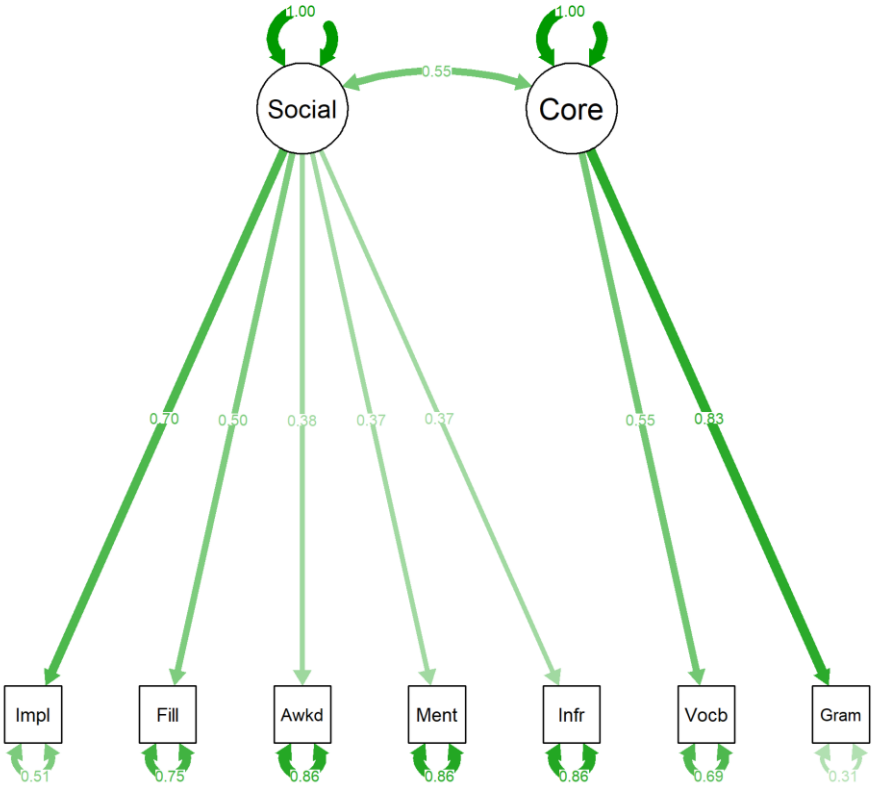


Figure 2.3. A two-factor correlated-traits model including a "social understanding" factor and a "core language" factor. The two factors are correlated moderately. Social = "social understanding" factor; Core = "core language" factor; Impl = Implicature; Fill = Fillers/Backchannels; Awkd = Awkward Dialogues; Ment = Mental state attribution on Frith-Happé Animations"; Infr = Narrative-based inferencing; Vocb = Receptive vocabulary; Gram = Receptive grammar.

Exploratory Analysis. In addition to the preregistered analysis, I ran a hierarchical multiple regression to explore predictors of implicature comprehension. In the first stage, the model was significant, $F(2, 109) = 6.41$, $p < .001$. Receptive vocabulary and receptive grammar explained 8.9% of the variance in implicature comprehension. The addition of the rest of the predictors in stage 2 significantly increased the amount of variance explained, $F(4, 105) = 5.60$, $p < .001$, and the full model remained significant, $F(6, 105) = 6.23$, $p < 0.001$. However, together the variables only explained 22.0% of the variance in implicature comprehension. See Table 2.7 for the significance of individual predictors. Note that

assumptions of multiple regression were checked: residuals were normally distributed and homoscedastic, and there were no influential observations (maximum Cook's distance = 0.07).

Table 2.7

Coefficients for multiple regression, with implicature scores as the criterion variable.

	Estimate	Standard Error	t-value	p-value
Stage 1				
Vocabulary	0.04	0.06	0.60	.553
Grammar	0.14	0.05	2.86	.005
Stage 2				
Vocabulary	0.01	0.06	0.21	.835
Grammar	0.09	0.05	1.83	.070
Inferencing	0.22	0.26	0.87	.387
Awkward Dialogues	0.04	0.02	1.64	.104
Filler/Backchannels	0.04	0.01	3.03	.003
Frith-Happé Animations	0.06	0.03	1.89	.061

Discussion

My key preregistered research question was whether the processing of implied meaning (or implicature) in conversation is distinct from core language abilities (i.e. an individual's grammar and vocabulary skills). The low correlations between the Implicature Comprehension Test (ICT) and the grammar and vocabulary tests support the hypothesis that core language skills and pragmatic processing of implicature are separable domains. As such, having well-developed core language skills does not necessarily mean that an individual will be adept at processing conversational implicature. I also set out a preregistered hypothesis that implicature comprehension and other aspects of social understanding would cluster together as a "social understanding" factor, distinct from a "core language" factor. This hypothesis was supported too, as data collected in the general adult population showed better fit to a two-factor model than a one-factor model. This means that rather than all the tests

reflecting a single general factor, there was evidence in the pattern of correlations that "core language" (grammar and vocabulary) and "social understanding" (including implicature comprehension) represented partially separable domains. While this was the case, the tests clustering under the "social understanding" factor showed only relatively low correlations with each other. Therefore, it would be most accurate to speak of these tests as only partly reflecting a general ability, with skills specific to the individual tests being most influential in determining how well people performed on them.

Although the preregistered hypotheses were supported, it should be noted that model fit was improved when one of the three tests set to load on the "core language" factor (narrative-based inferencing) was modeled as part of the other factor instead. This does raise questions about what this factor represents. The first three tests loading on this factor required individuals to interpret how conversational partners communicate implicitly, how they use fillers to negotiate conversational turns, and how they convey social discomfort/offence to each other. These tests all required sensitivity to a speakers' communicative intents in conversational contexts. In the fourth test, individuals needed to attribute "mental states" to abstract shapes interacting in short videos. We might expect these four tests to interrelate due to their shared demand on making inferences in contexts with quite explicit interpersonal interaction. However, in the best model these tests also clustered with a fifth test that involved narrative-based inferencing. In contrast to the other tests, it is less clear that narrative-based inferencing involves interpersonal interaction, as it simply requires the individual to integrate information coherently across sentences. Having said that, even in processing local-level coherence in a task such as this, we need to infer the narrator's intention to be relevant - i.e. we expect them to maintain coherence for us in an optimal way, and not, for instance, to change setting without telling us. This suggests that if the tests interrelate because they involve making inferences about interpersonal interactions, then we

should note that these interactions can be really quite implicit, such as that between a writer and an implied reader. We should be careful therefore not to identify this factor with understanding explicit interpersonal communication but rather with forming inferences more implicitly based on integrating information in context using heuristics about how people interact.

While core language skills and inferential skills are related (in the CFA, they were correlated at .55), the indication that a two rather than a one factor structure underpins performance on the test battery suggests that these two sets of skills are partially dissociable in the general population. This would be expected based on the small amount of evidence that core language predicts some of the variance in pragmatic language skills both cross-sectionally (Volden et al., 2009; Whyte & Nelson, 2015) and longitudinally (Bernard & Deleau, 2007; Greenslade et al., 2018; Hale & Tager-Flusberg, 2005; Miniscalco et al., 2014). However, these seem to be separable domains. This agrees with the longstanding clinical intuition that some individuals (such as some of those with autism or pragmatic language impairments) can have difficulties with conversational use and understanding of language in the relative absence of problems with grammar and vocabulary (Baird & Norbury, 2016). Later in this thesis, I address that issue directly, with a comparison between autistic and non-autistic adults in their performance on the same test battery. I expected that autistic adults would find the tests specifically clustered within the "social understanding" factor more difficult. Beside the theoretical interest of carrying out this analysis, it would also be useful in establishing whether the tests, especially the ICT, are useful in explaining real-life conversational difficulties. In the findings reported here, the tests generally showed low correlations, which might lead us to question whether they are really measuring general abilities that are relevant to everyday life. However, if the tests show group differences when

comparing those with and without communication challenges, then we can argue that they are sensitive to cognitive processes relevant to day-to-day communication.

It is worth dwelling for a moment on the relatively low correlations observed between the tests. Reliability was good, so low correlations between tests cannot be attributed to their being noisy. The only correlation showing a moderate effect size was between grammar and vocabulary, indicating that these tests hung together as measures of core language skills. All other tests were rather weakly correlated. This suggests that performance was influenced by a range of task-specific skills rather than dominated by domain-general abilities. It is particularly interesting that the Awkward Dialogues and Frith-Happé Animations were minimally correlated. Both tests would be assumed to tap advanced "theory of mind"/mentalising skills. However, the lack of correlation here indicates that individual differences in performance on these tests are not well accounted for by a general social cognition factor but rather by much more task-specific skills. This replicates previous findings that question the coherence of "theory of mind" as a single construct. In their review paper, Gernsbacher and Yergeau (2019) collate amounting evidence for poor convergent validity between several "theory of mind" tasks in autistic and non-autistic participants. Given the influence of "theory of mind" in accounts of cognitive differences in autism, it is important that future research explores the dimensionality of "theory of mind" in well-powered samples using factor analysis.

One objective of this study was to understand more about what the new Implicature Comprehension Test measured. Core language skills, such as vocabulary and grammar, accounted for only a small proportion of variance in scores on the test. This means that some individuals might be able to decode the basic "literal" meaning of an utterance, as encoded by the individual words and grammatical structure, while struggling to process an implied

meaning. Such individuals might be assumed to include autistic people, who have been reported as having difficulties forming inferences (Loukusa & Moilanen, 2009); diagnostic criteria often refer to problems with non-literal/implicit meanings (Baird & Norbury, 2016). The dissociation between implicature comprehension and core language skills is also in line with linguistic theories that describe implicature as context-dependent meaning that is not intrinsic to the linguistic code (see Grice's theories, Relevance Theory, etc. in Ariel, 2010). Also, I found that an individual's ability to make inferences, as measured in several of the tests, explained some variability in how effectively people process conversational implicature, even accounting for the role played by grammar and vocabulary skills. This suggests that there is some commonality between processing implicature in conversational interchanges and forming inferences in other contexts - and the contexts in the test battery were wide-ranging, including narratives, abstract cartoons and social dialogues.

However, the shared variance was relatively small, leaving a considerable proportion of the variability in scores on the Implicature Comprehension Test unexplained. What skills might explain why people varied in their scores on the test? There are no categorical answers to this, but we can bear a couple of things in mind. First off, it is likely that interpreting implied meanings is a complex process underpinned by multiple strategies; we may use formal logic when responding to test items requiring inferences to be made, but we may also be influenced in a more automatic way by what we feel other people might choose, i.e. by social norms. As such, there may be a combination of effortful processing and more intuitive "gut-based" responses. It should be noted, however, that items on this test were not correct simply by virtue of being selected by the most people. There was less consensus for some items on the ICT, and yet item-level accuracy tended to correlate well with test totals excluding that item. This suggests that there was some latent ability underpinning performance across items on the test. As such, if there are multiple strategies in processing

implicature, including formal reasoning and sensitivity to social norms, then these strategies likely combine as a unitary process. And what might this unitary process involve? I designed the test under the influence of Relevance Theory, and so an obvious answer might be sensitivity to the principle of communicative relevance (Sperber & Wilson, 1986). In the context of implicature, this is the expectation that an utterance should respond relevantly to the previous contribution in a conversation, and if it doesn't seem to, then we should be open to the possibility of the interlocutor intended us to pick up an implied meaning. It may be that some individuals are more active in seeking relevance and implied meaning, and this tendency may explain individual differences in how people detect implicature.

In summary, this chapter presents evidence that understanding language in its communicative context is not simply a matter of core language skills. In particular, I found that understanding implicated meanings in conversation is somewhat distinct from vocabulary knowledge and grammatical competence, and seems to cluster as part of a "social understanding/inferential processing" domain. This raises the question of whether individuals with social communication difficulties may have especial problems with this conversational understanding even if they perform at a typical level on tests of vocabulary and grammar. The next chapter will explore this question in relation to those on the autistic spectrum.

Chapter Three. Cognitive Differences between Autistic and Non-Autistic Adults in Processing Implicature[1]

Social communication difficulties are a diagnostic feature of autism (American Psychiatric Association, 2013). These difficulties are sometimes thought to be due to a core impairment in pragmatics (e.g. Baron-Cohen, 1988; Rapin & Dunn, 2003). Pragmatics refers to the role of context in shaping communicative meaning, such as when we interpret beyond the literal content of an utterance to understand what a person means in context (Ariel, 2010). This is, of course, especially important in the case of language that is clearly non-literal, such as metaphor, irony and jokes, which may pose difficulty for some autistic individuals, depending perhaps most importantly on their core language ability (Kalandadze et al., 2018). However, much of what is said in conversation is indirect, implied, incomplete or otherwise open to interpretation, and needs to be inferred pragmatically using features of the social-linguistic context (Sperber & Wilson, 1986). Given this aspect of communication, there are continual opportunities for conversational glitches if pragmatic processing works differently for autistic people, as suggested by qualitative observations that autistic people tend to use and understand language more literally than their non-autistic peers (Hobson, 2012).

In this chapter, I compare performance on my new Implicature Comprehension Test (ICT), described in the previous chapter. Here I test the hypothesis that autistic individuals would have lower implicature scores on the ICT, even controlling for core language ability (i.e. grammar and vocabulary skills). Following this hypothesis-testing, I present some

[1] This chapter has been published as Wilson, A. C., & Bishop, D. V. M. (2020). "Second guessing yourself all the time about what they really mean…": Cognitive differences between autistic and non-autistic adults in understanding implied meaning. *Autism Research.* https://doi.org/10.1002/aur.2345

exploratory analysis comparing two different response formats on the test to tease out the cognitive differences that might underpin performance on the test.

Methods

Ethical review for this study occurred in two stages, separately for the autistic and non-autistic participants, as the groups were recruited sequentially. The first stage of the study was granted ethical clearance in July 2018 (Refs. R57087/RE002) and the second in November 2018 (R59912/RE002) by the Medical Science Interdivisional Research Ethics Committee at Oxford University.

Participants

I recruited autistic adults through support and social groups, both organised independently and by the National Autistic Society (NAS), as well as through Autistica, a research-focused charity in the UK. Inclusion criteria for individuals giving informed consent to participate included: (1) an autism spectrum diagnosis by a clinical service, (2) native-level fluency in English, and (3) age of 18 years or over. Exclusion criteria included: (1) significant visual or hearing impairment, (2) history of neurological illness or head injury, and (3) no access to a computer with internet access and audio. Individuals were invited to participate regardless of other diagnoses, including ADHD, genetic syndromes or learning disabilities. I aimed to recruit at least 50 autistic individuals, in line with the power calculation described in the next chapter, but rather than prescribe a set sample size, I set a specific date as the stopping rule for recruitment: the study opened in 2019 and all individuals expressing interest in participating by 31/03/2019 were invited to do so.

The comparison sample of 120 non-autistic adults was recruited online via the participant platform, Prolific. They fulfilled similar inclusion criteria, apart from having no

autism diagnosis. This is the same sample reported in the previous chapter. Average age of the non-autistic participants was 30 years; 11 months (*SD* = 11 years; 3 months, min = 18 years, max = 64 years). 65 identified as women, 54 as men, and one person did not declare their gender. The majority of the sample described themselves as White (103 out of 120); four people identified as Mixed Race, four as Black, and eight as Asian. 34 people said they were currently students. Of the 86 individuals who reported not being students, highest level of education was given as high school/secondary school for 18 individuals, vocational training/college courses for 13, bachelor's degree for 53, and a higher degree for nine.

The autistic sample comprised 71 people. Of 80 individuals recruited into the study, 74 completed the battery of cognitive tasks which forms the basis of the analysis here, and of these 74 people, 71 reported a clinical diagnosis by appropriately trained professionals (clinical psychologists, psychiatrists, and specialist nurse practitioners trained in autism diagnosis) and mostly as part of multidisciplinary teams (MDTs) in National Health Service (NHS) settings. 45 individuals identified as female, 25 as male and one as non-binary. Average age was 38 years (*SD* = 14 years, min = 18 years, max = 70 years). Approximate average age at diagnosis was 31 years (*SD* = 18 years, 50 individuals were diagnosed as adults). The sample was unusual among samples in autism research for having a high proportion of female participants and individuals diagnosed in adulthood. As part of the research protocol, the majority of participants *(n* = 65) took part in an ADOS-2 assessment as noted below. 55 people scored above the ADOS-2 "autism spectrum" cut-off. Of the 10 who did not, all were female, which speaks to the notion that autistic females may mask some of their difficulties, especially in "safe" one-to-one situations such as during an ADOS-2 evaluation (Lai et al., 2017). Other neurodevelopmental diagnoses reported by participants included: dyslexia (*n* = 8), language disorder (*n* = 3), dyspraxia/developmental coordination disorder (*n* = 9), and ADHD (*n* = 7). Except for one Asian person, each autistic participant

was White. 15 people said they were currently students. Of the 56 individuals who reported not being students, highest level of education was given as at least some high school/secondary school for 8 individuals, vocational training/college courses for 16, bachelor's degree for 16, and a higher degree for 15 (one person gave no response to this question).

Procedure

For autistic participants, the study involved three sections, the first two of which took place online. In the first section, individuals were asked to complete questionnaire measures. In the second section, a set of seven cognitive tests were administered online at a time and place of the participant's choosing. This cognitive battery included the ICT. The tasks were supported by Gorilla, the online tool for behavioural experiments. In the third section, participants were either invited to the Department of Experimental Psychology, University of Oxford, or visited at home, for an in-person communication assessment using Module 4 of the diagnostic instrument, the Autism Diagnostic Observation Schedule Second Edition (ADOS-2; Lord et al., 2012), and also completed some further cognitive tasks under supervision, including the ICT—Adapted. Individuals were compensated with a £20 voucher as a thank you for their participation.

Non-autistic participants only completed the second section of the study, detailed above. They were compensated £5.

Measures

Implicature Comprehension Test (ICT). Participants watch a sequence of 55 videos, each consisting of a conversational adjacency pair, with two characters each producing a short utterance, followed by a comprehension question eliciting a response of "yes", "no" or

"don't know". For 36 items, participants need to process implicature to answer the question. In half the items "yes" is the correct answer, and in half "no" is. By "correct", I mean the answer that is chosen by the most people and/or correlates with total scores on the measure, suggesting that it taps a latent ability in inferring meaning. Full reliability analysis is presented in the previous chapter (and in the next, for the autistic group). Cronbach's alpha for implicature total score was .89 in the present sample. There are also 10 items where the answer is more explicit; these serve as positive control items. The main measured variable was sum of implicature items correctly answered (out of 36), and there was also a control variable, the sum of explicit-response items correctly answered (out of 10).

Synonyms Test. Participants are presented with a written target word on the screen, and select a synonym from a choice of four words; there are 25 trials. This is a timed task, with up to 12 s per item. There was one measured variable: the sum of items correctly answered (out of 25). Cronbach's alpha for the total score was .81 in the present sample.

Grammaticality Decision Test. Participants listen to 44 sentences and decide if they are well-formed and grammatical; there is a six second limit to listen and respond. There was one measured variable: the sum of items currently answered (out of 44). Cronbach's alpha for the total score was .94 in the present sample.

Implicature Comprehension Test—Adapted Version (ICT-Adapted). This test was structured like the ICT, with the exception that there was no "don't know" option. There are different reasons that a person may choose "don't know" when given the chance: perhaps they simply do not know and would be guessing otherwise, or perhaps they do not feel confident in their response but would tend to arrive at the same response as other people if constrained to give a response. This adapted version of the test was devised to tease out these two alternatives. There was one measured variable: the sum of implicature items correctly

answered (out of 36). This task was only completed by autistic participants, and was carried out during the third part of the research protocol with a researcher present. Cronbach's alpha for implicature total score was .85 in the present sample.

Data Analysis

Data were analysed in R (R Core Team, 2017). Data and scripts are accessible on OSF (*https://osf.io/mhfet/*).

Scores on the vocabulary and grammar tasks were processed using principal components analysis (PCA) with the prcomp() function. The first principal component was extracted as a measure of "core language ability"; this showed a correlation of .85 with the vocabulary and grammar variables. As nine participants had outlying scores on the grammar task, they were excluded at this stage so that their scores would not influence the PCA. For these participants, a predicted score for "core language ability" was assigned based on linear regression with their vocabulary score as predictor. Finally, to facilitate interpretation of results, core language scores were standardised to have a mean of 0 and SD of 1 in the sample.

In proceeding to the main analysis, participants with outlying scores on the control items of the ICT were excluded from analysis, as these individuals were unlikely to have engaged with the task very well. An outlying score was below 7 out of 10; this represented a score at least 2.2 times the interquartile range below the lower quartile (Hoaglin & Iglewicz, 1987).

I tested whether the autistic and non-autistic groups differed in their tendency to select incorrect and "don't know" compared to correct responses on the ICT, when also accounting for core language ability. I ran a multinomial mixed effects model using the MCMCglmm

package (Hadfield, 2010) with group membership (autistic or non-autistic) and core language ability as fixed effects, and participant and item as random effects. The model was fitted using Markov chain Monte Carlo (MCMC) sampling across 3500 samples. It estimated two logit equations, one comparing probability of a correct to incorrect response, and a second comparing a correct to "don't know" response, and assigned predicted log-odds to each comparison. I report odds ratios with 95% credible intervals (and approximate *p*-values calculated through the MCMC approach) for each fixed effect.

I followed up this analysis of group differences on the ICT with an analysis of differences by test condition. Autistic participants completed the ICT and the ICT-Adapted (without the "don't know" option) with a researcher. There were two hypotheses to explain why participants may choose the "don't know" option first time round. It may be due to an impairment in integrating information and making normative inferences. However, it may also be due to an aspect of cognitive style whereby a participant feels uncomfortable making inferences when they feel there is not enough information to say; however, when constrained to give an answer, they are likely to make the expected inference. In the first account, a "don't know" response first time round would translate to chance-level accuracy second time round, whereas in the second account, we might expect high accuracy. I ran a random effects logistic regression using the lme4 package (Bates et al., 2015) to differentiate between these accounts. Response first time round (correct, incorrect, "don't know") was fixed effect, and participant and item were random effects. Accuracy on a given implicature item second time round was dependent variable. I report probability of a correct response, along with 95% confidence intervals.

Finally, I assessed how well scores on the ICT discriminated between those with and without an autism spectrum diagnosis. I attempted to optimise the index extracted from the

test to separate the groups, using Area Under the Curve (AUC) in Receiver Operator Curve (ROC) analysis to select the best index. I tested total implicature score, number of "don't know" responses, and residualised versions of these indices. For residualised scores, I calculated the difference between a participant's actual score and the score expected from their performance on the control items; expected scores were computed based on linear regression in an unrelated sample of 120 non-autistic adults (Study One in Chapter Two). I also tested whether it was appropriate to exclude participants based on their score on the control items of the ICT (out of 10). After determining which index had the best AUC, I calculated the sensitivity and specificity of that index in the sample. R package ROCR (Sing et al., 2005) was used for this analysis.

Results

See Table 3.1 for descriptive statistics for each variable by group. Individuals are excluded from the dataset if they had an outlying score on the control items of the ICT: this included five autistic participants and two non-autistic. There was a group difference in implicature scores on the ICT, as hypothesised, with autistic participants scoring lower than non-autistic people, $t(114.09) = 7.83$, $p < .001$, Cohen's $d = 1.25$, but there was no difference in core language scores, $t(115.75) = 1.68$, $p = .096$. Total scores on the ICT did not differ by sex, $t(162.73) = 0.92$, $p = .360$, and nor by ADOS-2 outcome ("autism"/"autism spectrum" vs "non spectrum") among the autistic participants, $t(15.55) = 0.56$, $p = .581$.

Table 3.1

Descriptive Statistics.

	Autistic Adults						
	N	Mean	SD	Min	Max	Skew	Kurtosis
Implicature	66	22.30	6.46	10	35	0.10	-0.82
Control items for Implicature	66	9.20	0.93	7	10	-0.84	-0.41
Time-outs on Implicature	66	0.73	1.12	0	5	1.78	3.07
Core Language	66	0.21	1.06	-2.60	2.08	-0.64	0.01
	Non-autistic Adults						
	N	Mean	SD	Min	Max	Skew	Kurtosis
Implicature	118	28.92	4.49	8	36	-1.20	2.80
Control items for Implicature	118	9.64	0.64	7	10	-1.70	2.39
Time-outs on Implicature	118	0.21	0.61	0	4	3.64	15.30
Core Language	118	-0.05	0.89	-2.00	2.44	0.08	-0.38

I tested whether group membership (autistic or non-autistic) predicted response (correct, incorrect or "don't know") to any given implicature item, while also accounting for core language ability, in a multinomial mixed effects model. Autistic participants were 2.56 [1.76, 3.77] times more likely to select an incorrect rather than correct answer compared to non-autistic participants, and 6.19 [3.63, 10.39] times more likely to pick the "don't know" rather than the correct response, $p < .001$ in both cases. Lower core language ability was related to a greater likelihood of selecting incorrect responses compared to the correct answer: for each SD decrease in core language, odds of an incorrect response were 1.78 [1.47, 2.15] times higher, $p < .001$, whereas there was no effect for "don't know" responses, $p = .142$.

Clearly, autistic people had a much greater tendency to select the "don't know" response. To make sense of this, I compared how autistic participants scored on the original version of the test compared to an adapted version without a "don't know" option. 61 people completed both versions; four of these participants were outliers on the control items of the ICT (i.e. scored below 7), and therefore were excluded from the analysis. I used a mixed

effects logistic regression, with accuracy on the second version of the test as response variable and response on the original version (correct, incorrect, "don't know") as fixed effect. Probabilities of a correct response second time round were 97% [95%, 99%] when a correct response was originally given and 71% [56%, 83%] when an incorrect response was originally given. Crucially, a "don't know" response first time round was associated with 91% [85%, 95%] accuracy second time round. In this analysis, we see high accuracy for items when "don't know" was the response first time round, indicating that participants will typically make the normative inference when constrained to do so. Accuracy was substantially higher than when an incorrect response was given the first-time round, and almost as high as when a correct answer was originally given. This provides support to the second account outlined above that differences in cognitive style are likely to play a role in explaining group differences in scores on this test.

I assessed how well scores discriminated between those with and without an autism spectrum diagnosis. The AUCs of different indices were compared, including raw implicature total (AUC = .80), raw total of implicature items given a "don't know" response (AUC = .75), residualised implicature total (AUC = .80), and residualised total of implicature items given a "don't know" response (AUC = .78). I tested whether it was appropriate to drop participants who did not score well on the control items, but all thresholds (out of 10) were associated with slightly worse AUCs. Raw implicature score represented the best index. The optimal cut-off was 27 (with lower scores more common in the autistic group) which gave sensitivity of 76% and specificity of 76% for autism in the sample.

Discussion

On the Implicature Comprehension Test (ICT), autistic people were over twice as likely as non-autistic people to choose the "incorrect" inference, suggesting that they had

some difficulty with arriving at the normative interpretation of implied language. In everyday conversations, this might translate as being more likely to "get the wrong end of the stick" sometimes. This finding agrees with previous research that autistic people underperform on tests of inferencing (for a review, see Loukusa & Moilanen, 2009), and consolidates previous research since I rule out core language as an explanation for inferencing difficulties, by including core language as a covariate and controlling language demands in the task itself. Difficulties with making inferences about implied language would be predicted by leading accounts of the cognitive differences in autism. These include the "theory of mind" account (Baron-Cohen, 2000), which focuses on making inferences about other people's mental states, and the "central coherence" account (Frith, 1989), which focuses more generally on integrating information into a global whole. Both accounts predict differences in process (e.g. reduced tendency) and outcome (e.g. arrival at unusual interpretations) when making inferences about other people's implied meanings in conversation.

Response Tendency on the Implicature Comprehension Test in Autistic Adults

Perhaps, the most notable finding in the present research was the much greater tendency of autistic people to record "don't know" responses when given the opportunity. As such, in an ambiguous scenario where a person may or may not intend to communicate an indirect meaning, autistic people seem inclined to avoid making the normative inference that non-autistic people are likely to make, perhaps because "there's not enough information to say". Interestingly, when the test was given in a more constrained format that did not allow a "don't know" response, autistic participants typically processed implicature as expected. A "don't know" response first time round was associated with over 90% accuracy second time round – indeed, not substantially lower than accuracy on items that were answered correctly first-time round. This suggests that autistic participants *could* process inferences about

implied meaning, but their tendency to do so differed according to the circumstances. As all participants completed the ICT—Adapted after the ICT, we should not rule out a practice effect, not least because accuracy on items originally given an incorrect response first time round was 73%, i.e. significantly greater than chance. However, this is rather lower than 90%, which suggests that a practice effect does not fully explain the high accuracy when people originally said "don't know".

We can interpret the tendency to select "don't know" in several ways. Firstly, the more open nature of the ICT (in which "don't know" was an option) may have caused difficulty for the autistic participants. This task required the individual to decide both whether an inference was warranted and also what that inference might be. This introduces two levels of uncertainty, and makes the task more open-domain (i.e. with rules and structure needing to be imposed by the individual), which Klin et al. (2003) suggested is a major factor in activities that autistic people find more challenging. Another possibility is that it was less the nature of the task and more the nature of the information being presented that led to the distinctive response patterns of autistic individuals. In implicature items, a question was asked by one character and a response was given by a second character, and this response only seemed adequate if the listener looked for an implied meaning. As such, the frequent use of the "don't know" option may have reflected an autistic preference for more certain, explicit information.

Qualitative comments made by participants about the test would support the suggestion that they preferred certainty, as several independently expressed frustration and noted that there was just not enough information to make a judgement on the implicature items. This probably reflected an element of *cognitive preference* rather than just reduced inferencing ability, as individuals typically could select the normative answer when

constrained to, even if they may have preferred to say “don’t know”, as they did first time round. Perhaps, the drive towards certainty is based on a history of communication mishaps, with “don’t know” even representing an adaptive response when someone is aware of having experienced difficulty in the past when determining whether a person was implicating something or not. One participant offered the following take on these issues:

> “I can make a really good guess at what people mean but the anxiety surrounding all the possible meanings is so exhausting that like if they say something I’m 99% sure it means this but that 1% of but what about all the other things it could possibly mean... It’s just really, really exhausting and second guessing yourself all the time of 'was that thing the right thing?' … And people aren’t brilliant at giving feedback, so you don’t know if you’ve said the right thing … I think it’s much more the anxiety of not being sure if you’re understanding someone correctly than just outright getting it wrong … because there were so many times as a kid when I misunderstood and got it wrong and then if you get it wrong people react to you badly or they ostracise you … I think it’s an anxiety that’s built up over a lifetime of not quite getting it right enough of the time.” (This was their response to the question “What do you find harder about social situations?” during an ADOS-2 interview).

Pulling together the threads discussed here, there seemed to be cognitive differences in how autistic people responded to and made predictions based on uncertainty (see Sinha et al., 2014 for a review of the role of prediction in autism). In this respect, it is worthwhile noting that the trait “intolerance of uncertainty” has recently emerged from anxiety studies into the autism literature. Several studies indicate that autistic individuals show greater levels of this trait, compared to the general population, and that variability in the trait relates to core features of autism, including sensory sensitivities, insistence on sameness and repetitive

behaviors (e.g. Hwang et al., 2020; Vasa et al., 2018; Wigham et al., 2015). One thing to note is that "intolerance of uncertainty" relates to appraisals (i.e. negative, potentially catastrophic appraisals) and is highly linked to anxiety. However, it is possible that selection of "don't know" on the test was less a negative appraisal of uncertainty but instead reflected a lower criterion for deciding something was uncertain (i.e. the influence was more "cognitive" than emotional). Therefore, it remains to be seen whether "intolerance of uncertainty" offers a helpful explanation of the results (and if it did, we might expect a correlation with anxiety, though scores on an anxiety questionnaire given to the autistic adults did not relate to ICT scores).

Interestingly, as described in the previous chapter, there was no relationship between scores on the Autism-Spectrum Quotient (AQ; Baron-Cohen et al., 2001) and the ICT in a general population sample of adults. The AQ is taken to measure autism-related traits, including subtle social difficulties and a preference for routine and focused interests, as they vary continuously throughout the general population. Given that there was a robust difference on the test when comparing people with and without an autism diagnosis, it is striking that there was no relationship with "sub-clinical" autistic features. This suggests that the cognitive differences observed here between autistic and non-autistic people are more of a categoric difference (i.e. a different way of thinking) rather than a thinking style that varies continuously with autistic-like traits. In the general population sample reported in Study One of Chapter Two, average AQ scores were elevated above population norms and the range of scores was broad, so the lack of a relationship between autistic traits and implicature scores cannot easily be attributed to issues with the data distribution. The AQ may not tap continuous variation well perhaps, as there is some uncertainty about its appropriateness for this purpose (James et al., 2016), although there is continuous genetic influence on the measure across the general population and at the extreme (Robinson et al., 2013). Overall,

then, the existence of a characteristic response pattern on the ICT in autistic adults but no relationship between test scores and AQ scores suggests possible categoric rather than continuous differences in relation to autism.

Potential Clinical Applications and Implications

Given that there was a large effect size on the ICT, we may ask if it could be useful for clinicians in describing the cognitive profile of individuals during the diagnostic process, or indeed afterwards. In the present sample, a score of 27 on the test gave 76% sensitivity and specificity for autism. This sensitivity may be interpreted in relation to the sensitivity of the gold-standard tool for autism assessment, the ADOS-2, which was 85% in the current sample (55 of 65 adults with existing diagnoses scored above the autism spectrum cut-off); it is not possible to assess the specificity of the ADOS-2 as it was not administered to the non-autistic sample. It goes without saying that the ICT is not diagnostic, but it may complement standard clinical tools in characterising an individual's communication abilities, at least for those with a language level equivalent to or higher than a child in the middle primary years. Further work would be needed to test its validity with individuals who are less verbal and/or have severe learning disabilities, where limitations in understanding the basic vocabulary and grammar used in the test items may constrain performance.

Just as the ICT might have practical uses in assessment, the findings reported here also have implications for how clinicians might interact with autistic people. This study indicated that autistic people have some difficulty with "catching on" to implied responses in conversation. They may not take meaning as intended, and may not feel confident about whether implied meaning is present or not. This lack of confidence may be exacerbated by a history of communication difficulties, and it is important to bear in mind that difficulties could be seen even in those with high levels of vocabulary and grammar skills. Regarding

autism-friendly communication, the National Autistic Society currently offer the following advice: "Avoid using irony, sarcasm, figurative language, rhetorical questions, idioms or exaggeration. If you do use these, explain what you have said and be clear about what you really mean to say" (National Autistic Society, 2017). This study supports the advice, but also suggests that common communication strategies, such as hinting and indirect responses, may pose barriers for autistic people. Non-autistic people have a role to play in improving the communication experiences of autistic people by accommodating their preferences for explicit communication.

Chapter Four. Receptive Communication Skills in Autistic and Non-Autistic Adults: Group Comparisons and Heterogeneity

In this chapter, I extend analysis in previous chapters to compare performance between autistic and non-autistic people on the full language and communication test battery devised in this project. In Chapter Two, I reported on a factor analysis of several novel tests administered to a non-autistic sample, and found that understanding conversational implicature, a quintessentially pragmatic skill, is related to but somewhat separate from grammar and vocabulary skills. Moreover, understanding conversational implicature clustered with other tests requiring social understanding, in a two-factor model with "core language" and "social understanding" as the two factors. Finally, in Chapter Three, we saw that autistic adults scored lower than non-autistic adults on the Implicature Comprehension Test, with a large effect size, even controlling for any differences in core language ability.

Here I extend the factor analysis reported in Chapter Two to include autistic adults alongside the non-autistic sample, in an analysis preregistered on OSF; see below for details on the preregistration. I had three hypotheses:

(1) The tasks would dissociate into two factors – "social understanding" and "core language" – rather than all grouping under one general factor.

(2) Autistic adults would show differences in pragmatic understanding, reflected in lower scores on the "social understanding" factor, but no differences in "core language" ability.

(3) Variability in researcher- and self-rated communication skills among the autistic sample would relate to scores on the "social understanding" factor.

The first hypothesis involved a replication in autistic adults of the factor structure observed in non-autistic adults. In the second hypothesis, I aimed to account for possible differences in pragmatic understanding on the Implicature Comprehension Test through a domain-level difference in social understanding. The theorised "social understanding" factor might be understood as involving "theory of mind", and so this hypothesis is in line with the research literature that has extensively documented differences in the performance of autistic groups on "theory of mind" tasks compared to non-autistic people (Baron-Cohen, 2000). Under the "theory of mind" account, the social communication challenges of autistic individuals are thought to reflect differences in the processes by which we keep track of and understand other people's mental states (although see Gernsbacher & Yergeau, 2019 for a critical view of this account). Given the view of Relevance Theory that "theory of mind" is a broad domain incorporating pragmatics (Sperber & Wilson, 1986), it is sensible to hypothesise that differences in performance on specific tests with a pragmatic demand should be accounted for by overall differences at the level of the "social understanding" factor. This is supported by previous research has found a link between pragmatic comprehension of figurative language and theory of mind in autistic and non-autistic people (Happé, 1993). Finally, as structural language impairments are not a core feature of autism (American Psychiatric Association, 2013), I hypothesised that the "core language" factor would show no differences between autistic and non-autistic groups, although it should be noted that language outcomes are highly heterogeneous in autistic adults (Pickles et al., 2014). The final hypothesis focused on variability in autistic adults: I expected global measures of social communication (based on researcher-rated face-to-face interaction and self-report) to inter-correlate with scores on the "social understanding" factor, thereby indexing a single communication continuum along which people would vary.

Method

The protocol for this project can be viewed on OSF: *https://osf.io/3ecr4*. Note that an analysis of mental health was planned, and this is reported in the next chapter.

Ethical review for this study occurred in two stages, separately for the autistic and non-autistic participants, as the groups were recruited sequentially. The first stage of the study was granted ethical clearance in July 2018 (Refs. R57087/RE002) and the second in November 2018 (R59912/RE002) by the Medical Science Interdivisional Research Ethics Committee at Oxford University.

Power Calculation

The sample size rationale was based on power analysis. First, I simulated 1000 data-sets (N = 170) from a population model with a two-factor correlated traits structure. Two variables were set to load on a "core language" factor and four variables were set to load on a "social understanding" factor; factor loadings were .6 and .8 for the "core language" factor, and .7, .5, .4, and .4 for the "social understanding" one, based on the factor structure observed in the control participants. The two factors were set to correlate at .5. Autism diagnosis was included in the model as a dichotomous covariate set to correlate with the "social understanding" factor at .3. All simulated data-sets were fitted to two models, the two-factor model and a one-factor model in which all variables were set to load on a general factor. Model fit was compared using a chi-square test. A significant effect ($p < .01$) was found in all simulations (power for Hypothesis One). I then fitted each data-set to two-factor models which included diagnosis as a covariate: in one model, the "social understanding" factor was allowed to covary with diagnosis, and in the comparison model, it was not. A significant effect was found in all data-sets (power for Hypothesis Two).

Participants

I recruited autistic adults through support and social groups, both organised independently and by the National Autistic Society (NAS), as well as through Autistica, a research-focused charity in the UK. Inclusion criteria for individuals giving informed consent to participate included: (1) an autism spectrum diagnosis by a clinical service, (2) native-level fluency in English, and (3) age of 18 years or over. Exclusion criteria included: (1) significant visual or hearing impairment, (2) history of neurological illness or head injury, and (3) no access to a computer with internet access and audio. Individuals were invited to participate regardless of other diagnoses, including ADHD, genetic syndromes or learning disabilities. I aimed to recruit at least 50 autistic individuals, in line with the power calculation, but rather than prescribe a set sample size, I set a specific date as the stopping rule for recruitment: the study opened in 2019 and all individuals expressing interest in participating by 31/03/2019 were invited to do so.

The comparison sample of 120 non-autistic adults was recruited online via the participant platform, Prolific. They fulfilled similar inclusion criteria, with the exception of having no autism diagnosis. Average age of the non-autistic participants was 30 years; 11 months (*SD* = 11 years; 3 months, min = 18 years, max = 64 years). 65 identified as women, 54 as men, and one person did not declare their gender. Full details of this sample are presented in Chapter Two.

The autistic sample comprised 71 people. Of 80 individuals recruited into the study, 74 completed the battery of cognitive tasks which forms the basis of the analysis here, and of these 74 people, 71 reported a clinical diagnosis. 45 individuals identified as female, 25 as male and one as non-binary. Average age was 38 years (*SD* = 14 years, min = 18 years, max = 70 years). Approximate average age at diagnosis was 31 years (*SD* = 18 years, 50

individuals were diagnosed as adults). The majority of participants *(n* = 65) took part in an ADOS-2 assessment as noted below, and 55 people scored above the ADOS-2 "autism spectrum" cut-off. Full details of this sample are presented in Chapter Three.

Procedure

For autistic participants, the study involved three sections, the first two of which took place online. In the first section, individuals were asked to provide information about their diagnosis, and completed three self-report questionnaire measures including the Communication Checklist Self-Report (CC-SR; Bishop et al., 2009). In the second section, individuals completed a set of seven cognitive tests online at a time and place of their choosing. The tasks were supported by Gorilla, the online tool for behavioural experiments. In the third section, participants were either invited to the Department of Experimental Psychology, University of Oxford, or visited at home, for an in-person communication assessment using Module 4 of the Autism Diagnostic Observation Schedule Second Edition (ADOS-2; Lord et al., 2012), and completed further cognitive tasks under supervision. Individuals were compensated with a £20 voucher for their participation.

Non-autistic participants only completed the second section of the study, detailed above. They were compensated £5.

Measures

I briefly outline measures used in the factor analysis, which were given in the second section of the study protocol and are fully detailed in Chapter Two. Then I describe the global measures of social communication skills and further cognitive tasks used in follow-up analysis.

Tests Expected to Tap "Social Understanding".

Implicature Comprehension Test. Each item consists of a conversational adjacency pair, with two characters each producing a short utterance, followed by a comprehension question eliciting a response of 'yes', 'no' or 'don't know'. For 36 items, participants need to process implicature to answer the question; note that in the data collected from the non-autistic adults, one of these items was poorly functioning, so was excluded from the test. There are also 10 items where the answer is more explicit; these serve as positive control items. The main measured variable was sum of implicature items correctly answered (out of 36), and there was also a control variable, the sum of explicit-response items correctly answered (out of 10).

Test of Fillers and Backchannels. Participants watch videos in which a character makes a short utterance; a second character then produces a word before the audio cuts off. Participants need to select which of the two characters is likely to be speaking now; they can indicate that it's "very difficult to say" as a third option. The task includes 40 items. For half the items, the second character produces a backchannel continuer (mm-hmm, uh-huh, really, right, huh?) and for the other half, a filler claiming the floor (um, uh, yeah, oh, well). Where a backchannel continuer is used, the first character is assumed to be the person likely to speak, and the second character where a filler claiming the floor is used. There was one measured variable: the sum of items correctly answered (out of 40).

Awkward Dialogues. Participants listen to eight short dialogues with two characters each taking five turns. In five dialogues, a character says something socially awkward. In three, nothing awkward is said. Participants indicate if something awkward was said, and type a couple of sentences to explain what was awkward, why it was said and how the characters might have felt. For each of the five awkward dialogues, participants were given a score out of two by two independent markers who showed substantial agreement (Cohen's

kappa = .73), and discrepancies were resolved through discussion. After each dialogue, participants are asked a factual recall question. If the participant did not score two out of two for an awkward dialogue and did not answer the factual recall question correctly, that item was recorded as invalid, since the participant was unlikely to have sufficiently processed the basic details of the scenario, so would not be well-placed to judge it as awkward or not. If more than one item was invalid, the whole test was recorded as invalid. Raw scores on the five awkward dialogues were processed into factor scores using item response modelling as the measured variable for this test.

Frith-Happé Animations (Abell et al., 2000). In this test of mental state attribution, participants watch a sequence of animations showing two moving triangles that interact. Two item types were presented during this study: "theory of mind" (n = 5) and "goal-directed" animations (n = 4). In the "theory of mind" animations, the triangles interact as if they are trying to influence the thoughts or feelings of each other, whereas in the control "goal-directed" animations, they interact physically (e.g. they might chase each other). Participants provide a typed description of what they think happened in the animation. Each of these descriptions was scored out of three according to the original mark scheme by two independent markers who showed substantial agreement (Cohen's kappa = .70), and discrepancies were resolved through discussion. This gave nine separate scores: for the five "theory of mind" animations and the four "goal-directed" ones. The main measured variable was a factor score for the "theory of mind" animations, processed using item response modelling. I also processed total scores for the control animations for use in follow-up analysis.

Tests Expected to Tap "Core Language".

Synonyms Test (Receptive Vocabulary). This includes 25 items in which participants are presented with words on the screen, and choose which of four words is a synonym for the target word. This is a timed task (up to 12 s per item). There was one measured variable: the sum of items correctly answered (out of 25).

Grammaticality Decision Test (Receptive Grammar). Participants listen to sentences and decide if they are well-formed and grammatical; there is a six second limit to listen and respond. There are 44 items (although 50 items were included in the original version of the test, six were poorly functioning, so have been dropped). There was one measured variable: the sum of items currently answered (out of 44).

Global Communication Measures.

Communication Checklist Self-Report (CC-SR; Bishop et al., 2009). This questionnaire consists of 70 statements (50 reflecting communication difficulties and 20 reflecting communication strengths), for which participants rate on a 4-point Likert scale how frequently each applies to them. Factor analysis in the original validation sample indicated the scale could be summarised through three scores: a structural language composite, pragmatic language composite, and social engagement composite. In the analysis reported here, the pragmatic language composite was the measure of interest as it taps perceived difficulties with pragmatic/social communication.

Module 4 of the Autism Diagnostic Observation Schedule Version Two (ADOS-2; Lord et al., 2012). This involves several structured tasks, free conversation and interview-style sections, which the participant carries out with a trained researcher. Individuals are rated on a range of behaviours, and scores on an 11-item subset allow classification of individuals as ‘non spectrum’, ‘autism spectrum’ or ‘autism’. Scores on these 11 items are totaled into a communication and social interaction index; for simplicity, I refer to ADOS-2 total.

Further Cognitive Variables Used in Follow-up Analysis.

***Test of Local Textual Inference*.** This measure of narrative-based inferencing was administered during the second section of the research protocol to all participants. Participants read two brief sections of a short story and after each section they respond to ten questions with a word or short phrase. Participants need to make coherence inferences based on the short story to answer the questions. The text remains on the screen when the questions are asked. There was one measured variable: the sum of items correctly answered (out of 40).

International Cognitive Ability Resource (ICAR) Sample Test (Condon & Revelle, 2014). This test of general ability was administered in the third section of the research protocol, only to autistic participants. There are four items of four different types: matrix reasoning, verbal reasoning, three-dimensional rotation, and letter-number sequences. The ICAR sample test has been validated in a large online sample, has good internal consistency (alpha = .81) and correlated with the Shipley-2, a commercial IQ measure, at .81 when corrected for reliability and restriction of range. Data for the ICAR sample described above is freely accessible on Dataverse (http://dx.doi.org/10.7910/DVN/AD9RVY). As young college students are significantly over-represented in the data-set, population norms cannot be adequately generated from it, though we used the data as a rough comparison for the distribution of scores in the autistic adults involved in this study. There was one measured variable: the sum of items correctly answered (out of 16).

Data Analysis

Data were analysed in R (R Core Team, 2017). Data and scripts are accessible on OSF: *https://osf.io/b9t8a/*. R package SemPlot (Epskamp & Stuber, 2017) was used to make factor model diagrams and yarrr (Phillips, 2017) to make the pirate plot.

Preliminary Analysis. In Chapter Two, we saw that the tests used here were reliable in the non-autistic sample, and I assessed whether this was also the case among the autistic adults using the same approaches (see that chapter for a fuller description of methods). I report a summary of item-level statistics for the tests, as well as reliability coefficients: Cronbach's alpha (with 95% CIs), standard error of measurement, and Revelle's beta. I used item response (IRT) models to test the extent to which all items on a test tap a single latent ability (i.e. are unidimensional), and report the root mean square error of estimation (RMSEA) for each model. Using IRT modelling, I produced item characteristic curves to check item quality, and planned to exclude any items with low, flat curves reflecting chance-level accuracy across the ability spectrum. IRT models were also used to generate the measured variables for the Awkward Dialogues and the Frith-Happé Animations (as explained in Chapter Two). For this analysis, I used the R packages psych (Revelle, 2017) and mirt (Chalmers, 2012).

Outlying scores on the tests were identified using the method of Hoaglin and Iglewicz (1987), such that scores 2.2 times the interquartile range below the lower quartile were excluded from the dataset. On the Implicature Comprehension Test, I excluded any test scores where the individual was an outlier on the explicit-response items; and on the Awkward Dialogues, I excluded scores where at least one trial was invalid (see Measures for details); these exclusions were made to remove scores where the individual had not engaged well with the tasks. As the factor analyses could account for missing scores, exclusion of individual scores did not mean exclusion of that participant from the dataset, though I checked for sensitivity of the models where individuals had several scores excluded.

I inspected the data for normality and multivariate outliers using R package MVN (Korkmaz et al., 2014). Multivariate outliers were defined as individuals whose adjusted

Mahalanobis' distance was above the 97.5th percentile of the chi-distribution. Maximum likelihood estimation was used for the factor analysis. As this is based on the multivariate normal distribution, it was important to check whether the data departed from this distribution (Flora et al., 2012), although given that robust estimation was used in the factor analysis, assumptions of normality were relaxed (Finney & DiStefano, 2013). Nonetheless, in the case of multivariate outliers being present, I planned to transform variables using the Tukey ladder of power transformations using R package rcompanion (Mangiafico, 2018) to reduce skew, and then test for multivariate outliers again. Remaining outliers would be excluded. The sensitivity of the analysis to any data transformation or outlier exclusion was evaluated by comparing results based on transformed and non-transformed data with and without outliers included.

Confirmatory Factor Analysis: Hypothesis One (Identifying Factor Structure). Using confirmatory factor analysis, I aimed to replicate a two-factor correlated traits model of communication skills already established in non-autistic adults, in a larger sample including autistic adults alongside the other adults. For this model, four variables were set to load on the "social understanding" factor and two variables on the "core language" factor, as set out in Measures. In line with the pre-registered analysis described on OSF, I planned to test whether this model fitted the data better than a one-factor model with all tests loading on one general factor. However, as reported in Results, there seemed to be some mis-specification in the pre-registered models as the vocabulary measure seemed quite distinct from other variables. Therefore, a further one-factor model of "receptive communication" excluding vocabulary was also tested. All factor analyses were run using with R package lavaan (Rosseel, 2012) with full information likelihood estimation to deal with missingness and robust standard errors to allow for non-normality. Confirmatory fit indices (CFIs) and RMSEA with 90% confidence intervals are presented for all models.

Confirmatory Factor Analysis: Hypothesis Two (Testing for Group Differences). The plan pre-registered on OSF was to test for group differences on the two-factor model. However, as indicated above, there was some mis-specification in the two-factor model and a revised one-factor model of "receptive communication" accounted better for the relationships between tests. Therefore, this revised one-factor model was analysed here. I tested the hypothesis that autistic individuals had lower scores on the communication tests, by specifying a MIMIC (multiple indicators multiple causes) model, with autism diagnosis as a covariate at the factor level. I compared a model in which the regression path from the diagnosis covariate to the "receptive communication" factor was set to zero and a model in which the path was allowed to vary using a Satorra-Bentler scaled difference chi-square test. Essentially, this tested whether there was a significant difference between autistic and non-autistic people on the "receptive communication" factor. The magnitude of the standardised regression path gave an effect size for the difference. As part of this analysis, I also tested the measurement invariance of the two-factor model across autistic and non-autistic groups, i.e. whether the same factor structure was present across the groups. This involves modelling the two groups separately before testing for metric, scalar and strict measurement invariance by progressively fixing the factor loadings, indicator intercepts and residuals such that they are constrained to be the same across groups (Sass & Schmitt, 2013). As each of these more stringent models was run, a chi-square difference test was used to assess whether fit significantly deteriorated.

Confirmatory Factor Analysis: Hypothesis Three (Investigating Variability in the Autistic Group). I hypothesised that performance on the language and communication test battery would relate to self-reported and researcher-rated global communication skills in face-to-face interaction. For this analysis, I re-ran the "receptive communication" one-factor model but only used data collected from the autistic adults, with regression paths specified

between the factor and self-reported communication challenges (CC-SR pragmatic composite) and researcher-rated communication (ADOS-2 total). I report the magnitude of the standardised regression paths.

Results

The previous chapter indicated that the test battery reliably tapped underlying cognitive abilities in a general population sample of adults. The same seems to be true when used with adults who have clinical diagnoses of autism, as shown by the reliability coefficients and item-level statistics in Table 4.1. Table 4.1 also shows the RMSEA for IRT models, which indicate how well unidimensional models describe the tests. The Test of Fillers/Backchannels has a high RMSEA, suggesting that this test was multidimensional; however, this is in the context of good reliability coefficients, so it does not seem a cause for concern. I examined item characteristic curves for all items on all tests, and these showed steep curves, reflecting good item quality. As such, there was no reason to adapt the content of the tests.

Table 4.1

Reliability analysis, including Cronbach's alpha and 95% confidence intervals, standard error of measurement (SEm), Revelle's beta, and IRT RMSEA for a unidimensional model for each test. I also present item-level statistics summarised by quartiles for corrected item-total correlations (totals excluding the item) and item-level accuracy.

						Item-total correlations			Item-level Accuracy		
Test	Alpha	Alpha 95% CI	SEm	Beta	IRT rmsea	Q1	Q2	Q3	Q1	Q2	Q3
Implicature	.89	.87, .91	2.14	.83	.02	.35	.42	.48	0.62	0.75	0.85
Fillers	.85	.82, .88	3.05	.87	.11	.27	.32	.42	0.49	0.62	0.82
Grammar	.94	.93, .95	1.39	.86	.02	.40	.49	.56	0.70	0.84	0.89
Vocabulary	.81	.77, .85	2.32	.68	.03	.28	.35	.43	0.35	0.56	0.65
Inferencing	.78	.73, .82	1.94	.64	.02	.24	.35	.40	1.58	1.75	1.83

See Table 4.2 for descriptive statistics for each of the measures by group (autistic and non-autistic) and Table 4.3 for estimates of group difference (Cohen's *d*) on each test. I calculated group differences based on two different definitions of the autistic group: anyone reporting a clinical diagnosis in the first case, and only those meeting the 'autism spectrum' cut-off on the ADOS-2 in the second. The measure of general cognitive ability was only administered to the autistic adults. Mean score was 8.16 (*SD* = 3.80). This is essentially identical to the ICAR sample described above (*M* = 8.21, *SD* = 3.77), suggesting that the cognitive ability of the autistic group was similar to the general population.

Table 4.2

Descriptive Statistics.

	Autistic Adults						
	N	Mean	SD	Min	Max	Skew	Kurtosis
Implicature	66	22.30	6.46	10	35	0.10	-0.82
Control items for Implicature	66	9.20	0.93	7	10	-0.84	-0.41
Fillers/Backchannels	71	23.18	7.87	2	36	-0.46	-0.45
Awkward Dialogues	68	-0.08	0.76	-1.68	1.02	-0.18	-0.85
Frith-Happé Animations	71	-0.10	0.85	-2.90	1.53	-0.38	0.15
Control items for Animations	71	9.17	2.43	3	12	-0.99	0.17
Vocabulary	71	14.49	5.33	0	25	-0.48	-0.36
Grammar	63	34.08	5.69	19	41	-0.96	0.10
Narrative-based Inferencing	68	33.24	4.13	22	40	-0.82	0.06
	Non-autistic Adults						
	N	Mean	SD	Min	Max	Skew	Kurtosis
Implicature	118	28.92	4.49	8	36	-1.20	2.80
Control items for Implicature	118	9.64	0.64	7	10	-1.70	2.39
Fillers/Backchannels	120	26.84	5.93	7	37	-0.83	0.79
Awkward Dialogues	119	0.07	0.74	-1.87	1.02	-0.59	-0.46
Frith-Happé Animations	120	0.08	0.71	-1.86	1.53	-0.27	-0.27
Control items for Animations	120	9.69	2.16	2	12	-1.15	1.04
Vocabulary	120	12.10	4.45	3	24	0.31	-0.26
Grammar	119	35.04	4.45	19	44	-0.75	0.47
Narrative-based Inferencing	115	34.43	3.24	24	40	-0.73	0.30

The Cohen's *d* results presented in Table 4.3 show a substantial difference on the implicature items in favour of the non-autistic group, and small-medium differences on the

fillers/backchannels task (in favour of the non-autistic group) and vocabulary task (in favour of the autistic group). Other tests show smaller differences, though the trend is for higher scores in the non-autistic group.

Table 4.3

Cohen's d, showing magnitude of the difference between the autistic and non-autistic groups on each measure. I present effect sizes comparing all participants reporting a clinical diagnosis (N = 71) and the non-autistic group (N = 120) in the left-hand columns. On the right, the autistic group comprises those reporting a diagnosis who met ADOS-2 criteria for "autism spectrum" or "autism" (N = 55); the non-autistic group was the same (N= 120).

	Autistic group (diagnosis reported)			Autistic group (diagnosis AND ADOS-2 criteria met)		
	Estimate	Lower 95% CI	Higher 95% CI	Estimate	Lower 95% CI	Upper 95% CI
Implicature	-1.25	-1.58	-0.92	-1.16	-1.51	-0.80
Fillers/Backchannels	-0.54	-0.85	-0.24	-0.56	-0.89	-0.24
Awkward Dialogues	-0.20	-0.51	0.10	-0.27	-0.60	0.05
Frith-Happé Animations	-0.24	-0.54	0.05	-0.36	-0.69	-0.04
Control items for Animations	-0.23	-0.53	0.07	-0.25	-0.57	0.07
Vocabulary	0.50	0.20	0.80	0.51	0.18	0.83
Grammar	-0.20	-0.50	0.11	-0.24	-0.58	0.10
Narrative-based inferencing	-0.33	-0.63	-0.03	-0.20	-0.54	0.13

Confirmatory Factor Analysis: Hypothesis One (Identifying Factor Structure).

The one-factor model incorporating the various "social understanding" tests alongside grammar and vocabulary did not fit the data well, CFI = .75, RMSEA = .15 [.11, .20]. However, when running the two-factor model with separate "social understanding" and "core language" factors, the residual variance for the grammar variable was negative. This therefore represented a Heywood case, suggesting the model was mis-specified (Kolenikov & Bollen, 2012). On inspection of the correlation matrix (see Table 4.4), we see that grammar is inter-related with all the variables, whereas vocabulary is moderately related to grammar but little

else, suggesting the language and communication measures clustered together whereas vocabulary was rather distinct.

Table 4.4

Correlations between language tests included in the factor analysis using pairwise-complete observations. Variables have been transformed.

	Fillers/ Backchannels	Awkward Dialogues	Frith-Happé Animations	Inferencing	Grammar	Vocabulary
Implicature	.41	.30	.31	.30	.31	-.05
Fillers/ Backchannels		.21	.31	.25	.32	.12
Awkward Dialogues			.24	.29	.31	.16
Frith-Happé Animations				.25	.26	.22
Narrative-based Inferencing					.26	.16
Grammar						.46

Therefore, I specified a new one-factor model of "receptive communication" excluding vocabulary (which can be seen in Figure 4.1). This new one-factor model fitted the data excellently, CFI = 1, RMSEA = .00, 90% CI [.00, .05]. This analysis was run using variables transformed to better approximate normal distributions, as there were several cases that were multivariate outliers in the untransformed, but not the transformed, data. However, results were essentially the same when run with transformed or non-transformed data. Note that the "receptive communication" factor did not simply represent general cognitive ability; correlation between factor scores and general cognitive ability in the autistic group was .30.

Confirmatory Factor Analysis: Hypothesis Two (Testing for Group Differences). Diagnosis was added as a covariate to the one-factor "receptive communication" model. I compared models where the regression path from diagnosis to the factor was allowed to vary and where it was set to zero. The model with the free regression path fitted the data

significantly better, $\chi 2$ (1) = 12.44, $p < .001$, indicating that autistic and non-autistic groups did differ on the factor. However, the fit statistics of this model were relatively weak, CFI = .88, RMSEA = .10, 90% CI [.06, .13]. Examining the pattern of residuals for the model, it seemed that poor fit was because group differences were only specified at the factor level. As can be seen in Table 4.3 above, the pattern of differences across groups was not uniform across tests. While differences were generally subtle, there was a very considerable difference on the Implicature Comprehension Test. To better reflect this, I also added a regression path to the model between diagnosis and this test. Following the refinement, the model showed much improved fit statistics, CFI = 1, RMSEA = .00, 90% CI [.00, .05]. The magnitude of the standardised path between diagnosis and the "receptive communication" factor was small in size, β = -.26, p = .005, reflecting somewhat lower scores in the autistic group. In addition to this domain-level difference, there was a small-medium difference on the Implicature Comprehension Test, with non-autistic participants scoring higher, β = -.37, $p < .001$.

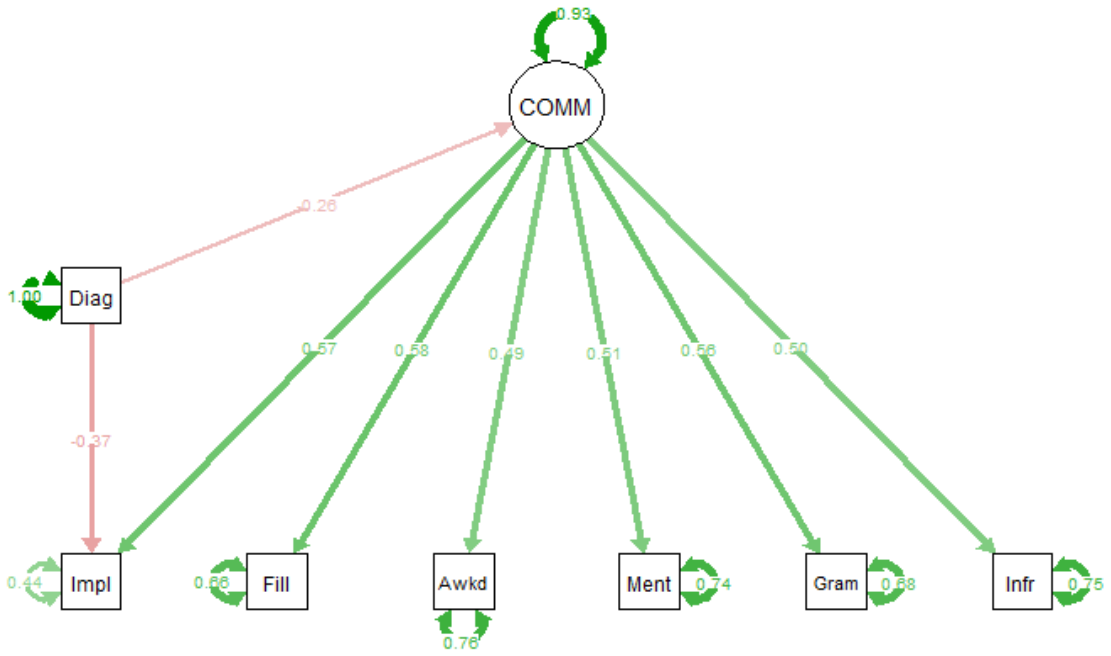


Figure 4.1. One-factor model of "receptive communication" skills across autistic and non-autistic adults, incorporating diagnosis (of autism) included as a covariate. COMM = "receptive communication" factor; Diag = Diagnosis of autism; Impl = Implicature; Fill = Fillers/Backchannels; Awkd = Awkward Dialogues; Ment = Mental state attribution on the Frith-Happé Animations; Gram = Receptive grammar; Infr = Narrative-based inferencing.

One assumption of testing for differences across a factor model is that factors are derived in the same way across groups, so I examined the extent to which this was the case across the autistic and non-autistic groups by progressively fixing loadings, intercepts and error terms of the indicators across groups. Clearly, we cannot expect the intercept for implicature to be the same, as there were differences in this test not accounted for by differences in factor scores. In the analysis for measurement invariance, I therefore allowed the intercept for this test to vary, but progressively fixed all other parameters. As can be seen in Table 4.5, fit did not significantly deteriorate as parameters were fixed, indicating that the factors were measured in a similar way across groups.

Table 4.5

Measurement invariance testing using multi-group confirmatory factor analysis across autistic and non-autistic groups.

	Model Df	Model chi-square	Chi-square difference	Df difference	P-value	Comparison	CFI	RMSEA
Configural	18	10.97					1	.00
Metric	23	12.11	1.12	5	.952	Configural	1	.00
Scalar	27	16.20	4.10	4	.393	Metric	1	.00
Strict	33	18.28	2.52	6	.866	Scalar	1	.00

To aid in interpretation of the results I considered (i) whether defining the autistic group as individuals meeting ADOS-2 criteria, and (ii) whether gender made any differences. With respect to the composition of the autistic group, I ran the same factor analysis described above but this time I only compared autistic adults meeting ADOS-2 criteria (N = 55) with the non-autistic comparison group (N = 120). Results were very similar for this model, CFI = 1, RMSEA = .00, 90% CI [.00, .08]. Regression pathways were also very similar between diagnosis and the "receptive communication" factor, β = -.30, p = .001, and implicature, β = -.27, p = .005. Moving on to gender, I tested whether autistic men and women scored differently on the "receptive communication" factor by running a two-way ANOVA on factor scores with group (autistic or non-autistic) and gender as between-subjects variables. There was no main effect of gender, $F(1, 186) = 0.88$, p = .350, but there was a marginally significant group by gender interaction, $F(1, 186) = 3.92$, p = .049. While this analysis is exploratory and the sample size was not powered for subgroup analysis by gender, inspection of Figure 4.2 would suggest a trend for autistic men to score lower on the "receptive communication" factor compared to autistic women, with no sex difference in the non-autistic group. As well as this difference on the test battery, it should also be noted that autistic men and women scored quite differently on the researcher-rated measure of global communication, the ADOS-2, $t(39.87) = 5.19$, $p < .001$, Cohen's d = 1.41, with women

showing fewer features characteristic of autism. There were, however, no sex differences in general ability, $t(44.75) = 0.18$, $p = .856$, or self-reported communication difficulties on the CC- SR pragmatic scale, $t(44.15) = 0.26$, $p = .795$. These gender-related communication trends will need to be investigated in future research.

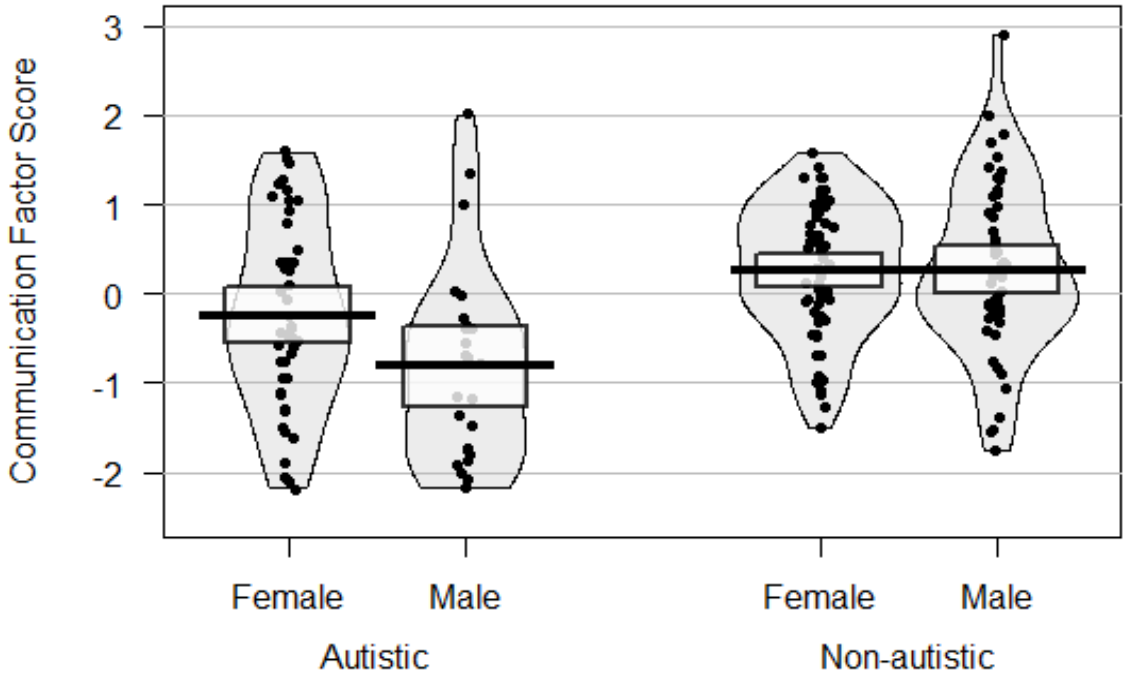


Figure 4.2. Pirate plot showing "receptive communication" factor scores by group (Autistic and Non-autistic) and gender. Central tendency is mean with 95% CI. Scores across the sample have been standardised to have a standard deviation of 1.

Confirmatory Factor Analysis: Hypothesis Three (Investigating Variability in the Autistic Group). For this final factor analysis, I moved from comparison of the autistic and non-autistic groups to assess whether scores on the test battery accounted for variability in global communication skills within the autistic adults. I included regression paths between the "receptive communication" factor and researcher-rated face-to-face communication on the ADOS-2 and self-reported communication challenges on the CC-SR pragmatic scale. Fit was excellent in this model run in the autistic group, CFI = 1, RMSEA = .00, 90% CI [.00, .08]. There were moderate-sized regression paths between the "receptive communication" factor and ADOS-2 score, $\beta = -.43$, $p < .001$, and CC-SR pragmatic score, $\beta = -.40$, $p = .001$.

This analysis suggests that the test battery did pick up difficulties related to face-to-face interaction. It is possible that relationships between these measures might have been simply due to general cognitive ability, so to rule out this possibility, I regressed ICAR Sample Test scores on the ADOS-2 and CC-SR pragmatic totals, extracted the residuals for these two variables, and re-ran the factor analysis using residualised scores (thereby controlling for general ability). Regression paths remained significant for the ADOS-2, β = -.37, p = .005, and CC-SR pragmatic scale, β = -.28, p = .047. Finally, it is interesting to note that there was no correlation between the ADOS-2 and CC-SR pragmatic scale, indicating that researcher-rated and self-reported communication skills represented entirely different constructs, p = .309. See a path diagram in Figure 4.3.

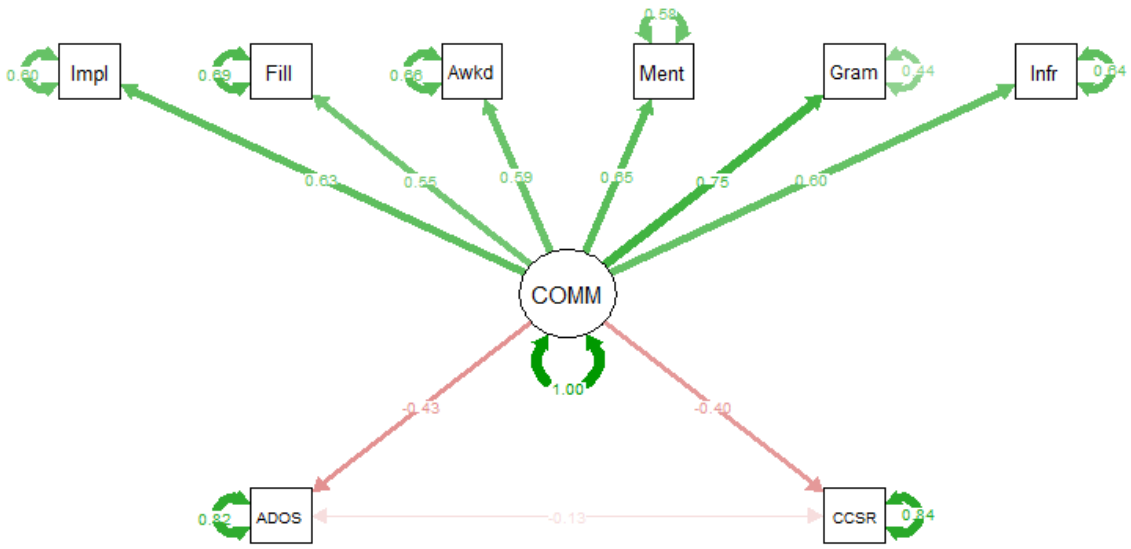


Figure 4.3. One-factor model of "receptive communication" skills in autistic adults, showing factor scores extracted from the test battery as a predictor of global communication ratings. COMM = "receptive communication" factor; Impl = Implicature; Fill = Fillers/Backchannels; Awkd = Awkward Dialogues; Ment = Mental state attribution on the Frith-Happé Animations; Gram = Receptive grammar; Infr = Narrative-based inferencing.; ADOS = ADOS-2 total; CCSR = CC-SR pragmatic composite.

Discussion

The results raise challenging questions about the nature of communication difficulties in autism. While I found that autistic people underperformed on the language and communication test battery, it was clear that the subtle group differences on the tests were *not* well accounted for by a specific "social understanding" factor relating to "theory of mind" as hypothesised. Instead, the root seemed to be much more domain-general processing differences. Also, I expected any difficulties on tests to reflect a single underlying difference between groups, but the reality was more complex. In addition to a domain-level processing difference, there seemed be further differences in cognitive style/processing present on one particular test, the Implicature Comprehension Test, as there was a much greater difference in scores between autistic and non-autistic people on this than on any other test, as shown in Figure 4.1. This supports the idea of multiple influences affecting the communication phenotype in autism. Finally, I expected that performance on the test battery would predict global communication difficulties in face-to-face scenarios, and while this was the case, there was an unexpected dissociation between observed and self-reported difficulties, and this adds further complexity to the communication phenotype in autism, as discussed below.

Comparing Autistic and Non-Autistic Adults: A Domain-Level Difference in Information-Processing

Firstly, it should be noted that the hypothesised structure of communication skills measured by the test battery was not supported. I expected the pattern of scores across the tests to cluster into two factors: "social understanding", representing more complex inferential skills required in processing social meaning and intentions, and "core language" representing formal aspects of language like grammar and vocabulary. Although vocabulary was distinct from other tests and only showed a clear relationship with grammar, all other

language and communication measures (including grammar) were inter-related. Therefore, the most appropriate factor analysis accounting for patterns across the test battery was a single-factor model representing a broad domain in "receptive communication" skills (but excluding vocabulary).[1] Autistic people underperformed on this domain.

However, what precisely did that domain represent? The test battery was devised to put special demand on our ability to make inferences about social information. This included inferences about implied meanings in conversation, social functions of conversational fillers, awkwardness in conversation, and social scenarios represented in abstract scenarios. However, these tests clustered with tests requiring the individual (a) to judge the grammaticality of isolated sentences, and (b) to process narrative meaning that carried from one sentence to the next (without the need for interpreting behavior or emotions). The common denominator in all these tests was less a requirement to make judgements about conversations and social situations, but more generally to make *normative judgements about novel communicative stimuli*. I say "communicative stimuli" rather than "language", since the Frith-Happé Animations involves nonverbal social situations. "Novel" is used as a qualifier because all the tests involved stimuli the participant had not encountered before; they could not rely on a learned response to a familiar stimulus. It should be noted that vocabulary did not cluster with these tests potentially because vocabulary skills do depend on learning and familiarity, i.e. crystalised word-level knowledge. Notably, autistic people sampled in this

[1] It should be noted that this factor model is somewhat different to the one tested in Chapter Two in non-autistic adults. The work presented in both chapters supports the view that vocabulary knowledge is a separable domain, but the present factor model suggests that grammaticality judgement is more closely linked to other tests involving linguistic/social judgements than is suggested in Chapter Two. However, we should not be concerned that the factor structure of communication skills is different in autistic and non-autistic adults, as measurement invariance testing, reported in Results above, indicated that a one-factor "receptive communication" factor model fitted data excellently in both groups.

study had a particular strength in this area. Finally, we should note that judgements were "normative", rather than right or wrong based on the application of clear-cut rules. For tests involving judgements about social situations or implied meaning, it is clear we are guided by our expectation of how other people might interpret the stimuli, but this is probably the case for all the tests. For example, in the grammar test, incorrect items were devised to sound clumsy rather than to violate explicit rules, and so consideration of the typical native speaker's reaction to items was likely a feature of the test.

As noted above, autistic people showed a reduced tendency to make normative judgements about the communicative stimuli included in the cluster of tests represented by the "receptive communication" factor, and their scores on this factor were a moderate predictor of both self-reported and researcher-rated communication difficulties. How might we understand these findings in the context of leading cognitive accounts of autism? We could look to information-integration accounts, such as the "weak central coherence" hypothesis that proposes autistic people favour local, detailed processing over global processing (Frith, 1989), and the "complex information-processing" model that proposes autistic people can struggle with assimilating information, especially when new or more subtle forms of organisation are needed (Minshew & Goldstein, 1998; Williams et al., 2015). Under these accounts, autistic people are less likely to integrate features in context, and this would predict the difficulties observed here on tests requiring inferences, e.g. about implied meaning or narrative, etc. (for a review of inferencing difficulties in autism, see Loukusa & Moilanen, 2009). These accounts are also consistent with the dissociation found in this study between word-level abilities (which were a strength for autistic people) and tasks that involved processing meaning in context at the sentence level and above (on which autistic people tended to underperform). My findings agree with studies of language comprehension influenced by "weak central coherence" theory, which have reported reduced processing of

overall coherence both "locally" at the sentence level and "globally" at the discourse level among autistic people (Jolliffe & Baron-Cohen, 1999, 2000).

No Evidence for an Isolated "Theory of Mind" Difficulty

In this study, I hypothesised that any difficulties on language tests could be attributed to "theory of mind", and on the face of it, we could link a difficulty in making normative judgements about communicative stimuli to the highly influential "theory of mind" account of autism (Baron-Cohen, 2000). However, there are a couple of reasons to be wary of this explanation. Firstly, it does not seem plausible that the "receptive communication" factor represented "theory of mind". While the Frith-Happé Animations and Awkward Dialogues could be viewed as "theory of mind" tests, it is difficult to see how tests of grammar and narrative-based inferencing placed a similar demand on "theory of mind"; all these tests showed similar loadings on the factor, so it seems unlikely it represented "theory of mind". Another argument against a "theory of mind" explanation of my results is that the account predicts autistic people would show a particular dip in performance on "theory of mind" tests compared to other measures of language/communication. This was clearly not the case. On the test that most explicitly targeted "theory of mind", the Frith-Happé Animations, autistic people performed very similarly on both the "theory of mind" and control items, and any group differences on either set of items only represented a small effect size. The upshot is that "theory of mind" does not provide a good explanation for my results, and we should look instead to domain-general information processing accounts such as those discussed above.

However, to linger a little on the Frith-Happé Animations, my results are surprising given the extremely substantial effect size reported in the first study using the task with autistic adults (Frith et al., 2002). Due to the discrepancy, I assessed how consistently other studies had found a difference between autistic and non-autistic adults on this task using

meta-analysis. See Appendix 2 for full details on the meta-analysis. Across over 400 individuals from eight studies, summary effect size for the "theory of mind" items was -0.73, 95% CI [-0.92, -0.55], and -0.50, 95% CI [-0.79, -0.22] for the control items. Two points stand out: (i) firstly there does tend to be a reliable difference between groups, though this is much less than in the first study, and (ii) there is not such a clear dissociation between performance on control and "theory of mind" items as suggested by the first study, undercutting the suggestion that group differences in performance on the Frith-Happé Animations can be explained by an isolated "theory of mind" processing difficulty. This is consistent with other studies that have failed to replicate a specific "theory of mind" difficulty among autistic people and/or have questioned the construct validity of "theory of mind" (for a review, see Gernsbacher & Yergeau, 2018). However, the meta-analysis does indicate that previous studies have found a reliable difference between autistic and non-autistic groups on the Frith-Happé Animations, so we should also wonder whether something about the present study led the two groups to perform more similarly than expected. Firstly, the methods were different: this study involved a remote online form of the Frith-Happé Animations, with participants giving written responses, whereas in previous studies the format was face-to-face and oral responses were recorded/transcribed. A study comparing children's performance on computerised and researcher-administered "theory of mind" tasks sheds an interesting light on these results (Chevallier et al., 2015). This study found that autistic and non-autistic groups showed similar scores with computerised testing, but when a researcher administered the test, the non-autistic group did much better, i.e. differences in task performance seemed to arise from the presence of an audience rather than the task itself. Difficulties experienced by autistic people are not necessarily a constant but vary with the context. In the present study, remote administration of the tests without an audience and perhaps less time constraint may have supported autistic people, leading to less pronounced differences compared to non-

autistic people. The second possible explanation is that my participant group was quite different to previous samples – in particular, overwhelmingly more female. Some previous studies using the Frith-Happé Animations only included men and all but one (Lugnegård et al., 2013) were male by a large majority. Notably, in the present sample, there was an important gender difference, with autistic women performing much better than autistic men on the "theory of mind" items, Cohen's d = 0.80, 95% CI [0.29, 1.31], with no difference on the control items. It seems, therefore, that we should not assume the social-cognitive skills assessed by this test are an area of relative weakness for all autistic people all the time. This speaks to the importance of considering the context of a task as well as gender and individual differences when accounting for the phenotype in autism.

In relation to this point about not assuming weaknesses, it is worth also drawing attention to a strength I observed in the autistic group relative to the non-autistic group: vocabulary. This was not predicted, and is perhaps a result of recruiting people to a study focused on language and communication: perhaps individuals with a special interest in words and language were motivated to volunteer, and such special interests may have been more common in the autistic group. Of course, this is merely speculation, but it is important not to erase autistic strengths to focus single-mindedly on what might be considered "weaknesses".

Multiple Influences on the Communication Phenotype in Autism

There was a spiky profile of group differences on the test battery. Autistic people scored higher on vocabulary but substantially lower than non-autistic people on the Implicature Comprehension Test in the context of relatively small differences in favour of non-autistic people on other tests. The large difference in performance on the Implicature Comprehension Test was partly explained at the factor level – autistic individuals scored lower on the "receptive communication" factor, which incorporated the implicature test.

However, when attempting to explain all group differences at the level of this factor, model fit was poor. Instead, I needed to allow for a further difference at the level of the implicature task too. What this shows is that multiple influences affected the performance of autistic people on the task. This suggestion agrees with the idea that features of autism are "fractionable" – that there are multiple distinct influences on the autism phenotype (Brunsdon & Happé, 2014). With respect to the Implicature Comprehension Test, some of the group difference on this task was "fractionated" at the general factor level, and in explaining the nature of that difference, I appeal to a domain-general information-integration account. In the previous chapter, I provided evidence that a cognitive preference for certainty and explicit communication played a role in performance on the task, so perhaps the group difference "fractionated" at the specific test-level reflected that cognitive preference. This is speculative, but suggests we ought to move away from a cognitive model that suggests there is one source of the communication phenotype in autism.

In addition to a comparison of communication in autistic and non-autistic people, I investigated variability in communication among autistic people using three methods: the test battery already discussed, face-to-face interaction with a researcher, and self-reported difficulties. I expected these measures to converge, with individuals falling along a single continuum of communication ability. This was not supported by the data; instead, communication is multifaceted, and it is clearly necessary to use multiple measures to characterise the communication profile of autistic people. Most individuals showed clinically-significant communication difficulty during one-to-one social interaction, i.e. over 80% scored 7 or above on the ADOS-2. This converts to a score below the 2nd percentile, based on general population epidemiological data reported in Brugha et al. (2011). However, the extent to which this researcher-rated difficulty co-occurred with (a) difficulties on the

language/communication test battery and (b) self-reported challenges in day-to-day communication was highly variable.

Factor scores extracted from the test battery showed moderate overlap with the self-reported and researcher-rated measures of global communication skills (with standardised coefficients of .40 and .43), indicating that it is possible to capture some of the difficulty experienced by autistic people in day-to-day conversation using computerised comprehension tests. However, there is clearly more to communication than what could be measured by these tests. That is, perhaps, unsurprising. What is more striking is that the two measures of global communication skills (ADOS-2 and CC-SR) were not correlated. This means that the extent of autistic communication behaviours observed by a researcher (or clinician) in a one-off interaction will say little about how a person experiences communication difficulty in day-to-day life. It is likely that the types of challenges that individuals are self-reporting on the CC-SR come and go; a communication mishap might happen once a day, perhaps quite unpredictably and more often in stressful situations, and this could be easily missed in an ADOS-2 assessment, which just samples a moment in time in a formal context. However, it is unlikely that the ADOS-2 simply "misses" difficulties; for one thing most individuals scored in the clinically-significant range on this measure, as they did on the CC-SR. Instead, we need to view self-reported and researcher-rated communication difficulties as different constructs. The ADOS-2 requires the administrator to make judgements about aspects of a participant's social communication that are characteristic of autism, guided by an awareness of how other researchers/clinicians would use the measure (as ADOS-2 training requires the administrator to demonstrate a high level of consensus with other users of the ADOS-2). The CC-SR, as a self-report questionnaire, measures an individual's perceptions of communication challenges they may have. While the CC-SR attempts to quantify such challenges objectively by asking participants how frequently a certain aspect of

communication poses a problem for them, the questionnaire nonetheless relies on an individual noticing a problem, conceptualising it, and reporting it as such, and an individual's perceptions of their communication experiences might or might not square with other people's observations. Beside frequency of actual communication challenges, scores on the CC-SR are going to be affected by an individual's insight, cognitive biases (such as overly negative appraisals), and reporting biases (such as social desirability, or tendency to report problems and seek help), etc. The relationship between observed difficulties and perceptions of difficulty is likely to be complex. This may explain some of the discrepancies between scores on the CC-SR and ADOS-2. For instance, an individual who is aware that they have difficulties with day-to-day communication, perhaps especially in group situations, may try to compensate for or "mask" those difficulties, and this may be especially effective in relatively structured one-to-one situations, as is the case with the ADOS-2 assessment. In this scenario, CC-SR scores would indicate impairment that would not show up on the ADOS-2. The opposite would occur if individuals had an elevated ADOS-2 score, but perhaps did not have insight into their difficulties. Similarly, an individual may have insight, but perhaps they tend to avoid social situations, live in circumstances that limit social contact, or only engage in a small range of scripted interactions. Such an individual may show communication features characteristic of autism when participating in an ADOS-2 assessment, but their frequency of communication difficulties, as measured by the CC-SR, would be low. Finally, it might be the case that an individual shows behaviours such as reduced eye contact and facial expression that are recorded on the ADOS-2, but perhaps the person doesn't perceive these as "difficulties", and so they go unreported on the CC-SR. As such, we needn't see the lack of correlation between ADOS-2 and CC-SR scores as reflecting issues with the validity of the measures, but instead showing the complexity of communication experiences for autistic individuals. The very fact that the communication measures did not indicate a single

continuum of communication skills suggests that difficulties do not occur in all contexts, which is important to bear in mind when making autism-positive environments.

Gender and Communication Skills in Autistic Adults

While I did not explicitly set out to test for differences between autistic men and women in communication skills, exploratory analysis suggested that gender might play an important role. Autistic women showed less communication difficulty, both on the cognitive test battery and during one-off interaction, but reported similar day-to-day communication challenges as men and were of comparable general cognitive ability. In the past, greater levels of impairment and lower cognitive ability tended to be reported for females on the spectrum (Lai et al., 2015), although this is likely because of diagnostic biases, with clinicians mainly identifying autistic presentations in males and only diagnosing autism in females with the most obvious difficulties. There is an increasing appreciation that females may be particularly likely to mask some of their difficulties using learned social strategies (e.g. Bargiela et al., 2016; Lai et al., 2017). As such, girls may show more reciprocal conversation during diagnostic assessments (Hiller et al., 2014) and women have been observed as showing more "normative" social behaviours on the ADOS-2 (Lai et al., 2011), which was replicated in the present study. There has been little investigation of gender differences in language abilities of autistic individuals. One small study (N = 52, with equal numbers of autistic and non-autistic boys and girls) found that autistic girls tended to have a subtly different language profile to boys (Sturrock et al., 2019). On tests of figurative language, inferencing and word associations, autistic children tended to underperform, with autistic girls scoring higher than boys, though this was in the context of a gender effect in the non-autistic children too. In the present project, autistic women tended to perform better on the "receptive communication" factor than autistic men, and this seemed to be specific to

autistic people rather than reflecting a gender difference also present in the general population. For autistic women, language skills represent an area of relative strength that might, for instance, be useful in compensating for social difficulties. However, the flipside is that clinicians may be expecting to see communication difficulties, and where these don't manifest during testing, there may be a tendency to underestimate the challenges that some autistic women experience in their day-to-day life. Having noted these average trends, it is also important to underscore the substantial variability found in autistic people of all genders.

Conclusion

In confirmatory factor analysis, I found that individuals with an autism spectrum diagnosis scored lower on a "receptive communication" factor comprising of six tests requiring the individual to make normative judgements about novel communicative stimuli. While a domain-level difference between autistic and non-autistic people was expected, this result was contrary to hypothesis as I expected these group differences would represent "social understanding" or "theory of mind", but it seemed this factor was rather more domain-general and involved making inferences and judgements about a full range of uses of language, including those that were not explicitly embedded in a live social context. Skills in this area seemed to have a real-world implication for autistic adults, as scores on the "receptive communication" factor predicted difficulties in face-to-face interaction. A final point to underscore is variability: the group difference between autistic and non-autistic adults was subtle, and there was a full range of variability in both groups. Adding to this picture of complexity was a lack of a relationship between self-reported and observed difficulties in social interaction among autistic people. This underscores the importance of any communication support for autistic people being tailored to their specific strengths and difficulties.

Chapter Five. Communication and Mental Health Challenges in Adults on the Autism Spectrum

In this chapter, I present some preliminary investigation of relationships between communication challenges and mental health in adults on the autism spectrum. Studies have consistently shown increased risk for anxiety disorders and depression in autistic people compared to the general population (Hollocks et al., 2019). Relatively little is known about factors associated with these mental health difficulties, although it is plausible that communication challenges may play a role in psychiatric presentations of verbally fluent autistic adults. There is some evidence that diagnostic features of autism, which include communication difficulties, are related to symptoms of anxiety and depression in autistic children (Mayes et al., 2011) and adults (Lever & Geurts, 2016). Likewise, a study of two nationally-representative twin cohorts totaling almost 30,000 children and adults found a monotonic increase in risk for depression and anxiety symptoms with increasing autistic traits, with a stronger effect for anxiety (Lundstrom et al., 2011). Turning to language ability, longitudinal studies consistently pinpoint language ability as strongly predictive of adult outcomes in children diagnosed with autism (Magiati et al., 2013). However, studies have generally not focused on mental health, and have included a significant proportion of minimally verbal individuals, so it is difficult to evaluate the importance of more fine-grained variability in language among the verbally fluent. For an insight into possible relationships, we need to rely on studies among the general population and individuals with language impairments. In large general population cohorts, weaker receptive language as a child predicts poorer mental health as an adult (Law et al., 2009; Schoon et al., 2010), although these studies relied on crude measures of language. Prospective studies have found poorer emotional and behavioural outcomes during later childhood and adolescence among those with pre-identified language impairments (Yew & Kearney, 2013). Studies including adult

follow-up are rare and the two largest offer conflicting results. One reported slightly greater emotional difficulties in adults diagnosed with language impairments as children (Botting et al., 2016), whereas the other found no increased risk when assessing mental health using psychiatric interview (Beitchman et al., 2014).

Based on these studies, we might tentatively suggest that language and communication difficulties may relate to poorer mental health in autistic people. As noted in the previous chapter, I used three diverse methods of measuring communication skills among autistic people: cognitive test battery, interview-based assessment and self-report checklist. Here I present an analysis of relationships between these measures and mental health.

Methods

The protocol for this study can be found on the OSF, titled "Identifying pragmatic difficulties in autistic adults" (*https://osf.io/3ecr4*). Ethical approval was granted in November 2018 by the Medical Science Interdivisional Research Ethics Committee at Oxford University (Ref: R59912/RE002).

Participants

Autistic adults were recruited through support and social groups, and Autistica, a research-focused charity in the UK. Individuals were included in the analysis if (i) they had native-level fluency in English, (ii) were aged at least 18, (iii) joined the project before 31/03/2019, (iv) completed at least the first two sections of the research protocol, and (v) had an autism spectrum diagnosis by a clinical service. 74 people completed the first two sections, of whom 71 reported a clinical diagnosis by appropriately trained professionals (clinical psychologists, psychiatrists, and specialist nurse practitioners trained in autism diagnosis) and mostly in National Health Service (NHS) settings. Exclusion criteria included:

(1) significant visual or hearing impairment, (2) history of neurological illness or head injury, and (3) no access to a computer with internet access and audio.

45 people identified as female, 25 as male and one as non-binary. Average age was 38 years (*SD* = 14 years, min = 18 years, max = 70 years). General cognitive ability was in line with population norms as measured by the ICAR Sample Test. The majority of participants took part in a standardised communication assessment using the ADOS-2 as part of the research protocol. 55 of 65 people scored above threshold on the ADOS-2 for autism spectrum. Other neurodevelopmental diagnoses reported by participants included: dyslexia (*n* = 8), language disorder (*n* = 3), dyspraxia/developmental coordination disorder (*n* = 9), and ADHD (*n* = 7). See full details about the sample in Chapter Three.

Procedure

The study involved three sections. The first two took place online, supported by Gorilla, the cloud-based tool for surveys and behavioural experiments. The third section involved a face-to-face session either in the Department of Experimental Psychology, University of Oxford, or at home. Full details can be found in previous chapters. Participants gave informed consent at each stage of the research protocol. Individuals were compensated with a £20 voucher as a thank you for their participation.

Measures

Measures are grouped below based on the section of the research protocol during which they were administered.

Section 1. Mental health symptoms were measured using the Generalized Anxiety Disorder 7-item (GAD-7) scale (Spitzer et al., 2006) and the Patient Health Questionnaire 9-item (PHQ-9) scale for depression (Spitzer et al., 1999). Participants are presented with

statements reflecting DSM criteria for anxiety and depression and rate on a 4-point Likert scale the frequency of symptoms over the past two weeks. On the PHQ-9, individuals need to endorse five or more items (including at least either the low mood or anhedonia items) as applying at least “more than half the days”, with the suicidal ideation/self-harm item counting if endorsed at all. This algorithm had 73% sensitivity and 96% specificity for major depression in the original validation study. On the GAD-7, a score of at least 10 is indicative of generalised anxiety disorder (89% sensitivity and 82% specificity in the original study). There were two continuous variables from these questionnaires: GAD-7 anxiety total (highest possible: 21) and PHQ-9 depression total (highest possible: 27). There also were two dichotomous variables: positive screen for major depression and positive screen for anxiety.

In addition, individuals were asked whether they had ever received a psychiatric diagnosis, including depression and/or anxiety.

Self-reported difficulties with communication were measured using the Communication Checklist Self-Report (CC-SR; Bishop et al., 2009). This questionnaire consists of 70 statements (50 reflecting communication difficulties and 20 reflecting communication strengths), and participants rate on a 4-point Likert scale how frequently each applies to them; the scale is from “less than once a week (or never)” to “several times a day (or all the time)”. Two composite scores were extracted: “pragmatic language”, which focuses on social use of language, and “structural language”, which focuses on fundamental language skills, like articulation and grammar. Example items include: “People tell me that I don’t use proper sentences” (structural language) and “People tell me that I ask the same question over and over” (pragmatic language). Where items were unanswered or unrecorded, scores were prorated for each composite for up to two items, as suggested by the manual (also, note that two structural language items were not recorded due to a technical error).

Where responses were unavailable for more than two items for either composite, that score was recorded as missing. There were two measured variables: pragmatic language composite and structural language composite.

Section 2. Participants completed a cognitive battery of language and communication tasks. Results were processed using factor analysis and factor scores were extracted for "receptive communication" from six tests. See Chapter Four for full details of the measures and data processing.

Section 3. Participants completed a communication assessment using Module 4 of the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012). This involves structured communication tasks, free conversation and interview-style questions during interaction with a trained researcher. Participants' communication and social interaction are rated using a coding scheme, and a total score is derived from an 11-item subset. Ratings on these items allow classification of individuals as "non spectrum", "autism spectrum" or "autism". This assessment was used to establish whether participants currently met autism criteria as evaluated using this test, and to generate a continuous variable indexing researcher-rated global communication skills.

Data Analysis

Data were analysed in R (R Core Team, 2017). Data and the analysis script are accessible on OSF: *https://osf.io/7n85h/*.

Descriptive statistics are shown for all measured variables, including z-scores on the CC-SR and ADOS-2. Whereas all analysis reported below used raw totals, z-scores are presented to characterise the extent of communication difficulties relative to the general population. CC-SR z-scores have been computed based on norms in the test manual, and

ADOS-2 Module 4 z-scores are based on epidemiological data collected in the Adult Psychiatric Morbidity Survey (National Centre for Social Research, 2017; see OSF for information on calculation of these norms).

In analysis preregistered on OSF, it was hypothesised that performance on the language and communication test battery would predict mental health among autistic adults. The factor structure of this test battery was expected to be two-factor, although analysis better supported a one-factor structure as reported in the previous chapter. Therefore, the analysis reported below was carried out using scores from this one-factor model. This represented a departure from pre-registered plans, but otherwise analysis proceeded according to the protocol on OSF. Factor scores were extracted from a confirmatory factor analysis of six cognitive tasks described in Chapter Four for a factor labelled "receptive communication". This factor analysis showed very good fit when fitted separately in autistic (N = 71) and non-autistic adults (N = 120), CFI = 1, RMSEA = .00, and factor scores showed a moderate relationship with self-reported and researcher-rated communication difficulties in face-to-face interaction. Factor scores extracted from this model were z-transformed based on the distribution of scores in the non-autistic adults, and assessed as a predictor of mental health (PHQ-9 total and GAD-7 total) using multivariate multiple regression. Pillai's trace and the associated p-value are reported as a test of significance. R package car (Fox & Weisberg, 2019) was used for this analysis.

Correlations between the full range of communication and mental health measures are presented, and follow-up exploratory analysis of relationships between these variables was carried out. The rationale and details of these analyses are reported below.

Results

Table 5.1 presents descriptive statistics for all measures of language/communication and mental health. Table 5.2 presents correlation coefficients. There were no gender differences in total anxiety symptoms on the GAD-7, t (41.75) = 0.98, p = .330, or total depressive symptoms on the PHQ-9, t (43.64) = .28, p = .783. Sample sizes for the CC-SR scales are smaller than for other measures because (i) a technical error meant that two structural language items were not presented, and (ii) some participants were reluctant to respond to some questions. Composite scores could not be calculated where more than two items were missing on the scale.

Table 5.1

Descriptive Statistics.

	N	Mean	SD	Min	Max	Skew	Kurtosis
"Receptive communication" z-score	71	-0.84	1.32	-2.94	2.11	0.28	-0.87
ADOS-2 raw score	65	9.65	3.91	1	21	0.34	0.28
ADOS-2 z-score	65	-2.30	0.52	-3.43	-0.45	0.77	2.61
CC-SR pragmatic language composite	61	23.76	14.60	1	58	0.47	-0.83
CC-SR pragmatic n unanswered items	71	1.30	3.20	0	20	3.97	17.61
CC-SR pragmatic z-score	61	-2.27	1.63	-5.60	1.70	0.19	-0.58
CC-SR structural language composite	49	14.10	10.02	0	47.78	0.88	0.81
CC-SR structural n unanswered items	71	1.07	2.48	0	14	3.18	11.11
CC-SR structural z-score	49	-1.30	1.39	-3.50	2.60	0.50	-0.20
PHQ depression total	71	12.10	7.75	0	25	0.25	-1.34
GAD anxiety total	71	11.46	6.38	0	21	0.00	-1.33

Table 5.2

Correlations between mental health and communication measures with 95% CIs.

	"Receptive communication" factor	ADOS-2 raw score	CC-SR pragmatic composite	CC-SR structural composite
PHQ-9 depression	-.16 [-.38, .08]	.07 [-.17, .31]	.47 [.24, .64]	.30 [.02, .54]
GAD-7 anxiety	-.12 [-.34, .12]	.05 [-.19, .29]	.56 [.36, .71]	.39 [.13, .61]

The “receptive communication” factor extracted from the test battery was evaluated as a predictor of total depressive and anxiety symptoms in a multivariate multiple regression. There was no effect, Pillai’s trace = 0.02, $p = .427$.

Exploratory Analysis

Relationships between the global communication measures (ADOS-2 and CC-SR) and mental health were investigated.

In a multivariate regression with PHQ-9 and GAD-7 totals as response variables, there was a significant effect of CC-SR pragmatic language composite, Pillai’s trace = 0.18, $p = .014$, but not CC-SR structural language composite, Pillai’s trace = 0.01, $p = .739$. As Table 5.2 shows, there were significant correlations between the mental health and CC-SR measures, but this analysis indicates that there was a stronger, more specific relationship between mental health and self-reported pragmatic difficulties compared to structural language difficulties. As the PHQ-9 and GAD-7 totals were highly correlated ($r = .82$), these were collapsed into a single “mental health symptoms” variable through principal components analysis using R function prcomp(). Mental health symptoms showed a large correlation with CC-SR pragmatic composite, $r = .54$, 95% CI [.33, .69], and a moderate correlation with CC-SR structural composite, $r = .36$, 95% CI [.09, .58].

Next, the relationship with self-reported communication difficulties was assessed when mental health was defined as a dichotomous variable. A significant number of people scored above clinical thresholds on the mental health questionnaires: 28 of 71 people (39%) were above-threshold for major depression on the PHQ-9 and 41 (58%) for generalized anxiety on the GAD-7. While these figures suggest considerable mental health challenges in the sample, we cannot be certain about clinical significance because full interview-based assessments of mental health were not undertaken. Therefore, in defining “clinically

significant mental health status" in the following analysis, I required individuals also to have an existing mental health diagnosis as a means of validating the questionnaire responses. As such, individuals were identified as having a clinically-significant mental health status if they (1) met GAD-7 and/or PHQ-9 cut-offs for anxiety and/or depression and (2) reported a diagnosis of anxiety and/or depression. 37 people (52%) met these criteria for clinically-significant mental health status. There was a strong relationship between CC-SR pragmatic composite and clinically-significant mental health status, t (47.17) = 3.50, p = .001, with individuals with greater mental health needs reporting more pragmatic difficulties, Cohen's d = 0.93, 95% CI [0.38, 1.47]. The relationship with CC-SR structural composite was non-significant, t (42.29) = 1.73, p = .091, Cohen's d = 0.50, 95% CI [-0.09, 1.09]. As noted above, there were some missing data for the CC-SR due partly to a technical error and also reluctance of some participants to endorse a response on some items. To check whether this influenced the results, the analysis was re-run using raw non-prorated totals on the CC-SR scales, firstly including all participants, then excluding those with one missing item, then two missing items, etc., and scores were compared between those with and without a clinically-significant mental health status. Results remained very similar.

Finally, turning to the relationship with researcher-rated communication skills, there was no relationship between ADOS-2 score and the "mental health symptoms" variable, r = .06, p = .594. As the ADOS-2 score is a global measure that may miss intricacies of communication with a meaningful link to mental health, specific items that might be implicated in mental health were extracted. These were five items previously linked to *alexithymia*, i.e. an individual's reduced awareness and ability to communicate about their emotions (Hobson et al., 2018). Given reasonable internal consistency (Cronbach's alpha = .73), these items were combined as a scale, and the correlation with the "mental health symptoms" variable was computed. The result was null.

Discussion

Autistic adults recruited into this project reported considerable mental health challenges. 39% of individuals screened positively for major depression on the PHQ-9 and 66% for generalised anxiety on the GAD-7. These values are about tenfold higher than prevalence rates based on general population screening using the PHQ-9 (3.8% in Martin et al., 2006; 3.3% in Patten & Schopflocher, 2009) and GAD-7 (6% in Löwe et al., 2008). This increased risk for anxiety and depression in autistic adults has been previously reported in multiple studies; for example, a meta-analysis pooling data on over 26,000 autistic adults showed high prevalence of current depression or anxiety at 23% and 27% respectively (Hollocks et al., 2019). Of course, these rates are rather lower than those observed in the present sample. It should be noted that this sample was not recruited through clinical services; there was no mention of mental health in materials promoting the project (as it primarily focused on communication); and participants were asked to travel and meet with a researcher (which may have been less motivating for individuals with increased mental health needs). These factors speak against the idea that there was some systematic sampling bias favouring those with low mood or anxiety. The extent of mental health needs found in the current study is therefore likely to be representative for autistic adults in the community.

The aim of this study was to assess whether language and communication difficulties related to mental health challenges. The preregistered hypothesis was that scores on the battery of language and social understanding tests would relate to anxiety and depression symptoms. Scores did not. Mental health also did not relate to researcher-observed communication difficulties during social interaction on the ADOS-2. On the other hand, self-reported communication challenges on the CC-SR *did* show a strong relationship with mental health. Therefore, mental health *is* related to communication difficulties, but only insofar as

these difficulties are *perceived as such* by participants. In addition, the CC-SR items endorsed by autistic people with low mood and/or anxiety were especially likely to relate to pragmatic language problems (i.e. for social use of language in conversation) compared to structural language problems (e.g. with grammar or speech), with the former perhaps more likely to impact on social relationships. In addition, we should bear in mind that CC-SR items are often framed as perceived criticism, e.g. "I am told that I keep talking about things that others are not interested in". Pulling these features together, it seems that autistic adults with co-occurring psychiatric problems are likely to perceive day-to-day challenges with social aspects of language use and to feel others are tuned into and critical of these difficulties/differences in communication style. Crucially, autistic individuals with mental health difficulties were *not* more likely to show language impairments or to show greater social communication difficulties in "objective" assessments compared to other autistic adults. It is possible that such assessments (such as the ADOS-2 as used here, and widely used in clinical settings as a diagnostic tool) may only detect a subset of difficulties, perhaps a subset not relevant to mental health. If the types of communication difficulties influencing mental health are unpredictable and sporadic, then observation in a standardised, safe situation might miss them. However, it is also possible that observable difficulties are less important to mental health than perceptions of difficulty. Perhaps the most important thing is a person's tendency to notice themselves having communication problems, to conceptualise them as such, and to report them. If we rely on the types of difficulties that we can observe, we may miss the language and communication challenges that are associated with mental health. Instead, we must listen to the self-reported difficulties of autistic individuals to understand when communication barriers may pose problems for mental health. This is likely to be important in therapy as well as psychological formulation, given that communication is

central to cognitive and behavourial interventions. This supports the view that adjustments may be helpful in CBT for autistic people (Spain & Happé, 2019).

There are several issues to bear in mind regarding the results reported here. Firstly, it could be argued that the relationship observed between the CC-SR and mental health questionnaires is due to shared methods inflating the correlation. As these are self-report measures, individuals may have developed a reporting tendency that applied on all measures, i.e. they might have indiscriminately reported extreme difficulties, chosen not to disclose any problems, or developed a response set such as selecting a middle of the road response on all items. However, this does not offer a satisfactory explanation of the results, as they remained unchanged when comparisons were made between individuals with and without existing diagnoses of depression or anxiety. Therefore, we should not write off the relationship between mental health and self-reported communication difficulties as simply due to a reporting bias.

Secondly, mental health was measured by screening questionnaires rather than full psychiatric interview, and also by questionnaires that have not been devised specifically for autistic people. It should be noted that the psychometric properties of these measures are sound, as described in Methods, though it is plausible that autism-specific mental health symptoms could be missed or symptoms common in non-autistic people might not be linked to distress in individuals on the spectrum. However, given that participants' existing psychiatric diagnoses corroborated their scores on the mental health questionnaires, we can be confident that scores validly reflected mental health. We should also bear in mind that some participants found the CC-SR somewhat difficult to complete, as indicated by the missing responses on this measure. This could be due to difficulties reporting on one's own experiences, or issues with the measure. Difficulties with self-report would occur if autistic

people sometimes experience less self-insight (Frith & Happé, 1999), although this assumption has been contested (Gernsbacher et al., 2017; Schriber et al., 2014). On the other hand, participants may have missed items because they thought the measure was not always appropriate to their experiences. For instance, more than one participant felt the CC-SR overestimated their social abilities due to its focus on frequency of communication difficulties (which might be low due to limited social contact) rather than extent of difficulties whenever they do occur (which may be severe). The impact of having missing data on the CC-SR was assessed by sensitivity analysis, which indicated that results did not depend on how missing data were handled. Therefore, it is unlikely that limitations with the CC-SR are a serious concern here, but we should aim to optimise questionnaires in collaboration with autistic people in the future so that measures are as relevant to the experiences of autistic people as possible.

Next, we should bear in mind that the results only apply to verbally fluent autistic adults. The link between mental health and self-reported communication, but not observed communication skills, might differ in those who are minimally verbal. As having fluent language is a strong predictor of a range of adult outcomes (Magiati et al., 2013), this might also apply to aspects of mental health, although the situation is complex. For one thing, those who are minimally verbal may be less likely to meet criteria for a mental health diagnosis, which is the case for children on the autism spectrum (e.g. Kerns et al., 2015; Lerner et al., 2018; Sukhodolsky et al., 2008). Individuals who are minimally verbal might experience distress but be less well-placed to report on their experiences, and therefore be missed by mental health services. On the other hand, it is possible that autistic individuals with well-developed language skills may experience quite specific mental health challenges mediated by social factors. For instance, they may be more likely to live independently, but the barriers to employment and the social disadvantage that they face (e.g. Howlin, 2013; NAS, 2016)

may bring significant distress. We should therefore be cautious about generalising the results reported here to individuals without fluent language.

Finally, the current research was limited by its cross-sectional and exploratory nature. Longitudinal data are required to understand the direction of influence between communication and mental health. The relationship could run in either direction, with communication mishaps affecting mental health, or vice versa. As communication differences are a core aspect of autism, we might expect communication to have a causal role, but the opposite is also plausible: mental health difficulties might lead to increased communication difficulties, perhaps affecting how individuals interpret communication challenges when they do occur. In addition to exploring causality, future research might also investigate possible cognitive and social mediators of these relationships. Cognitive mediators may include negative mental imagery or negative interpretation biases that develop through a history of communication and mental health difficulties. Socially, there may be an important role played by negative social experiences, such as bullying and ostracism, in linking communication and mental health.

Self-reported communication difficulties are closely linked to increased mental health needs in autistic individuals. While measures of *observed* language and social communication skills showed no relationship with mental health, *perceptions* of communication difficulty were influential in explaining the highly elevated mental health needs in autistic adults.

Chapter Six. An Attempt to Tease Apart Core Language and Pragmatics in Children Using a Novel Online Assessment

Previous chapters have examined pragmatic processing in adults. In this and the next chapter, we turn our focus to pragmatics in children. In Chapter Two, I presented evidence that pragmatic processing of implied meaning in conversation is somewhat distinct from aspects of core language such as vocabulary and grammar skills. The purpose behind the work reported in this chapter was to test whether there is a dissociation between pragmatic processing and core language skills in children.

Implied Meaning and Core Language in Children

As with the adult sample, I expected to find some relationship between implicature comprehension and core language in children. More advanced vocabulary and grammatical skills are likely to mean a child will more easily process the "explicit" content of an utterance, which is therefore more fully available for "reading between the lines" to grasp what the speaker implicitly means in context. Indeed, implicature comprehension and core language may be more interrelated in children than in adults. In adults, the linguistic code may be stored in a more self-encapsulated way, as it represents "crystallised" knowledge". However, given that we enter the linguistic environment without a lexicon, children may depend on more of a continual interaction between pragmatic and core language skills, meaning that there may be few situations in which they rely on one domain in relative exclusion of the other. It is likely that this interaction runs in both directions: from pragmatics to core language, and vice versa. For instance, a child with well-developed core language may understand more of what they hear, giving more opportunity for learning to infer from context. In a reciprocal way, a child who is more sensitive to the importance of context in communication may be more resourceful in interpreting unfamiliar language, thereby

accelerating their acquisition of core language skills. On the other hand, it is possible that pragmatics and core language show a similar level of dissociation in children as adults, if these different processes are "modular"; certainly, pragmatic processing of communicative stimuli is seen as modular by Relevance Theory (Carston, 2002).

However, existing studies would tend to support the view that there is interaction between domains. In approximately 95 typically developing children aged 4 to 12 years, a composite score for sentence comprehension and vocabulary skills was moderately correlated (r = .46) with implicature comprehension (including manner, scalar and metaphor implicatures) (Antoniou & Katsos, 2017). Similarly, research indicates that while children with communication difficulties perform lower on tests of figurative/nonliteral/ambiguous language, this is typically at a level that directly relates to their core language skills rather than reflecting a specific pragmatic problem (for reviews, see Andrés-Roqueta & Katsos, 2017; Kalandadze et al., 2018). For an alternative view, see the review by Loukusa and Moilanen (2009) which emphasises the importance of other aspects of cognition, including theory of mind and central coherence, in making inferences in a range of language tasks, although it is arguable whether enough attention is given in the review to the role played by core language abilities. It is worth noting that core language demand is not usually controlled in language tasks involving inferencing and implicature, making it difficult to assess the extent to which processing "explicit" and more implicit language relies on different skills. It is also rare to find studies that present language in conversational contexts – i.e. the kind of contexts where fast-paced, context-dependent pragmatic processing might be most important. Even with these provisos, though, it is unreasonable to suggest that a "pure" test of language pragmatics, if such a thing were possible, would be entirely distinct from core language.

Implicature as part of Broader Domains of Language Development

Implicature is, of course, just one pragmatic phenomenon. To more fully sample a child's developing ability in pragmatics, we need to administer a range of pragmatic subtests before extracting a general factor of pragmatic ability from them. In previous chapters, I used factor analysis on data collected from adults to test whether implicature comprehension clustered as part of a more general communication factor revolving around inference/linguistic judgement. Chapters Two and Four showed this to be the case in autistic and non-autistic adults. Although overall factor structure was somewhat different in the two chapters, it was clear that performance on the Implicature Comprehension Test related to other tests requiring judgements about social and linguistic meanings, with vocabulary knowledge being quite a different domain to this more open/inferential aspect to language comprehension. Broadly speaking, this analysis supported the prediction of Relevance Theory that language comprehension involves an interaction between decoding (i.e. explicit language knowledge) and inference. One qualification to bear in mind is that tests typically showed relatively weak relationships with each other, suggesting that specific skills may be as important as domain-level ability for performance on these tests with an inferential/judgement-based aspect.

In the work reported in the current chapter, I administered a test battery to a sample of children to see whether a pragmatic language factor emerged as distinct from core language. To complement the Implicature Comprehension Test, I devised two new pragmatic language tests; one involving the detection of pragmatic violations and a second involving sensitivity to conversational overtures. For the pragmatic violations task, I devised short conversational interchanges in which a character violated the principle of optimal relevance; either they gave insufficient or redundant information in response to a question, or they drifted from the topic.

As well as being motivated by Relevance Theory, these categories of violation represent a range of pragmatic infelicities made by children with and without language and communication difficulties (Bishop & Adams, 1989). The other new pragmatic task involved conversational overtures; a character produces utterances such as “I had a really bad journey today” or “I’m going to close the windows”, and the participant needs to decide if the character is trying to start a conversation with the other character or just providing a brief comment. As such, this is a test requiring us to read communicative intentions based on a sentence in a social context. This task is clinically relevant as responses to conversational overtures are evaluated in the ADOS-2 (Lord et al., 2012), the widely used tool for assessing communication skills in people with autism. Alongside the implicature test and these two new pragmatic tasks, tests of receptive grammar and vocabulary were administered as measures of core language.

I had two hypotheses:

(1) Comprehension of conversational implicature would only be weakly related to grammar and vocabulary skills after controlling for age.

(2) In factor analysis, a model with pragmatic and core language skills as separable factors would provide a better fit to the data than a single factor model.

I also planned some follow-up analyses. In the case of pragmatics and core language skills being related, this may be because they are both reflections of general cognitive ability rather than because they are both linguistic, so it was important to assess whether any relationship was accounted for by nonverbal ability. I also aimed to assess the relationship between the language factors and sex and age.

Methods

This study was a partial replication of the work reported in Chapter Two, which involved a factor analysis of different language and social/communication skills in typical adults; the present study tested the factor structure of a similar set of tasks in typically-developing children. The study was granted ethical clearance in November 2018 by the Medical Sciences Interdivisional Research Ethics Committee at the University of Oxford: Ref. R59912/RE001.

Power Calculation

To determine the minimum sample size, I used an adapted form of the power calculation from Chapter Two. This method involved creating simulated datasets with known factor structure, and repeatedly running the analysis with different sample sizes to establish the sample size at which the factor structure could be recovered. Using the R package simsem (Jorgensen et al., 2018), I simulated datasets conforming to a two-factor model at the population level, with two indicators loading on one factor and three indicators loading on the other factor. Factor loadings in the population model were set to .7 and factor correlation to .6. In 10,000 simulations of samples of 120 individuals, 9980 datasets showed a significantly better fit to a two-factor compared to a one-factor model, when compared using a chi-square test with the alpha level set to .05. Thus, this sample size gives 99.8% power to detect the improved fit of a two-factor model at the .05 level.

Participants

I aimed to recruit at least 120 child participants aged between 7;0 and 13;11, by inviting all the year 3 to 8 children in several schools to take part. I used an opt-out approach to recruitment; schools were asked to circulate information to parents/carers two weeks

before research sessions, and if parents/carers preferred their child not to be involved, they returned an opt-out form to school. Children who were happy to take part in the research were asked to complete an assent form, and I informally checked ongoing assent throughout the research sessions. For any child preferring not to take part, teachers discretely arranged for them to complete other computer-based activities that were part of the normal week-to-week learning. The data from four children identified by their teachers as having an autism diagnosis were excluded from the analysis.

To maintain anonymity of the children involved in our project, I did not collect personally-identifying information. Teachers were simply asked against each child's research ID to indicate their year group (i.e. years 3-8), gender, whether they were diagnosed with autism, and whether they spoke English as an additional language.

The sample consisted of 390 children. See Table 6.1 for details of the sample, including year group, sex, English-speaking status and missing data. Children speaking English as an additional language tended to perform as well on the tasks as native speakers; it was only on vocabulary that a small advantage emerged for native speakers, who had a mean of 28.78 compared to 27.39, t (75.89) = 2.26, p = 0.027. Due to the practicalities of working in busy schools over several testing sessions, and IT difficulties, there is some missing data. Of the primary-age children (i.e. years 3 to 6, which comprise 7- to 11-year-olds), just over 30% did not complete at least one of the tests. There is considerable missing data for the youngest children; this was not a consequence of any particular challenges these children experienced with testing but instead due to IT difficulties that occurred as a matter of coincidence with two year 3 classes. As for the secondary-age children (years 7 and 8, comprising 11- to 13-year-olds), I did not set out to collect full data from any one child,

owing to limited timetabling opportunities. With this group, the order of tasks was randomised and children completed as many tasks as possible.

Table 6.1

Sample Characteristics. Teachers reported the sex of children in their classes, as well the number who spoke English as an additional language (EAL). Here we also indicate the number of children for whom there are complete data.

	N children	Girls	Boys	Sex not reported	EAL	Complete data
Year 3	54	29	25	0	15	21
Year 4	77	23	27	27	16	54
Year 5	83	42	41	0	8	72
Year 6	81	32	25	24	13	56
Year 7	51	24	27	0	0	0
Year 8	44	24	20	0	1	0

Procedure

For the primary school classes (years 3 to 6), two group-based testing sessions were arranged in school with each class. Children completed a battery of newly devised tasks during each session. Sessions lasted not longer than 50 minutes. Children worked on a sequence of computerised tasks, listening to audio through headphones and supervised by myself and a member of the school teaching staff. The tasks were supported by Gorilla. In the first session, children completed the Implicature Comprehension Test, Receptive Vocabulary task, Pragmatic Violations and Animal Matrices, in that order. In the second session, children completed the Social Overtures, Receptive Grammar task, Children's Test of Local Textual Inference, and one of the tasks from the first session again.

For the secondary school classes (years 7 and 8), children took part in one group-based testing session lasting 50 minutes. The order of tasks was randomised, and children completed as many tasks as they could manage in the time.

Measures

1. Pragmatic Language Tasks. In the factor analysis, these tasks were set to load on the "pragmatic language" factor. Further details of the tasks are presented in Appendix D and information on accessing the full assessments can be found here (*https://osf.io/xn48e/*).

Implicature Comprehension Test. See Chapter Two for a full description. Participants watch a series of cartoon videos, in each of which two characters produce a short utterance one after the other. Together the utterances form a conversational adjacency pair; in most cases, this is a question and answer. Afterwards, participants hear a comprehension question to which they are asked to give a yes-no-don't know response. For 36 items, participants need to process implicature to answer the question. There are also 10 items where the answer is more explicit; these serve as positive control items. From these items, there were two measured variables: sum of implicature items correctly answered (out of 33; three items were excluded, as mentioned in Results) and sum of explicit-response items correctly answered (out of 10).

Pragmatic Violations. Participants watch a sequence of 35 videos, in which one character produces a question and a second character answers the question. In 21 videos, the second character violates a pragmatic norm; in the others, the answer is good. Here is an example of a good item: "Why did the dog bark?" "Someone came to the door. They made a lot of noise. It scared the dog." There are four types of pragmatic violation: redundancy (n = 6), missing information (n = 6), topic drift (n = 6) and unresponsiveness (n = 3). The unresponsive items are designed to be trivially easy, as the answer is simply "er". In the redundant items, the second character states information that can be readily assumed from preceding sentences or world knowledge; e.g. "Where did you go yesterday?" "Yesterday I went to the swimming pool. I wanted to go swimming yesterday. I had to take my towel with

me." In the topic drift items, the second character shifts the topic under discussion in a tangential manner; e.g. "Where is the dog?" "I let the dog outside into the garden. I didn't give it any toys. It chewed up a teddy bear last week." In the missing information items, the second character does not provide the full context/background to their answer; e.g. "Did you go to the cinema?" "I missed the film. He said sorry." Individuals are asked to rate how good the answer is by placing a mark on an analogue scale showing a sad face at one end and a happy face at the other. The mark can be placed at seven increments along the line. Although numbers are not shown on the scale, scores between 1 and 7 are directly assigned for participants' responses according to where they put the mark on the line. The measured variable is the total number of pragmatic violations correctly identified. To calculate a participant's total score, I first identified the score given to the 21 items the participant thought were the worst (as 21 items were devised to include violations). Items with a pragmatic violation given a score below that level were given one point. For items with a pragmatic violation that were given a score at that level, I gave one point divided by the number of items given that score (to allow for tied items). During piloting with 30 non-autistic adults, three of the items did not seem to hang well with the others (1 missing information item and 2 redundancy items). To improve the reliability of the measure, these items were excluded from the total score. Cronbach's alpha calculated from the adult pilot data was .81, 95% CI [.72, .89].

Social Overtures. Participants hear 24 utterances spoken by a character to a conversational partner. Twelve are social overtures that attempt to engage the partner in a conversation (e.g. "I can't believe what happened today.") and twelve are not conversational bids (e.g. "I'm going to have a shower now."). Participants click yes-no buttons to indicate whether the character wants to have a conversation or not. There was one measured variable:

the sum of items correctly identified as a social overture or not (out of 23; one item was excluded, as mentioned in Results).

2. Core Language Tasks. In the factor analysis, these tasks were set to load on the "core language" factor. Further details of the tasks are presented in Appendix D and information on accessing the full assessments can be found here (*https://osf.io/xn48e/*).

Receptive Vocabulary. This includes 40 items in which participants choose which of four pictures is related to a word. Words are presented over audio and include nouns, verbs and adjectives. The words vary in approximate age of acquisition from 5 to 12, with similar numbers of easy and harder words; two experienced teachers independently rated the ages at which they would expect 50% and 90% of children in a typical class to be familiar with the word. There was one measured variable: the sum of items correctly answered (out of 39; one item was excluded, as mentioned in Results).

Receptive Grammar. Participants listen to sentences and decide if they are good or have mistakes/sound odd. There are 50 items: four sentences are entirely muddled and should be easily rejected, 20 items are borrowed from McDonald (2006) and showed high accuracy in primary school children, and 26 items are a subset of our adult version of this test (see Chapter Two); these latter items were chosen on the basis of high accuracy and high item-total correlations. Excluding the four muddled sentences, 23 items are grammatical and 23 are not. There was one measured variable: the sum of items currently answered (out of 50).

3. Further Cognitive Measures.

Animal Matrices. This non-verbal reasoning task is an adapted version of the Animalogica multiple choice test (Stevenson et al., 2016). There are 18 items. Each item is a 2x2 matrix presented on the computer screen. In three of the boxes of each matrix, there are

cartoon pictures of animals, and the fourth box is empty. The animals in the three boxes vary along six dimensions: species, colour, size, number, direction faced, and position in the box. There are systematic relationships between the three animals, and participants need to deduce which of five options fits in the empty box. For example, the top two boxes may show red lions, one big and one small, and the bottom left box may show a big yellow horse; the correct option to fill the empty box would be a small yellow horse. There was one measured variable: the sum of items correctly answered (out of 16; two items was excluded, as mentioned in Results).

Children's Test of Local Textual Inference. This is based on the adult version of this test (see Chapter Two). Participants hear two brief sections of a short story (about 90 words per part). After each section, they hear ten questions and four possible answers for each one. "We don't know" is an answer option for every question, and is the correct answer to four questions. Participants click the correct option on the screen. As well as auditory presentation of all materials, everything is shown in text-based form on the screen; the short story sections remain on the screen while participants are answering questions about that section. Participants need to make coherence inferences based on the short story to answer the questions. There was one measured variable: the sum of items correctly answered (out of 20).

Vocabulary and Matrix Reasoning. 20 children aged 9 to 10 completed these two subtests from the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) so that the validity of the novel tasks could be evaluated in comparison to these standardised assessments.

Data Analysis

Data were analysed in R (R Core Team, 2017). Data and scripts are accessible on OSF: *https://osf.io/xya3j/*.

As the tests were novel, I began by assessing item functioning and test reliability to establish the quality of the tests. I used classical test theory (CTT) and item response theory (IRT; Embretson & Reise, 2000) approaches, implemented with R packages psych (Revelle, 2017) and mirt (Chalmers, 2012). For each item in each test, I inspected accuracy and the correlation between item-accuracy and total-accuracy on the test with that item excluded. Items were identified as poor if they had low accuracy and a low item-total correlation. I also inspected item characteristic curves (ICCs) produced using IRT analysis, which determines difficulty and discrimination parameters for each item. ICCs are useful in identifying poor items, as they show the probability that an individual of a certain ability scores at a particular level on an item; low flat curves indicate items that are ambiguous or too difficult for the participant group, with no consensus answer at any point along the ability spectrum. I excluded any items showing this pattern. I then computed the reliability of each of the tests using CTT coefficients, Cronbach's alpha (with 95% confidence intervals) and Revelle's beta. Revelle's beta is the worst split-half reliability – i.e. the correlation between the halves of a test when it is split in such a way that the correlation is minimised (Zinbarg et al., 2005). Beta was used to give an estimation of general factor saturation that removes the influence of any subgroup factors. Where there is a large discrepancy between alpha and beta, the measure is heterogenous in its content. I also carried out IRT analysis, modelling test performance as a function of item characteristics and participant ability. ICCs were inspected, as described above, when assessing item functioning, and generated estimates of reliability by ability level. IRT reliability plots were created using R package ggplot2 (Wickham, 2009). As an indication of the unidimensionality of the tests, I report the root mean square error of estimation (RMSEA) for a unidimensional IRT model for each test.

To test the relationship between implicature comprehension on the one hand and core language skills and narrative-based inferencing on the other, I ran a hierarchical multiple

regression using data from individuals who completed the four tests relevant to this analysis. In the first stage of the regression, year group was included as a predictor of implicature comprehension; in the second stage, vocabulary and grammar were included as predictors; and in the third stage, narrative-based inferencing was added. Residual plots were inspected to check whether assumptions of regression were met (Altman & Krzywinski, 2016). Adjusted R-squared was the effect size, with 95% confidence intervals estimated through bootstrapping (using R package boot; Canty & Ripley, 2019).

For the second hypothesis, I used confirmatory factor analysis to test whether implicature comprehension forms part of a "pragmatic language" factor that is distinct from "core language". First, I inspected the data for normality and multivariate outliers using R package MVN (Korkmaz et al., 2014). Multivariate outliers were defined as individuals whose adjusted Mahalanobis' distance was above the 97.5th percentile of the chi-distribution. Maximum likelihood estimation was used for the factor analysis. As this is based on the multivariate normal distribution, it was important to check whether our data departed from this distribution (Flora et al., 2012), although given that robust estimation was used in the factor analysis, assumptions of normality were relaxed (Finney & DiStefano, 2013). Nonetheless, in the case of multivariate outliers being present, I planned to transform variables using the Tukey ladder of power transformations using R package rcompanion (Mangiafico, 2018) to reduce skew, and then test for multivariate outliers again. Remaining outliers would be excluded. The sensitivity of the analysis to any data transformation or outlier exclusion was evaluated by comparing results based on transformed and non-transformed data with and without outliers included.

For the confirmatory factor analysis, I used R package lavaan (Rosseel, 2012), and semtools (Jorgensen et al., 2019) for the visual representations of our analysis. I specified a

two-factor correlated-traits model with a “core language” factor and a “pragmatic language” factor. Two indicators were set to load on the “core language” factor: receptive grammar and receptive vocabulary. Three indicators were set to load on the “pragmatic language” factor: implicature comprehension, sensitivity to pragmatic violations, and sensitivity to social overtures. The two factors were allowed to freely correlate. The comparison model was a one-factor model in which all five indicators were set to load on the same factor. In both models, year group was included as a covariate, regressing each indicator onto year group to control for age differences in test scores. I used full information maximum likelihood estimation to account for missing data, and computed robust standard errors in case of data not being normally distributed. A chi-square test with Satorra-Bentler correction was used to evaluate whether the two-factor model fitted better. I report confirmatory fit indices (CFIs) and root mean square error of estimation (RMSEA) with 90% confidence intervals. R package SemPlot (Epskamp & Stuber, 2017) was used to make a visualisation of the factor model.

Some follow-up analyses were planned. The first involved adding nonverbal reasoning (measured by the Animal Matrices) to the two-factor model as a covariate, with regression paths from age to nonverbal reasoning (to control for age effects) and from nonverbal reasoning to the two language factors. The aim was to establish the strength of the relationship between core and pragmatic language when controlling for general nonverbal ability. The other follow-up analyses assessed for age and sex effects. I considered whether the relationships between the indicators and factors were the same when sub-grouping children by age and gender, by testing for measurement and structural invariance (Sass & Schmitt, 2013). Measurement invariance is satisfied when the factors are derived from the indicators in the same way across sub-groups. Separate models are defined for each sub-group, and parameters are progressively constrained to be the same across sub-groups. The

first step is to test for configural invariance (that the same basic structure with five indicators loading on two factors exists across sub-groups), then metric invariance (that indicators load to the same extent on their factors across sub-groups), scalar invariance (that indicator intercepts are equivalent across sub-groups, i.e. that any sub-group differences in individual test scores are fully accounted for by underlying factors), and finally strict invariance (that indicator error terms are the same across sub-groups). As each of these more stringent models is run, a chi-square difference test is used to assess whether fit significantly deteriorates. If there is a significant change in fit, it might be appropriate to allow certain parameters to freely differ, indicating partial invariance across sub-groups. Having established the level of measurement invariance, I planned to assess for structural invariance (that factors themselves are equivalent across sub-groups). This involves comparing the most stringent model supported during the assessment for measurement invariance across sub-groups, with models where factor means are fixed across sub-groups, and where factor correlation is fixed. This would indicate whether core and pragmatic language abilities, and the relationship between them, differed by age or sex. This analysis was run twice: in the first, sub-groups were defined by age (years 3 to 5 and years 6 to 8), and in the second by sex.

Results

Reliability Analysis

First, I assessed whether the tests reliably tapped an underlying cognitive ability. I supply classical test theory coefficients and item-level statistics in Tables 6.2 and 6.3.

As can be seen here, the tests showed reasonably high internal consistency, so scores are subject to relatively little measurement error. The beta reliability coefficients are somewhat lower than the alpha coefficients. As the beta coefficient (worst split half reliability) reflects general factor saturation more accurately than the alpha coefficient

(average split half), the discrepancies observed between the two types of reliability coefficients suggest that items in individual tests are not homogeneous. This is because item-total correlations are not universally high – this is likely due to item-level accuracy being variable, and we would not, for instance, expect to find strong correlations between total scores and accuracy on items with ceiling effects.

Table 6.2

Reliability Analysis, including Cronbach's alpha and 95% confidence limits, standard error of measurement (SEm), and Revelle's beta. I also provide RMSEA when running IRT models, which gives an indication of unidimensionality of the test.

Test	Alpha	Alpha 95% CI	SEm	Beta	IRT rmsea
Implicature	.79	.76, .82	2.22	.68	.040
Pragmatic Violations	.73	.69, .77	1.57	.48	.073
Social Overtures	.82	.79, .85	1.51	.59	.086
Vocabulary	.79	.76, .82	1.99	.63	.024
Grammar	.89	.87, .91	2.05	.79	.059
Inferencing	.71	.66, .75	1.44	.60	.028
Matrices	.81	.78, .84	1.60	.64	.031

Table 6.3

Item-level statistics for each test: corrected item-total correlations (totals excluding the item) and item-level accuracy.

	Item-total correlations			Item-level Accuracy		
Test	Lower quartile	Median	Upper quartile	Lower quartile	Median	Upper quartile
Implicature	.25	.29	.32	.69	.77	.84
Pragmatic Violations	.24	.32	.35	.52	.69	.78
Social Overtures	.27	.36	.47	.81	.85	.89
Vocabulary	.17	.25	.31	.57	.89	.96
Grammar	.27	.35	.45	.80	.89	.92
Comprehension	.21	.28	.34	.89	.92	.94
Inferencing	.20	.29	.33	.77	.88	.92
Matrices	.36	.40	.47	.37	.56	.71

I also ran reliability analysis using the item response theory (IRT) approach, as this is especially appropriate when item difficulty varies. The IRT RMSEA values indicate that a unidimensional model fitted each test at least adequately according to the criteria of 0.05 for 'close fit' and 0.089 for 'adequate fit' set out by Maydeu-Olivares and Joe (2014). Figure 6.1 below shows curves for test reliability by differing level of participant ability. These curves show that the tests produce highly precise scores in the below-average ability range, meaning that they should be effective in identifying individuals with difficulties.

The reliability analysis reported above was based on final versions of the tests. A few items were dropped from the tests because they were not reliable. They had flat ICCs (corresponding to chance-level accuracy and low item-total correlations) and so did not reliably tap a latent ability across the age-span. The weak items included one overtures item, four implicatures items, one vocabulary item, and two matrices items.

I assessed the extent to which our tests measured a stable ability by calculating test-retest correlations across two administrations 2 weeks apart. Correlations were as follows: Receptive Vocabulary .77 (n = 34), Animal Matrices .88 (n = 26), Implicature Comprehension .77 (n = 77), and Pragmatic Violations .52 (n = 55). The tests all showed reasonable test-retest reliability with the exception of the Pragmatic Violations. As the tests were novel, it was important to assess their convergent validity with existing standardised measures, so a small subset of children aged 9 to 10 (n = 20) were given two WASI sub-tests. Correlation between the novel vocabulary task and WASI Vocabulary was .69, and between the Animal Matrices and WASI Matrix Reasoning was .70.

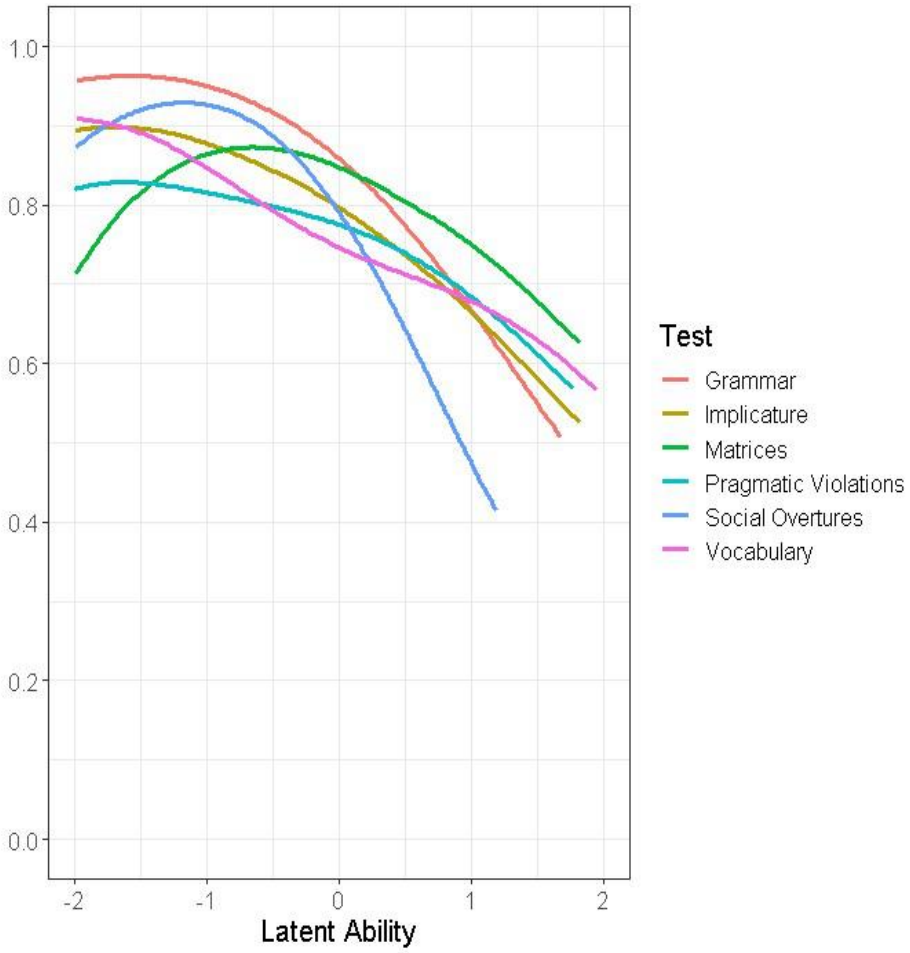


Figure 6.1. Plot showing reliability computed through item response theory (IRT) modelling for the tests.

Hypothesis-Testing

See Table 6.4 for descriptive statistics for each measure, and Table 6.5 for a correlation matrix. The tests showed small to moderate correlations.

Table 6.4

*Descriptive Statistics. Variables included in the factor analysis are in **bold** type.*

	N	Mean	SD	Min	Max	Skew	Kurtosis
Implicature	330	25.20	4.85	8	33	-0.68	0.18
Control items for Implicature	328	8.68	1.76	2	10	-1.60	2.13
Pragmatic Violations	320	11.90	3.03	0.48	18	-0.95	1.35
Social Overtures	297	19.60	3.57	8	23	-1.18	0.56
Vocabulary	338	29.60	4.35	10	39	-0.58	0.78
Grammar	291	41.40	6.17	14	49	-1.43	2.00
Narrative-based Inferencing	294	16.40	2.66	6	20	-1.32	2.17
Matrices	313	8.86	3.68	0	16	-0.31	-0.63

My first hypothesis was that implicature comprehension would be relatively distinct from grammar and vocabulary skills, and also distinct from the ability to draw inferences in narrative. For this hypothesis, I ran a hierarchical multiple regression. The first stage of the regression, which included year group as a predictor of implicature comprehension, was marginally non-significant, $F(1, 210) = 3.12$, $p = .079$. Incorporating receptive grammar and receptive vocabulary as predictors significantly improved the model in the second stage, $F(2, 208) = 12.80$, $p < .001$, and predicted 11% of the variance in implicature scores. In the third stage, narrative-based inferencing also improved the model, $F(1, 207) = 14.25$, $p < .001$. See Table 6.6 for the significance of individual predictors at each stage of the regression. The final model was significant, $F(4, 207) = 11.29$, $p < .001$, and explained 16.3% of the variance, bootstrapped 95% CI [5.4%, 24.4%]. Compared to the sample of typical adults reported in Chapter Two, this is a similar proportion of variance; in adults, comparable tests of vocabulary, grammar and inferencing predicted 11.1% of variance in implicature comprehension, bootstrapped 95% CI [0%, 21.4%]. The results supported my hypothesis that implicature comprehension was only related to core language skills and narrative-based inferencing to a limited degree

Table 6.5

Correlations between language tests using pairwise complete observations.

	Pragmatic Violations	Social Overtures	Grammar	Vocabulary	Narrative Inferencing	Matrices	Year Group
Implicature	.29	.43	.33	.36	.36	.38	.23
Prag. Violations		.34	.22	.38	.12	.30	.15
Social Overtures			.43	.53	.30	.40	.40
Grammar				.57	.36	.44	.53
Vocabulary					.32	.45	.47
Narr. Inferencing						.42	.25
Matrices							.47

Table 6.6

Coefficients for Multiple Regression, with Implicature scores as the criterion variable.

	Estimate	Standard Error	t-value	p-value
Stage 1				
Year group	0.54	0.30	1.77	.079
Stage 2				
Year group	0.00	0.31	-0.02	.984
Vocabulary	0.15	0.14	1.06	.289
Grammar	0.08	0.02	3.76	<.001
Stage 3				
Year group	-0.11	0.30	-0.36	.719
Vocabulary	0.06	0.14	0.41	.680
Grammar	0.06	0.02	3.28	.001
Inferencing	0.14	0.04	3.78	<.001

In my second hypothesis, I stated that implicature comprehension would cluster with other tests that theoretically involve pragmatic processing but would be relatively distinct from core language skills. To test this hypothesis, I compared a two-factor model of language processing with a one-factor model. The two-factor model fitted the data well, CFI = 1, RMSEA = 0, 90% CI [.00, .06] whereas the one-factor model was weaker, CFI = .97, RMSEA = .08, 90% CI [.04, .12]. The difference in fit between the models was significant, $\chi 2$ (1) = 20.63, $p < 0.001$. In the two-factor model, the correlation between the "core language" and "pragmatic language" factors was .72, 95% CI [.58, .87]. This indicates that, while a two-factor model does show a better fit, the factors are highly correlated in typically developing children. It is possible that where children found it more challenging to engage with the basic language level of the pragmatic tasks, this may have inflated the correlation between domains. Therefore, I also ran an analysis in which I excluded children scoring less than 8 out of 10 on the control items of the Implicature Comprehension Test; this score represents the threshold used for outlier exclusion in the adult data reported in Chapter Two, and equated approximately to the 20th percentile in the present data. Carrying out these exclusions did not alter the outcome of the analysis, with a two-factor model being preferred and a similar correlation between the factors (r = .71).

Note that the factor analyses were run with (i) all data and only complete cases, and (ii) non-transformed and transformed data, to evaluate the sensitivity of the results to the data processing. Without transformations, there were 23 multivariate outliers, and with transformations, there were none. All factor analysis results were very similar regardless of processing. See Figure 6.2 for a visual representation of the factor model, and see Figure 6.3 for a plot of "core language" against "pragmatic language" factor scores extracted from the model. This plot shows a cloud of points positively correlated, with no obvious outlying

individuals – i.e. there are no clear cases of children having well-developed core language but weaker pragmatics, or vice versa.

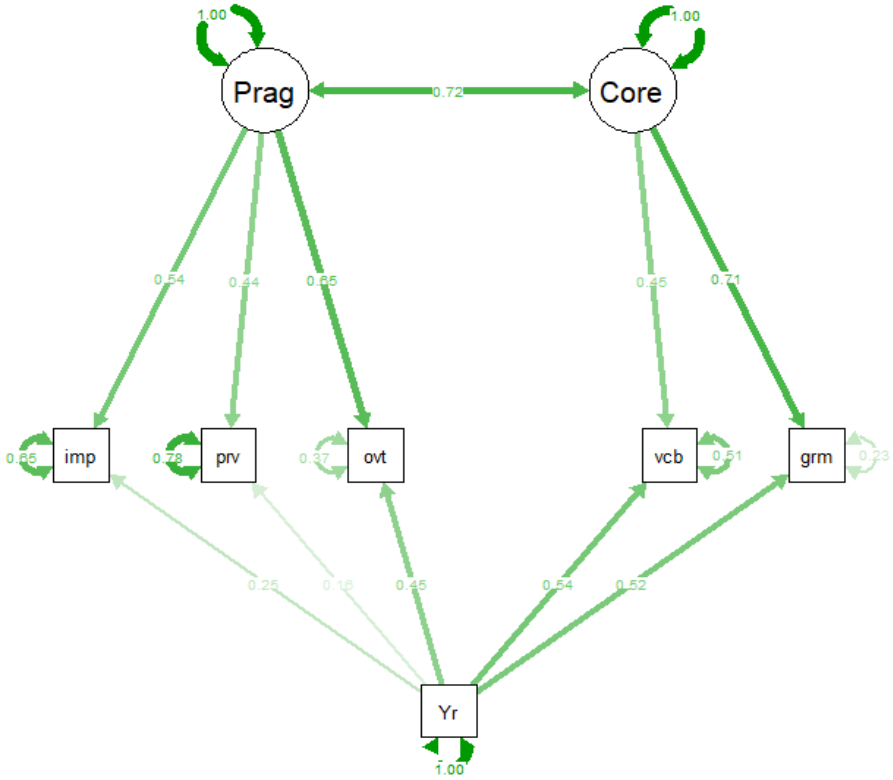


Figure 6.2. A two-factor correlated-traits model including a "pragmatic language" factor and a "core language" factor. Year group has been included as a covariate at the level of the indicators to control for age effects. Prag = "pragmatic language" factor; Core = "core language" factor; imp = Implicature Comprehension Test; prv = Pragmatic Violations; ovt = Social Overtures; vcb = Receptive Vocabulary; grm = Receptive Grammar; Yr = year group.

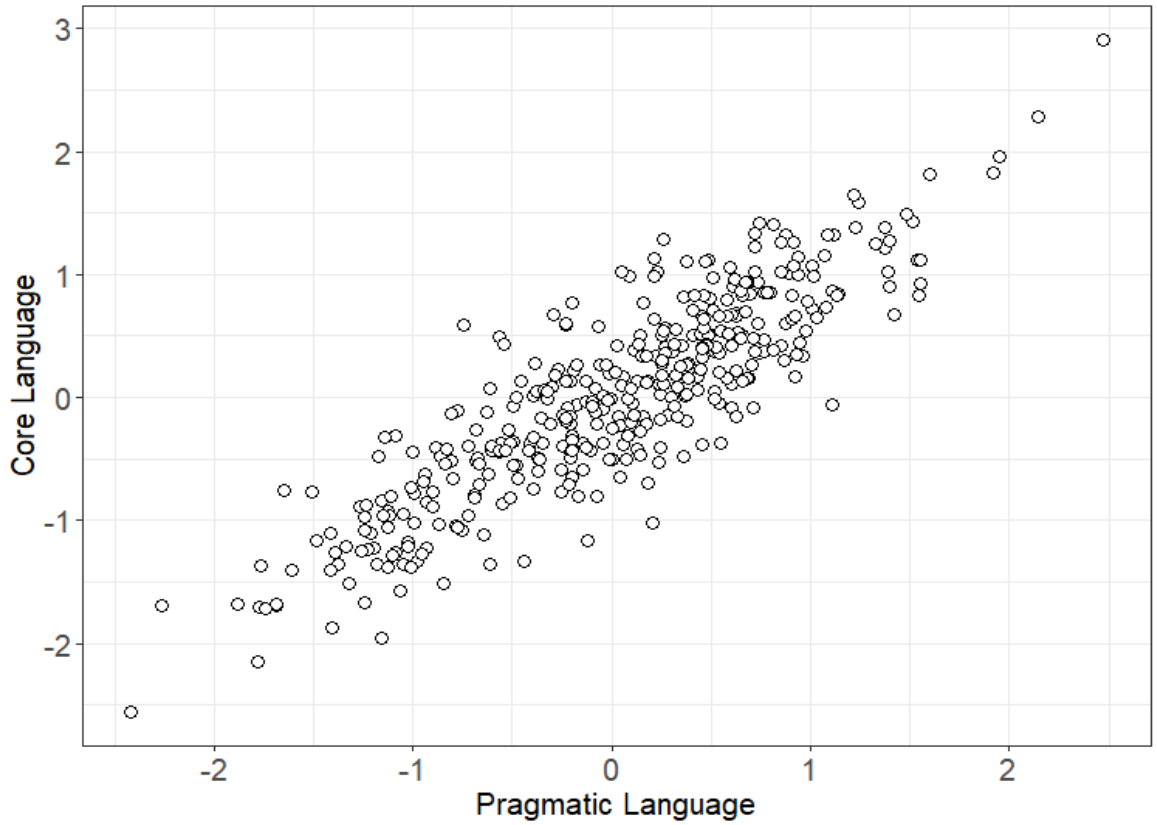


Figure 6.3. Plot showing "pragmatic language" factor scores against "core language" factor scores.

Follow-up Analysis

Given that the two-factor model fitted the data well, I asked three further questions. Firstly, what was the relationship between nonverbal reasoning and the two language factors? I added nonverbal reasoning to the two-factor model as a covariate. Regression paths were included between year group and nonverbal reasoning, and between nonverbal reasoning and the two factors. Nonverbal reasoning showed moderate relationships with "pragmatic language", .53, 95% CI [.41, .66], and "core language", .48, 95% CI [.32, .63]. The inclusion of nonverbal reasoning as a covariate had very little effect on the correlation between the "core language" and "pragmatic language" factors, which was now .68, 95% CI [.51, .85].

The second question related to possible differences by age; was the same structure, with five indicators loading on two factors, present across year groups? We examined the measurement invariance of the model by progressively fixing loadings, intercepts and residuals of the indicators to be the same across two sub-groups: children in years 3 to 5 (7 to 10-year-olds) and years 6 to 8 (10 to 13-year-olds). Fit deteriorated slightly when testing for strict invariance, although the fit statistics shown in Table 6.7 indicate this was not a major cause for concern. It seems that the factors were measured in a similar way across age groups.

Table 6.7

Testing for measurement invariance by age when comparing younger (years 3 to 5) and older children (years 6 to 8), using multi-group confirmatory factor analysis.

	Model Df	Chi-square Statistic	Chi-square Diff	Diff Df	P-value	CFI	RMSEA
Configural	8	10.97				.99	.05
Metric	10	14.47	2.85	2	.241	.98	.05
Scalar	13	17.83	3.09	3	.378	.98	.04
Strict	18	32.41	19.34	5	.002	.94	.07

Next, I tested whether the correlation between the factors differed when dividing children into these two age-groups. With loadings, intercepts and residuals still fixed (except for the Pragmatic Violations factor loading), I constrained the factor correlation to be the same across the two sub-groups. Model fit was not significantly affected, $\chi 2$ (1) = 0.36, p = .551. This indicates that the relationship between core and pragmatic aspects of language remained constant, at least when comparing these two broad age-groups.

The third, and final question considered sex effects. This analysis also began with an assessment for measurement invariance. As Table 6.8 shows, the factors were measured identically across males and females.

Table 6.8

Testing for measurement invariance by sex when comparing boys and girls using multi-group confirmatory factor analysis.

	Model Df	Chi-square Statistic	Chi-square Diff	Diff Df	P-value	CFI	RMSEA
Configural	8	3.39				1	.00
Metric	11	5.21	1.93	3	.586	1	.00
Scalar	14	9.46	4.45	3	.217	1	.00
Strict	19	12.73	3.74	5	.587	1	.00

When further constraining the model with strict invariance so that factor means were the same, fit was marginally poorer, $\chi 2$ (2) = 7.36, p = 0.025, CFI = 1, RMSEA = .01. Figure 6.4 presents mean score (with 95% CIs) on the "pragmatic language" and "core language" factors for males and females. I also present Figure 6.5 which shows factor means by year group; the factor scores represented here have been derived from the same two-factor model reported above, except that the year group covariate at the level of each indicator was removed. Figure 6.5 is on the same scale as Figure 6.4 to allow comparison of the small sex effect with the substantial age effect. Females have slightly higher scores on average, Cohen's d = 0.34 for "pragmatic language" and 0.35 for "core language".

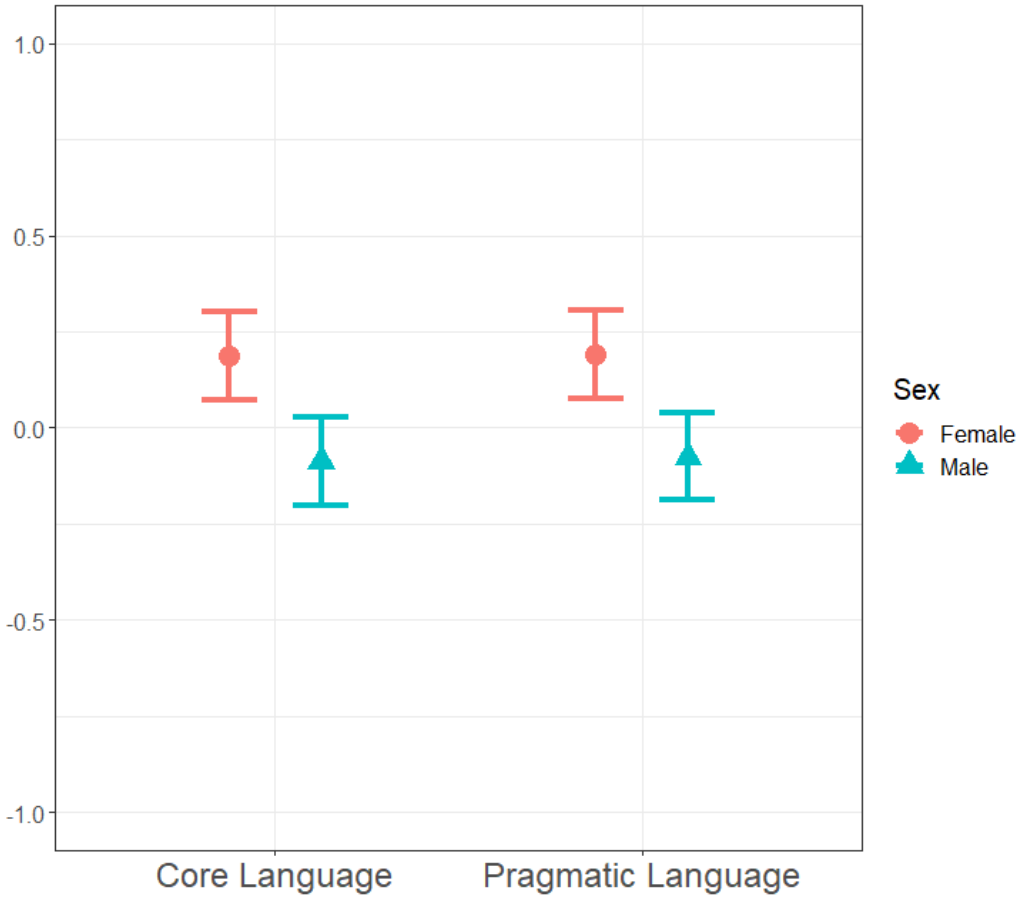


Figure 6.4. Plot showing mean scores (with 95% CIs) for the "pragmatic language" and "core language" factors by sex.

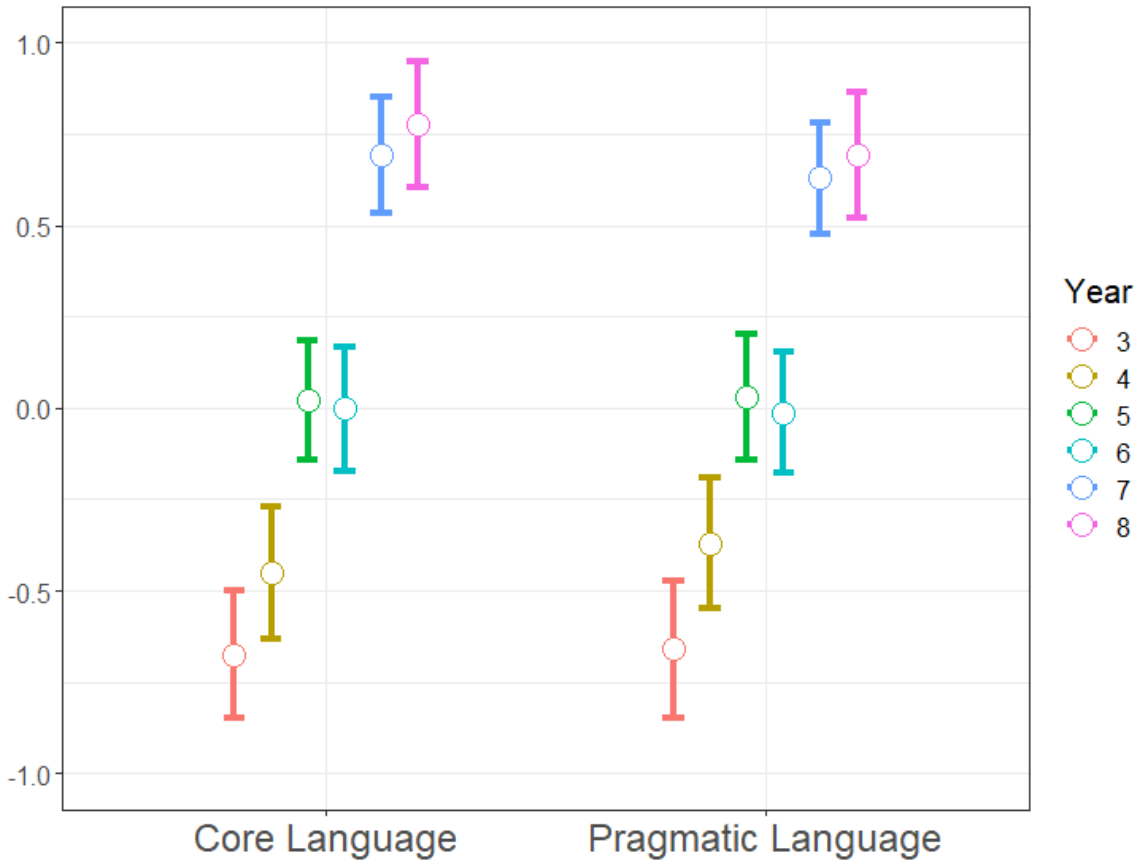


Figure 6.5. Plot showing mean scores (with 95% CIs) for the "pragmatic language" and "core language" factors by year group. Note that in the original factor analysis, year group was included as a covariate to control for age effects in individual test scores. Clearly, this would mean that effects of age would not be expressed at the factor level, so the covariate was removed. This reveals a substantial effect of age on language abilities, which can be compared with the small sex difference shown in Figure 6.4 (where age is controlled).

Discussion

My analysis in this chapter showed that a two-factor model of language comprehension fitted data collected in almost 400 children aged 7 to 13 years better than a single factor model, indicating that core and pragmatic aspects of language comprehension could be measured distinctly across several novel language tests. Nevertheless, these language domains were strongly related, and this was the case across the age range. The magnitude of the relationship (r = .72) is in line with previous research, even though the tests set out to measure pragmatics more distinctly than has generally been attempted before

(Matthews et al., 2018). The findings support the view that well-developed language abilities in one area support development of those in the other. The theoretical and practical implications of my analysis run in perhaps rather different directions, so I discuss these separately below.

Theoretical Implications

As I was able to measure core and pragmatic language as separable – albeit highly related – factors, the findings support linguistic theories that language comprehension is not unitary but depends on twin processes of (i) decoding words and grammatical structures based on stored knowledge and (ii) inferring intended meanings in a flexible, context-dependent way (Cummings, 2007; Sperber & Wilson, 1986). Linguistic theories of pragmatics typically agree on a broad distinction between decoding and inference when contrasting semantics ("dictionary meaning" for want of a better term) and pragmatics (context-dependent meaning), as discussed in a book-length treatment of the definition of pragmatics (Ariel, 2010). While the factor analysis does support the view that there are somewhat separate cognitive domains underpinning language comprehension, this should not be overstated as there is substantial overlap. Interestingly, this overlap is not accounted for by general cognitive ability. As part of the analysis, I included nonverbal reasoning in the model as a covariate. On the one hand, it was moderately associated with both language factors, as would be expected based on the idea that the g-factor (or general intelligence) affects all cognitive tests (Jensen, 1998). However, when controlling for nonverbal reasoning ability, the relationship between the two language factors remained strong. This suggests that core and pragmatic language abilities are related because of their linguistic nature rather than because they are simply reflections of general cognitive ability.

There is a view that some individuals may have difficulties with implied meaning in the absence of a core language impairment. For instance, the communication profile of verbal autistic individuals may involve well-developed vocabulary/grammar but differences in ability and/or tendency to make inferences (Loukusa & Moilanen, 2009) and a preference for unusually literal language (Hobson, 2012). Based on these observations, we might hypothesise that processing implicature involves more than just core language skills, and our analysis supports that view. In the analysis, vocabulary and grammatical skills were only a modest predictor of children's implicature scores. It is not surprising that these core language skills should play some role, as we must understand the "explicit content" of an utterance to have any chance of picking up an implied meaning, but clearly vocabulary and grammatical skills are not sufficient by themselves. Narrative-based inferencing ability was also a predictor of children's implicature scores, but again only explained a modest amount of variance. Narrative-based inferencing requires individuals to form coherence inferences in a story to fill in "blanks" (Garnham & Oakhill, 1992). I expected to find some overlap between scores on this measure and the implicature comprehension test, since both require the integration of information across more than one sentence. However, the modest amount of overlap suggests that making inferences across narrative is quite different to picking up implied meanings in conversation, where we need to detect whether a response might have provided an indirect answer to a question.

Practical Implications

While a two-factor model was preferred in the analysis, we should not downplay the large correlation between core and pragmatic aspects of language function. Indeed, we might debate how meaningful it is to think of a real-world distinction between the two language domains, at least as observed in typically developing children. There is the longstanding

intuition that core and pragmatic abilities may be dissociable, owing to clinical observation of individuals with communication difficulties that apparently affect pragmatics independently of core language; such individuals might be labelled as having pragmatic language impairment (PLI; Bishop, 1998). While this label was not formalised in any diagnostic manual, DSM-5 has incorporated a new diagnosis which is not dissimilar to PLI: social (pragmatic) communication disorder (American Psychiatric Association, 2013). This requires social communication difficulties in the absence of both (i) structural language problems and (ii) repetitive and restrictive behaviours and interests characteristic of autism. However, this diagnosis has been the subject of controversy, not least because researchers question the validity of an isolated pragmatic impairment (see Norbury, 2014; Swineford et al., 2014 for reviews). For instance, attempts to differentiate children with pragmatic language impairments from those with core language impairments on the basis of core language have not been very successful (Botting & Conti-Ramsden, 2003; Gibson et al., 2013). Similarly, children said to have core language impairments are frequently found to have social communication impairments too upon standardised assessment (Leyfer et al., 2008). As pragmatic and core language impairments often co-occur, it is perhaps not surprising to see little evidence in the present data for children who fit the profile of a specific pragmatic impairment (i.e. a low "pragmatic language" factor score and a significantly higher "core language" factor score). Looking at Figure 6.3, which plots factor scores against each other, it is difficult to pick out any child with this profile on visual inspection. To take a more formal approach, we could operationalise criteria as a "pragmatic language" factor score at least 1*SD* below the mean and a discrepancy of at least 1*SD* between "pragmatic language" and "core language" factor scores. Using these criteria, there are just two children (of 390) who may fit the profile of a specific pragmatic difficulty. As the only information available about these children is their test scores, it is not possible to evaluate the extent to which they might

genuinely be presenting with primary pragmatic difficulties in their day-to-day settings (or indeed whether other children not identified by the tests may show these types of difficulties). It is clear that specific issues with pragmatics, at least as assessed by my pragmatic tests, are not common; instead, pragmatic difficulties are more than likely to co-exist with impairments in core language.

There are a couple of caveats to bear in mind. Firstly, the ability to detect pragmatic difficulties may be limited by the measures used here. The nature of pragmatics is flexibility – the ability to adapt our use and understanding of language to the social context. This is inherently tricky to operationalise in a structured testing situation, and communication difficulties may only emerge in naturalistic conversation. Secondly, it is worth making a distinction between skills and domains. The factor analysis described here aimed to measure broad domains of pragmatics and core language by drawing upon what is common across multiple tests. However, as can be seen in Table 6.2, the correlations between tests were small to moderate, indicating that the tests reflect a combination of domain-level language abilities and also narrower test-specific skills too. It is, therefore, entirely plausible that some children may show a spiky profile, in which some test-specific skills are stronger than others. In this regard, we might reflect again on the relatively low relationship observed between the implicature and narrative tasks, as discussed above. Both tests required the child to draw inferences, but clearly there were somewhat different skills involved, and some children might have found one task more challenging than the other. It is important, therefore, that where educationalists and clinicians are assessing children's ability with implied and inferred meanings that they don't only assess inferences in spoken or written narratives, as inferencing is multifaceted. Conversational implicature should also be considered, and given that participation in conversation can be more tricky for children with communication difficulties, this may be particularly relevant to "real-world" difficulties. Overall, the message might be

that children will typically show similar levels of ability at the domain level across pragmatics and core language, but they might show some variability in task-specific skills relevant to different aspects of communication.

I present evidence for a two-factor model of language processing in typically developing children, with a "pragmatic language" and "core language" factor. These factors were highly related, however, so we should be careful not to imply there is a clear dissociation between these language domains. At least among typically developing children, it is not plausible to find substantial discrepancies between their pragmatic and core language skills. In the next chapter, I will explore the extent to which this also applies to children diagnosed with autism.

Chapter Seven. Testing for Pragmatic Differences between Autistic and Non-Autistic Children

In this chapter, I report a comparison between autistic and non-autistic children on the language test battery described in the previous chapter. As noted previously, the battery was devised to tease apart pragmatic and core language skills. Therefore, I hypothesised that a two-factor model with separate "pragmatic language" and "core language" factors would fit performance across the battery better than a one-factor model. This is an extension of the analysis reported in the previous chapter, which found that a two-factor model of language skills was preferred in a large sample of non-autistic 7- to 13-year-olds tested in several mainstream schools. In this chapter, I extend that analysis to include a sample of autistic children alongside the non-autistic sample. Given that autistic individuals are described as having specific difficulties with understanding and using language in its social context, I hypothesised that the autistic subgroup would score lower on the "pragmatic language" factor compared to non-autistic individuals.

Methods

I report below how I determined the sample size, all data exclusions, all manipulations, and all measures in the study. The study was granted ethical clearance in November 2018 by the Medical Sciences Interdivisional Research Ethics Committee at the University of Oxford: Ref. R59912/RE001.

Power Calculation

To determine the minimum sample size, I used a similar power calculation as in Chapter Four. Using the R package simsem (Jorgensen et al., 2018), I simulated datasets conforming to a two-factor model at the population level, with two indicators loading on a

"core language" factor and three indicators loading on a "pragmatic language" factor. Factor loadings in the population model were set to .6 and factor correlation to .7. I specified a correlation between group membership (autistic or non-autistic) and the pragmatic language factor of .3. In 10,000 simulations of samples of 170 individuals (50 autistic), 9681 datasets showed a significantly better fit to a two-factor compared to a one-factor model, and a significant relationship between group and pragmatic language factor score was found in 9998 samples. The alpha level was set to .025 (i.e. the conventional .05 level adjusted for two statistical tests). Thus, this sample size gives over 96% power to detect the improved fit of a two-factor model, compared to one-factor model, and over 99% power to detect a significant effect of group.

Participants

The sample included 55 children aged between 7;0 and 14;11 with an autism spectrum diagnosis, alongside the 390 non-autistic children whose data contributed to the analysis in Chapter Six. As five children did not complete a full assessment, 55 children were recruited in total, rather than the 50 specified in the power calculation so that 50 full data-sets would be available for analysis. Families of children with autism were recruited through Autistica, a national UK network for individuals and families interested in autism research, as well as through social media, special schools, and support groups organised privately and through the National Autistic Society. Exclusion criteria included: (1) severe and uncorrected sensory impairment, (2) history of neurological illness or brain injury, (3) nonverbal or word/phrase-level speech, and (4) English spoken as an additional language. To screen for limited language, families were asked "We would like to know whether your child can speak in sentences such as 'I didn't go to the park because it rained' and 'I think the film was really good'. Your child may make mistakes when using sentences, but can they at least sometimes

use sentences like these?" Families needed to say 'yes' in response to this question. Children were not excluded on the basis of other neurodevelopmental diagnoses or learning disabilities. Parents/guardians reported diagnoses of dyslexia (n = 6), language disorder/specific language impairment (n = 4), dyspraxia/ developmental co-ordination disorder (n = 6), attention deficit hyperactivity disorder (n = 9), anxiety disorder (n = 8), and depression (n = 3). 30 children had autism as a sole diagnosis; 17 had one other diagnosis and eight had multiple co-occurring diagnoses. All children had received an autism diagnosis through a clinical service, and were also assessed using the 3Di-sv family interview as part of the research protocol (described below). Nine children (eight boys) did not meet 3Di-sv criteria for autism, although they were included in the analysis reported below.

Procedure

There were three stages to the research protocol. Firstly, parents/guardians were asked to complete some online questionnaires taking around 20 minutes. These questionnaires included the Pediatric Symptom Checklist (a measure of general psychosocial adjustment; Jellinek et al., 1986), and the Children's Communication Checklist (a measure of communication difficulties; Bishop, 1998). Parents/guardians were also asked about their child's autism diagnosis and developmental history. In the second stage, parents/guardians took part in a telephone interview taking around 40 minutes, about their child's development. This interview consisted of the short version of the Developmental, Diagnostic and Dimensional Interview (3Di-sv; Santosh et al., 2009), a clinical assessment for autism. In the third stage, children were asked to complete an online battery of newly devised language tasks. Alongside these tasks, children also completed the Animal Matrices, a test of nonverbal reasoning (see Chapter Six for details). Families were offered three formats for this assessment session: with myself at home/school, with myself in the Dept of Experimental

Psychology at Oxford University, or in the family's own time at home without a researcher present. If I met with families, I also collected a language sample by engaging the child in around 20 minutes of conversation using a story-book and picture as prompts. All online aspects of the study were supported by Gorilla. Consent from parents/guardians was sought at the beginning of each stage: on an online form in stage one, using an oral consent script in stage two, and an online/paper form in stage three. Children gave assent to take part in stage three, and ongoing assent was checked throughout all in-person sessions.

Measures

The language tasks have been fully described in the previous chapter, so only brief descriptions are presented here.

1. Pragmatic Language Tasks. In the factor analysis, these tasks were set to load on the "pragmatic language" factor.

Implicature Comprehension Test. Participants watch a series of cartoon videos, in each of which two characters produce a short utterance one after the other. Afterwards, participants hear a comprehension question to which they are asked to give a yes-no-don't know response. For 36 items, participants need to process implicature to answer the question. There are also 10 items where the answer is more explicit; these serve as positive control items. As three implicature items were identified as poorly functioning in the data collected from the non-autistic children, these were excluded from analysis. There were two measured variables: sum of implicature items correctly answered (out of 33) and sum of explicit-response items correctly answered (out of 10).

Pragmatic Violations. Participants watch a sequence of 35 videos, in which one character produces a question and a second character answers the question. In 21 videos, the

second character violates a pragmatic norm; in the others, the answer is good. Individuals are asked to rate how good the answer is by placing a mark on an analogue scale with 7 increments. Participant responses are converted to scores between 1 and 8 based on placement of the mark on the line. The measured variable was the total number of pragmatic violations correctly identified. To calculate totals, I identified the score given to the 21 items the participant thought were the worst (as 21 items were devised to include violations). Items with a pragmatic violation given a score below that level were given one point. For items with a pragmatic violation that were given a score at that level, participants were awarded one point divided by the number of items given that score (to allow for tied items). During piloting with adults, three items were identified as poorly functioning, and so were excluded from analysis, as described in Chapter Six. There was a highest possible total of 18.

Social Overtures. Participants hear 24 utterances spoken by a character to a conversational partner. Twelve are social overtures that attempt to engage the partner in a conversation (e.g. “I can’t believe what happened today.”) and twelve are not conversational bids (e.g. “I’m going to have a shower now.”). Participants click yes-no buttons to indicate whether the character wants to have a conversation or not. One item was poorly functioning in the data collected from the non-autistic children, and this was excluded from analysis. There was one measured variable: the sum of items correctly identified as a social overture or not (out of 23).

2. Core Language Tasks. In the factor analysis, these tasks were set to load on the “core language” factor.

Receptive Vocabulary. This includes 40 items in which participants choose which of four pictures is related to a word presented over audio. One item was poorly functioning in

the data collected from the non-autistic children, and this was excluded from analysis. There was one measured variable: the sum of items correctly answered (out of 39).

Receptive Grammar. Participants listen to 50 sentences and decide if they are good or have mistakes/sound odd. There was one measured variable: the sum of items correctly answered (out of 50).

3. Further Measures. The Animal Matrices was administered to all children but data on all other measures listed below were only collected for autistic children.

Animal Matrices. This test of non-verbal reasoning includes 18 matrix puzzles. Each item is a 2x2 matrix, with three boxes showing cartoon pictures of animals and the fourth box is empty. Participants need to deduce which of five options fits in the empty box. Two items were poorly functioning in the data collected from the non-autistic children, and these were excluded from analysis. There was one measured variable: the sum of items correctly answered (out of 16).

Children's Communication Checklist (CCC; Bishop, 1998). This questionnaire includes 70 statements about communication difficulties. Caregivers indicate whether each statement "definitely applies", "somewhat applies" or "does not apply" to their child. There is also a "can't judge" option. The items are grouped into nine subscales: A. speech, B. syntax, C. inappropriate initiation, D. coherence, E. stereotyped conversation, F. use of context, G. rapport, H. social relationships and I. interests. Where caregivers left up to 25% of items incomplete for any single subscale, these were prorated; where more than 25% of items were incomplete, the subscale was recorded as missing for that child. The questionnaire was developed to tap communication features characteristic of children with "pragmatic difficulties" that were not well-captured in existing standardised language tests; subscales C to G were devised for this purpose and contribute to a "pragmatic language composite" (in

this sample, these subscales showed an average correlation of .54 with the composite when excluding that subscale). In line with the theoretical basis of the questionnaire, a "structural language composite" was summed from the first two subscales (which showed a correlation of .63) and an "autistic features composite" from the final two (which showed a correlation of .43). It should be noted that this earlier version of the final published questionnaire was used in this study, as it is integrated into the 3Di-sv.

Pediatric Symptom Checklist (Jellinek et al., 1986). This consists of 35 items about psychosocial difficulties that caregivers report as applying to their child "never", "sometimes" or "often" (scored as 0, 1 or 2 respectively). A total score of 28 or above is taken as indicating psychosocial impairment, with a sensitivity of 95% and specificity of 68% (Jellinek et al., 1988). Factor analysis of the scale validated identification of three subscales that represented attentional (5 items), internalising (5 items) and externalising problems (7 items), with cut-offs of 7, 5 and 7 indicating clinically-significant difficulties in these areas (Gardner et al., 1999).

Short version of the Developmental, Diagnostic and Dimensional Interview (3Di-sv; Santosh et al., 2009). This is a 53-question parent-report interview for assessing and diagnosing autism, which was administered over the phone in approximately 30 to 40 minutes for each child. The original validation study showed 93% agreement in caseness for autism with the Autism Diagnostic Interview-Revised (ADI-R).

Data Analysis

Data were analysed in R (R Core Team, 2017).

In Chapter Six, we saw that the tests used here were reliable in the non-autistic children, and I assessed whether this was also the case among the autistic children using the

same approaches (see that chapter for a fuller description of methods). I report a summary of item-level statistics for the tests, as well reliability coefficients: Cronbach's alpha (with 95% CIs), standard error of measurement, and Revelle's beta. I used item response (IRT) models to test the extent to which all items on a test tap a single latent ability (i.e. are unidimensional), and report the root mean square error of estimation (RMSEA) for each model. Using IRT modelling, I produced item characteristic curves to check item quality, and planned to exclude any items with low, flat curves reflecting chance-level accuracy across the ability spectrum. For this analysis, I used the R packages psych (Revelle, 2017) and mirt (Chalmers, 2012).

Two hypotheses were tested: firstly, that patterns in performance across the language tests would reflect two domains ("core language" and "pragmatic language") rather than just reflecting one general language domain; and secondly, that there would be a group difference in scores on the "pragmatic language" factor, specifically that the autistic children would score lower. As in previous chapters, I began by inspecting the data for normality and multivariate outliers, before applying transformations to variables where necessary, using the Tukey ladder of power transformations implemented with R package rcompanion (Mangiafico, 2018). The sensitivity of the analysis to the data processing was evaluated by comparing results based on transformed and non-transformed data. The hypotheses were tested with confirmatory factor analysis, using R package lavaan (Rosseel, 2012), and semtools (Jorgensen et al., 2019) for a visual representation of the analysis. All factor analyses were run using full information maximum likelihood estimation to account for missing data, and robust standard errors were computed in case of data not being normally distributed. All models were compared using chi-square difference tests with a Satorra-Bentler correction.

For the first hypothesis, I tested whether a two-factor model accounted for variability in language skills better than a one-factor model. Effectively, this extended the analysis in the previous chapter, which covered only the non-autistic children, to the full sample of autistic and non-autistic children. I specified the same two-factor correlated-traits model as previously, with a "core language" factor and a "pragmatic language" factor. Receptive grammar and receptive vocabulary were set to load on the "core language" factor; and implicature comprehension, sensitivity to pragmatic violations, and sensitivity to social overtures were set to load on to the "pragmatic language" factor. The comparison model was a one-factor model in which all five indicators were set to load on the same factor. In both models, year group was included as a covariate at the level of the indicators, to control for age differences in test scores.

Next, I set about testing for group differences in the factor analysis. Firstly, I checked whether the factors were derived from the indicators in the same way across groups – i.e. I tested for measurement invariance, which is a prerequisite for making group comparisons at the factor level. Separate models were defined for the autistic and non-autistic groups, and parameters were progressively constrained to be the same across the two groups. First, I tested for configural invariance (that the same basic structure with five indicators loading on two factors exists across groups), then metric invariance (that indicators load to the same extent on their factors across groups), then scalar invariance (that indicator intercepts are equivalent across groups – i.e. that any group differences in individual test scores are fully accounted for by underlying factors), and finally strict invariance (that there is the same amount of measurement error across groups). Fit was compared as increasingly constrained models were run. After assessing for measurement invariance, I constrained factor means to be the same across groups. This amounted to testing whether autistic and non-autistic children showed any differences in the core and pragmatic language domains.

Finally, I carried out some finer-grained analysis of the Implicature Comprehension Test to enable comparison with the data collected in adults. I tested whether the autistic and non-autistic groups differed in their tendency to select incorrect and 'don't know' compared to correct responses, when also accounting for core language ability. I ran a multinomial mixed effects model using the MCMCglmm package (Hadfield, 2010) with group membership (autistic or non-autistic) and core language ability (factor score from the two-factor model, z-transformed to have a mean of 0 and *SD* of 1) as fixed effects, and participant and item as random effects. The model was fitted using Markov chain Monte Carlo (MCMC) sampling across 3500 samples. It estimated two logit equations, one comparing probability of a correct to incorrect response, and a second comparing a correct to 'don't know' response, and assigned predicted log-odds to each comparison. I report odds ratios with 95% credible intervals for each fixed effect.

Results

Reliability Analysis

In the previous chapter, we saw that the tests showed good reliability in the non-autistic children. In Table 7.1, I supply reliability coefficients that indicate this was also the case for the data collected in the autistic group.

Hypothesis-Testing

See Table 7.2 for descriptive statistics for each of the measures by group (autistic and non-autistic), and Table 7.3 for a correlation matrix. Visual inspection of Table 7.2 suggests that the two groups performed very similarly on all tests, and Table 7.3 indicates that there were small to moderate correlations between tests.

Table 7.1

Reliability Analysis, including Cronbach's alpha and 95% confidence limits, standard error of measurement (SEm), corrected item-total correlations (totals excluding the item), and item-level accuracy. Item-level statistics are summarised as quartiles.

				Item-total correlations			Item-level Accuracy		
Test	Alpha	Alpha 95% CI	SEm	Q1	Q2	Q3	Q1	Q2	Q3
Implicature	.79	.70, .86	2.22	.19	.30	.38	.64	.77	.86
Pragmatic Violations	.79	.70, .86	1.59	.31	.37	.42	.55	.70	.78
Social Overtures	.83	.76, .89	1.34	.30	.39	.51	.86	.90	.92
Vocabulary	.88	.83, .92	1.92	.30	.42	.49	.59	.80	.95
Grammar	.90	.86, .94	1.98	.17	.42	.55	.77	.88	.95
Matrices	.76	.66, .85	1.71	.29	.37	.40	.51	.62	.71

Table 7.2

*Descriptive Statistics by group. Tests included in the factor analysis are in **bold**.*

	Autistic group					Non-autistic group				
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Implicature	54	24.89	4.84	13	32	333	25.22	4.85	8	33
Control items for Implic.	54	8.69	1.72	3	10	331	8.67	1.75	2	10
Pragmatic Violations	52	12.52	3.46	2.50	17.20	324	11.91	3.02	0.48	18
Social Overtures	48	20.54	3.25	9	23	297	19.60	3.57	8	23
Vocabulary	54	30.81	5.54	13	38	341	29.60	4.34	10	39
Grammar	48	41.69	6.26	25	49	291	41.45	6.17	14	49
Matrices	48	10.79	3.50	1	18	316	8.84	3.67	0	16

Table 7.3

Correlations between language tests using pairwise complete observations.

	Pragmatic Violations	Social Overtures	Grammar	Vocab	Matrices	Year Group
Implicature	.29	.41	.32	.39	.32	.25
Pragmatic Violations		.38	.28	.38	.33	.21
Social Overtures			.47	.54	.43	.42
Grammar				.59	.47	.56
Vocabulary					.43	.47
Matrices						.46

First, I tested whether a two-factor model fitted the language test data better than a one-factor model. The two-factor model fitted the data well, CFI = 1, RMSEA = .00, 90% CI [.00, .05], whereas the one-factor model was weaker, CFI = .98, RMSEA = .07, 90% CI [.03, .11]. The difference in fit between the models was significant, $\chi 2$ (1) = 16.53, $p < 0.001$. In the two-factor model, the correlation between the "core language" and "pragmatic language" factors was .75, 95% CI [.62, .89]. This indicates that, while a two-factor model does show a better fit, pragmatic and core language skills, at least as assessed by the tests used here, are closely related. Note that the factor analyses were run with non-transformed and transformed data, to evaluate the sensitivity of the results to the data processing. Without transformations, there were several multivariate outliers, and with transformations, there were none. All factor analysis results were very similar regardless of processing. See Figure 7.1 for a visual representation of the factor model.

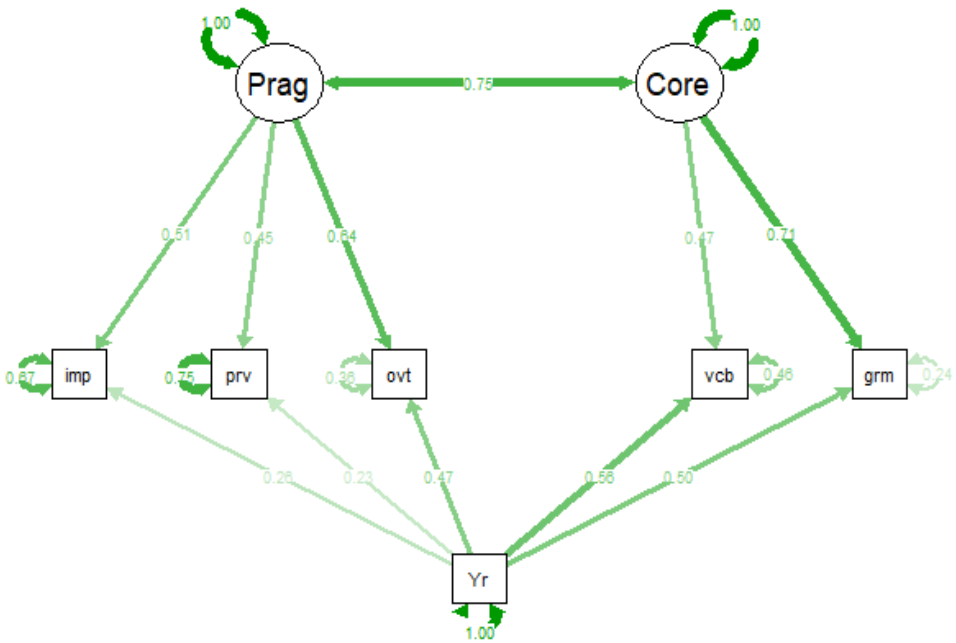


Figure 7.1. A two-factor correlated-traits model of data collected from autistic and non-autistic children. Year group has been included as a covariate at the level of the indicators to control for age effects. Prag = "pragmatic language" factor; Core = "core language" factor; imp = Implicature Comprehension Test; prv = Pragmatic Violations; ovt = Social Overtures; vcb = Receptive Vocabulary; grm = Receptive Grammar; Yr = year group.

Table 7.4

Measurement invariance testing across autistic and non-autistic children using multi-group confirmatory factor analysis.

	Model Df	Model chi-square	Chi-square difference	Df difference	P-value	Comparison	CFI	RMSEA
Configural	8	9.64					1	.03
Metric	11	10.43	0.49	3	.920	Configural	1	.00
Scalar	14	13.04	2.62	3	.454	Metric	1	.00
Strict	19	15.29	4.89	5	.430	Scalar	1	.00

Next, I tested for measurement invariance across the two groups. As shown by the results in Table 7.4, there is strong evidence that the factors were derived in the same way across groups, and so it was valid to make comparisons across the autistic and non-autistic groups. I further constrained the model with strict measurement invariance (i.e. with all indicator-level parameters fixed across groups) to have factor means forced to be the same across groups. Inclusion of these additional constraints did not lead to any deterioration in fit, $\chi 2$ (2) = 1.71, p = .425. The fit statistics of the model with fixed means were strong, CFI = 1, RMSEA = .00, 90% CI [.00, .05]. This indicates, contrary to hypothesis, that autistic and non-autistic children did not differ in their language skills, neither on the "core language" or "pragmatic language" factor. Excluding the 9 children who did not meet 3Di-sv criteria for autism did not affect the results. See Figure 7.2 for a plot of factor scores. As results were unexpected, I examined relationships between factor scores and parent-reported language difficulties to gain an insight into precisely what the factors measured. CCC structural language difficulties correlated with the "core language", r = .58, 95% CI [.37, .73], and "pragmatic language" factors, r = .54, 95% CI [.32, .70]. However, CCC pragmatic language difficulties did not correlate with either "core language", r = .00, or "pragmatic language", r

= .03. It seems, therefore, that the factors were sensitive to co-occurring language impairment but not to the communication features characteristic of autism.

Next, I completed some more detailed analysis of one of the pragmatic tasks, the Implicature Comprehension Test, to allow comparison with adult performance on this test. Here I compared the odds of incorrect and 'don't know' compared to correct responses by group and core language ability. Controlling for core language ability, autistic children were 1.51 [1.05, 2.18] times more likely to choose an incorrect compared to correct response, but not more likely to select 'don't know' than non-autistic children. Meanwhile a *SD* decrease in core language ability was associated with a 1.84 [1.63, 2.11] times greater likelihood of selecting an incorrect and 1.22 [1.01, 1.49] greater likelihood of selecting 'don't know' compared to a correct response.

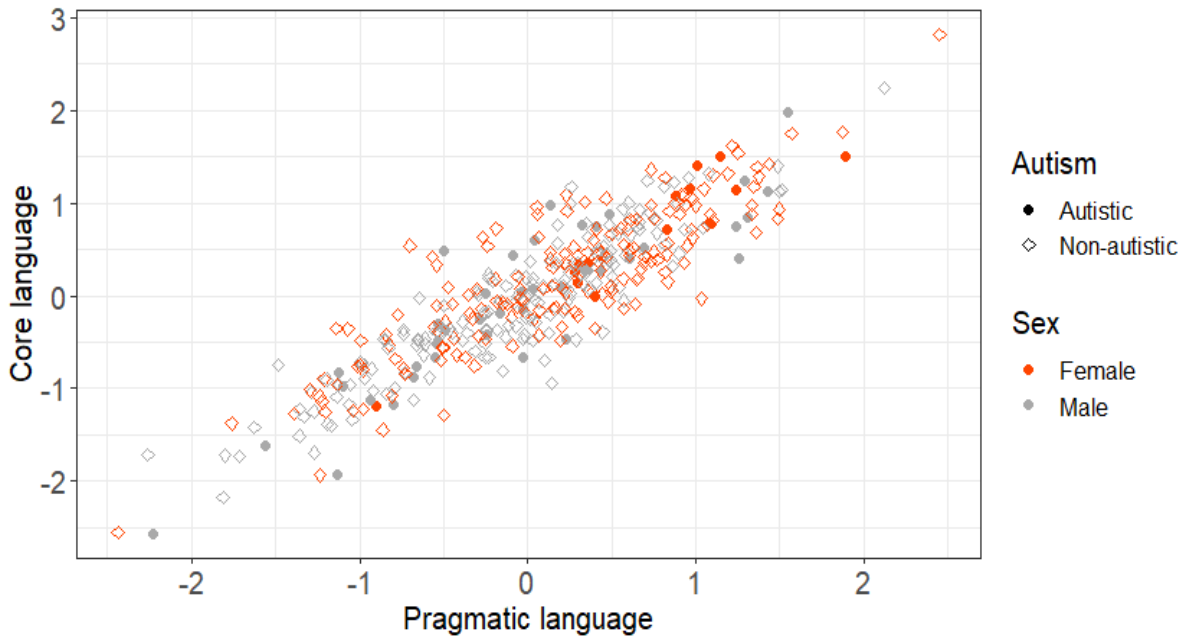


Figure 7.2. A scatter plot showing "pragmatic language" factor scores against "core language" factor scores. Autistic children perform very similarly to non-autistic children, though there is evidence for a sex difference favouring girls.

Sex seemed to play a role in the data. As noted in the analysis of the non-autistic sample in Chapter Six, girls tended to have slightly stronger language skills than boys, and this relationship was also found in the autistic group, but it was more marked. As can be seen in Figure 7.2, the autistic girls (marked in orange dots) seem to have been particularly proficient in their language abilities, though it should be noted that the group of autistic girls was small (n = 14). Table 7.5 displays descriptive statistics by sex among the autistic children for language scores, nonverbal reasoning (on the Animal Matrices), autistic features (3Di-sv scales), psychosocial adjustment (PSC scales), and caregiver-reported communication skills (CCC composites). Beside the substantial differences in language ability, sex appears to play a role in social impairment, reflected in the elevated 3Di-sv social scores among girls, and in general psychosocial difficulties, with caregivers reporting more challenges for girls on the PSC, especially on the internalising scale. However, we should be cautious about these results, as the female sample was small. Otherwise, the overall picture is similar in girls and boys, and it is particularly interesting that girls and boys performed similarly on the Animal Matrices, suggesting that the linguistic strengths girls demonstrated on the test battery were specific to language, rather than reflecting higher overall cognitive ability.

Table 7.5

Descriptive Statistics for autistic children grouped by sex. Cohen's d (and 95% CIs) are presented to quantify sex differences on the measures. Where test developers provide clinical thresholds for indicating impairment, these are after the name of the measure. Measures with a sex difference are in ***bold****.*

Test, clinical threshold		Autistic girls			Autistic boys			Sex differences		
		N	Mean	SD	N	Mean	SD	Cohen's d	Lower 95% CI	Upper 95% CI
Core language	*****	**14**	**0.65**	**0.74**	**41**	**-0.08**	**0.90**	**0.85**	**0.21**	**1.49**
Pragmatic language	*****	**14**	**0.70**	**0.65**	**41**	**-0.06**	**0.84**	**0.95**	**0.31**	**1.60**
Nonverbal reasoning	*	12	9.92	3.00	36	11.08	3.64	-0.33	-1.01	0.34
CCC pragmatic lang.	*	14	115.28	20.99	41	120.55	13.18	-0.34	-0.97	0.28
CCC structural lang.	*	14	63.07	4.57	41	61.58	6.32	0.25	-0.37	0.87
CCC autistic features	*	14	51.17	6.41	40	53.41	7.86	-0.30	-0.92	0.33
PSC total	**28**	**14**	**41.07**	**11.17**	**41**	**34.93**	**8.15**	**0.68**	**0.05**	**1.32**
PSC attention	7	14	6.93	1.94	41	6.88	1.82	0.03	-0.59	0.65
PSC internalising	**5**	**14**	**7.14**	**2.66**	**41**	**5.51**	**2.56**	**0.63**	**0.00**	**1.26**
PSC externalising	5	14	6.50	3.28	41	5.41	2.84	0.37	-0.26	0.99
3Di soc. interaction	**10**	**14**	**17.97**	**5.04**	**41**	**14.15**	**4.80**	**0.79**	**0.15**	**1.42**
3Di communication	8	14	15.47	3.45	41	15.31	3.81	0.04	-0.58	0.67
3Di RRBI	3	14	7.21	2.58	41	7.29	2.65	-0.03	-0.65	0.59

**Lower scores indicate more difficulties on these measures.*
For other measures, a higher score indicates greater difficulty.

Discussion

In this study, I set out to test whether autistic children showed a specific difficulty with pragmatic language processing. The findings reported here did not support that hypothesis, as factor analysis showed no differences between autistic and non-autistic children in "pragmatic language" and "core language" domains. It should be noted, though, that when completing finer-grained analysis of the Implicature Comprehension Test, the pragmatic test that has been the focus of this thesis, there was a group difference, with autistic children more likely to select non-normative responses. This is a similar result as found in autistic adults, and supports the view that picking up implied meanings in conversation is an area of difficulty for people on the spectrum. However, when pragmatics was measured through a combination of three tests, there was no group difference suggesting that the communication difficulties in autism do not arise in all situations and are subject to the context.

The lack of a clear relationship between autism and the "pragmatic language" domain as investigated in this study is surprising, given the intuition among clinicians and researchers that autism involves a core cognitive difference in pragmatics. What might account for the findings? There could be an issue with the sample or the tasks. Alternatively, the theory that autism involves a core impairment in pragmatics might need revision. The least plausible explanation is the nature of the sample, so I will address that first. While inclusion of children in the sample was dependent on a caregiver-reported diagnosis, the vast majority of children did meet criteria for autism when the 3Di-sv (Santosh et al., 2009) was administered as part of the research protocol. In the current sample, the 3Di-sv had a sensitivity for autism of 84% (if we take existing diagnoses as the clinical benchmark) and this is identical to the sensitivity reported previously in a large clinically-referred sample (Slappendel et al., 2016). Also, many

children included in the sample screened as having psychosocial impairment on the PSC (80% scored at 28 or above) and some had co-occurring diagnoses, so it is likely that the average child in the sample had clinically-significant challenges in their day-to-day lives. These points should lead us to be skeptical of any explanation of the results that suggests the sample was unrepresentative of verbally-fluent children on the autism spectrum.

Moving on to consider the nature of the tasks, it is difficult to argue that they did not tap pragmatics. They required the test-taker to interpret particularised conversational implicature, read social intent into conversational overtures, and detect violations of conversational norms. If we take the linguistic definition that pragmatics involves inferring intended meaning from context based on the use of conversational heuristics (such as the expectation that people communicate *relevantly* in the context), it seems that the tasks had good face validity as pragmatic measures. We should also note that they showed good reliability, did not have issues of restricted variance, and were intercorrelated (as measures of a single underlying pragmatic ability). Therefore, if autistic children have pragmatic difficulties with making judgements about conversational language, then we would expect to see these difficulties on the tasks administered here. Of course, a subtle difficulty did emerge in making inferences on the Implicature Comprehension Test; autistic children were more likely to select non-normative interpretations of implied meaning than non-autistic children. However, more broadly across the language battery, we did not see convincing evidence for a domain-level difference in pragmatic processing among children on the spectrum.

Perhaps, the most plausible explanation of the results is that the communication challenges experienced by autistic children do not closely map onto "pragmatics" – at least as defined by linguists and as measurable in the context of structured cognitive tasks. Caregivers reported communication difficulties on behalf of their children on the CCC, with scores on

the pragmatic composite in line with other children with an autism spectrum diagnosis (Bishop & Baird, 2001). However, scores on this composite, which tapped communication features characteristic of autism (such as repetitive speech or difficulties maintaining conversation), showed no relationship with the "pragmatic language" factor extracted from the three tasks devised to tap linguistic pragmatics. Interestingly, the "pragmatic language" factor did show a moderate to large correlation with a structural language composite comprising the syntax and speech subscales of the CCC, indicating that the "pragmatic language" factor was sensitive to general language impairment. However, the lack of a relationship with the CCC pragmatic language composite suggests that the cognitive underpinnings of communication behaviours associated with autism are not linguistic but have a quite different basis.

One thing to note about communication is that it is open and flexible. While a previous turn in conversation provides some constraints over what an appropriate next turn might be, there are invariably many possible turns the next speaker could take and many possible meanings. This openness is inherently difficult to operationalise in the parameters of a cognitive task. While aspects of the structure and content of some conversations may be partly predefined in advance (e.g. if we deliberately meet up to discuss some issue), in many ways a conversation is quite the opposite to a task, as there is no structure or outcome that is externally imposed. Instead, the topic of discussion is an organic thing flexibly co-constructed by the interlocutors. If it is precisely the open nature of communication that autistic people sometimes find challenging, then cognitive tasks are going to struggle to capture those difficulties. Indeed, assessing pragmatic ability to interpret language within the constraints of a cognitive task is likely to cater to the strength of autistic people if having a clear structure and instructions facilitates performance. Perhaps, this explains the discrepancy

between parent-reported difficulties and the absence of clear-cut difficulties on the pragmatic tasks presented in this project.

Reflecting on Inconsistencies between Child and Adult Performance on the Test Batteries

How come autistic adults underperformed on a test battery compared to their non-autistic peers, whereas analysis reported in this chapter indicated no differences in children? In the adult sample, autistic individuals underperformed across several tests representing a "receptive communication" factor (involving judgements of novel communicative stimuli, as reported in Chapter Four). Scores on this "receptive communication" factor predicted self-reported and researcher-rated communication challenges with moderate effect sizes. These results seem starkly different to the findings among children. Firstly, it should be noted that the children and adults completed quite different test batteries. The only test completed by both adults and children was the Implicature Comprehension Test. On this test, autistic adults and children shared a tendency to interpret implied meaning in a non-normative way. In addition, autistic adults showed a considerably greater likelihood than non-autistic adults of choosing a "don't know" option on this test – a tendency that was not observed among autistic children. As autistic adults were typically able to respond correctly when the test was given in a more restricted format without a "don't know" response, this tendency to select "don't know" was interpreted as a cognitive preference for literal communication rather than reflecting an impairment per se. Children did not tend to select "don't know", perhaps because they felt this was a wrong answer and/or had been encouraged during their schooling not to say "don't know" in response to challenges. Putting aside these differing tendencies to select the "don't know" option, there did seem to be consistency between autistic adults and children on this test, in that both groups were more likely to endorse non-normative responses than their non-autistic peers. This reflected some subtle difficulty with interpreting implied

meaning, even controlling for core language skills. Beside the Implicature Comprehension Test, it is difficult to compare performance on the adult and child test batteries as these were devised to tap somewhat different skills. In adults, the focus on linguistic pragmatics was "muddier", including broader demands on social cognition and expressive language as well, with requirements to interpret social situations, awkwardness in conversation, implied meanings, etc., often in free responses (what I call "making normative judgements about novel communicative stimuli" in Chapter Four). The focus in children was narrower: I aimed to isolate linguistic pragmatics without a need for complex expressive language skills, so kept answer formats more constrained with no option for free responses. One of the pragmatic tests given to the children required test-takers to detect social overtures by indicating whether a speaker wanted a conversation or not. In this task, children needed to process social intention, which was also a requirement in many of the tests in the adult battery. However, it is possible that simply requiring a yes-no answer facilitated performance among children on the spectrum, as it was highly structured without a need to generate a response to the overture.

It is notable that, in adults, test scores seemed to account for the communication phenotype in autism to some degree, i.e. scores predicted researcher-rated and self-reported face-to-face communication difficulties. However, in children there was no relationship between test scores and parent-reported communication difficulties. Appealing to the different nature of the test batteries is one possible explanation for this discrepancy, although it does remain the case that the demand on pragmatics was an overlapping feature of the batteries, so we should entertain other explanations too. One suggestion consistent with these differences between children and adults would be that the cognitive underpinnings of communication difficulties in autism vary developmentally. Consider, for instance, if differences in social motivation account for early differences between autistic and non-

autistic people (Chevallier et al., 2012). This might predict difficulties with social interaction in children but might not necessarily impact a person's ability to make judgements about linguistic stimuli in the controlled context of a cognitive task. However, thinking developmentally, if an autistic individual has profited less from social learning experiences perhaps through reduced motivation or a history of social difficulties, they may have underdeveloped receptive communication skills and feel less confident in those skills, meaning that over time a gap between autistic and non-autistic performance is likely to become evident. Therefore, differences in social motivation/attention become embedded at the cognitive level with implications on the phenotype in autism. This is all hypothetical but would account for the differences between adults and children observed in this project. Certainly, the tendency that autistic adults had to select "don't know" on the Implicature Comprehension Test makes sense if we see this tendency as a reaction to knowing that you've sometimes struggled with implied meanings in the past.

Beside these differences, a similarity that did emerge between autistic children and adults was the importance of considering sex in the autism phenotype. In both the child and adult samples, autistic females outperformed autistic males on the language/communication test battery, and the extent of the sex difference seemed greater than in the general population. There is relatively little existing research about sex differences in language skills in autism, though one small study showed that in tests of figurative language, inferencing and word associations, autistic girls tended to outperform autistic boys, but this was in the context of a similar sex difference in non-autistic children (Sturrock et al., 2019). In the current study, it was also the case that parents/carers reported more psychosocial difficulties, especially of an internalising nature, and more impairment in the core social domain on behalf of autistic girls. Whereas the sex differences in language test scores were similar across adults and children, these parent-reported differences were not corroborated by the data collected from

autistic adults. In terms of psychosocial and mental health difficulties, autistic men and women gave very similar results on all self-report measures. In terms of previous research, autistic girls may possibly be more prone to internalising problems, although evidence is sparse and the jury remains out (Rubenstein et al., 2015). There has been greater focus on possible sex differences in the core features of autism, and meta-analysis has previously established no sex difference on measures of social interaction (Wijngaarden-Cremers et al., 2014). Indeed, autistic females might be especially likely to show *less* social impairment on standard diagnostic tools, such as the ADOS-2 (Lai et al., 2011), which was the case in the adult sample reported in this thesis. Bearing in mind these inconsistencies, we ought to be cautious about the sex differences found in the present study, not least because the number of girls included in the sample was small. One possible issue is that there may be different dynamics at play for parents/carers when reporting on difficulties in girls compared to boys. For instance, Sturrock et al. (2020) found greater discrepancies between parent-reported and clinician-observed difficulties in girls compared to boys. This may be due to boys and girls having different difficulties in different contexts. Alternatively, parents/carers may be mindful that expectations of "typical" social and emotional behaviour vary across the genders and that autism has often been less recognised in girls, which may affect reporting.

Conclusion

In this chapter, I presented evidence disconfirming the hypothesis that autistic children show a core impairment in pragmatic language processing. Autistic and non-autistic children showed no domain-level differences in either core or pragmatic language. Similarly, the lack of a relationship between language test scores and parent-reported pragmatic difficulties suggests that these communication features are unlikely to have a linguistic basis in children. However, there is some evidence that autistic children, as with autistic adults, have subtle

difficulties with interpreting implied meanings in conversation. Given that there was no evidence of impairment at the domain-level, we need to be mindful that communication difficulties in autism are likely to depend on the nature of the context in which they occur.

Chapter Eight. Concluding Reflections

Let's begin by recalling the key aims of this thesis. My aims were threefold: to isolate pragmatic aspects of language processing, to test for a dissociation between pragmatic understanding and core language skills, and to assess the extent to which autism is associated with a specific difficulty with pragmatic understanding. More broadly, I intended to contribute to our knowledge of the cognitive basis of the social communication challenges experienced by autistic people.

In Chapter One, I started with the idea that social aspects of language use might be dissociable from core aspects of language, such as grammar and vocabulary skills. Two terms are relevant to this notion of "social aspects of language use": social communication and pragmatics. Social communication involves a complex interaction of diverse socio-cognitive-linguistic skills and is the emergent property of interaction between multiple people. As this project focused on investigating dissociations and the specific nature of difficulties in autism, we are unlikely to get far by focusing on something as general as social communication: clearly, social communication is a broad umbrella that incorporates core language skills, and it is by definition an area of difficulty for people with communication impairments. On the other hand, pragmatic processing is more specific. As a subcomponent of social communication, pragmatic processing involves using context to infer meaning over and above what is explicitly stated. In this thesis, I took pragmatics as my focus. Linguistic theories stipulate that pragmatics is separable from core language processes (Ariel, 2010; Sperber & Wilson, 1986), although empirical studies have typically found that pragmatics and core language are closely related (Matthews et al., 2018). Researchers have argued that autism involves a core impairment in pragmatics (e.g. Baron-Cohen, 1988; Rapin & Dunn, 2003), although empirical reviews focusing on pragmatics have been equivocal about the

extent and basis of pragmatic difficulties in autism (Kalandadze et al., 2018; Loukusa & Moilanen, 2009). These uncertainties are only bolstered by issues around the construct and discriminant validity of pragmatic assessments, as discussed in Chapter One.

In order to operationalise pragmatic processing in a format that could be assessed, I focused on implicature. Implicature is meaning that can only be accessed by integrating an utterance in its broader communication context, and therefore depends on pragmatic processing. While pragmatics might not be sufficient for processing implicature (i.e. vocabulary knowledge might be a limiting factor where language is complex), it is nonetheless the case that pragmatics is necessary for understanding implied meaning, as such meaning cannot be accessed simply by decoding the dictionary meaning of the words. As such, I devised the Implicature Comprehension Test to target pragmatics, keeping vocabulary and grammatical structures simple and test items short to minimise other linguistic and cognitive demands. Furthermore, control items were included, and the test was presented as part of a battery including core language tests to test the specificity of pragmatic difficulties. Studies using the Implicature Comprehension Test indicated that (1) performance on this test was separable from core language ability in children and adults with and without communication challenges and (2) that autistic adults and children underperformed on the task independently of any role played by core language ability. These results suggest this thesis has achieved its three aims: to isolate, dissociate and test for specific difficulties in pragmatics.

On the other hand, as part of this thesis, I also investigated implicature comprehension in the context of broader domains using confirmatory factor analyses, and the results complicate these conclusions, both in terms of the extent to which skills are dissociable and the extent of differences between autistic and non-autistic people. In Chapter Two, I carried

out a factor analysis of several language and communication tests that clustered into two factors: a core language factor (comprising vocabulary and grammaticality judgement tests) and a second factor including implicature comprehension as well as other tests requiring inferences/judgements to be made about various communicative stimuli (including conversations, social situations, a short story, and so on). While this factor structure is in line with predictions based on Relevance Theory (i.e. comprehension depends on twin processes of decoding vocabulary/grammar and making inferences about intended meaning; Sperber & Wilson, 1986), the two domains were quite highly correlated. In children, these pragmatic and core language domains were even more highly correlated (though we should be cautious about directly comparing the adult and child samples as the adult tests were devised to tap a more diverse set of socio-cognitive skills). Previous research has suggested large correlations between pragmatics and other social-linguistic domains (Matthews et al., 2018). My studies support that view, and challenge my earlier assertion that pragmatic processing is dissociable from core language.

However, it is important to stress that throughout this project, it has proven difficult to confidently measure pragmatics at the domain-level. Individual tests showed good reliability and face validity but typically correlations between tests were modest; grammar and vocabulary consistently correlated at around .5 but correlations between tests devised to tap pragmatics and social aspects of communication have never reached that magnitude. This indicates that these tests involve quite substantial task-specific demands and scores are less determined by domain-level ability. Therefore, although the domain-level ability extracted from these tests is closely related to core language ability, a lot of variance in performance on tests with a pragmatic/inferential demand is peculiar to those tasks, so there is a lot of room for people to show a spiky profile of performance across the tasks. As such, the tests are typically dissociable from core language, but also dissociable from tests which purportedly

measure similar social-linguistic skills. Is this because of the artificial nature of tasks, with people developing task-specific strategies that don't necessarily generalise between tasks (or to everyday life)? Or because social aspects of language use are a highly splintered array of skills, with different skills specialised to different contexts? Essentially, these are questions of task validity, and are difficult to answer in the context of the work presented here. The main insight we have regarding external validity is that the tests did predict self-reported and researcher-rated difficulties in face-to-face interaction among autistic adults, with moderate effect sizes, suggesting that the tasks did tap skills relevant to day-to-day communication.

An alternative take on the modest correlations is that perhaps the tasks did not offer a pure measure of communication ability, but might also have reflected certain cognitive processing tendencies too. For example, on the Implicature Comprehension Test, there was a considerably greater difference in scores between autistic and non-autistic people than on other tests, but this was only partly due to ability-based differences and was also influenced by a cognitive preference among autistic people for more definite information, as explored in Chapter Three. Such cognitive tendencies might have been quite specific to individual tasks, so the correlations between tasks are not going to be large. How might this apply to other tests? Certainly, the Frith-Happé Animations and Awkward Dialogues seemed to draw somewhat on domain-level ability (i.e. in judging social/communicative stimuli), but there were perhaps cognitive tendencies that uniquely affected test performance too (so perhaps, a tendency to anthropomorphise or impose a narrative in the Frith-Happé Animations, and a tendency to screen out or overthink awkwardness in the Awkward Dialogues). We need not think of these cognitive tendencies as unwanted confounds in the tests, but as indicators of potentially meaningful biases in how information is processed, and these may have implications for pragmatic processing. Certainly, pragmatics is all about context-dependent interpretations, and clearly these sorts of biases may influence interpretations, but perhaps

only in certain contexts. Indeed, as pragmatics is all about specificity and context, it is perhaps difficult to measure these skills in a way that necessarily generalises across tests.

Another issue that emerges when viewing the Implicature Comprehension Test in the context of other tasks administered in this project is that autistic and non-autistic people performed surprisingly similarly to each other on a range of tasks with a pragmatic/social aspect. The question is why autistic people, who might experience considerable difficulty in face-to-face interaction, show relatively minimal difficulty on pragmatic/social reasoning tasks. In this project, autistic adults performed on average at around the 2nd percentile on the ADOS-2 communication assessment (in relation to norms reported in Brugha et al., 2011) and self-reported considerable communication difficulties, and while they underperformed on the test battery, there was not a large margin between the performance of autistic and non-autistic people on the cognitive tasks. Only on the Implicature Comprehension Test was there a large effect size between autistic and non-autistic adults, and this was probably due in part to differences in cognitive preferences rather than ability, as previously mentioned. A small domain-general difference in making judgements about novel communicative stimuli was apparent across tests, and while this could be equated to a subtle impairment relevant to pragmatics, there was certainly no evidence for a marked pragmatic disability on the tests administered to the autistic adults. For the child samples, the situation was starker, with no differences on the communication tasks but significant day-to-day challenges reported on behalf of the autistic children on the 3Di-sv autism assessment. It has been highlighted before that autistic people may perform on tasks similarly to controls but nonetheless experience considerable day-to-day difficulties related to their autism (e.g. Happé & Frith, 1996; Klin et al., 2003; Sturrock et al., 2020). It is possible that pragmatics (i.e. context-dependent interpretation of language) is not as relevant to communication in autism as has been assumed in the past (e.g. Baird & Norbury, 2016). Certainly, there was no relationship between

pragmatic test scores and parent-reported communication struggles in the child sample. On the other hand, the data collected from the autistic adults did show a moderate relationship between test scores and the overall communication phenotype. In many ways, this inconsistency agrees with previous findings that performance in "theory of mind" tasks also does not often correlate well with performance in naturalistic interactions (see Gernsbacher & Yergeau, 2019 for a review). What might influence this discrepancy?

In a task, there are explicit rules and expectations whereas day-to-day communication is more noisy, open, and typically does not have the same externally imposed structure as a task. In a task, extraneous variables are controlled and a variable of interest is optimised. Recent theories that take a Bayesian and/or predictive coding approach to autistic cognition might offer a useful insight into the discrepancy between performance in tasks and naturalistic interaction. This approach considers the role of expectations, new experience and uncertainty in autism – all of which are relevant to the differences between tasks and naturalistic interaction. The first description of a Bayesian approach to autistic cognition suggested that autistic people might show reduced use of prior expectations to structure and interpret new experience (Pellicano and Burr, 2012). However, it has been noted that autistic people clearly can construct and use informative prior expectations, especially in tasks, where performance might be facilitated compared to in other contexts (e.g. van der Cruys et al., 2017). Instead, van der Cruys et al. (2017) refined the hypothesis to suggest that it is the mismatch between expectations and incoming information that is important, with autistic people being less likely to downregulate the weight they give to this mismatch when it is adaptive to do so (e.g. van der Cruys et al., 2017). Where structure and expectations are explicit and there is relatively little noise, such as in a cognitive task, this is unlikely to bring too many problems. However, in situations where we need to adapt to noise and variability in a context-dependent way, and where many mismatches between expected and actual

experience are irrelevant, the individual is likely to experience difficulties. Free social interaction is such a situation. In the here and now of social interaction, the person is likely to have difficulties screening out noise and irrelevancies to develop a topic of discussion with the interlocutor, and over time they are likely to struggle to develop generalisable skills if they are overly sensitive to noise specific to individual contexts.

Where autistic adults underperformed on the test battery administered in this project, that too can be explained by appealing to this idea that autistic people are highly sensitive to mismatches between expectations and incoming information. Autistic adults performed better than non-autistic adults on a measure of vocabulary knowledge (i.e. where they simply needed to access stored knowledge), however there was some difficulty wherever they needed to respond to novel stimuli for which they couldn't rely on stored knowledge. In these situations, perhaps they gave additional weight to random noise and variation in some of the linguistic/social stimuli – e.g. irrelevant words or extraneous details in a situation. Or perhaps they sometimes found it difficult when the specific stimulus did not seem to correspond fully and exactly to the question in hand – e.g. when meaning was not stated in full. As reported in Chapter Three, autistic people often avoided giving a definite answer about implied meanings, although they usually could give a correct answer when constrained. Within the mini dialogues making up the Implicature Comprehension Test, autistic adults may have experienced a mismatch between their expectation for a full, explicit response and the actual implied response that was given. On the one hand, they seemed to construct an appropriate expectation (including relevant background knowledge) as suggested by the tendency to respond accurately when constrained. However, the mismatch between the explicit response they expected and the implied response they actually heard was strong enough that autistic adults tended to go for a "don't know" option much more frequently than non-autistic adults, whenever that option was given. By contrast, non-autistic participants were perhaps more

likely to “tune down” this error signal and focus on how the implied response might relevantly address the question. The inclusion of a “don’t know” response and items that could legitimately be answered by “don’t know” allowed the Implicature Comprehension Task to more closely approximate day-to-day communication in that it allowed for greater noise and uncertainty. Future research might explore further how greater noise and uncertainty can be incorporated in a graded way within tasks to explore their impact on autistic people’s ability and confidence when processing meaning in naturalistic conversation.

In summary, this thesis has explored whether processing implied meaning dissociates from core language skills and whether people on the autism spectrum experience a specific difficulty with implied meaning. This hypothesis was supported in children and adults, although the situation became more complex when looking at implied meaning in the context of other socio-cognitive-linguistic skills. Domain-level abilities in different aspects of communication tended to show high correlations and domain-level differences between autistic and non-autistic people were subtle at most, even though autistic people were observed and reported themselves as having considerable day-to-day communication difficulties. I suggest that the differences observed in adults reflected domain-level information processing differences in judging the meaning of novel stimuli in the context of uncertainty and a preference for explicit information. However, the subtle scale of any differences observed supports the view that communication difficulties in autism vary according to the context, with an important contextual factor being the level of structure and uncertainty inherent in the situation.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16. https://doi.org/*10.1016/S0885-2014(00)00014-9*

Adams, C. (2002). Practitioner review: The assessment of language pragmatics. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *43*(8), 973–987. https://doi.org/*10.1111/1469-7610.00226*

Adams, C. (2005). Social communication intervention for school-age children: Rationale and description. *Seminars in Speech and Language*, *26*(3), 181–8. https://doi.org/*10.1055/s-2005-917123*

Adams, C., Clarke, E., & Haynes, R. (2009). Inference and sentence comprehension in children with specific or pragmatic language impairments. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *44*(3), 301–318. https://doi.org/*10.1080/13682820802051788*

Altman, N., & Krzywinski, M. (2016). Regression diagnostics. *Nature Methods*, *13*(5), 385–386. https://doi.org/*10.1038/nmeth.3854*

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

Andrés-Roqueta, C., & Katsos, N. (2017). The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders. *Frontiers in Psychology*, *8*, 996. https://doi.org/*10.3389/fpsyg.2017.00996*

Antoniou, K., & Katsos, N. (2017). The effect of childhood multilingualism and bilectalism on implicature understanding. *Applied Psycholinguistics*, *38*(4), 787–833. https://doi.org/*10.1017/S014271641600045X*

Arcara, G., & Bambini, V. (2016). A test for the Assessment of Pragmatic Abilities and Cognitive Substrates (APACS): Normative data and psychometric properties. *Frontiers in Psychology, 7*, 70. https://doi.org/*10.3389/fpsyg.2016.00070*

Ariel, M. (2010). *Defining pragmatics*. Cambridge University Press.

Baird, G., & Norbury, C. F. (2016). Social (pragmatic) communication disorders and autism spectrum disorder. *Archives of Disease in Childhood, 101*(8), 745–751. https://doi.org/*10.1136/archdischild-2014-306944*

Baron-Cohen, S. (1988). Social and pragmatic deficits in autism: Cognitive or affective? *Journal of Autism and Developmental Disorders*, *18*(3), 379–402. https://doi.org/*10.1007/BF02212194*

Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohe (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience* (pp. 3–20). Oxford University Press.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418. https://doi.org/*10.1023/A:1023035012436*

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning

autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/*10.1023/A:1005653411471*

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/*10.18637/jss.v067.i01*

Beitchman, J. H., Brownlie, E. B., & Bao, L. (2014). Age 31 mental health outcomes of childhood language and speech disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(10), 1102–1110. https://doi.org/*10.1016/j.jaac.2014.07.006*

Bernard, S., & Deleau, M. (2007). Conversational perspective-taking and false belief attribution: A longitudinal study. *British Journal of Developmental Psychology*, *25*(3), 443–460. https://doi.org/*10.1348/026151006X171451*

Bishop, D. V. M. (1998). Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(6), 879–891. https://doi.org/*10.1017/S0021963098002832*

Bishop, D. V. M. (2003a). *The Children's Communication Checklist - Second Edition (CCC-2)*. Pearson.

Bishop, D. V. M. (2003b). *The Test for Reception of Grammar - Version 2 (TROG-2)*. Pearson.

Bishop, D. V. M., & Adams, C. (1989). Conversational characteristics of children with semantic-pragmatic disorder. II: What features lead to a judgement of inappropriacy?

British Journal of Disorders of Communication, *24*(3), 241–263. https://doi.org/*10.3109/13682828909019890*

Bishop, D. V. M., & Adams, C. (1992). Comprehension problems in children with specific language impairment: Literal and inferential meaning. *Journal of Speech, Language, and Hearing Research*, *35*(1), 119–29. https://doi.org/*10.1044/jshr.3501.119*

Bishop, D. V. M., & Baird, G. (2001). Parent and teacher report of pragmatic aspects of communication: Use of the children's communication checklist in a clinical setting. *Developmental Medicine and Child Neurology*, 43(12), 809–818. https://doi.org/10.1017/s0012162201001475

Bishop, D. V. M., Whitehouse, A. J. O., & Sharp, M. (2009). *Communication Checklist - Self Report (CC-SR).* Pearson.

Bornkessel-Schlesewsky, I., Krauspenhaar, S., & Schlesewsky, M. (2013). Yes, you can? A speaker's potency to act upon his words orchestrates early neural responses to message-level meaning. *PLoS ONE*, *8*(7), e69173. https://doi.org/*10.1371/journal.pone.0069173*

Botting, N., & Adams, C. (2005). Semantic and inferencing abilities in children with communication disorders. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *40*(1), 49–66. https://doi.org/*10.1080/13682820410001723390*

Botting, N., & Conti-Ramsden, G. (1999). Pragmatic language impairment without autism: The children in question. *Autism*, *3*(4), 371–396. https://doi.org/*10.1177/1362361399003004005*

Botting, N., & Conti-Ramsden, G. (2003). Autism, primary pragmatic difficulties, and specific language impairment: Can we distinguish them using psycholinguistic markers? *Developmental Medicine & Child Neurology*, *45*(8), 515–524. https://doi.org/10.1017/s0012162203000963

Botting, N., Durkin, K., Toseeb, U., Pickles, A., & Conti-Ramsden, G. (2016). Emotional health, support, and self-efficacy in young adults with a history of language impairment. *British Journal of Developmental Psychology*, *34*(4), 538–554. *https://doi.org/10.1111/bjdp.12148*

Brunsdon, V. E. A., & Happé, F. (2014). Exploring the 'fractionation' of autism at the cognitive level. *Autism*, *18*(1), 17–30. https://doi.org/*10.1177/1362361313499456*

Canty, A., & Ripley, B. (2019). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-24. *https://cran.r-project.org/web/packages/boot/boot.pdf*

Carrow-Woolfolk, E. (2017). *Comprehensive Assessment of Spoken Language - Second Edition (CASL-2)*. Western Psychological Services.

Carston, R. (2002). Linguistic meaning, communicated meaning and cognitive pragmatics. *Mind & Language*, *17*(1-2), 127–148. https://doi.org/*10.1111/1468-0017.00192*

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *13*(2), 314–325. https://doi.org/*10.4324/9780203496190*

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/*10.18637/jss.v048.i06*

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences, 16*(4), 231–239. *https://doi.org/10.1016/j.tics.2012.02.007*

Chevallier, C., Parish-Morris, J., Tonge, N., Le, L., Miller, J., & Schultz, R. T. (2015). Susceptibility to the audience effect explains performance gap between children with and without autism in a theory of mind task. *Journal of Experimental Psychology General*, *143*(3), 972–979. https://doi.org/10.1037/a0035483

Chevallier, C., Wilson, D., Happé, F., & Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*(9), 1104–17. https://doi.org/*10.1007/s10803-010-0960-8*

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/*10.1016/S0010-0277(02)00017-3*

Collins, A., Lockton, E., & Adams, C. (2014). Metapragmatic explicitation ability in children with typical language development: Development and validation of a novel clinical assessment. *Journal of Communication Disorders*, *52*, 31–43. https://doi.org/*10.1016/j.jcomdis.2014.07.001*

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64. https://doi.org/*10.1016/j.intell.2014.01.004*

Conti-Ramsden, G., Crutchley, A., & Botting, N. (1997). The extent to which psychometric differentiate subgroups of children with SLI. *Journal of Speech, Language, and Hearing Research*, *40*(4), 765–77. https://doi.org/*10.1044/jslhr.4004.765*

Cordier, R., Munro, N., Wilkes-Gillan, S., Speyer, R., Parsons, L., & Joosten, A. (2019). Applying item response theory (IRT) modeling to an observational measure of childhood pragmatics: The Pragmatics Observational Measure-2. *Frontiers in Psychology*, *10*, 408. https://doi.org/*10.3389/fpsyg.2019.00408*

Cummings, L. (2007). Clinical pragmatics: A field in search of phenomena? *Language and Communication*, *27*(4), 396–432. https://doi.org/*10.1016/j.langcom.2007.06.001*

Cummings, L. (2014). *Pragmatic disorders*. Springer.

Deliens, G., Papastamou, F., Ruytenbeek, N., Geelhand, P., & Kissine, M. (2018). Selective pragmatic impairment in autism spectrum disorder: Indirect requests versus irony. *Journal of Autism and Developmental Disorders*, *48*(9), 2938–2952. https://doi.org/*10.1007/s10803-018-3561-6*

Drummond, K., & Hopper, R. (1993). Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on Language and Social Interaction*, *26*(2), 157–177. https://doi.org/*10.1207/s15327973rlsi2602*

Dunn, L. M., Dunn, L. M., & Whetton, C. (1982). *British Picture Vocabulary Scale*. NFER-Nelson.

Egorova, N., Shtyrov, Y., & Pulvermüller, F. (2013). Early and parallel processing of pragmatic and semantic information in speech acts: Neurophysiological evidence. *Frontiers in Human Neuroscience*, *7*, 86. https://doi.org/*10.3389/fnhum.2013.00086*

Elleseff, T. (2015). Assessing social communication abilities of school-aged children. *Perspectives on School-Based Issues*, *16*(3), 79–86. https://doi.org/*10.1044/sbi16.3.79*

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

Epskamp, S., & Stuber, S. (2017). semPlot: Path diagrams and visual analysis of various SEM packages' output. R package version 1.1. *https://cran.r-project.org/package=semPlot*

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching. structural equation modeling: A second course* (pp. 439–492). Information Age Publishing.

Flora, D. B., Labrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology, 3*, 55. https://doi.org/*10.3389/fpsyg.2012.00055*

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd. ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Frith, U. (1989). *Autism: Explaining the enigma*. Blackwell.

Frith, U., Castelli, F., Frith, C., & Happé, F. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*(8), 1839–1849. *https://doi.org/10.1093/brain/awf189*

Frith, U. and Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind & Language, 14*(1), 82-89. https://doi.org/*10.1111/1468-0017.00100*

Frizelle, P., Thompson, P., Duta, M., & Bishop, D. V. M. (2019). Assessing children's understanding of complex syntax: A comparison of two methods. *Language Learning*, *69*(2), 255–291. https://doi.org/*10.1111/lang.12332*

Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., McInerny, T. K., Wasserman, R. C., Nutting, P., & Chiappetta, L. (1999). The PSC-17: A brief pediatric symptom checklist with psychosocial problem subscales. A report from PROS and ASPN. *Ambulatory Child Health*, *5*(3), 225–236.

Garnham, A., & Oakhill, J. (1992). Discourse processing and text representation from a "mental models" perspective. *Language and Cognitive Processes*, *7*(3), 193–204. https://doi.org/*10.1080/01690969208409384*

Gernsbacher, M. A., Stevenson, J. L., & Dern, S. (2017). Specificity, contexts, and reference groups matter when assessing autistic traits. *PLoS ONE*, *12*(2), e0171931. https://doi.org/10.1371/journal.pone.0171931

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology*, *7*(1), 102–118. https://doi.org/*10.1037/arc0000067*

Geurts, B., Kissine, M., & Tiel, B. van. (2019). Pragmatic reasoning in autism. In K. Morsanyi & R. Byrne (Eds.), *Thinking, reasoning, and decision making in autism*. Routledge.

Geurts, H. M., Hartman, C., Verte, S., Oosterlaan, J., Roeyers, H., & Sergeant, J. A. (2009). Pragmatics fragmented: The factor structure of the Dutch Children's Communication Checklist (CCC). *International Journal of Language & Communication Disorders /*

Royal College of Speech & Language Therapists, *44*(5), 549–574. https://doi.org/*10.1080/13682820802243344*

Gibson, J., Adams, C., Lockton, E., & Green, J. (2013). Social communication disorder outside autism? A diagnostic classification approach to delineating pragmatic language impairment, high functioning autism and specific language impairment. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *54*(11), 1186–1197. https://doi.org/*10.1111/jcpp.12079*

Greenslade, K. J., Utter, E. A., & Landa, R. J. (2018). Predictors of pragmatic communication in school-age siblings of children with ASD and low-risk controls. *Journal of Autism and Developmental Disorders*, *35*(3), 454–452. https://doi.org/*10.1007/s10803-018-3837-x*

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). Academic Press.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22. *http://www.jstatsoft.org/v33/i02/*

Hale, C. M., & Tager-Flusberg, H. (2005). Social communication in children with autism. *Autism*, *9*(2), 157–178. https://doi.org/*10.1177/1362361305051395*

Happé, F.G.E. (1993) Communicative competence and theory of mind in autism: A test of Relevance theory. *Cognition, 48*(2), 101-119. https://doi.org/10.1016/0010-0277(93)90026-r

Happé, F., & Frith, U. (1996). The neuropsychology of autism. *Brain, 119*(4), 1377–1400. *https://doi.org/10.1093/brain/119.4.1377*

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*(400), 1147–1149. https://doi.org/*10.1080/01621459.1987.10478551*

Hobson, H., Westwood, H., Conway, J., McEwen, F., Colvert, E., Catmur, C., Bird, G., & Happé, F. (2018). The impact of alexithymia on autism diagnostic assessments. *PsyArXiv*. https://doi.org/*10.31234/osf.io/8na9q*

Hobson, R. P. (2012). Autism, literal language and concrete thinking: Some developmental considerations. *Metaphor and Symbol*, *27*(1), 4–21. https://doi.org/*10.1080/10926488.2012.638814*

Hochstein, L., Bale, A., & Barner, D. (2018). Scalar implicature in absence of epistemic reasoning? The case of autism spectrum disorder. *Language Learning and Development*, *14*(3), 224–240. https://doi.org/*10.1080/15475441.2017.1343670*

Hollocks, M. J., Lerh, J. W., Magiati, I., Meiser-Stedman, R., & Brugha, T. S. (2019). Anxiety and depression in adults with autism spectrum disorder: A systematic review and meta-analysis. *Psychological Medicine*, *49*(4), 559–572. https://doi.org/*10.1017/S0033291718002283*

Howlin, P. (2013). Social disadvantage and exclusion: Adults with autism lag far behind in employment prospects. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(9), 897–899. https://doi.org/*10.1016/j.jaac.2013.06.010*

Hwang, Y. I., Arnold, S., Srasuebkul, P., & Trollor, J. (2020). Understanding anxiety in adults on the autism spectrum: An investigation of its relationship with intolerance of uncertainty, sensory sensitivities and repetitive behaviours. *Autism*, *24*(2), 411–422. https://doi.org/*10.1177/1362361319868907*

James, R. J. E., Dubey, I., Smith, D., Ropar, D., & Tunney, R. J. (2016). The latent structure of autistic traits: A taxometric, latent class and latent profile analysis of the adult Autism Spectrum Quotient. *Journal of Autism and Developmental Disorders*, *46*(12), 3712–3728. https://doi.org/*10.1007/s10803-016-2897-z*

Jellinek, M. S., Murphy, M., & Burns, B. J. (1986). Brief psychosocial screening in outpatient pediatric practice. *The Journal of Pediatrics, 109*(2), 371–378. https://doi.org/*10.1016/S0022-3476(86)80408-5*

Jellinek, M. S., Murphy, J. M., Robinson, J., Feins, A., Lamb, S., & Fenton, T. (1988). Pediatric Symptom Checklist: screening school-age children for psychosocial dysfunction. *The Journal of Pediatrics, 112*(2), 201–209. https://doi.org/1016/s0022-3476(88)80056-8

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Jolliffe, T., & Baron-Cohen, S. (1999). A test of central coherence theory: Linguistic processing in high-functioning adults with autism or Asperger syndrome: Is local coherence impaired? *Cognition*, *71*(2), 149–185. https://doi.org/*10.1016/S0010-0277(99)00022-0*

Jolliffe, T., & Baron-Cohen, S. (2000). Linguistic processing in high-functioning adults with autism or Asperger's syndrome. Is global coherence impaired? *Psychological Medicine, 30*(5), 1169–87. https://doi.org/*10.1017/S003329179900241X*

Jorgensen, T. D., Pornprasertmanit, S., Miller, P., & Schoemann, A. (2018). simsem: Simulated structural equation modeling. R package version 0.5-14. *https://cran.r-project.org/package=simsem*

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2019). semTools: Useful tools for structural equation modeling. R package version 0.5-2. https://CRAN.R-project.org/package=semTools

Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. *ACL/COLING Workshop on Discourse Relations and Discourse Markers*, 114–120. *http://aclweb.org/anthology/W98-0319*

Kalandadze, T., Norbury, C. F., Nærland, T., & Næss, K. B. (2018). Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism, 22*(2), 99–117. https://doi.org/*10.1177/1362361316668652*

Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81. https://doi.org/*10.1016/j.cognition.2011.02.015*

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*, *20*(4), 442–462. https://doi.org/*10.1177/1362361315588200*

Kerns, C. M., Kendall, P. C., Zickgraf, H., Franklin, M. E., & Miller, J. (2015). Not to be overshadowed or overlooked: Functional impairments associated with comorbid anxiety disorders in youth with ASD. *Behavior Therapy*, *46*(1), 29–39. https://doi.org/*10.1016/j.beth.2014.03.005*

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2003). The enactive mind, or from actions to cognition: Lessons from autism. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *358*(1430), 345–360. https://doi.org/*10.1098/rstb.2002.1202*

Klusek, J., Martin, G. E., & Losh, M. (2014). A comparison of pragmatic language in boys with autism and Fragile X syndrome. *Journal of Speech, Language, and Hearing Research, 57*(5), 1692–1707. https://doi.org/*10.1044/2014_JSLHR-L-13-0064*

Kolenikov, S. & Bollen, K. A. (2012). Testing negative error variances: Is a Heywood case a symptom of misspecification? Sociological Research & Methods, 41*(1), 124–167.* https://doi.org/*10.1177/0049124112442138*

Korkmaz S, Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, *6*(2), 151–162. *https://journal.r-project.org/archive/2014/RJ-2014-031/RJ-2014-031.pdf*

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30, 000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/*10.3758/s13428-012-0210-4*

Lai, M., Baron-Cohen, S. & Buxbaum, J.D. (2015). Understanding autism in the light of sex/gender. *Molecular Autism, 6,* 24. https://doi.org/10.1186/s13229-015-0021-4

Lai, M. C., Lombardo, M. V., Pasco, G., Ruigrok, A. N., Wheelwright, S. J., Sadek, S. A., Chakrabarti, B., MRC AIMS Consortium, & Baron-Cohen, S. (2011). A behavioral comparison of male and female adults with high functioning autism spectrum conditions. *PloS ONE*, *6*(6), e20835. https://doi.org/10.1371/journal.pone.0020835

Lai, M. C., Lombardo, M. V., Ruigrok, A. N. V., Chakrabarti, B., Auyeung, B., Szatmari, P., Happé, F., Baron-Cohen, S., & Consortium, M. R. C. A. (2017). Quantifying and exploring camouflaging in men and women with autism. *Autism*, *21*(6), 690–702. https://doi.org/*10.1177/1362361316671012*

Lam, Y. G., & Yeung, S. S. S. (2012). Towards a convergent account of pragmatic language deficits in children with high-functioning autism: Depicting the phenotype using the pragmatic rating scale. *Research in Autism Spectrum Disorders*, *6*(2), 792–797. https://doi.org/*10.1016/j.rasd.2011.08.004*

Language and Reading Research Consortium, & Muijselaar, M. M. L. (2018). The dimensionality of inference making: Are local and global inferences distinguishable? *Scientific Studies of Reading*, *22*(2), 117–136. https://doi.org/*10.1080/10888438.2017.1371179*

Lavi, A. (2019). *Clinical Assessment of Pragmatics (CAPs)*. Western Psychological Services.

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling developmental language difficulties from school entry into adulthood: Literacy, mental health, and employment outcomes. *Journal of Speech, Language, and Hearing Research*, *52*(6), 1401–1416. https://doi.org/*1092-4388/09/5206-1401*

Leinonen, E., Ryder, N., Ellis, M., & Hammon, C. (2003). Use of context in pragmatic comprehension by specifically language-impaired and control children. *Linguistics*, *41*(2), 407–423. https://doi.org/*10.1515/ling.2003.014*

Lerner, M. D., Mazefsky, C. A., Weber, R. J., & Transue, E. (2018). Verbal ability and psychiatric symptoms in clinically referred inpatient and outpatient youth with ASD. *Journal of Autism and Developmental Disorders*, *48*(11), 3689–3701. https://doi.org/*10.1007/s10803-017-3344-5*

Lever, A. G., & Geurts, H. M. (2016). Psychiatric co-occurring symptoms and disorders in young, middle-aged, and older adults with autism spectrum disorder. *Journal of*

Autism and Developmental Disorders, *46*(6), 1916–1930. https://doi.org/*10.1007/s10803-016-2722-8*

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Leyfer, O. T., Tager-Flusberg, H., Dowd, M., Tomblin, J. B., & Folstein, S. E. (2008). Overlap between autism and specific language impairment: Comparison of Autism Diagnostic Interview and Autism Diagnostic Observation Schedule scores. *Autism Research, 1*(5), 284–296. https://doi.org/*10.1002/aur.43*

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012). *Autism Diagnostic Observation Schedule - Second Edition (ADOS-2)*. Western Psychological Services.

Loukusa, S., Leinonen, E., Jussila, K., Mattila, M. L., Ryder, N., Ebeling, H., & Moilanen, I. (2007). Answering contextually demanding questions: Pragmatic errors produced by children with Asperger syndrome or high-functioning autism. *Journal of Communication Disorders*, *40*(5), 357–381. https://doi.org/*10.1016/j.jcomdis.2006.10.001*

Loukusa, S., Leinonen, E., Kuusikko-Gauffin, S., Jussila, K., Mattila, M. L., Ryder, N., Ebeling, H., & Moilanen, I. (2007). Use of context in pragmatic language comprehension by children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *37*(6), 1049–1059. https://doi.org/*10.1007/s10803-006-0247-2*

Loukusa, S., Mäkinen, L., Kuusikko-Gauffin, S., Ebeling, H., & Leinonen, E. (2018). Assessing social-pragmatic inferencing skills in children with autism spectrum

disorder. *Journal of Communication Disorders, 73*, 91–105. https://doi.org/*10.1016/j.jcomdis.2018.01.006*

Loukusa, S., & Moilanen, I. (2009). Pragmatic inference abilities in individuals with Asperger syndrome or high-functioning autism. A review. *Research in Autism Spectrum Disorders, 3*(4), 890–904. https://doi.org/*10.1016/j.rasd.2009.05.002*

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care*, *46*(3), 266–274. https://doi.org/*10.1097/MLR.0b013e318160d093*

Lucas, R., & Norbury, C. F. (2014). Levels of text comprehension in children with autism spectrum disorders (ASD): The influence of language phenotype. *Journal of Autism and Developmental Disorders*, *44*(11), 2756–2768. https://doi.org/*10.1007/s10803-014-2133-7*

Lugnegård, T., Unenge, M., Hjärthag, F., & Gillberg, C. (2013). Social cognition impairments in Asperger syndrome and schizophrenia. *Schizophrenia Research, 143*(2-3), 277–284. https://doi.org/*10.1016/j.schres.2012.12.001*

Lundström, S., Chang, Z., Kerekes, N., Gumpert, C. H., Råstam, M., Gillberg, C., Lichtenstein, P., & Anckarsäter, H. (2011). Autistic-like traits and their association with mental health problems in two nationwide twin cohorts of children and adults. *Psychological Medicine, 41*(11), 2423–2433. https://doi.org/*10.1017/S0033291711000377*

Magiati, I., Tay, X. Y., & Howlin, P. (2014). Cognitive, language, social and behavioural outcomes in adults with autism spectrum disorders: A systematic review of

longitudinal follow-up studies in adulthood. *Clinical Psychology Review*, *34*(1), 73–86. https://doi.org/*10.1016/j.cpr.2013.11.002*

Mangiafico, S. (2018). rcompanion: Functions to support extension education program evaluation. R package version 1.13.2. *https://cran.r-project.org/package=rcompanion*

Martin, A., Rief, W., Klaiberg, A., & Braehler, E. (2006). Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *General Hospital Psychiatry*, *28*(1), 71–77. https://doi.org/*10.1016/j.genhosppsych.2005.07.003*

Matthews, D., Biney, H., & Abbot-Smith, K. (2018). Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition, and executive functions. *Language Learning and Development*, *14*(3), 186–223. https://doi.org/*10.1080/15475441.2018.1455584*

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/*10.1080/00273171.2014.911075*

Mayes, S. D., Calhoun, S. L., Murray, M. J., & Zahid, J. (2011). Variables associated with anxiety and depression in children with autism. *Journal of Developmental and Physical Disabilities*, *23*(4), 325–337. https://doi.org/*10.1007/s10882-011-9231-7*

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*(3), 381–401. https://doi.org/*10.1016/j.jml.2006.06.006*

Miniscalco, C., Rudling, M., Gillberg, C., & Johnels, J. Å. (2014). Imitation (rather than core language) predicts pragmatic development in young children with ASD: A preliminary

longitudinal study using CDI parental reports. *International Journal of Language and Communication Disorders*, *49*(3), 369–375. https://doi.org/*10.1111/1460-6984.12085*

Minshew, N. J., & Goldstein, G. (1998). Autism as a disorder of complex information processing. *Mental Retardation and Developmental Disabilities Research Reviews*, *4*(2), 129–136. *https://doi.org/10.1002/(SICI)1098-2779(1998)4:2<129::AID-MRDD10>3.0.CO;2-X*

National Autistic Society. (2016). The autism employment gap: Too much information in the workplace. *https://www.autism.org.uk/get-involved/campaign/employment.aspx*

National Autistic Society. (2017, May). Communicating. *National Autistic Society*. *https://www.autism.org.uk/about/communication/communicating.aspx*

National Centre for Social Research, University of Leicester (2017). Adult psychiatric morbidity Survey, 2007 (4th. ed.) [data collection]. https://doi.org/*10.5255/UKDA-SN-6379-2*

Norbury, C. F. (2014). Practitioner review: Social (pragmatic) communication disorder conceptualization, evidence and clinical implications. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *55*(3), 204–216. https://doi.org/*10.1111/jcpp.12154*

Norbury, C. F., & Bishop, D. V. M. (2002). Inferential processing and story recall in children with communication problems: A comparison of specific language impairment, pragmatic language impairment and high-functioning autism. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *37*(3), 227–251. https://doi.org/*10.1080/13682820210136269*

Norbury, C. F., & Nation, K. (2011). Understanding variability in reading comprehension in adolescents with autism spectrum disorders: Interactions with language status and decoding skill. *Scientific Studies of Reading*, *15*(3), 191–210. https://doi.org/*10.1080/10888431003623553*

Norbury, C. F., Nash, M., Baird, G., & Bishop, D. V. M. (2004). Using a parental checklist to identify diagnostic groups in children with communication impairment: A validation of the Children's Communication Checklist-2. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *39*(3), 345–364. https://doi.org/*10.1080/13682820410001654883*

Parsons, L., Cordier, R., Munro, N., Joosten, A., & Speyer, R. (2017). A systematic review of pragmatic language interventions for children with autism spectrum disorder. *PLoS ONE*, *12*(4), e0172242. https://doi.org/*10.1371/journal.pone.0172242*

Pastor-Cerezuela, G., Yllescas, J. C., González-Sala, F., Montagut-Asunción, M., & Fernández-Andrés, M. I. (2018). Comprehension of generalized conversational implicatures by children with and without autism spectrum disorder. *Frontiers in Psychology*, *9*, 272. https://doi.org/*10.3389/fpsyg.2018.00272*

Patten, S. B., & Schopflocher, D. (2009). Longitudinal epidemiology of major depression as assessed by the Brief Patient Health Questionnaire (PHQ-9). *Comprehensive Psychiatry*, *50*(1), 26–33. https://doi.org/*10.1016/j.comppsych.2008.05.012*

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in cognitive sciences*, *16*(10), 504–510. *https://doi.org/10.1016/j.tics.2012.08.009*

Phelps-Gunn, D., & Phelps-Terasaki, T. (2007). *Test of Pragmatic Language - Second Edition* (TOPL-2). PRO-ED.

Phillips, N. (2017). yarrr: A companion to the e-book "YaRrr!: The pirate's guide to R". R package version 0.1.6. *www.thepiratesguidetor.com*

Pickles, A., Anderson, D. K., & Lord, C. (2014). Heterogeneity and plasticity in the development of language: A 17-year follow-up of children referred early for possible autism. *Journal of Child Psychology and Psychiatry, 55*(12), 1354–1362. https://doi.org/*10.1111/jcpp.12269*

Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J. P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *39*(4), 607–18. https://doi.org/*10.1007/s10803-008-0661-8*

Prutting, C. A., & Kirchner, D. M. (1987). A clinical appraisal of the pragmatic aspects of language. *Journal of Speech and Hearing Disorders*, *52*(2), 105–19. https://doi.org/*10.1044/jshd.5202.105*

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rapin, I. (1996). Practitioner review: Developmental language disorders: A clinical update. *Journal of Child Psychology and Psychiatry*, *37*(6), 643–655. https://doi.org/*10.1111/j.1469-7610.1996.tb01456.x*

Rapin, I., & Dunn, M. (2003). Update on the language disorders of individuals on the autistic spectrum. *Brain and Development*, *25*(3), 166–172. https://doi.org/*10.1016/S0387-7604(02)00191-2*

Reichow, B., Salamack, S., Paul, R., Volkmar, F., & Klin, A. (2008). Pragmatic assessment in autism spectrum disorders. *Communication Disorders Quarterly*, *57*(6), 2162–73. https://doi.org/*10.1177/1525740108318697*

Revelle, W. (2017). psych: Procedures for personality and psychological research. R package version 1.7.8. *https://cran.r-project.org/package=psych*

Robinson, E., Koenen, K. C., Mccormick, M. C., Munir, K., Hallett, V., Happé, F., & Ronald, A. (2013). Evidence that autistic traits show the same etiology in the general population and at the quantitative extremes (5%, 2.5% and 1%). *Archives of General Psychiatry*, *68*(11), 1113–1121. https://doi.org/10.1001/archgenpsychiatry.2011.119

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. *http://www.jstatsoft.org/v48/i02/.*

Rubenstein, E., Wiggins, L. D., & Lee, L. (2015). A review of the differences in developmental, psychiatric, and medical endophenotypes between males and females with autism spectrum disorder. *Journal of Developmental and Physical Disabilities*, *27*(1), 119–139. https://doi.org/*10.1007/s10882-014-9397-x.A*

Russell, R. L., & Grizzle, K. L. (2008). Assessing child and adolescent pragmatic language competencies: Toward evidence-based assessments. *Clinical Child and Family Psychology Review*, *11*(1-2), 59–73. https://doi.org/*10.1007/s10567-008-0032-1*

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: A systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism*, *6*, 2. https://doi.org/*10.1186/2040-2392-6-2*

Ryder, N., & Leinonen, E. (2014). Pragmatic language development in language impaired and typically developing children: Incorrect answers in context. *Journal of Psycholinguistic Research*, *43*(1), 45–58. https://doi.org/*10.1007/s10936-013-9238-6*

Ryder, N., Leinonen, E., & Schulz, J. (2008). Cognitive approach to assessing pragmatic language comprehension in children with specific language impairment. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *43*(4), 427–447. https://doi.org/*10.1080/13682820701633207*

Sacco, K., Angeleri, R., Bosco, F. M., Colle, L., Mate, D., & Bara, B. G. (2008). Assessment Battery for Communication - ABaCo: A new instrument for the evaluation of pragmatic abilities. *Journal of Cognitive Science*, *9*, 111–157. https://doi.org/*http://dx.doi.org/10.17791/jcs.2008.9.2.111*

Santosh, P. J., Mandy, W. P., Puura, K., Kaartinen, M., Warrington, R., & Skuse, D. H. (2009). The construction and validation of a short form of the developmental, diagnostic and dimensional interview. *European Child & Adolescent Psychiatry*, *18*(8), 521–524. https://doi.org/*10.1007/s00787-009-0004-0*

Sass, D. A., & Schmitt, T. A. (2013). Testing measurement and structural invariance. In T. Teo (Ed.), *Handbook of quantitative methods for educational research*. SensePublishers.

Schoon, I., Parsons, S., Rush, R., & Law, J. (2010). Children's language ability and psychosocial development: A 29-year follow-up study. *Pediatrics*, *126*(1), e73. https://doi.org/*10.1542/peds.2009-3282*

Schriber, R. A., Robins, R. W., & Solomon, M. (2014). Personality and self-insight in individuals with autism spectrum disorder. *Journal of Personality and Social Psychology, 106*(1), 112–130. https://doi.org/10.1037/a0034950

Scott-Phillips, T. C. (2017). Pragmatics and the aims of language evolution. *Psychonomic Bulletin & Review, 24*(1), 186–189. https://doi.org/*10.3758/s13423-016-1061-2*

Simmons, Elizabeth, S., Paul, R., & Volkmar, F. (2014). Assessing pragmatic language in autism spectrum disorder: The Yale in vivo Pragmatic Protocol. *Journal of Speech, Language, and Hearing Research, 57*(6), 2162–2173. https://doi.org/*10.1044/2014*

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics, 21*(20), 3940–3941. https://doi.org/*10.1093/bioinformatics/bti623*

Sinha, P., Kjelgaard, M. M., Gandhi, T. K., Tsourides, K., & Cardinaux, A. L. (2014). Autism as a disorder of prediction. *PNAS*, *111*(42), 15220–15225. https://doi.org/*10.1073/pnas.1416797111*

Slappendel, G., Mandy, W., Ende, J. V. D., & Verhulst, F. C. (2016). Utility of the 3Di short version for the diagnostic assessment of autism spectrum disorder and compatibility with DSM-5. *Journal of Autism and Developmental Disorders*, *46*(5), 1834–1846. https://doi.org/*10.1007/s10803-016-2713-9*

Sobhani Rad, D. (2014). A review on adult pragmatic assessments. *Iranian Journal of Neurology, 13*(3), 113–118.

Spain, D., & Happé, F. (2019). How to optimise cognitive behaviour therapy (CBT) for people with autism spectrum disorders (ASD): A delphi study. *Journal of Rational-*

Emotive and Cognitive-Behavior Therapy. https://doi.org/10.1007/s10942-019-00335-1

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Blackwell.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language, 17*(1-2), 3–23. https://doi.org/*10.1111/1468-0017.00186*

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA, 282*(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Lo, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine, 166*(10), 1092–1097. https://doi.org/*10.1001/archinte.166.10.1092*

Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (2016). Dynamic testing of analogical reasoning in 5- to 6-year-olds: Multiple-choice versus constructed-response training items. *Journal of Psychoeducational Assessment, 34*(6), 550–565. https://doi.org/*10.1177/0734282915622912*

Sturrock, A., Marsden, A., Adams, C., & Freed, J. (2020). Observational and reported measures of language and pragmatics in young people with autism: A comparison of respondent data and gender profiles. *Journal of Autism and Developmental Disorders, 50*(3), 812–830. https://doi.org/*10.1007/s10803-019-04288-3*

Sturrock, A., Yau, N., Freed, J., & Adams, C. (2019). Speaking the same language? A preliminary investigation, comparing the language and communication skills of

females and males with high-functioning autism. *Journal of Autism and Developmental Disorders*. https://doi.org/*10.1007/s10803-019-03920-6*

Sukhodolsky, D. G., Scahill, L., Gadow, K. D., Arnold, L. E., Aman, M. G., Mcdougle, C. J., McCracken, J. T., Tierney, E., White, S. W., Lecavalier, L., & Vitiello, B. (2008). Parent-rated anxiety symptoms in children with pervasive developmental disorders: Frequency and association with core autism symptoms and cognitive functioning. *Journal of Abnormal Child Psychology*, *36*(1), 117–128. https://doi.org/*10.1007/s10802-007-9165-9*

Surian, L., Baron-Cohen, S., & Van Der Lely, H. (1996). Are children with autism deaf to Gricean maxims? *Cognitive Neuropsychiatry*, *1*(1), 55–72. https://doi.org/*10.1080/135468096396703*

Swineford, L. B., Thurm, A., Baird, G., Wetherby, A. M., & Swedo, S. (2014). Social (pragmatic) communication disorder: A research review of this new DSM-5 diagnostic category. *Journal of Neurodevelopmental Disorders*, *6*(1), 41. https://doi.org/*10.1186/1866-1955-6-41*

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*(4), 401–413. https://doi.org/*10.1075/prag.10.4.01tom*

Trouvain, J., & Truong, K. P. (2012). Comparing non-verbal vocalisations in conversational speech corpora. In L. Devillers, B. Schuller, A. Batliner, P. Rosso, E. Douglas-Cowie, R. Cowie, & C. Pelachaud (Eds.), *Proceedings of the 4th international workshop on corpora for research on emotion sentiment and social signals (es3 2012)* (pp. 36–39). European Language Resources Association.

van Berkum, J. J. A., Brink, D. V. D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591. https://doi.org/*10.1162/jocn.2008.20054*

van de Cruys, S., van der Hallen, R., & Wagemans, J. (2017). Disentangling signal and noise in autism spectrum disorder. *Brain and cognition, 112*, 78–83. *https://doi.org/10.1016/j.bandc.2016.08.004*

van Wijngaarden-Cremers, P. J. M., van Eeten, E., Groen, W. B., van Deurzen, P. A., Oosterling, I. J., & van der Gaag, R. J. (2014). Gender and age differences in the core triad of impairments in autism spectrum disorders: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders*, *44*(3), 627–635. https://doi.org/*10.1007/s10803-013-1913-9*

Vasa, R. A., Kreiser, N. L., Keefer, A., Singh, V., & Mostofsky, S. H. (2018). Relationships between autism spectrum disorder and intolerance of uncertainty. *Autism Research*, *11*(4), 636–644. https://doi.org/*10.1002/aur.1916*

Volden, J., Coolican, J., Garon, N., White, J., & Bryson, S. (2009). Brief report: Pragmatic language in autism spectrum disorder: Relationships to measures of ability and disability. *Journal of Autism and Developmental Disorders*, *39*(2), 388–393. https://doi.org/*10.1007/s10803-008-0618-y*

Volden, J., & Phillips, L. (2010). Measuring pragmatic language in speakers with autism spectrum disorders: Comparing the Children's Communication Checklist-2 and the Test of Pragmatic Language. *American Journal of Speech-Language Pathology*, *19*(3), 204–12. https://doi.org/*10.1044/1058-0360(2010/09-0011)*

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. The Psychological Corporation.

Whyte, E. M., & Nelson, K. E. (2015). Trajectories of pragmatic and nonliteral language development in children with autism spectrum disorders. *Journal of Communication Disorders, 54*, 2–14. https://doi.org/*10.1016/j.jcomdis.2015.01.001*

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wigham, S., Rodgers, J., & South, M. (2015). The interplay between sensory processing abnormalities, intolerance of uncertainty, anxiety and restricted and repetitive behaviours in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 45*(4), 943–952. https://doi.org/*10.1007/s10803-014-2248-x*

Williams, D. L., Minshew, N. J., & Goldstein, G. (2015). Further understanding of complex information processing in verbal adolescents and adults with autism spectrum disorders. *Autism*, *19*(7), 859–867. https://doi.org/*10.1177/1362361315586171*

Wilson, A. C., & Bishop, D. V. M. (2019). "If you catch my drift…": Ability to infer implied meaning is distinct from vocabulary and grammar skills [version 3; peer review: 3 approved]. *Wellcome Open Research, 4*(68). https://doi.org/10.12688/wellcomeopenres.15210.3

Wilson, A. C., & Bishop, D. V. M. (2020). "Second guessing yourself all the time about what they really mean…": Cognitive differences between autistic and non-autistic adults in understanding implied meaning. *Autism Research.* https://doi.org/10.1002/aur.2345

Yew, S. G. K., & Kearney, R. O. (2013). Emotional and behavioural outcomes later in childhood and adolescence for children with specific language impairments: Meta-

analyses of controlled prospective studies. *Journal of Child Psychology & Psychiatry*, *54*(5), 516–524. https://doi.org/*10.1111/jcpp.12009*

Young, E. C., Diehl, J. J., Morris, D., Hyman, S. L., & Bennetto, L. (2005). The use of two language tests to identify pragmatic language problems in children with autism spectrum disorders. *Language Speech and Hearing Services in Schools*, *36*(1), 62–72. https://doi.org/*10.1044/0161-1461(2005/006)*

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/*10.1007/s11336-003-0974-7*

Appendix A. Existing Tests with a Pragmatic Component (Supplementary Material for Chapter One)

As noted in Chapter One, I collated all existing tests devised to have a pragmatic component, or which are marketed as involving pragmatics in some form. Tests were collated from five review papers as indicated in the chapter (Adams, 2002; Elleseff, 2015; Norbury, 2014; Russell & Grizzle, 2008; Sobhani Rad, 2014). Tests have been excluded if (1) they were devised purely for preschool children, (2) they were unpublished or out of publication, or (3) details were not openly available regarding them.

Measure	Authors	Age	Description
Checklists completed by families, teachers and individuals; norm-referenced			
Children's Communication Checklist-2 (CCC-2)	Bishop, 2003	4 – 17	Family/teacher rates frequency of 70 communication behaviours on a 4-point scale. Ten subscales arranged as language content/form (speech, syntax, semantics, coherence), pragmatics (inappropriate initiation, use of context, scripted language, nonverbal communication), and autistic features (social relations, interests).
Communication Checklist-Self Report (CC-SR)	Bishop, Sharp, & Whitehouse, 2009	10 – Adult	Informant rates frequency of 70 communication behaviours on a 4-point scale. Three subscales: (1) structural language, (2) pragmatics, and (3) social engagement.
Communication Checklist-Adult (CC-A)	Whitehouse & Bishop, 2009	Adult	Individual rates frequency of 70 communication behaviours on a 4-point scale. Three subscales: (1) structural language, (2) pragmatics, and (3) social engagement.
Clinical Evaluation of Language Fundamentals, Fifth Edition (CELF-5) Pragmatics Profile	Wiig & Secord, 2014	5 – 21	Family/teacher rates the frequency of 50 communication behaviours on a 4-point scale. Three subscales: (1) rituals and conversational skills, (2) asking for, giving, and responding to information, and (3) nonverbal communication skills.

Pragmatic Language Skills Inventory (PLSI)	Gilliam & Miller, 2006	5 – 12	Teacher rates the effectiveness of a child's overall communication skills in comparison with other children on a 9-point scale. There are 45 items split into three subscales: (1) personal interaction skills, (2) social interaction skills, and (3) classroom interaction skills.
Pragmatic Language Observation Scale (PLOS)	Newcomer & Hammil, 2009	8 – 17	Teacher rates the effectiveness of a child's overall communication skills in the classroom in comparison with other children on a 5-point scale. There is one scale comprising 30 items.
Checklists completed by families, teachers and individuals; not norm-referenced			
Adolescent Pragmatics Screening Scale (APSS)	Brice, 1992	11 – 18	Teacher rates appropriacy of 36 communication behaviours on a 5-point scale. Six subscales: (1) directive speech acts, (2) self-expression, (3) greetings, (4) maintenance of conversation, (5) active listening, and (6) ability to cue topic shifts.
Teacher Assessment of Student Communicative Competence (TASCC)	Smith, McCauley, & Guitar, 2000	6 – 11	Teacher rates frequency of 50 communication behaviours on 5-point scale. Five subscales: (1) approach/avoidance attitude, (2) intelligibility, (3) clarification/repair, (4) appropriateness of speech, and (5) nonverbal communication.
Conversational Effectiveness Profile	Kowalski, 2005	School-age	Teacher rates communication behaviour for appropriacy on a 3-point scale. There are six subdomains: (1) social interaction, (2) social communication, (3) academic communication, (4) nonverbal communication, (5) perspective taking, and (6) social-emotional.
Interviews and mixed method approaches; descriptive and not norm-referenced			
The Pragmatics Profile of Everyday Communication in Children, Revised Edition; The Pragmatics	Dewart & Summers, 1995; 1996	Preschool – Adult	Three structured interviews conducted with families of pre-schoolers, school-age children (approx. 5-10), and older children/adults; a self-report version also exists for older children/adults. The aim is to gain a qualitative description of the individual's communication. Each

Profile of Everyday Communication in Adults			interview is divided into four sections: (1) communicative functions, (2) response to communication, (3) interaction and conversation, and (4) contextual variation.
Functional Communication Profile-Revised (FCP-R)	Kleiman, 2003	3 – Adult	Examiner indicates whether individual shows normal or impaired behaviour/communication over a number of items in eleven domains: (1) motor/sensory, (2) behaviour, (3) attention, (4) receptive language, (5) expressive language, (6) pragmatics/social language, (7) speech, (8) voice, (9) oral, (10) fluency, and (11) nonverbal communication. Examiner is free to use direct assessment, informal observation and self/informant report to complete checklist.
Communication Activities of Daily Living, Third Edition (CADL-3)	Holland, Wozniak, & Fromm, 2018	Adult	The individual takes part in a structured interview with an examiner involving questions and simulated situations. Their responses are rated on a 3-point scale for appropriacy. There are 50 items across seven areas: (1) r*eading, writing and using numbers, (2) social Interactions, (3) contextual communication, (4) nonverbal communication, (5) sequential relationships, (6) humour, metaphor and absurdity, and (7) internet basics.*
American Speech-Language-Hearing Association Functional Assessment of Communication Skills for Adults, revised edition (ASHA FACS)	Frattali, Thompson, Holland, Wohl, Wenck, Slater, & Paul, 2017	Adult	The individual's skills are rated across 43 items divided into four domains: (1) social communication, (2) communication of basic needs, (3) reading, writing and number concepts, and (4) daily planning. The independence with which the person can carry out each item is rated on a 7-point scale, and each of the four domains is rated on a 5-point scale for adequacy, appropriateness, promptness, and communicative sharing. The ratings may be based on direct assessment, informal observation or informant/self report.
Rating scales completed by clinician/researcher following observation; not norm-referenced			

Clinical Evaluation of Language Fundamentals (CELF-5) Pragmatic Activities Checklist	Wiig & Secord, 2014	5 – 21	Examiner completes 32-item checklist after carrying out three simple communication activities with the examinee as a springboard for conversation. This is a criterion-referenced measure, where a positive score on at least 9 items merits further investigation. The checklist targets two areas: (1) atypical nonverbal pragmatics (e.g. gaze, gesture, etc.) and (2) atypical verbal pragmatics (e.g. relevance, quantity of communication, etc.).
Targeted Observation of Pragmatics in Children's Conversations (TOPICC)	Adams, Gaile, Freed, & Lockton, 2010	6 – 11	Examiner takes part in a conversation with the child, and rates the extent to which certain communication behaviours impact on the interaction; impact is rated on a 4-point scale. Ratings are given to 17 conversational parameters, divided into 7 subdomains: (1) reciprocity, (2) turn-taking, (3) taking account of listener, (4) verbosity, (5) topic management, (6) discourse style, and (7) response problems.
Analysis of Language Impaired Children's Conversation (ALICC)	Bishop & Adams, 1989	4 – Adult	A conversation with the individual is transcribed and coded at the utterance level. Codes are converted into quantitative indices for: discourse participation, conversational dominance, assertiveness, verbosity, responsiveness, and meshing (quality of fit between conversational turns).
Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)	Lord, Rutter, DiLavore, Risi, Gotham, Bishop, Luyster, & Guthrie, 2012	1 – Adult	A sequence of structured communication tasks and free conversations are facilitated by an examiner, and performance is rated against a number of parameters relating to communication, social interaction, imagination and restricted and repetitive behaviour. This gives quantitative indices, and there are cut-offs for communication difficulty consistent with an autism diagnosis.
Yale Pragmatic Protocol	Schoen & Paul, 2009	9 – 17	A semi-structured conversation task is administered by a professional involving 23 probes given by the examiner

			covering five conversational subdomains: (1) discourse management, (2) communicative function, (3) conversational repair, (4) presupposition, and (5) register variation.
Pragmatic Observational Measure (POM)	Cordier, Munro, Wilkes-Gillan, Speyer, & Pearce, 2014	5 – 11	The examiner observes the child interacting with a peer in a play situation, and the consistency with which 27 communication behaviours are observed is rated on a 4-point scale. There are five subscales: (1) introduction/responsiveness, (2) nonverbal communication, (3) social-emotional atonement, (4) executive function, and (5) negotiation.
Pragmatic Rating Scale (PRS)	Landa, Piven, Wzorek, Gayle, Chase, & Folstein 1992; Landa, 2011	School-age – Adult	In the original version of the checklist, a conversation with the examiner was rated against 19 communication behaviours on a 3-point scale for appropriacy. The school-age version was expanded to 34 items, divided into four subscales: (1) presupposition, (2) discourse management, (3) suprasegmental characteristics, and (4) nonverbal communication.
Pragmatic Protocol	Prutting & Kirchner, 1987	5 – Adult	30 communication parameters are rated based on a one-off observation by an examiner. Communication parameters are subdivided into verbal (speech acts, topic, turn-taking, lexical selection/use across speech acts, stylistic variation), nonverbal, and paralinguistic aspects. Items are judged as appropriate, inappropriate or not observed.
Profile of communicative appropriateness	Penn, 1985	Adult	The appropriacy of 45 communication behaviours is rated on a 5-point scale by an examiner based on an elicited conversation sample. Six subdomains rated: (1) response to interlocutor, (2) topic control, (3) cohesion, (4) fluency, (5) sensitivity to sociolinguistic context, and (6) nonverbal communication.

Social thinking dynamic assessment protocol	Winner, 2008	8 – Adult	This is an interview-style assessment with several tasks: examiner and examinee interview each other; then examinee interprets some pictures, sequences other pictures into a story, makes judgements about examiner's gaze, and interprets social scenarios. Examiner completes ratings for each task to give a qualitative measure of individual's communication.
Formal norm-referenced language tests, with a focus on pragmatics or described as having a pragmatic aspect/subcomponent			
Test of Pragmatic Language, Second Edition (TOPL-2)	Phelps-Terasaki & Phelps-Gunn, 2007	6 – 18	Individual is asked a comprehension question or to produce socially appropriate responses across 43-items involving situational and narrative prompts. Organised around six aspects of communication: physical setting, audience, topic, purpose (speech acts), visual-gestural cues, and abstraction.
Comprehensive Assessment of Spoken Language, Second Edition (CASL-2)	Carrow-Woodfolk, 2017	3 – 21	Includes a pragmatic domain, comprising one subtest: individual is presented with social scenarios, and must judge the pragmatic appropriacy of a statement or identify an appropriate thing to say (from age 3). Also includes a supralinguistic domain relevant to pragmatics. Four subtests: (1) nonliteral language (from age 7), (2) double meanings (from age 9), (3) use of context to infer meaning of an unfamiliar word (from age 7), and (4) use of world knowledge to infer full meaning (from age 3).
Diagnostic Evaluation of Language Variation – Norm-Referenced Edition (DELV-NR)	Seymour, Ropier, & de Villiers, 2005	4 – 9	Includes a pragmatic domain, assessed by three subtests: (1) producing a contextually appropriate response to a situational probe, (2) narrative generation, and (3) formulating questions to seek specific information.
Clinical Evaluation of Language Fundamentals Fifth Edition (CELF-5) Metalinguistics (superseding Test of	Wiig & Secord, 2014	9 – 18	Four subtests: (1) making inferences, (2) manipulating language given a context, (3) understanding multiple meanings; and understanding figurative language.

Language Competence, TLC)			
Assessment of Comprehension and Expression 6 – 11 (ACE 6 – 11)	Adams, Coke, Crutchley, Hesketh, & Reeves, 2001	6 – 11	General language battery involving five core subtests: (1) sentence comprehension, (2) inferential comprehension, (3) naming, (4) syntactic formulation and (5) semantic decision. In addition, there are two supplementary subtests: (1) nonliteral language, and (2) narrative retelling. Inferential comprehension and the two supplementary subtests are relevant to pragmatics.
Social Language Development Test– Elementary: Normative Update (SLDT – E:NU)	Bowers, Huisingh, & LoGiudice, 2016	6 – 11	Four subtests: (1) drawing mental-state inferences using visual cues, (2) suggesting solutions to peer conflicts, (3) forming multiple Interpretations of a picture, and (4) supporting peers with "white lies".
(SLDT-A: NU) Social Language Development Test–Adolescent: Normative Update	Bowers, Huisingh, & LoGiudice, 2017	12 – 17	Five subtests: (1) drawing mental-state inferences using visual cues, (2) interpreting social language such as idioms and nonverbal communication, (3) offering solutions to social problems, (4) supporting others in social interactions, and (5) interpreting ironic statements.
Test of Problem Solving – 3:Elementary (TOPS-3:E)	Bowers, Huisingh, & LoGiudice, 2005	6 – 12	Six subtests: (1) inferences, (2) sequencing, (3) explaining why certain events wouldn't/shouldn't occur, (4) problem solving, (5) predicting, and (6) determining causes.
Test of Problem Solving – 2:Adolescent (TOPS-2:A)	Bowers, Huisingh, & LoGiudice, 2007	12 – 17	Five subtests: (1) inferences, (2) determining solutions, (3) problem solving, (4) interpreting perspectives, and (5) transferring insights.
Social Emotional Evaluation, Second Edition (SEE)	Wiig, 2008	6 – 12	Five subtests: (1) recalling facial expressions, (2) identifying common emotions, (3) recognising emotional reactions, (4) understanding social gaffes, and (5) understanding conflicting messages.
Test of Narrative Language, Second Edition	Gillam & Pearson, 2017	4 – 15	Six subtests, comprising three comprehension tasks and three narration tasks. The comprehension tasks involve literal and inferential questions based on narratives told by the examiner. For the production tasks, the examinee

			retells a story told by the examiner, tells the story shown by a set of pictures, and makes up a story from a visual prompt.
Bus story	Renfrew, 1995	3 – 8	Individual retells a story told by the examiner using a picture book. Scores can be extracted for story content, sentence length, complexity and level of cueing required.
Expression, Reception and Recall of Narrative Instrument (ERRNI)	Bishop, 2003	4 – Adult	Individual relates a story based on a picture-book, recalls it after a delay and responds to comprehension questions.
Language tests without norms that have been identified as having a pragmatic aspect			
Discourse Comprehension Test, Revised Edition	Brookshire & Nicholas, 1997	Adult	10 short stories presented in spoken format followed by yes/no questions assessing understanding of directly stated and implied details and main ideas.
Right Hemisphere Language Battery, Second Edition (RHLB)	Bryan, 1994	Adult	General language battery consisting of seven subtests, most of which have relevance to pragmatics: (1) lexical semantic comprehension, (2) spoken metaphor appreciation, (3) written metaphor appreciation, (4) verbal humour appreciation, (5) comprehension of inference, (6) production of emphatic stress, and (7) comprehensive discourse analysis.
Strong Narrative Assessment Procedure	Strong, 1998	7 – 12	Individual hears a story and retells it and answers comprehension questions. The story retelling is scored for content, length and retelling errors.
The Listening Skills Test (LIST)	Lloyd, Peers, & Foster, 2001	3 – 7	Four subtests: (1) selecting visual referents that go with a statement, (2) judging the truth status of a statement in relation to a picture, (3) mapping a route based on spoken instructions, and (4) monitoring statements for inconsistent, incorrect or incomplete information.
Understanding Ambiguity	Rinaldi, 1996	8 – 13	Assesses ambiguity introduced by words with multiple meanings and inconsistency between verbal message and emotional expression

Appendix B. Instructions for Adult Assessments and Examples of Test Items

In this appendix, I provide greater detail about the assessments reported in Chapters Two to Four. Full assessments are not included here to safeguard materials for future research and clinical uses. However, I am happy to share assessments and associated scoring with researchers wishing to replicate and/or extend the work reported in this thesis; please see *https://doi.org/10.17605/OSF.IO/XN48E* for information on gaining access.

Implicature Comprehension Test

Instructions are shown on the first page:

In this task, you will see some short videos. Each video shows two people, Tom and Sally, talking. A robot will also appear in the videos. This robot is trying to learn what people mean when they speak.

When the robot is not sure what Tom and Sally mean, it will ask you a question. You can answer the question with Yes or No. On your keyboard, press the C key for Yes and the N key for No. Sometimes you might not be able to answer the question. If that happens, press the Spacebar for Don't Know.

In the videos, you will see a question mark and hear a beep after the robot has asked a question. Wait for these before you press a button for your answer. You will have up to 5 seconds to answer.

There are now some practice videos. Press Continue to see them.

For each of the three practice videos, participants click the video to start it. Each video consists of a short conversational exchange followed by a comprehension question. After the video, participants see an answer page showing a question mark and the following text:

On the keyboard, press the C key for Yes, the N key for No, or the Spacebar for Don't Know.

Scripts for the practice videos are shown in the item list below. After the three practice videos, there is a transition page, with the following text:

You can now try the real videos. The whole task will take about ten minutes.

Don't forget the C key is Yes, N is No, and the Spacebar is Don't Know. You will have up to 5 seconds to give each response.

Press Continue to start.

As with the practice videos, the test videos are started by manually clicking them. For examples of the test videos, see the list below. Each video consists of a conversational exchange and comprehension question, and after the video, they see an answer page. Participants' responses to the comprehension question are timed from onset of this answer screen, which shows a question mark and the following text:

On the keyboard, press the C key for Yes, the N key for No, or the Spacebar for Don't Know.

Stimulus List: Practice Items

Character 1: Have you fed the dogs yet?
Character 2: I gave them both their breakfast.
Do you think he has fed the dogs yet?

Character 1: Have you finished packing?
Character 2: Not yet, I'll finish soon.
Do you think she has finished packing?

Character 1: Have we packed the gloves?

Character 2: I don't know about the gloves.

Do you think they have packed the gloves?

Stimulus List: Sample of Test Items

The item type to which each item belongs is indicated; item types include implicature, explicit-response and "don't know".

Character 1: Did you see the policemen earlier on?

Character 2: I saw them standing on the platform.

Do you think she saw the policemen?

Explicit-response

Character 1: Did the police speak to anyone else?

Character 2: I wasn't watching them much.

Do you think the police spoke to anyone else?

Don't know

Character 1: Could you hear what the police said?

Character 2: There were lots of trains going past.

Do you think he heard what the police said?

Implicature

Character 1: Does the temperature need turning up in here?

Character 2: I just took off my jumper.

Do you think the temperature needs turning up?

Implicature

Character 1: Have you tried the new tea?

Character 2: I needed a hot drink after skiing.

Do you think he tried the new tea?

Implicature

Character 1: Did you buy the milk this morning?

Character 2: I went out for a walk earlier on.

Do you think she bought the milk this morning?

Implicature

Test of Fillers/Backchannels

Instructions are presented on the first screen:

Two characters will talk to each other in some videos. In each video, the character on the left will say a sentence. Afterwards, the character on the right will say a word or a sound. Then the video will suddenly stop.

You need to imagine how the conversation would probably continue. Who would be likely to speak? Maybe the second speaker would just keep speaking, or maybe the first speaker would talk again.

Your job is to decide who will <u>probably</u> speak by clicking the button showing the right face. If it's very difficult to know, click the question mark.

Press Continue to see a practice video.

Participants are presented with a practice video on the next page. They click the video to start it. Once the video has finished, they see the answer screen. This shows pictures of the two speakers, as well as a question mark. At the top of the screen, they see the question *Who is likely to be speaking now?*

After clicking their answer, participants see a feedback page:

In that video, it is likely that the man with blond hair would be speaking now.

Remember that the man with brown hair first said, "She told me about it last week." Then the man with blond hair said, "She was." It is likely that the man with blond hair would continue speaking to finish his sentence.

To start the rest of the task, press Continue.

Next, participants see a transition screen for 7 seconds. The screen shows the following text:

The task is about to begin. Please note that you will hear many of the words and sentences more than once in this task.

Participants are then presented with 40 trials. On the first screen of each trial, they click the video to play the short conversational fragment. Then the answer page is presented. On this page, participants see pictures of the two speakers, as well as a question mark, and they click who they think will speak. At the top of the screen, they see the question *Who is likely to be speaking now?* There is no time limit.

Stimulus List: Sample of Test Items

Character 1: I said we'd meet near the shops.
Character 2: Ahuh—

Character 1: There wasn't time to stop over.
Character 2: Well—

Character 1: I was sad when I heard.
Character 2: Mhmm—

Character 1: They went swimming last week.

Character 2: Uh—

Awkward Dialogues

Instructions are presented on the first page:

You will hear some short conversations between two people. In some of the conversations, something awkward will be said. By "awkward", we mean something embarrassing that might make the other person feel uncomfortable.

After listening to each conversation, you will need to say whether anything awkward was said. You'll be asked to explain what was awkward, why it was said, and how the people might have felt. You should use a sentence or two for your answer.

You'll also be asked a more general question about what happened in the conversation. When you are ready to begin, press Continue.

For each dialogue, participants are shown a short description of the situation. This piece of text remains on the screen as they listen to the dialogue and answer the questions; see an example of a dialogue and the accompanying text below. Participants can start the audio for each dialogue in their own time. Once they have listened to the dialogue, they move on to a page which asks *Was something awkward said?* They click *YES* or *NO* to give their answer. If they click *Yes*, participants are presented with the following text: *Please explain: <u>What</u> was awkward? <u>Why</u> the people said the awkward thing(s)? How the people probably <u>felt</u>?* Participants type their response in a text box and click *Continue* when they are finished. Whether or not they click *YES* or *NO* to the first question, participants are asked a factual recall question displayed in the table below. They type their response and then move on to the next dialogue. Dialogues are presented in a set order.

Example dialogue

Participants are presented with the following text on the screen:

Helen mentions to Sarah that it is Clara's party tomorrow. Sarah is not going to the party.

Here is the script of the audio recording for this dialogue:

- *Helen, I think you've got a tear in your dress, there.*
- *Have I? I like this dress.*
- *Shouldn't be too difficult to sew up.*
- *Yeah, I should be able to do that by tomorrow.*
- *You going somewhere tomorrow?*
- *It's Clara's birthday. Are you not going?*
- *Probably not a good idea. We had a bad row.*
- *Sorry, I do know. I was so sad to hear what happened.*
- *To be honest, I'm surprised you're going to Clara's birthday, Helen. She was very nasty to me.*
- *Well, I already said I'd go.*

The factual recall question that accompanies this dialogue is *What does Sarah say about Clara?* Examples of correct answers to this question include: she was nasty/hadn't been very nice to her/they'd had a row, etc.

Marking Protocol

Participants are scored on their ability to detect social awkwardness. They are marked out of 2 for their response to each of the five awkward dialogues. Markers use the following generic mark scheme.

NA	Participant did not answer the factual recall question correctly AND did not provide a response to the social awkwardness meriting 2 marks (this means we can't be confident that the participant processed the explicit content of the dialogue, so the item is unlikely to be a valid indication of their ability to process implicit awkwardness).
Level 0	In response to the question *Was something awkward said?* the participant said *NO.*
Level 1: recognises awkwardness	Identifies an awkward moment, but the explanation is incomplete or unusual. The explanation may focus on factual details rather than mental states; it may be overly general; it may be distracted by irrelevance; or it may be an odd interpretation.
Level 2: mental state explanation	Explains how a moment is awkward by specific reference to either (1) how/why one character is trying deliberately to influence the mental states of the other character, or (2) how the incorrect assumptions of one character lead them to make a socially inappropriate statement.

Marking is not based on the length of response or the sophistication of written communication, but the validity of the explanation. As such, an answer to the example dialogue of "She is surprised her friend is going to Clara's party because she was nasty to her" would be scored 2 marks. On the other hand, an answer of "it is awkward when Sarah thinks she should not go to the party" would be given 1 mark, as it is incomplete. Some indicative content was devised for each dialogue to guide marking.

Frith-Happé Animations

Instructions are presented on the first page:

You will see some videos showing two moving triangles. The triangles will interact with each other. Sometimes they might interact as if they are aware of each other's thoughts and feelings.

After each video, you will be asked to describe what the triangles were doing. Write a sentence or two for your answer.

Please note that you might find it helpful to start describing to yourself what is happening while the video is still playing.

Press Continue to see a practice video.

On the next screen, participants watch the practice video. They click to start it when they are ready.

When the video has finished, the see the following question:

What do you think the two triangles were doing in the clip? Write one or two sentences.

They type their answer in a text box below the question. Then participants click *Continue* to move on to the transition page, which shows the following the text:

You can now have a go at the main task.

After each video, you will be asked to write about what happened.

Press Continue to start.

Then participants watch five theory of mind and four goal-directed animations in a pseudo-random order that is the same for all participants. After each video, they see the following question:

What do you think the two triangles were doing in the clip? Write one or two sentences.

Participants type their answer in a text box below the question. There is no time limit on the response.

The animations are drawn from Abell et al. (2000), and can be requested from the authors: *https://sites.google.com/site/utafrith/animations*. The marking guidelines of Castelli et al. (2000) were used for marking. Please note that the theory of mind practice item was repurposed as a test item in the present work, and animations were slightly shortened to 30 seconds or less to speed up the task.

Test of Local Textual Inference

Instructions are presented on the first page:

You will see a story in two parts. Each part is around 100 words. There will be some questions after each part. For some questions, the answer might be obvious and for others it might not be. Just type what you think the answer is likely to be. In some cases, there may not be any relevant information, and you can type "dk" -- don't know.

First of all, you will see the story text. Make sure you read the text in full. Then click the button saying "Read Questions". Please note that you will have 25 seconds to answer each question and a clock will show how much time you have left. Just use a word or phrase for your answer, and don't worry about spelling.

Participants click *Continue* to move on to the next page, which shows the following instruction and the first part of the story:

Read the story carefully. Then click "See Questions".

Unfortunately, the family couldn't go swimming. The sea was rougher and colder than expected. Instead, Billy spent the whole morning playing a ballgame with his sister, Susie….
[The story continues.]

Once participants click *See Questions*, they are presented with the first ten questions one after the other. Each question is presented visually below the story text. They have up to 25 seconds to type a short answer to each question in a text box, and see a timer on the screen throughout this time.

After they have completed the first set of questions, participants are presented with the second part of the story:

Read the story carefully. Then click "See Questions".

[The second part of the story is shown.]

Once participants click *See Questions*, they are presented with the second set of ten questions one after the other. Each question is presented visually below the story text. They have up to 25 seconds to type a short answer to each question in a text box, and see a timer on the screen throughout this time.

Stimulus List: Sample of Test Items

What was the weather like?

What had Billy planned to do?

Answers to each question are scored out of 2. Where the participant makes the key point in response to a question, a mark of 2 is given. A mark of 1 out of 2 is awarded for an incomplete answer where it is not entirely clear the participant has made the necessary inference for a question. A mark of 1 out of 2 is also given for an otherwise correct answer that additionally includes information that can't be reliably inferred from the passage.

Synonyms Test

Written instructions are presented on the first screen:

This task is about the meanings of words. On each screen, you will see a word. Below it you will see four possible meanings for that word. You need to click on the best meaning.

You have 12 seconds to give each answer. A clock will count down for the last 6 seconds.

Press Continue to try a practice item.

Participants see the word *puzzle* in the middle of the screen. There are four possible answers below: *emotion*, *mystery*, *story* & *weather*.

After clicking their answer, participants see a transition screen showing the following text:

The correct answer for the practice item was "mystery", as this was the word with the closest meaning to "puzzle".

Now you can have a go at the real items. Try and be as accurate as possible. Press Continue to start.

Participants are then presented with the 25 items in a randomised sequenced.

Stimulus List: Sample of Test Items

Target Word	Option 1	Option 2	Option 3	Option 4
renounce	refuse	give up	repeat	condemn
Ploy	perform	illusion	manoeuvre	deception

Grammaticality Decision Test

Written instructions are presented on the first screen:

You will hear some sentences and you need to decide if they are well-formed. Listen out for grammatical mistakes that you would not expect native speakers to make.

When you hear a sentence, press C if it is well-formed or N if there are mistakes. You will not have long to respond, so answer quickly.

To try two practice items, press Continue.

Participants now see a transition screen presented for 5 seconds. The following text is presented on the screen:

Keep your fingers close to the keys so you can respond quickly. For each sentence, press C if it is a good, well-formed sentence without mistakes. Press N if it is not.

Participants now complete two practice items presented over audio. Practice item 1 is *The dog chases the cat*. Practice item 2 is *The dog chasing the cat is*. Prior to each trial, a fixation cross is presented on the screen for 500 milliseconds, then participants hear the stimulus and have 6 seconds to respond from the onset of the stimulus. The following text is presented on the screen throughout:

Press C for a good well-formed sentence. Press N for a badly formed sentence with mistakes.

Following the two practice trials, participants see the following text:

Well done! That's the end of the practice. Press Continue to move on the main task.

Once they click *Continue*, participants see the following text during a 5-second transition:

The task is about to begin. Keep your fingers close to the keyboard so you can respond quickly. For each sentence, press C if it is a good, well-formed sentence without mistakes. Press N if it is not.

The format of each trial is the same as the practice trials. Prior to each trial, a fixation cross is presented on the screen for 500 milliseconds, then participants hear the stimulus and have 6

seconds to respond from the onset of the stimulus. The following text is presented on the screen throughout:

Press C for a good well-formed sentence. Press N for a badly formed sentence with mistakes.

50 sentences are presented in a randomised sequence.

Stimulus List: Sample of Test Items

They may come but I've not checked for sure.

Sam gave Ann a bottle of juice from the box.

The woman put in the fridge the milk.

What did he ask her what she bought?

Appendix C. Meta-Analysis of Performance of Autistic and Non-Autistic Adults on the Frith-Happé Animations (Supplementary Material for Chapter Four)

In this meta-analysis, I aimed to quantify the group difference between autistic and non-autistic adults on performance on the Frith-Happé Animations (Abell et al., 2000). To be included in the meta-analysis, studies needed to fulfil the following criteria: (1) autistic and non-autistic adults were asked to produce verbal descriptions of the "theory of mind" animations, and (2) these descriptions were rated on their "appropriateness" based on the mark scheme by the original authors.

We extracted sample sizes, means and SDs for appropriateness ratings on the "theory of mind" animations, and the "goal-directed" control items too, if administered in the study. We collated papers analysed by Bliksted et al. (2016) in their meta-analysis of performance by clinical groups on this task, and also checked studies that either (1) cited any of the original papers using the task (Abell et al., 2000; Castelli et al., 2000; Frith et al., 2002), or (2) cited a more recent paper that developed a multiple-choice format for the task (White et al., 2011). The final search was carried out in November 2019.

The effect size of interest was Hedge's g (standardised group difference with a correction for small sample size), and this was calculated for each comparison between groups of autistic and non-autistic adults on either the "theory of mind" or control animations. Individual effect sizes were aggregated in a random effects meta-analysis for each type of animation using the R package metafor (Viechtbauer, 2010).

The meta-analysis for the "theory of mind" animations included 237 autistic adults and 236 non-autistic adults; for the control animations, overall sample size was 217 and 216 respectively. Participants were well-matched across studies for age and verbal intelligence. Across studies, average age varied between 23 and 37 years for autistic adults and 23 and 39

years for non-autistic adults, and average verbal IQ was invariably in the average to high average range. Most samples were predominantly men: 75%+, with the exception of one sample (N = 103) that included 58 women (Lugnegård et al., 2013). Across studies, autistic adults tended to underperform on the task compared to non-autistic adults with a moderate effect size, but this was not specific to the "theory of mind" animations. Hedge's g [95% CI] for the "theory of mind" animations was -0.73 [-0.55, -0.92], and for the control animations was -0.50 [-0.22, -0.79]. See Figure B.1 for a forest plot of the results.

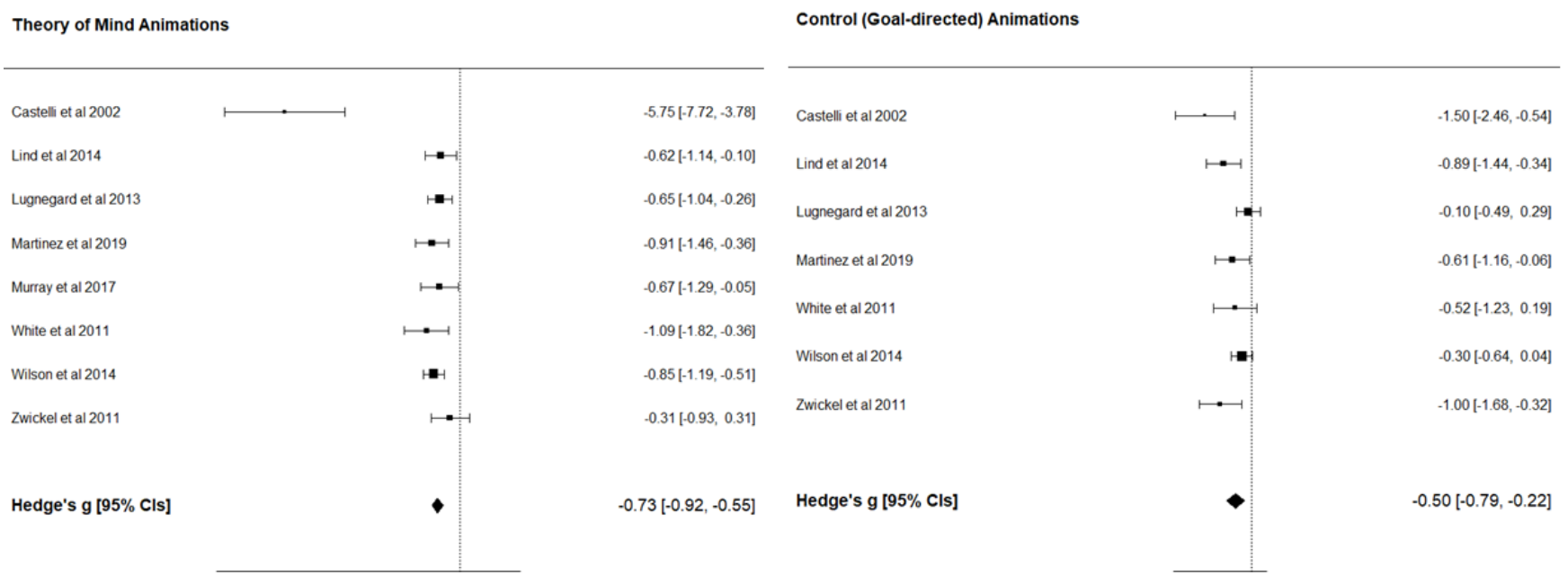


Figure B.1. Forest plot showing meta-analytic results for a comparison between performance by autistic and non-autistic adults on the Frith-Happé Animations.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development, 15*(1), 1–16. https://doi.org/*10.1016/S0885-2014(00)00014-9*

Bliksted, V., Ubukata, S., & Koelkebeck, K. (2016). Discriminating autism spectrum disorders from schizophrenia by investigation of mental state attribution on an on-line mentalizing task: A review and meta-analysis. *Schizophrenia Research, 171*(1-3), 16–26. *https://doi.org/10.1016/j.schres.2016.01.037*

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage, 13*(2), 314–325. https://doi.org/*10.4324/9780203496190*

Frith, U., Castelli, F., Frith, C., & Happé, F. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*(8), 1839–1849. *https://doi.org/10.1093/brain/awf189*

Lind, S. E., Williams, D. M., Bowler, D. M., & Peel, A. (2014). Episodic memory and episodic future thinking impairments in high-functioning autism spectrum disorder: an underlying difficulty with scene construction or self-projection? *Neuropsychology, 28*(1), 55–67. *https://doi.org/10.1037/neu0000005*

Lugnegård, T., Unenge Hallerbäck, M., Hjärthag, F., & Gillberg, C. (2013). Social cognition impairments in Asperger syndrome and schizophrenia. *Schizophrenia Research, 143*(2-3), 277–284. *https://doi.org/10.1016/j.schres.2012.12.001*

Martinez, G., Mosconi, E., Daban-Huard, C., Parellada, M., Fananas, L., Gaillard, R., Fatjo-Vilas, M., & Krebs, M. (2019). "A circle and a triangle dancing together": Alteration of social cognition in schizophrenia compared to autism spectrum disorders. *Schizophrenia Research*, 210, 94–100. *https://doi.org/10.1016/j.schres.2019.05.043*

Murray, K., Johnston, K., Cunnane, H., Kerr, C., Spain, D., Gillan, N., Hammond, N., Murphy, D., & Happé, F. (2017). A new test of advanced theory of mind: The "Strange Stories Film Task" captures social processing differences in adults with autism spectrum disorders. *Autism Research*, *10*(6), 1120–1132. *https://doi.org/10.1002/aur.1744*

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. *https://www.jstatsoft.org/v36/i03/*

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: a quick and objective test of Theory of Mind for adults with autism. *Autism Research*, *4*(2), 149–154. *https://doi.org/10.1002/aur.174*

Wilson, C. E., Happé, F., Wheelwright, S. J., Ecker, C., Lombardo, M. V., Johnston, P., Daly, E., Murphy, C. M., Spain, D., Lai, M. C., Chakrabarti, B., Sauter, D. A., MRC AIMS Consortium, Baron-Cohen, S., & Murphy, D. G. (2014). The neuropsychology of male adults with high-functioning autism or Asperger syndrome. *Autism Research, 7*(5), 568–581. *https://doi.org/10.1002/aur.1394*

Zwickel, J., White, S. J., Coniston, D., Senju, A., & Frith, U. (2011). Exploring the building blocks of social cognition: spontaneous agency perception and visual perspective taking in autism. *Social Cognitive and Affective Neuroscience*, *6*(5), 564–571. https://doi.org/10.1093/scan/nsq088

Appendix D. Instructions for Child Assessments and Examples of Test Items

In this appendix, I provide greater detail about the assessments reported in Chapters Six and Seven. Full assessments are not included here to safeguard materials for future research and clinical uses. However, I am happy to share assessments and associated scoring with researchers wishing to replicate and/or extend the work reported in this thesis; please see *https://doi.org/10.17605/OSF.IO/XN48E* for information on gaining access.

Implicature Comprehension Test – Children's Version

Participants hear the following instructions. Animations of Robot, Tom and Sally are also shown on the screen throughout:

Welcome to the Conversations Game!

This is Robot. She is trying to understand what people mean when they talk to each other.

And this is Tom and Sally. They talk a lot, and Robot wants to understand what they're talking about. We're going to listen to Tom and Sally in some short videos. When Robot is not sure what Tom and Sally mean, she will ask you a question.

There will be buttons on the screen, like this, and you can click YES or NO for your answer. Sometimes you might not be able to answer the question. If that happens, you can press DON'T KNOW.

Let's hear an example.

After you've heard Tom and Sally speak, listen out for my question. Here goes…

Participants then hear Tom say, "*We're going to the mountains*," and Sally say, "*We're going skiing for our holiday*". Then Robot says, "*Do you think Tom and Sally are going skiing in the mountains?*"

Participants click their answer. Then Robot continues:

The answer to my question was yes, they're going skiing.

Okay, we will now hear some practice videos. Press GO to start.

Participants now complete the three practice items. Each trial includes two pages: the first page where participants view the video and the second page where they give their response to the comprehension question. Participants click videos to start them, and they are reminded of this on the first page of each trial, which shows the instruction *Click the video to start* below the video space. Each video consists of a short conversational exchange followed by Robot's comprehension question. Once the participant has watched the video, a second page is shown, which includes the three answer options *YES*, *NO* and *DON'T KNOW*. There are also some instructions as reminders. The vertical order of objects on this second page is: the video space, the instruction *Click your answer*, the answer options presented horizontally across the screen, and finally the instruction *Click the video if you want to hear it again*. Videos may be replayed once, and there is no time-limit for the response.

Scripts for the practice videos are shown in the item list below. After the three practice videos, participants hear the instruction:

We can now try the rest of the videos. Let's start!

Then the test items are presented one by one, in the same way as the practice items. See the list of items below.

Stimulus List: Practice Items

Character 1: Have you fed the dogs yet?

Character 2: I gave them both their breakfast.

Do you think he has fed the dogs yet?

Character 1: Have you finished packing?

Character 2: Not yet, I'll finish soon.

Do you think she has finished packing?

Character 1: Have we packed the gloves?

Character 2: I don't know about the gloves.

Do you think they have packed the gloves?

Stimulus List: Sample of Test Items

The item type to which each item belongs is indicated; item types include implicature, explicit-response and "don't know".

Character 1: Did you see the policemen earlier on?

Character 2: I saw them standing on the platform.

Do you think she saw the policemen?

Explicit-response

Character 1: Did the police speak to anyone else?

Character 2: I wasn't watching them much.

Do you think the police spoke to anyone else?

Don't know

Character 1: Could you hear what the police said?

Character 2: There were lots of trains going past.

Do you think he heard what the police said?

Implicature

Character 1: Does the temperature need turning up in here?

Character 2: I just took off my jumper.

Do you think the temperature needs turning up?

Implicature

Character 1: Have you tried the new tea?

Character 2: I needed a hot drink after skiing.

Do you think he tried the new tea?

Implicature

Character 1: Did you buy the milk this morning?

Character 2: I went out for a walk earlier on.

Do you think she bought the milk this morning?

Implicature

Children's Test of Local Textual Inference

Participants hear the following instructions:

Welcome to the story task.

You are going to hear a short story. The story will be told in two parts. [*STORY IN TWO PARTS* shown on the screen.]

After each part, you will hear some questions. [*STORY IN TWO PARTS* and *SOME QUESTIONS* shown on the screen.]

For each question, there will be four answers to choose from. [*STORY IN TWO PARTS, SOME QUESTIONS* and *FOUR ANSWERS TO CHOOSE FROM* shown on the screen.]

You just need to click on the right answer. [*STORY IN TWO PARTS, SOME QUESTIONS, FOUR ANSWERS TO CHOOSE FROM* and *CLICK THE RIGHT ANSWER* shown on the screen.]

The story will stay up on the screen throughout to help you. [*STORY WILL STAY UP ON THE SCREEN* shown on the screen.]

When you are ready to hear the story, click GO.

After clicking *GO*, there is a 5-second countdown and then participants hear the first part of the short story:

Unfortunately, the family couldn't go swimming. The sea was rougher and colder than expected. Instead, Billy spent the whole morning playing a ballgame with his sister, Susie. [The short story continues.]

While they listen to the story, participants see the full text on the screen throughout. As each sentence is read out, it is highlighted in larger font.

After the story, participants hear the following instructions:

There are now some questions. For each one, you need to click the right answer. Let's start.

Participants click *GO* to move on the questions. There is a 5-second countdown, and then the first ten questions are presented one by one. The story text remains on the screen throughout. Underneath the story, participants see the question, which they also hear spoken aloud. Underneath the question, four answer options appear horizontally one by one, and each is

spoken aloud as it appears. There is 500-millisecond interval between auditory presentation of each option. Participants need to wait until the final option has been presented before their response will be registered with a mouse click. After completing these questions, participants hear the following:

You're doing really well. We're now going to listen to the second part of the story. There will then be some questions, just like before. Ready?

Participants click *GO* to hear the next part of the story. There is a 5-second countdown before the story starts. As they listen to the story, participants see the full text on the screen throughout. As each sentence is read out, it is highlighted in larger font.

After the story, participants hear "*And here are the questions!*"

Participants click *GO* to move on the questions. There is a 5-second countdown, and then the second set of ten questions is presented in the same way as the first set.

Stimulus List: Sample of Test Items

What was the weather like?

Really bad *A bit cold* *We don't know* *Sunny*

What had Billy planned to do?

We don't know *Walk along the sea* *Go swimming* *Play a ballgame*

Pragmatic Violations

Participants hear the following instructions:

This activity is called Tom and Robot Chat. You will hear some conversations between Tom and Robot. For each video, you need to decide if Robot gives a good answer. Let's hear an example. Click the video to play it.

Participants watch a short video.

T: When did you arrive?
R: I arrived yesterday.

An analogue scale appears on the screen.

After you've watched each video, you'll see a line on the screen. Click on the line. You'll see a little pointer appear. Now click along the line. You'll see the pointer move. Have a go at moving the pointer. Then click GO.

You need to use the line to tell us how good Robot's answer was.

A sad face appears at the left end of the line.

Click on this end of the line if Robot's answer was really bad.

A happy face appears at the right end of the line.

And click on this end of the line if Robot's answer was really good.

A neutral face appears in the middle of the line.

And click in the middle of the line if Robot's answer was okay. You click along the line to tell us how good Robot's answer was. We're now going to hear Robot and Tom's conversation again. Click the video to play it.

Participants hear the sample video again:

T: When did you arrive?

R: I arrived yesterday.

Participants hear the following guidance:

So how good was Robot's answer? She answered Tom's question. And she kept to the point. So this was a pretty good answer. You need to click on the line near the smiley face. Try this now. Then click GO.

Participants now move into the practice phase of the task.

Let's have a go at a couple more examples. Click the video to play it.

T: How are you?

R: I'm very well, thanks. How are you?

Participants register their response on the analogue scale, and then hear the following feedback:

That was another good answer, wasn't it? Robot answered Tom's question well. Let's hear another video.

T: How old are you?

R: I am 15 years old. I have been 15 for 10 months. This means I am almost 16.

Participants register their response on the analogue scale, and then hear the following feedback:

Mmm, that was an unusual answer. Most people would give a shorter answer. They would not give so much detail. They might just say that they're 15. You might have dragged the

pointer towards the sad face for that one. When you're ready to start the rest of the videos, click GO.

Participants now start the main task. For each trial, participants start the video by clicking it. Above the video space is the instruction *Click the video to play it.* After the video has played, participants are presented with an answer screen. Under the video space is the analogue scale on which participants record their answer. Immediately above the scale are the words *ROBOT'S ANSWER WAS* and below the scale are schematic faces: a happy face at the far right and a sad face at the far left, and a neutral face in the middle. *REALLY BAD* is written below the sad face and *REALLY GOOD* is written below the happy face. At the bottom of the screen, in the middle, is the *GO* button, which participants click to move to the next trial. There are 35 trials, details of which are given in the stimulus list below. Trials are presented in three blocks, with a short "robot dancing" sequence in the two intervals.

Stimulus List: Sample of Test Items

What was the weather like?
It rained a lot this morning. Then it got sunny this afternoon. It was nice.

Where did you go?
Er

Did you go to the cinema?
I missed the film. He said sorry.

Why did the dog bark?
Someone came to the door. They made a lot of noise. It scared the dog.

Where is the dog?

I let the dog outside into the garden. I didn't give it any toys. It chewed up a teddy bear last week.

What did you do yesterday?
I went to the swimming pool. I wanted to go swimming yesterday. I had to take my towel with me.

Social Overtures

Participants hear the following instructions:

This activity is called Starting conversations.
There are different reasons why we say things to other people. Sometimes, we want to start a conversation. We want the other person to ask us questions and say lots of things to us. Other times we just want to tell the other person something very quickly. We don't always want to start a long conversation.

You're going to hear a man saying some things to a woman. Sometimes he'll want to start a conversation with her. Other times he won't. You need to decide if he does or he doesn't.

Let's try some practice videos.

Participants now click *GO* for the practice videos. *Click to start practice* is shown on the screen.

There is a 5-second countdown and then the first practice trial plays. For the two practice trials, a short video plays in which a man says a sentence. After the video has finished, participants hear the question *Does he want a conversation with her?* This question is also shown as text on the screen, and underneath the question, there are two answer options, *YES*

and *NO*, presented horizontally. In the first video, the man says *Did you have a good day?* After participants click their answer, they hear the following feedback:

The man asked the woman if she had a good day. The two people are likely to have a bit of a conversation about what they've been doing today. This might lead on to other topics. This means YES is the answer, he does want to have a conversation with her.

Now there is a 3-second countdown for the second practice trial. This time the man says *I'm leaving now. I'll see you later.* After participants click their answer, they hear the following feedback:

The man said that he's leaving now. Maybe he's going to work or going to the shops. He probably doesn't want to have a conversation, so NO is the answer.

You can now have a go at some more videos on your own.

Participants click *GO* to move on. There is a 5-second countdown. For each trial, the video starts by itself, and shows a man saying something to a woman. After the video has finished, the response format shows on the screen. Below the video space is the question *Does he want a conversation with her?* Under this question, there are *YES* and *NO* options. At the top of the screen, there is the instruction *Click the video to hear it again*. Participants can replay the video once. Note that, unlike in the practice, the question is not spoken aloud; only the stimulus itself is presented auditorily.

Stimulus List: Sample of Test Items

I need to go and get my coat.

I can't believe what happened today.

I'd really like to get a dog.

I'm going to have a shower.

Receptive Vocabulary

Participants hear the following instructions:

We're now going to play a word game. You will hear some words. Each time you hear a word, there will be four pictures on the screen. You need to choose which picture goes best with the word.

Let's have a go at two examples. Listen to the word. Then click the picture that is linked to the word.

Animal. [Pictures of a bed, kite, fan and fox are shown on the screen.]

Participants click a picture. The following feedback is then provided:

A fox is an animal, so you needed to click on the fox. Here's another word.

Drive. [Pictures of a witch, pen, car and leaf are shown on the screen.]

Participants click a picture. The following feedback is then provided:

People can drive cars, so you needed to click on the car. We can now listen to the rest of the words. Click GO to begin.

There is a 5-second countdown. Then the words are presented; see the stimulus list below. Each word plays by itself. Pictures are shown in the four quadrants of the screen. In the middle of the screen is a play button and at the bottom of the screen is the instruction:

Click play to hear the word again.

Participants can replay a word twice (i.e. they can hear each word up to three times).

Stimulus List: Sample of Test Items

Target word	**Picture top left**	**Picture top right**	**Picture bottom left**	**Picture bottom right**
eat	bread	wheel	wall	bell
magical	arm	witch	snow	thief
punctual	watch	pen	scarf	sword
illuminate	swing	lamp	snow	eye