



## RESEARCH ARTICLE

# **REVIS** "If you catch my drift...": ability to infer implied meaning is distinct from vocabulary and grammar skills [version 3; peer review: 3 approved]

Alexander C. Wilson , Dorothy V.M. Bishop

Department of Experimental Psychology, University of Oxford, Oxford, Oxfordshire, OX2 6GG, UK

**v3** **First published:** 15 Apr 2019, 4:68 (<https://doi.org/10.12688/wellcomeopenres.15210.1>)  
**Second version:** 10 Jul 2019, 4:68 (<https://doi.org/10.12688/wellcomeopenres.15210.2>)  
**Latest published:** 30 Aug 2019, 4:68 (<https://doi.org/10.12688/wellcomeopenres.15210.3>)

## Abstract

**Background:** Some individuals with autism find it challenging to use and understand language in conversation, despite having good abilities in core aspects of language such as grammar and vocabulary. This suggests that pragmatic skills (such as understanding implied meanings in conversation) are separable from core language skills. However, it has been surprisingly difficult to demonstrate this dissociation in the general population. We propose that this may be because prior studies have used tasks in which different aspects of language are confounded.

**Methods:** The present study used novel language tasks and factor analysis to test whether pragmatic understanding of implied meaning, as part of a broader domain involving social understanding, is separable from core language skills. 120 adult participants were recruited online to complete a 7-task battery, including a test assessing comprehension of conversational implicature.

**Results:** In confirmatory analysis of a preregistered model, we compared whether the data showed better fit to a two-factor structure (including a "social understanding" and "core language" factor) or a simpler one-factor structure (comprising a general factor). The two-factor model showed significantly better fit.

**Conclusions:** This study supports the view that interpreting context-dependent conversational meaning is partially distinct from core language skills. This has implications for understanding the pragmatic language impairments reported in autism.

## Keywords

Autism, pragmatic language, social communication, implicature, conversation, psychometric, online research

## Open Peer Review

**Reviewer Status**

	Invited Reviewers		
	1	2	3
<b>version 3</b> published 30 Aug 2019			
<b>version 2</b> published 10 Jul 2019	 report	 report	
<b>version 1</b> published 15 Apr 2019	 report	 report	 report

- Danielle Matthews** , University of Sheffield, Sheffield, UK
- Lauren Swineford**, Washington State University, Pullman, USA
- Katie Alcock** , Lancaster University, Lancaster, UK

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Alexander C. Wilson ([alexander.wilson2@psy.ox.ac.uk](mailto:alexander.wilson2@psy.ox.ac.uk))

**Author roles:** **Wilson AC:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Bishop DVM:** Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by the Wellcome Trust [082498] and the European Research Council [694189].  
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2019 Wilson AC and Bishop DVM. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Wilson AC and Bishop DVM. "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 2019, 4:68 (<https://doi.org/10.12688/wellcomeopenres.15210.3>)

**First published:** 15 Apr 2019, 4:68 (<https://doi.org/10.12688/wellcomeopenres.15210.1>)

**REVISED Amendments from Version 2**

The funding statement has been amended to include funding from an Advanced Grant from the European Research Council [694189].

**Any further responses from the reviewers can be found at the end of the article**

**Introduction: Theoretical underpinnings**

Observations of people with autism and social communication difficulties suggest that it is possible to have problems with conversational language in the relative absence of impairments in aspects of “core language”, such as vocabulary knowledge and grammatical competence (e.g. [Baird & Norbury, 2016](#); [Lam & Yeung, 2012](#)). However, attempts to demonstrate this dissociation using objective tests have mostly failed. Two separate reviews have concluded that conversational skills are closely related to core language abilities ([Andrés-Roqueta & Katsos, 2017](#); [Matthews et al., 2018](#)). However, these findings are based on language tests that may not act as pure indicators of particular language skills. Our goal was to devise novel language tests to give a more convincing answer to the question of whether core language abilities and sensitivity to social aspects of language are separable sets of skills in the general adult population.

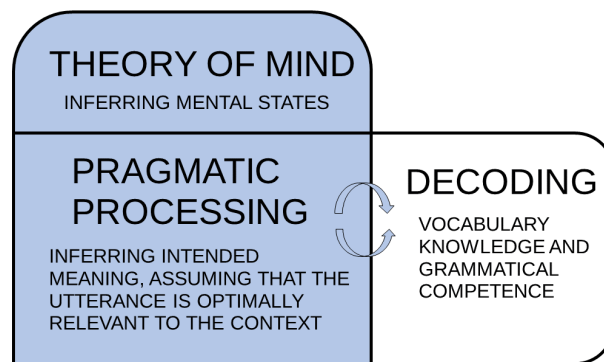
While standardised tests of core language skills are in abundance, only a limited range of tests focus on social aspects of language ([Norbury, 2014](#)). Social communication is, of course, a broad construct. It has been defined as the combination of “social interaction, social cognition, pragmatics (verbal and nonverbal) and receptive and expressive language processing” that supports conversation ([Adams, 2005](#), p. 128). The broad nature of this construct makes it difficult to study with precision. By contrast, *pragmatics* is a specific facet of language processing that can potentially be more easily operationalised for assessment - although, from the offset, we should note that the term has frequently been overextended in the field of communication disorders to essentially mean all social aspects of communication ([Cummings, 2007](#)). In linguistics, pragmatics is defined as “a process of reasoning based on features of context” ([Cummings, 2007](#), pp. 425–426). The pragmatic aspect of language comprehension mediates between dictionary meaning and a speaker’s communicative meaning). As an example of pragmatics in action, consider the utterance “it’s cold here”. Descriptively, the speaker tells us about the temperature of a place. Through pragmatic processing of the utterance in its particular context, we might also infer that the speaker is implying they want to close a window, for instance, or to go inside. As such, pragmatics involves using context to “read between the lines”.

Existing tests of communication skills are not sufficiently focused to measure pragmatic processing. Communication is often measured globally, through observation of semi-naturalistic conversation (e.g. using the Pragmatic Protocol, [Prutting & Kirchner, 1987](#)) or through behavioural checklists completed

by informants (e.g. the Children’s Communication Checklist, [Bishop, 2003](#)). These assessments are likely to conflate pragmatics, social interaction and core language skills. A further drawback is that they focus on social communication “behaviours” rather than 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 failure to produce contingent turns in conversation? This was a problem for autistic children assessed using the Yale *in vivo* Pragmatic Protocol, a semi-naturalistic conversational assessment including “specific pragmatic probes” for eliciting target behaviours ([Schoen Simmons et al., 2014](#)). A lack of contingent turns may arise from difficulty inferring 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, problems in finding words, perseveration on one’s own topic of interest, or other issues besides. These are not pragmatic problems.

To quantify pragmatic skills, psychometric assessment is necessary. However, traditional standardised language measures do not pick up pragmatic difficulties ([Conti-Ramsden et al., 1997](#)), and tests designed for this purpose have been quite uninformative. The Test of Pragmatic Language-2 (TOPL-2) is perhaps the most well-known pragmatic test. It requires the individual to produce situationally-appropriate speech acts, and can detect pragmatic impairment in groups with autism ([Young et al., 2005](#)). However, it is not consistently sensitive ([Reichow et al., 2008](#); [Volden & Phillips, 2010](#)). The same problem exists with the pragmatic judgement subtest of the Comprehensive Assessment of Spoken Language, which is a similar, commonly used test ([Klusek et al., 2014](#)). These tests correlate highly with core language skills (e.g. [Akbar et al., 2013](#)), and they do not attempt to minimise demands on vocabulary/grammar or expressive language ability, so we should question their specificity as tests of pragmatics. In addition, both these tests focus on politeness and knowledge of social rules, which is different from the linguistic definition of pragmatics, as inference of meaning from context. Another “pragmatic” test, the new Pragma test ([Loukusa et al., 2018](#)), discriminates well between autistic and neurotypical individuals, but is not a pure test of pragmatics, as it contains subscales tapping rather different skills, such as theory-of-mind and emotion recognition. While this kind of test may be useful clinically for picking up communicative problems, it is not well-suited for our purposes, i.e., pinpointing the underlying nature of those problems.

We need to be clear what we should measure in a pragmatic test. In reviewing definitions of pragmatics, [Ariel \(2010\)](#) stresses the need to contrast semantics and pragmatics. While semantics involves decoding conventional “dictionary” meaning, pragmatics is all about inference: we use context to infer further non-codified meaning. Please see [Figure 1](#) for a visual representation of our model of language processing. We follow Relevance Theory in assuming that language processing is a combination of (1) decoding the “literal meaning” of an utterance using our



**Figure 1. Our model of language processing, as borrowed from Relevance Theory.** Language processing is an interaction between decoding and pragmatic inference. This pragmatic processing is seen as a sub-domain of our more general “theory-of-mind” capacity.

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 (Sperber & Wilson, 1986). In Relevance Theory, inferring communicative meaning is assumed to be a subdomain of our general ability to attribute mental states to others (our “theory-of-mind” or capacity for “mind-reading”). By “subdomain”, we mean that pragmatic processing is thought to happen through a domain-specific heuristic – “communicative relevance” – which, as the name suggests, is specifically activated by communicative behaviour. This means that we infer what an utterance means based on the automatic assumption that interlocutors will communicate in a way that is optimally relevant to the context (see Carston & Powell, 2008).

Relevance Theory assumes that we make inferences at two levels. On the one hand, we need to infer the full “explicit” meaning of particular words and phrases; this is explicature. Partly this is a semantic process, as we need to access the dictionary meaning for the word/phrase; however context is important in determining the appropriate meaning, so this is also a pragmatic process. In addition to explicature, we may also infer further “implicit” meaning, or implicature, from a global understanding of the utterance in context. This involves “reading between the lines” and understanding what has not been directly stated. This is a purer pragmatic process, as it depends wholly on inference. For illustration, consider the following dialogue:

SPEAKER ONE. Can you pick Sally up from the station?

SPEAKER TWO. I’m working all day.

In terms of explicature, “I’m working all day” might be taken to mean “Today I will be doing work tasks at my workplace during my working hours”. Here we make local-level inferences about what the speaker’s individual words mean in context. 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 ONE’s request to pick Sally up. In other contexts, these words would not communicate anything about picking a person up from the station: the implicated, indirect meaning is only expressed here because we assume that an utterance in this

context must relevantly address SPEAKER ONE’s question, and so we actively seek an implied meaning.

In designing a theoretically-motivated test of pragmatics, implicature is a good focus, as implicated meaning can only be interpreted through context-dependent inference - i.e. pragmatic processing. Existing research has not explored conversational implicature as discussed above. Generally, empirical research investigating implicature has focused on generalised conversational implicature, especially scalar implicature (where we infer, for instance, that not all the apples are moldy in the sentence “Some of these apples are moldy”). However, with generalised conversational implicature, the implied meaning is invariably present whenever this sort of language is used (e.g. Carston, 1998). Therefore, it is very different from the more particularised conversational implicature discussed above, which is much more dependent on the communicative context.

While research into implicature has generally not been strongly influenced by Relevance Theory, a set of papers by Leinonen, Loukusa and others is an exception (Leinonen *et al.*, 2003; Loukusa *et al.*, 2007a; Loukusa *et al.*, 2007b; Ryder & Leinonen, 2014; Ryder *et al.*, 2008). These researchers report reduced ability in children with language impairments or autism to make inferences on language tests designed based on concepts from Relevance Theory. However, these studies did not focus on implicature in conversational contexts; they focused on making elaborative inferences to fill in semantic gaps in short story material. This focus is likely to miss the more social aspects of implied meanings, which are important for understanding pragmatic problems affecting everyday communication.

## Introduction: Empirical work

Our objective was to develop a test of pragmatic language processing that removes as much as possible the effect of grammar/vocabulary skills on scores by (a) using simple language and (b) incorporating control items (see below). We focused on implicature comprehension, as linguists are unanimous in viewing implicature as a complex pragmatic phenomenon (e.g. Ariel, 2010). As well as allowing us to test how far pragmatic skill is separable from other language abilities, the test was 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 goals, we were mindful of making our Implicature Comprehension Test (ICT) sufficiently child-friendly, while avoiding ceiling effects in adults.

In a pilot study (see Methods), we found that the ICT was internally consistent, and that scores did not correlate significantly with semantic knowledge (measured by tests of vocabulary and recognition of conversational phrases/idioms). This provided some evidence that implicature comprehension dissociates from basic linguistic decoding. In the present study, we set out to collect some more normative data on the ICT in typical adults to replicate that finding, and also to test for a dissociation between implicature comprehension and grammatical ability too. This was our key question: is pragmatic processing (as measured by the ICT) separable from core language skills (grammar and vocabulary)? Our 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 above, and shown in Figure 1, the capacity to infer intended meanings from context – pragmatic comprehension – is thought to be a subdomain of our more general “theory-of-mind”. Conceptually, then, we would expect our 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, we developed a test battery including two tasks that we expected to require sensitivity to conversational contexts as well as a test requiring the individual to infer mental states in different scenarios.

We hypothesised that a two-factor model (“core language competence” and “social understanding”) would account for individuals’ performance across our 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 above, which may inflate the correlations. However, the possibility that individuals can have relatively specific difficulties with pragmatics (Lam & Yeung, 2012) led us 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” (Wilson & Bishop, 2018): <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. We report below how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

## Participants

We recruited 120 adults online via the participant recruitment platform, Prolific, who all gave informed consent to participate.

Inclusion criteria included: (i) age over 18, (ii) private access to a computer with a good internet connection, (iii) no significant visual or hearing impairment, (iv) native-level fluency in English, and (v) no participation in the pilot study of this project. Of the 120 participants, 8 were excluded based on poor performance on at least one test; see our exclusion criteria under Data analysis below.

Our sample size was based on a power calculation using simulations. We used a p-value of 0.05 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, we simulated data conforming to a two-factor correlated traits model with a core language factor relating to three tests and a conversation comprehension factor relating to four tests. Factor loadings were set at 0.7 and the factors were correlated at 0.2 (a low correlation was used, as the semantic tests did not correlate with the ICT in the pilot study). 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.

In an open response format, we asked participants to give their age, gender and race/ethnicity. 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 1 person did not declare their gender. The majority of the sample described themselves as White (103 out of 120); 4 people identified as Mixed Race, 4 as Black, and 8 as Asian. We also asked participants whether they were currently studying. 34 of the participants said they were, of whom 28 reported they were completing undergraduate studies and 6 indicated they were post-graduates. 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 9.

## Procedure

This was an online study supported by Gorilla and all data were collected via Prolific on 2<sup>nd</sup> August 2018. 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. See Extended data (Wilson & Bishop, 2019) for examples of test items.

**Implicature Comprehension** was measured using the Implicature Comprehension Test (ICT). The task involves watching a series of 57 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 middle primary school level. 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”-“don’t know” response to the question, recorded by keyboard presses. There is a 5-second limit for responding. Note that there were few instances of time-outs in this study. 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 4, and 94% of participants missed one or fewer trials (83% did not miss any). Five example items (one of the practice trials and four implicature items) are available as Extended data (Wilson & Bishop, 2019).

In 27 or the 57 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 10 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 95.4% (SD= 9.95%), 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 10 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 legitimize “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. These trials functioned well in getting participants to use the “don’t know” response; mean accuracy on these items was 87.5% (SD= 20.1%).

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 10 “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 the different item types. Mental state words are also avoided, to remove the “theory-of-mind” demand of these. 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).

**Receptive vocabulary** was measured by a Synonyms Test devised for this study. This includes 25 items. 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).

**Receptive grammar** was measured by a Grammaticality Decision Test devised for this study. In this task, 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 an 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 6 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 currently answered (out of 50).

**Sensitivity to social awkwardness** was measured using the Awkward Dialogues. This task is based loosely on the Faux Pas Recognition Test (Baron-Cohen *et al.*, 1999). Individuals needed to 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 by the first author. 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 0.78 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 our measure of participant ability on this test. We chose to use factor scores rather than sum totals, since it was not clear that the intervals in the mark-scheme (i.e. between 0, 1 and 2) 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 (1 mark for each of 5 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.

**Comprehension of fillers/backchannel continuers** was measured using a Test of Fillers and Backchannels devised for this study. We expected this 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. We 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 4 backchannel continuers (mm-hmm, uh-huh, really,

right) and 5 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). We therefore use prosodic information to disambiguate "yeah" (we do the same for "oh", since this is likely to function similarly to "yeah"). There was one measured variable: the sum of items correctly answered (out of 40).

**Narrative-based inferencing** was measured by a Test of Local Textual Inference devised for this study. In this study, we were interested in narrative-based inferencing, as the relationship between this kind of inferencing and other language skills was of theoretical interest. 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). We expected narrative-based inferencing to be quite a different process to interpreting an implicated meaning in a two-way conversational context. For the latter, a key feature is that we need to understand what is expected of a speaker such that their turn is relevant at the particular point in the conversation. As such, we expected to find a dissociation between narrative-based inferencing and comprehension of conversational implicature. Since Relevance Theory stipulates that all language is underdetermined and requires inferences to be made at the local level to enrich and disambiguate the utterance, we expected this type of inferencing to reflect core language competence.

In this task, 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.

**Mental state attribution** was measured using the Frith-Happé Animations (Abell *et al.*, 2000). Each animation shows 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; we did not use those animations depicting random movement. 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. We gave the original goal-directed practice item as our 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 3 by the first author 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 0.79 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 our measured variable for this task.

### Pilot study

Prior to our main study, we wanted to establish the reliability of our Implicature Comprehension Test and the distribution of scores expected in the general adult population. We also planned to assess how implicature processing relates to other language skills and to the broad autism phenotype.

We conducted an online study supported by Gorilla, using Prolific as a platform for recruiting participants in May 2018. We recruited 120 adults who reported speaking English as a first language and living in the UK. Individuals were excluded if they reported a significant uncorrected hearing or visual impairment. For our 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 completed four language tests: the Implicature Comprehension Test (ICT) and Synonyms Test, which were also administered in the main study, and also tests of syntax and idiom recognition (see below). In the pilot study, there was no time limit for responses on the ICT. Participants also completed the Autism-Spectrum Quotient (AQ), a self-report measure of autistic traits ([Baron-Cohen \*et al.\*, 2001](#)).

**Receptive syntax processing** was measured by the Test of Complex Syntax-Electronic (TESC-E; [Frizelle \*et al.\*, 2017](#); see <https://osf.io/5ntvc/> for full details). 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 we incorporated some adaptations to make it more appropriate

for adults. In the original version, accuracy was the variable of interest, though we expected a ceiling effect for accuracy, so we instead focused on mean reaction time. In making this change, we needed to present the sentence at the end of the video rather than during it, which was the original format; this meant we could time an individual’s response from the onset of the sentence. Furthermore, we 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 memory 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.

**Idiom recognition** was measured by an Idiom Decision Test 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. Our variable of interest was total accuracy (1 point per correct answer).

See [Table 1](#) for descriptive statistics and the reliability of our measures.

For each variable, we excluded scores according to the criteria of [Hoaglin & Ingleswicz \(1987\)](#) 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, we used listwise deletion of any cases with an outlying score. See [Table 2](#) for Spearman’s correlations between the five test variables.

The correlation matrix suggests that the synonym 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. Implicature scores on the ICT were not well-correlated with scores on any other tests. We cannot attribute this to noisy tests, as reliability was consistently good across the tests, as indicated by the Cronbach’s alpha values. Similarly, there was



**Table 1. Descriptive statistics, excluded values, and reliability coefficients (Cronbach's alpha and standard error of measurement).** ICT - Implicature Comprehension Test, RT – response time.

	N	Mean	SD	Min	Max	Skew	Kurtosis	Exclusions	Alpha	SEm
ICT, Accuracy on explicit items	118	9	0.63	8	10	-1.38	0.72	6, 1		
ICT, Accuracy on implicatures	118	27.2	4.08	14	35	-0.79	0.37	29, 2	0.80	1.82
Accuracy on synonyms	120	13.4	5.34	3	25	0.12	-1.03		0.85	2.07
Accuracy on idioms	113	35.8	3.19	26	40	-1.31	1.23	16, 18, 24, 23, 22, 20, 17	0.83	1.32
Accuracy on syntax	114	19.3	0.80	17	20	-1.01	0.46	16, 20, 17, 16, 13, 11		
Mean RT on syntax (ms)	114	3320	447	2188	4721	0.77	0.81	2780, 5375, 6131, 4334, 3572, 1231	0.84	178.64
Total score (out of 50) on AQ	118	20.7	6.92	5	39	0.33	-0.23	NA, NA	0.79	3.17

**Table 2. Correlations between language tests and Autism-Spectrum Quotient (AQ).** There has been listwise deletion of cases with outlying scores (N = 108). RT – response time.

	Accuracy on synonyms	Accuracy on idioms	Mean RT on syntax	Total score on AQ
Accuracy on implicatures	-0.04	-0.10	0.01	-0.09
Accuracy on synonyms		0.49	0.25	0.04
Accuracy on idioms			0.24	0.12
Mean RT on syntax				0.02

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 skills measured by the test battery. Likewise, the AQ was not closely related to the other tests, suggesting that the broad autism phenotype was not associated with language skills tested here. This was unexpected, as we thought AQ scores would be negatively related to ICT scores, as autism has been related to difficulties processing implied and indirect meanings. As noted in the main body of the paper, the lack of correlations between the ICT and other tests left us unsure what the test was measuring: very task-specific skills or a more general competence in conversational understanding? This question motivated the follow-up work reported in the main study.

#### Data analysis

Data were analysed using R version 3.4.0 (R Core Team, 2017); We used the following R packages as part of our data analysis: *psych* version 1.7.8 (Revelle, 2017), *mirt* version 1.28 (Chalmers, 2012), *MVN* version 5.5 (Korkmaz *et al.*, 2014), *rcompanion* version 1.13.2 (Mangiafico, 2018), and *lavaan* version 0.6.4.1348 (Rosseel, 2012). Plots were produced using the following

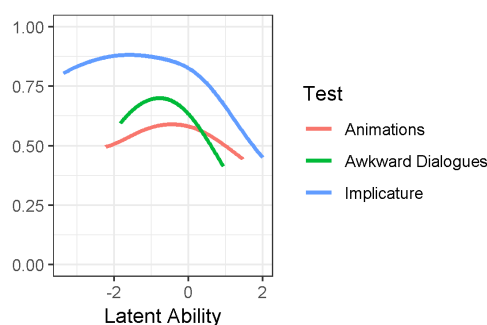
packages: *ggplot2* version 2.2.1 (Wickham, 2009) and *SemPlot* version 1.1 (Epskamp & Stuber, 2017). *knitr* version 1.21 (Xie, 2017), *papaja* version 0.1.0.9842 (Aust & Barth, 2018), *weights* version 0.85 (Pasek, 2016), and *htmlTable* version 1.11.2 (Gordon *et al.*, 2018) were used to produce the *R*mark-down report for this project. Our data are accessible on OSF (Underlying data (Wilson & Bishop, 2019)).

We assessed item functioning and the reliability of our measures using classical test theory (CTT) and item response theory (IRT; Embretson & Reise, 2000). The purpose of this analysis was to reduce the amount of measurement error in our tests, so that scores reflected as far as possible participants' "true scores" in the particular domain being measured. This was particularly important as our tests were novel, and not standardised measures with established psychometric properties. We therefore needed to establish that they were good quality measures. (Note that while there are standardised measures of vocabulary and grammar available, these generally can't be replicated in an online format due to copyright – and indeed, are not validated for online use.) For each item in each test, we inspected accuracy and the correlation between item-accuracy and total-accuracy on the test with that item excluded (all item-total correlations

calculated in this study used test totals with the item excluded). Items were identified as poor if they had low accuracy and a low item-total correlation. We 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, with no consensus answer at any point along the ability spectrum. We excluded any items showing this pattern. We then computed the reliability of our measures using CTT coefficients (Cronbach's alpha and standard error of measurement). We report these below, alongside a summary of the item-total correlations and item-level accuracy for each test; we summarise these as median and upper and lower quartiles. Please note that this CTT reliability analysis was not done for the Awkward Dialogues or Frith-Happé Animations, as we derived factor scores from IRT models for these tests, and therefore used IRT reliability analysis (explained with Figure 2 below).

Any individual with at least one outlying score on any test was excluded from the dataset. Outliers were defined according to Hoaglin & Iglewicz (1987) as 2.2 times the interquartile range below the lower quartile. We 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 Task (4/5), as these were taken to be control variables identifying individuals who did not engage well with the tasks. Then we inspected the data for univariate normality and multivariate outliers. Multivariate outliers were defined as individuals whose adjusted Mahalanobis' distance was above the 97.5th percentile of the chi-distribution. Where necessary, we transformed variables using the Tukey ladder of power transformations to reduce skew.

We had one preregistered analysis for this study: a confirmatory factor analysis (CFA). We specified a two-factor correlated-traits model with a "core language" factor and a "social understanding" factor; see Table 3 for which variables were set to load on which factor. We 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



**Figure 2. Reliability curves based on item-response theory (IRT) modelling.** These curves show test reliability across the ability spectrum for implicature comprehension, sensitivity to social awkwardness and mental state attribution.

better than a one-factor model on which all seven variables were set to load on a single factor. The conventional alpha level of 0.05 was used to indicate statistical significance, as we only preregistered one statistical test. We report confirmatory fit indices (CFIs) and root mean square error of estimation (RMSEA) with 90% confidence intervals.

We carried out one exploratory analysis. This was intended as a test of the extent to which implicature comprehension overlapped with core language skills (with which we expected minimal overlap) and other tests involving social understanding/inferential skills (with which we expected more overlap). Essentially, we were looking at whether the ability measured by the implicature comprehension test was more specific or more general, as a means of understanding more about the construct we were measuring. As such, we quantified the proportion of variance in implicature scores predicted by our tests tapping social understanding/inferential skills over and above the effect of core language skills. We ran a hierarchical multiple regression, entering receptive vocabulary and receptive grammar as predictors of implicature comprehension in the first stage. In the second stage, comprehension of fillers/backchannels, sensitivity to social awkwardness, narrative-based inferencing, and mental state attribution were added to the model. We report F-statistics, p-values, and adjusted R-square values for the stages.

## Results

Table 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 6.61 out of 10 (SD= 2.51).

## Reliability analysis

When assessing item functioning, we found a few weak items in the grammaticality 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. We did not detect any issues with item functioning in any other test.

The final versions of the tests were reliable, as indicated by the CTT reliability coefficients shown in Table 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.

See Figure 2 for reliability plots of the tests that we analysed using IRT modeling. IRT models compute standard error of

**Table 3. Study variables for each factor.**

Test	Study Variable	Details
<b>Social Understanding</b>		
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
FrithHappé Animations	Mental State Attribution	Factor score derived from IRT model
<b>Core Language</b>		
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)

**Table 4. Descriptive Statistics.** We report all excluded values for each study variable. (One excluded value for Sensitivity to Social Awkwardness is NA, as the participant provided no responses to any dialogue.)

	N	Mean	SD	Minimum	Maximum Achieved	Maximum Possible	Skew	Kurtosis	Exclusions
Implicature	117	28.9	4.50	8	36	36	-1.20	2.79	13 20 27
Fillers/Backchannels	120	26.8	5.93	7	37	40	-0.83	0.79	
Social Awkwardness	118	0.01	0.75	-1.83	0.94	0.94	-0.57	-0.57	-1.1 NA
Mental State Attribution	120	0.00	0.74	-2.23	1.47	1.47	-0.38	-0.18	
Inferencing	120	34.4	3.24	24	40	40	-0.73	0.30	3 12 19 19 20
Grammar	118	35.2	4.21	25	44	44	-0.52	-0.37	15 19
Vocabulary	120	12.1	4.45	3	24	25	0.31	-0.26	

**Table 5. Reliability Analysis, including Cronbach's alpha, standard error of measurement (SEm), corrected item-total correlations (totals excluding the item), and item-level accuracy.**

Study Variable	Alpha	SEm	Item-total correlations			Item-level Accuracy		
			Lower quartile	Median	Upper quartile	Lower quartile	Median	Upper quartile
Implicature	0.80	2.01	0.23	0.30	0.38	0.68	0.86	0.91
Fillers/Backchannels	0.80	2.65	0.2	0.27	0.35	0.51	0.67	0.85
Inferencing	0.75	2.63	0.36	0.27	0.36	1.6	1.76	1.86
Grammar	0.79	1.93	0.15	0.27	0.36	0.76	0.89	0.93
Vocabulary	0.76	2.18	0.22	0.31	0.35	0.28	0.49	0.58

measurement as a function of participant ability, and so we can use this to estimate reliability [ $SEm = SD \cdot \sqrt{1 - \text{reliability}}$ ]. We present reliability curves for the Awkward Dialogues and Frith-Happé Animations, and also include the IRT curve for the ICT, as we were particularly interested in how this task functioned across the ability spectrum. The low number of items (only five) seems to have limited the reliability of the Awkward Dialogues and Frith-Happé Animations. This IRT analysis is only valid insofar as a unidimensional model fits the data, so we report fit indices for these models: Awkward Dialogues CFI= 1.00, RMSEA= 0.06; Frith-Happé Animations CFI= 0.84, RMSEA= 0.00; ICT CFI= 1.00, RMSEA= 0.01.

### Confirmatory factor analyses

As shown in Table 2, a few scores were identified as outliers. We 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. Therefore, the six variables were transformed. Please note that the CFAs reported below were also run using the non-transformed data, and results were very similar. We decided to transform the data because several multivariate

outliers were identified in the non-transformed data. Following transformation, there were no multivariate outliers.

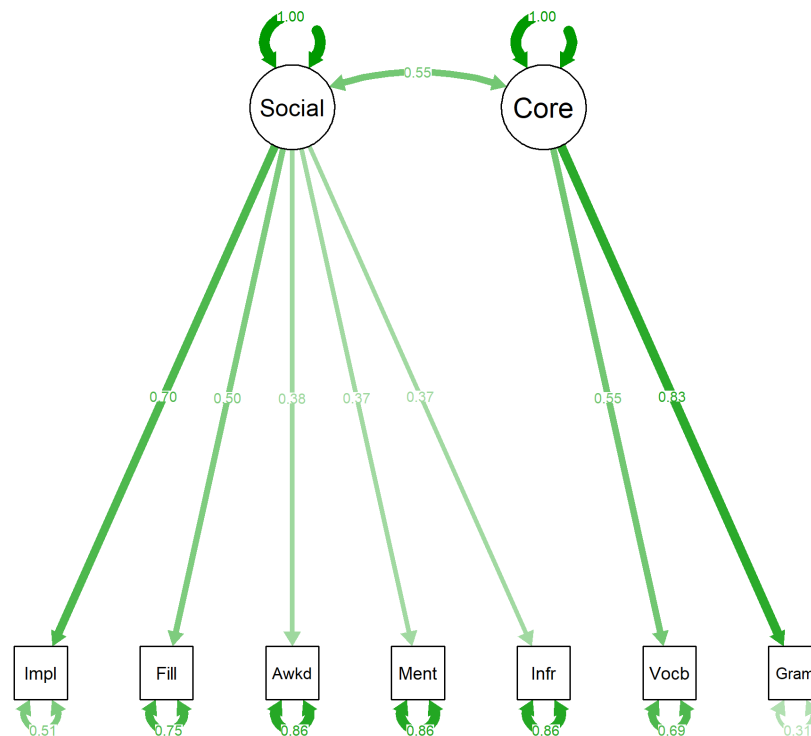
**Table 6** presents the correlation matrix for the transformed language variables.

We ran two CFAs comparing one and two factor models. A two-factor model fitted the data well (CFI = 0.95, RMSEA [90% CIs] = 0.05 [0, 0.11]), whereas a one-factor model did not (CFI = 0.87, RMSEA [90% CIs] = 0.07 [0, 0.13]). The Satorra-Bentler scaled chi-square difference test showed that the two-factor model gave significantly better fit,  $\chi^2(1) = 5.48$ ,  $p = 0.02$ .

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 inferencing and the pragmatic/social communication tests, indicating that these relationships were not well accounted for by the model. Therefore, we 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 = 1.00, RMSEA [90% CIs] = 0 [0, 0.08]). 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 3** for a visualisation of this two-factor model.

**Table 6. Correlations between Study Variables.** Variables have been transformed and there has been listwise deletion of cases to show only the data to be modelled in the Confirmatory Factor Analysis (N = 112).

	Fillers/Backchannels	Social Awkwardness	Mental State Attribution	Inferencing	Grammar	Vocabulary
Implicature	0.38	0.26	0.27	0.22	0.32	0.20
Fillers/Backchannels		0.11	0.19	0.16	0.27	0.11
Social Awkwardness			0.10	0.27	0.22	0.13
Mental State Attribution				0.20	0.09	0.20
Inferencing					0.13	0.13
Grammar						0.46



**Figure 3.** Final two-factor model, with one factor representing social understanding/inferencing and the other factor representing core language skills.

## Exploratory analysis

In addition to our preregistered analysis, we ran a hierarchical multiple regression to explore predictors of implicature comprehension. In the first stage, the full model was significant,  $F(2, 109) = 6.41$ ,  $p = 0.002$ . The adjusted R-square value indicated that receptive vocabulary and receptive grammar explained 8.89% of the variance in implicature comprehension. The addition of the rest of the predictors in stage 2 significantly increased the amount of variance explained,  $F(4, 105) = 5.60$ ,  $p < 0.001$ , and the full model remained significant,  $F(6, 105) = 6.23$ ,  $p < 0.001$ . However, together the variables only explained 22.04% of the variance in implicature comprehension. See [Table 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).

## Discussion

Our key preregistered research question was whether the processing of implied meaning (or implicature) in conversation was distinct from core language abilities (i.e. their grammar and vocabulary skills). The low correlations between our 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. We 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 our 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, we found evidence in the pattern of correlations that core language skills (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 our 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/offense 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, 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 based on heuristics about how people interact.

**Table 7. Coefficients for Multiple Regression, with Implicature Comprehension scores as the criterion variable.**

	Estimate	Standard Error	t-value	p-value
<b>Stage 1</b>				
Vocabulary	0.03	0.06	0.60	0.553
Grammar	0.14	0.05	2.86	0.005
<b>Stage 2</b>				
Vocabulary	0.01	0.06	0.21	0.835
Grammar	0.09	0.05	1.83	0.070
Inferencing	0.22	0.26	0.87	0.387
Social Awkwardness	0.04	0.02	1.64	0.104
Filler/Backchannels	0.04	0.01	3.03	0.003
Mental State Attribution	0.06	0.03	1.89	0.061

While core language skills and inferential skills are related (in our CFA, they were correlated at 0.55), the indication that a two rather than a one factor structure underpins performance on our 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., 2019; Hale & Tager-Flusberg, 2005; Miniscalco et al., 2014](#)). However, these are 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 language in the relative absence of problems with grammar and vocabulary ([Baird & Norbury, 2016](#)). Indeed, in follow-up work, we plan to administer the same test battery to autistic adults, with the expectation that they will find the tests specifically clustered within the “social understanding” factor more difficult. This



work will be useful in establishing whether our tests, especially the ICT, are useful in explaining real-life conversational difficulties. In our 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 our 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.

Alongside this follow-up work with autistic adults, we also plan to administer a similar test battery to children, as we are interested in how pragmatic processing (as measured by the ICT) and grammar and vocabulary skills might develop in a separable or co-dependent way. Given that we enter the linguistic environment without a lexicon, it may be that we depend on a continual interaction between pragmatic and core language skills, meaning that there may be few situations in which we rely on one domain in relative exclusion of the other. Our aim in creating the ICT was to isolate as much as possible pragmatic processing by limiting the demand on core language skills; this may be possible in adults, as the linguistic code may be stored in a more self-encapsulated way as it represents “crystallized” knowledge. However, in children there may be continual interaction between core language, pragmatics, other social-cognitive skills, logical reasoning etc., to allow for the greater possibility that they might encounter unfamiliar language. This potentially differing architecture of mind may mean that pragmatic processing in the ICT is less dissociable from grammar and vocabulary skills in children. It all depends on how “modular” we take the different functions to be; certainly, pragmatic processing of communicative stimuli is seen as modular by Relevance Theory (Carston & Powell, 2008), so with regard to the ICT, for instance, it may be that the tendency to search for relevant implicit meanings is dissociable from core language skills from an early age. This has repercussions for our understanding of developmental conditions involving a pragmatic impairment, like autism. If we assume that pragmatic and core language skills highly interact during development, then we might expect problems with pragmatics to have knock-on effects on acquisition of the linguistic code, but that there might be a piggybacking of skills in one domain on the other. This would mean that pragmatic skills would likely develop slowly but along a normal course in people with autism, especially those with well-developed core language skills. However, if pragmatic processing really is modular, then there may be more fundamental differences for those with a pragmatic impairment in how they process language involving implied meanings. There may be some compensation in these individuals which allows them to process language (as all language involves pragmatic interpretation, according to Relevance Theory), but this may involve effortful, error-prone or otherwise atypical processing. This remains speculative, and our future work may shed some light on these questions.

An unexpected finding in this study was the relatively low correlations between the language tests. Reliability of the tests was good, so this 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”/mentalizing skills. However, the lack of correlation here indicates that, at least in the general population, individual differences in performance on these tests are not accounted for by a general social cognition factor but rather by much more task-specific skills. Our findings agree with research in children and adolescents, which has found low correlations between several advanced “theory-of-mind” measures (Hayward & Homer, 2017). This questions the coherence of “theory-of-mind” as a single construct, and it would be worthwhile exploring this issue in future research using factor analysis.

One objective of this study was to understand more about what our test of implicature comprehension measured. Core language skills, such as vocabulary and grammar, accounted for only a small proportion of variance in scores on our Implicature Comprehension 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, without processing an implied meaning. Such individuals would be assumed to include those with autism and related conditions, who are found to have 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). We also found that an individual’s ability to make inferences, as measured in several of our 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 our 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 implicature scores unexplained. What skills might explain why people varied in their scores on the test? We have no categorical answers to this, but there are a couple of things to bear 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 more effortful processing involved in the latter case and also more intuitive “gut-based” responses in the latter. 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? We 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. We hope to explore this question in future research through assessing the relationships between implicature comprehension and novel tasks that involve sensitivity to the principle of relevance. One possible task might involve having participants make judgements on how relevant conversational turns are – e.g. whether the turn provides too much or insufficient information in the context of the conversation. We might expect individuals who are sensitive to utterances that are optimally relevant in their context to also be adept at picking up implied meanings suggested by the context.

In summary, this study presents evidence that understanding language in its communicative context is not simply a matter of core language skills. In particular, we found that understanding implicated meanings in conversation is somewhat distinct from vocabulary knowledge and grammatical competence. This raises the question of whether individuals with autism and 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. Our future work will explore this question.

## Data availability

### Underlying data

Open Science Framework: Structural and pragmatic language processing in adults. <https://doi.org/10.17605/OSF.IO/XN48E> (Wilson & Bishop, 2019)

This project contains the following underlying data:

- Data
  - alldataAnalyse.R (script for collating data, and running confirmatory factor analyses and exploratory regressions)
  - AwkwardConvo.csv (data for the Awkward Dialogues)
  - Backchannel.csv (data for the Test of Fillers and Backchannels)
  - Data\_dictionary.xlsx (spreadsheet detailing contents of each data file; one sheet per file)
  - Grammar.csv (data for the Grammaticality Decision Test)

- Implicature.csv (data for the Implicature Comprehension Test)
- Implicature\_itemcodes.csv (spreadsheet detailing item-type for each item in the Implicature Comprehension Test)
- Inferencing.csv (data for the Test of Local Textual Inference)
- TOMAnimations.csv (data for the Frith-Happé Animations)
- Vocab.csv (data for the Synonyms Test)

### Extended data

Materials for tests devised for this study are being developed as an assessment tool that we hope will be sensitive to pragmatic impairments in individuals with social communication difficulties, but we still need to establish validity of the tests in clinical groups before they are made available. We are also concerned that open availability of the materials may reduce their usefulness if participants have already viewed them prior to testing. However, we are happy to share our materials with other researchers wishing to use them. Please contact the corresponding author, with an explanation of why access is sought.

Information for researchers wishing to gain access to the Frith-Happé Animations is available here: <https://sites.google.com/site/utafirth/animations>. Any researchers wishing to use these animations should contact Sarah White ([s.white@ucl.ac.uk](mailto:s.white@ucl.ac.uk)).

Open Science Framework: Structural and pragmatic language processing in adults. <https://doi.org/10.17605/OSF.IO/XN48E> (Wilson & Bishop, 2019)

This project contains the following extended data:

- MaterialsGUIDE-MATERIALS.odt (this document provides a description of the files available under Materials, and provides item-level statistics for some test items.)
- Awkward Dialogues
  - AwkwardDialogue.mp3 (example item for the Awkward Dialogues)
- Implicature Comprehension Test
  - Practice\_Item\_1.mp4 (Example practice trial of the Implicature Comprehension test)
  - ICT1.mp4 (Example implicature item)
  - ICT2.mp4 (Example implicature item)
  - ICT3.mp4 (Example implicature item)
  - ICT4.mp4 (Example implicature item)
- Test of Fillers/Backchannels
  - Filler1.mp4 (Example item)

- Filler2.mp4 (Example item)
- Filler3.mp4 (Example item)
- Filler4.mp4 (Example item)
- Filler5.mp4 (Example item)

- Filler6.mp4 (Example item)

- Test of Local Textual Inference
  - TestLocalTextualInference.docx (Full materials for the narrative inferencing test)

## References

- Abell F, Happé F, Frith U: **Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development.** *Cogn Dev.* 2000; 15(1): 1–16.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Adams C: **Social communication intervention for school-age children: rationale and description.** *Semin Speech Lang.* 2005; 26(3): 181–188.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Adams C, Clarke E, Haynes R: **Inference and sentence comprehension in children with specific or pragmatic language impairments.** *Int J Lang Commun Disord.* 2009; 44(3): 301–318.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Akbar M, Loomis R, Paul R: **The interplay of language on executive functions in children with ASD.** *Res Autism Spectr Disord.* 2013; 7(3): 494–501.  
[Publisher Full Text](#)
- Andrés-Roqueta C, Katsos N: **The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders.** *Front Psychol.* 2017; 8: 996.  
[PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
- Ariel M: **Defining pragmatics.** Cambridge, UK: Cambridge University Press. 2010.  
[Publisher Full Text](#)
- Aust F, Barth M: **papaja: Prepare reproducible APA journal articles with R Markdown.** R package version 0.1.0.9842. 2018.  
[Reference Source](#)
- Baird G, Norbury CF: **Social (pragmatic) communication disorders and autism spectrum disorder.** *Arch Dis Child.* 2016; 101(8): 745–751.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Baron-Cohen S, O'Riordan M, Stone V, *et al.*: **Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism.** *J Autism Dev Disord.* 1999; 29(5): 407–418.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Baron-Cohen S, Wheelwright S, Skinner R, *et al.*: **The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians.** *J Autism Dev Disord.* 2001; 31(1): 5–17.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Bernard S, Deleau M: **Conversational perspective-taking and false belief attribution: a longitudinal study.** *Br J Dev Psychol.* 2007; 25(3): 443–460.  
[Publisher Full Text](#)
- Bishop DV: **Children's Communication Checklist (CCC-2).** Version 2, London: Psychological Corporation. 2003.  
[Publisher Full Text](#)
- Bishop DV, Adams C: **Comprehension problems in children with specific language impairment: literal and inferential meaning.** *J Speech Hear Res.* 1992; 35(1): 119–129.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Botting N, Adams C: **Semantic and inferencing abilities in children with communication disorders.** *Int J Lang Commun Disord.* 2005; 40(1): 49–66.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Carston R: **Informativeness, relevance and scalar implicature.** In Carston R & Uchida S (Eds.), *Relevance theory: Applications and implications.* Amsterdam: John Benjamins. 1998; 179–236.  
[Publisher Full Text](#)
- Carston R, Powell G: **Relevance Theory—New Directions and Developments.** In Lepore E & Smith BC (Eds.), *The Oxford Handbook of Philosophy of Language.* Oxford: Oxford University Press. 2008; 341–361.  
[Publisher Full Text](#)
- Castelli F, Happé F, Frith U, *et al.*: **Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns.** *NeuroImage.* 2000; 12(3): 314–325.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Chalmers RP: **mirt: A Multidimensional Item Response Theory Package for the R Environment.** *J Stat Softw.* 2012; 48(6): 1–29.  
[Publisher Full Text](#)
- Clark HH, Fox Tree JE: **Using uh and um in spontaneous speaking.** *Cognition.* 2002; 84(1): 73–111.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Conti-Ramsden G, Crutchley A, Botting N: **The extent to which psychometric tests differentiate subgroups of children with SLI.** *J Speech Lang Hear Res.* 1997; 40(4): 765–777.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Cummings L: **Clinical pragmatics: a field in search of phenomena?** *Lang Commun.* 2007; 27(4): 396–432.  
[Publisher Full Text](#)
- Drummond K, Hopper R: **Back channels revisited: acknowledgment tokens and speakership incipency.** *Res Lang Soc Interac.* 1993; 26(2): 157–177.  
[Publisher Full Text](#)
- Embretson SE, Reise SP: **Item response theory for psychologists.** 2000; Mahwah NJ: Lawrence Erlbaum Associates Publishers.  
[Reference Source](#)
- Epskamp S, Stuber S: **semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output.** R package version 1.1. 2017.  
[Reference Source](#)
- Frizelle P, Thompson PA, Duta M, *et al.*: **The understanding of complex syntax in children with Down syndrome.** 2017.  
[Publisher Full Text](#)
- Garnham A, Oakhill J: **Discourse processing and text representation from a "mental models" perspective.** *Lang Cognitive Proc.* 1992; 7(3): 193–204.  
[Publisher Full Text](#)
- Gordon M, Gragg S, Konings P: **htmlTable: Advanced Tables for Markdown/HTML.** R package version 1.11.2. 2018.  
[Reference Source](#)
- Greenslade KJ, Utter EA, Landa RJ: **Predictors of Pragmatic Communication in School-Age Siblings of Children with ASD and Low-Risk Controls.** *J Autism Dev Disord.* 2019; 49(4): 1352–1365.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Hale CM, Tager-Flusberg H: **Social communication in children with autism: the relationship between theory of mind and discourse development.** *Autism.* 2005; 9(2): 157–178.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Hayward EO, Homer BD: **Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence.** *Br J Dev Psychol.* 2017; 35(3): 454–462.  
[PubMed Abstract](#) [Publisher Full Text](#)
- Hoaglin DC, Iglewicz B: **Fine-tuning some resistant rules for outlier labeling.** *J Am Stat Assoc.* 1987; 82(400): 1147–1149.  
[Publisher Full Text](#)
- Jurafsky D, Shriberg E, Fox B, *et al.*: **Lexical, prosodic, and syntactic cues for dialog acts.** *ACL/COLING Workshop on Discourse Relations and Discourse Markers.* 1998; 114–120.  
[Reference Source](#)
- Klusek J, Martin GE, Losh M: **A comparison of pragmatic language in boys with autism and Fragile X syndrome.** *J Speech Lang Hear Res.* 2014; 57(5): 1692–1707.  
[PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
- Korkmaz S, Goksuluk D, Zararsiz G: **MVN: An R Package for Assessing Multivariate Normality.** *The R Journal.* 2014; 6(2): 151–162.  
[Publisher Full Text](#)
- Lam YG, Yeung SSS: **Towards a convergent account of pragmatic language deficits in children with high-functioning autism: depicting the phenotype using the pragmatic rating scale.** *Res Autism Spectr Disord.* 2012; 6(2): 792–797.  
[Publisher Full Text](#)
- Leinonen E, Ryder N, Ellis M, *et al.*: **The use of context in pragmatic comprehension by specifically language-impaired and control children.** *Linguistics.* 2003; 41(2): 407–423.  
[Publisher Full Text](#)

- Loukusa S, Leinonen E, Jussila K, *et al.*: **Answering contextually demanding questions: pragmatic errors produced by children with Asperger syndrome or high-functioning autism.** *J Commun Disord.* 2007a; **40**(5): 357–381.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Loukusa S, Leinonen E, Kuusikko S, *et al.*: **Use of context in pragmatic language comprehension by children with Asperger syndrome or high-functioning autism.** *J Autism Dev Disord.* 2007b; **37**(6): 1049–1059.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Loukusa S, Mäkinen L, Kuusikko-Gauffin S, *et al.*: **Assessing social-pragmatic inferencing skills in children with autism spectrum disorder.** *J Commun Disord.* 2018; **73**: 91–105.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Loukusa S, Moilanen I: **Pragmatic inference abilities in individuals with Asperger syndrome or high-functioning autism. A review.** *Res Autism Spectr Disord.* 2009; **3**(4): 890–904.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lucas R, Norbury CF: **Levels of text comprehension in children with autism spectrum disorders (ASD): the influence of language phenotype.** *J Autism Dev Disord.* 2014; **44**(11): 2756–2768.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mangiafico S: **rcompanion: Functions to Support Extension Education Program Evaluation.** R package version 1.13.2. 2018.  
[Reference Source](#)
- Matthews D, Biney H, Abbot-Smith K: **Individual differences in children's pragmatic ability: a review of associations with formal language, social cognition, and executive functions.** *Lang Learn Dev.* 2018; **14**(3): 186–223.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Miniscalco C, Rudling M, Råstam M, *et al.*: **Imitation (rather than core language) predicts pragmatic development in young children with ASD: a preliminary longitudinal study using CDI parental reports.** *Int J Lang Commun Disord.* 2014; **49**(3): 369–375.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Norbury CF: **Practitioner review: Social (pragmatic) communication disorder conceptualization, evidence and clinical implications.** *J Child Psychol Psychiatry.* 2014; **55**(3): 204–216.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Pasek J: **weights: Weighting and Weighted Statistics.** R package version 0.85. 2016.  
[Reference Source](#)
- Prutting CA, Kirchner DM: **A clinical appraisal of the pragmatic aspects of language.** *J Speech Hear Disord.* 1987; **52**(2): 105–119.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- R Core Team: **R: a language and environment for statistical computing.** Vienna, Austria: R Foundation for Statistical Computing. 2017.  
[Reference Source](#)
- Reichow B, Salamack S, Paul R, *et al.*: **Pragmatic Assessment in Autism Spectrum Disorders: A Comparison of a Standard Measure With Parent Report.** *Commun Disord Q.* 2008; **29**(3): 169–176.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Revelle W: **psych: Procedures for Personality and Psychological Research.** R package version 1.7.8. 2017.  
[Reference Source](#)
- Rosseel Y: **lavaan: An R Package for Structural Equation Modeling.** *J Stat Softw.* 2012; **48**(2): 1–36.  
[Publisher Full Text](#)
- Ryder N, Leinonen E: **Pragmatic language development in language impaired and typically developing children: incorrect answers in context.** *J Psycholinguist Res.* 2014; **43**(1): 45–58.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ryder N, Leinonen E, Schulz J: **Cognitive approach to assessing pragmatic language comprehension in children with specific language impairment.** *Int J Lang Commun Disord.* 2008; **43**(4): 427–447.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Schoen Simmons E, Paul R, Volkmar F: **Assessing pragmatic language in autism spectrum disorder: the Yale in vivo Pragmatic Protocol.** *J Speech Lang Hear Res.* 2014; **57**(6): 2162–2173.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sperber D, Wilson D: **Relevance: communication and cognition.** Oxford, UK: Blackwell. 1986.  
[Reference Source](#)
- Trouvain J, Truong KP: **Comparing non-verbal vocalisations in conversational speech corpora.** In Devillers L, Schuller B, Batliner A, Rosso P, Douglas-Cowie E, Cowie R, & Pelachaud C, (Eds.), *Proceedings of the 4th international workshop on corpora for research on emotion sentiment and social signals (es3 2012)*. Paris, France: European Language Resources Association (ELRA). 2012; 36–39.  
[Reference Source](#)
- Volden J, Coolican J, Garon N, *et al.*: **Brief report: pragmatic language in autism spectrum disorder: relationships to measures of ability and disability.** *J Autism Dev Disord.* 2009; **39**(2): 388–393.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Volden J, Phillips L: **Measuring pragmatic language in speakers with autism spectrum disorders: Comparing the children's communication checklist-2 and the test of pragmatic language.** *Am J Speech Lang Pathol.* 2010; **19**(3): 204–212.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Whyte EM, Nelson KE: **Trajectories of pragmatic and nonliteral language development in children with autism spectrum disorders.** *J Commun Disord.* 2015; **54**: 2–14.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wickham H: **ggplot2: Elegant Graphics for Data Analysis.** New York, NY: Springer-Verlag. 2009.  
[Publisher Full Text](#)
- Wilson A, Bishop DVM: **Structural and pragmatic language processing in adults.** 2019.  
<http://www.doi.org/10.17605/OSF.IO/XN48E>
- Wilson A, Bishop DVM: **Structural and pragmatic language processing in adults.** 2018.  
[Publisher Full Text](#)
- Xie Y: **knitr: A General-Purpose Package for Dynamic Report Generation in R.** R package version 1.17. 2017.  
[Reference Source](#)
- Young EC, Diehl JJ, Morris D, *et al.*: **The use of two language tests to identify pragmatic language problems in children with autism spectrum disorders.** *Lang Speech Hear Serv Sch.* 2005; **36**(1): 62–72.  
[PubMed Abstract](#) | [Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status:   

---

## Version 2

Reviewer Report 29 July 2019

<https://doi.org/10.21956/wellcomeopenres.16739.r35947>

© 2019 Swineford L. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Lauren Swineford**

Washington State University, Pullman, WA, USA

I appreciate the changes made in response to comments from all reviewers. No further changes are requested.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Social communication and language development.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 19 July 2019

<https://doi.org/10.21956/wellcomeopenres.16739.r35946>

© 2019 Matthews D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Danielle Matthews** 

Department of Psychology, University of Sheffield, Sheffield, UK

Thanks for addressing my questions.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Pragmatic development.



I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

---

**Version 1**

Reviewer Report 21 June 2019

<https://doi.org/10.21956/wellcomeopenres.16598.r35476>

© 2019 Alcock K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Katie Alcock** 

Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK

This is an interesting and useful paper examining the development and properties of a set of tests of pragmatic language and associated skills.

The introduction sets out the difficulty in assessing pragmatic language skills separate from other aspects of language ability, and other cognitive and social abilities, in people with autism (PWA). It also helpfully outlines the definition of pragmatics and how this definition is sometimes expanded by authors testing other social aspects of language. This topic is of interest both to those with a linguistics background and those from a cognitive framework and hence the definition of some basic terms is in order.

The methods are clearly designed but like another reviewer, I would really like more information (perhaps in an Appendix) about some of the stimuli. I get the point that participants might perform differently if pre-exposed to the tests but sample items (that are not in the actual tests) would be helpful to follow the thread of the paper.

Table 1 would benefit from presentation of the maximum (and minimum, where relevant) scores possible as well as the maximum achieved, to give an additional piece of information relevant to ceiling effects.

There are some very interesting implications in the Discussion for the coherence of Theory of Mind theory. It might be interesting to expand this issue a little bit more and/or say something about its coherence in children as well.

As the overall findings were that none of the other measures (except Filler/Backchannels) uniquely predicted variability in Implicature, it would be very helpful to know how the authors may intend to proceed to further unpack what abilities *are* related to Implicature.

In addition, the finding that tasks did not intercorrelate but were influenced by task-specific skills leads one to wonder if they are measuring things that are generalisable to everyday life? A little more discussion on this would be welcome.

A few minor points:

Pg 4 paragraph 1. I think the third sentence should read “I’m working all day” implicitly turns down SPEAKER TWO’s request to pick Sally up.’

Regarding participant characteristics:

The authors appear to be using “is/are” and “identify as” interchangeably. Participants can, in studies like this, self-declare more or less anything without proof (I am assuming no proof is asked for). The wording of questions is therefore quite important, especially since participants may decline to answer or answer differently depending on how a question is worded. In one study we asked younger parents about the wording of age-band questions in a study of infant language, for example, and younger parents did not always want to state their exact age due to feelings of being judged as a younger parent. One can envisage that a participant may *identify* as a student without in fact being one, and some individuals find it impossible or uncomfortable to “identify as” male or female as their sex is a matter of fact not identity. It is therefore helpful to know the precise source/wording of these questions (and also, for future research, to consider whether the wording used will get accurate and complete answers).

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Developmental psychology, language development, vocabulary development.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 11 Jul 2019

**Alexander Wilson**, University of Oxford, Oxford, UK

Thank you very much for your review. We respond below to any comments that have not been addressed in our responses to the other reviewers.

- “Table 1 would benefit from presentation of the maximum (and minimum, where relevant) scores possible as well as the maximum achieved, to give an additional piece of information relevant to ceiling effects.”

We assume you mean Table 4, which presents descriptive statistics for the main study. We have now included this information.

- “There are some very interesting implications in the Discussion for the coherence of Theory of Mind theory. It might be interesting to expand this issue a little bit more and/or say something about its coherence in children as well.”

We agree that there are implications for theory of mind. It should be noted, however, that the Awkward Dialogues was not designed specifically to measure theory of mind, so we need to be tentative here. Our discussion already refers to a study finding low correlations between different theory of mind measures in children. We have added a suggestion that future work could more systematically investigate the factor structure of theory of mind tasks.

- “As the overall findings were that none of the other measures (except Filler/Backchannels) uniquely predicted variability in Implicature, it would be very helpful to know how the authors may intend to proceed to further unpack what abilities *are* related to Implicature.”

We have no categorical answers to this question, but elaborate on this issue at the end of our Discussion.

- “In addition, the finding that tasks did not intercorrelate but were influenced by task-specific skills leads one to wonder if they are measuring things that are generalisable to everyday life? A little more discussion on this would be welcome.”

This is a fair criticism, and we appreciate that there are questions of ecological validity in relation to our tests. We would argue that the issue of the tests’ usefulness hinges on whether there are group differences between autistic and non-autistic groups in our follow-up work. If there are, that is a strong argument for the relevance of our tests.

- “Pg 4 paragraph 1. I think the third sentence should read “‘I’m working all day” implicitly turns down SPEAKER TWO’s request to pick Sally up.”

Thanks for pointing out the error in this sentence. We have amended SPEAKER TWO to SPEAKER ONE.

- Regarding participant characteristics: “The authors appear to be using “is/are” and “identify as” interchangeably. Participants can, in studies like this, self-declare more or less anything without proof (I am assuming no proof is asked for). The wording of questions is therefore quite important, especially since participants may decline to answer or answer differently depending on how a question is worded. In one study we asked younger parents about the wording of age-band questions in a study of infant language, for example, and younger parents did not always want to state their exact age due to feelings of being judged as a younger parent. One can envisage that a participant may *identify* as a student without in fact being one, and some individuals find it impossible or uncomfortable to “identify as” male or female as their sex is a matter of fact not identity. It is therefore helpful to know the precise source/wording of these questions (and also, for future research, to consider whether the wording used will get accurate and complete answers).”

Our demographic information was collected in response to open questions asking participants what their age, gender and race/ethnicity were. We hoped that this format would give participants the opportunity to identify in ways that felt most appropriate to them. We chose to avoid the term “identify” in the questions themselves, as this is likely to mean different things to different people – indeed, as you suggest. Thank you for pointing out that we use “are/is” and “identify as” interchangeably in describing our participants; this was motivated by a view that people are how they identify. However, for consistency, in our report, we have amended all uses of “is/are” to language around “identifying as”/“reporting”, as this is the nature of the information we have.

Regarding student status, this information is derived from a question asking participants whether they were currently studying. The response options were: “yes (at high school/secondary school)”, “yes (at university as an undergraduate)”, “yes (at university as a postgraduate)”, “yes (undertaking vocational training)” and “no”. The motivation behind this question was to see what proportion of our sample was students – due to the student bias in psychology samples.

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 23 May 2019

<https://doi.org/10.21956/wellcomeopenres.16598.r35325>

© 2019 Swineford L. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Lauren Swineford**

Washington State University, Pullman, WA, USA

This study examined the independence of pragmatic skills from core language skills in a sample of 120 adults. This study used newly developed language tasks including the Implicature Comprehension Test designed to measure comprehension of conversational implicature. This paper was well written, and the amount of detail provided across sections was much appreciated.

The introduction to this study provides a thorough overview of the theoretical background relevant to designing an assessment to solely measure pragmatic processing and provides sufficient rationale for why assessment of comprehension of conversational implicature could be used as a measure of pragmatic abilities that could be separable from core language abilities. The authors follow the theoretical discussion with a brief introduction into their study including clear hypotheses.

The information provided in the methods section is detailed and describes all important aspects for the reader; this paper is a great model of a complete methods section with the inclusion of statistical code. The proposed statistical plan is appropriate for the data and the research hypothesis. One concern I had in regard to the methods was related to the validity of some tasks to measure other aspects of language. For example, assessments such as the Synonyms Test or the Grammaticality Decision Test were noted to be devised for this study. The description of these assessments provides adequate detail about these tasks/assessments, but when examining the independence/relationship between different aspects of language (core vs. pragmatic), more well-known, validated assessments would strengthen the analyses and results. I would not expect the authors to recollect data using different assessments at this point as I do not see this as a fatal flaw, especially since they examine and report the item level functioning and reliability of measures, but perhaps including a discussion of why these newer assessments were used over currently available (with strong psychometrics) tools. Also, more information would be useful on what the authors believe the other assessments set to load onto the pragmatics factor measure as the goal was to test how well the ICT measures solely pragmatics rather than broader social communication (mental state, TOM, etc) which seem to be measured by other assessments included. Perhaps providing all item level information of the assessments in a supplement, rather than just a few examples in the methods, would help to interpret the factor loadings and correlations. The detail (including whether or not

assumptions for statistical analyses were met) was appropriate. Standing alone, it is not clear from Figure 2 what the different shades of color/types of lines are.

I appreciate the authors discussion of their findings, especially regarding what the pragmatic factor in their final two-factor model may measure given the loading of narrative-based inferencing. The authors do a nice job of discussing the results in light of the theoretical discussion laid out in the introduction. I would love to hear more discussion about the next steps based on the findings in the discussion.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Social communication and language development.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 11 Jul 2019

**Alexander Wilson**, University of Oxford, Oxford, UK

Thank you for your comments. Below we address any suggestions not captured in our response to the first reviewer.

- “One concern I had in regard to the methods was related to the validity of some tasks to measure other aspects of language. For example, assessments such as the Synonyms Test or the Grammaticality Decision Test were noted to be devised for this study. The description of these assessments provides adequate detail about these tasks/assessments, but when examining the independence/relationship between different aspects of language (core vs. pragmatic), more well-known, validated assessments would strengthen the analyses and results. I would not expect the authors to recollect data using different assessments at this point as I do not see this as a fatal flaw, especially since they examine and report the item



level functioning and reliability of measures, but perhaps including a discussion of why these newer assessments were used over currently available (with strong psychometrics) tools.”

As we were collecting data online, it was not possible to use existing measures due to copyright issues in presenting them online. It should also be noted that such measures would typically have been validated for face-to-face administration on a one-to-one basis, and so we cannot assume that they would have identical properties if given online. We have mentioned these issues in our Data Analysis section.

- “Also, more information would be useful on what the authors believe the other assessments set to load onto the pragmatics factor measure as the goal was to test how well the ICT measures solely pragmatics rather than broader social communication (mental state, TOM, etc) which seem to be measured by other assessments included. Perhaps providing all item level information of the assessments in a supplement, rather than just a few examples in the methods, would help to interpret the factor loadings and correlations.”

Issues pertaining to this comment are more thoroughly discussed in our response to the first reviewer under points 3 and 4. However, to briefly respond to this comment, we have clarified what skills we intended to measure, and how these reflect our model of language processing, in the Introduction, and we provide more examples of items at different levels of difficulty in Extended Data. As noted in our paper, we have ongoing work, so we prefer not to make our tests fully available in case participants see them prior to testing. However, we hope that showing more examples will help the reader conceptualise what the tests are measuring. Please note that Table 5 presents some useful item-level statistics.

- “I would love to hear more discussion about the next steps based on the findings in the discussion.”

We briefly discuss related work with autistic adults and with children in our Discussion.

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 29 April 2019

<https://doi.org/10.21956/wellcomeopenres.16598.r35322>

© 2019 Matthews D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Danielle Matthews**

Department of Psychology, University of Sheffield, Sheffield, UK

I enjoyed reading this paper. Pre-registration and access to the data and R script was a plus. With the exception of the request for more details about the stimuli/items (point 4 – the most important point for me), the following comments are mainly points for discussion (that could be accommodated with clarification in the discussion section or with minor revision to the introduction) not suggestions for substantial revision. I’ve commented on the paper with a focus on the logic of the arguments put forward. While I have looked at the R script and the data, I haven’t run the script myself. I haven’t commented on the stats beyond a couple of questions (about the relation between test development and factor analysis, and about how to interpret the factor analysis).

The goal of this study was “to devise novel language tests to give a more convincing answer to the question of whether core language abilities and sensitivity to social aspects of language are separable sets of skills in the general adult population.” A definition of pragmatics is given and the need to contrast this with semantics discussed. “While semantics involves decoding conventional “dictionary” meaning, pragmatics is all about inference: we use context to infer further non- codified meaning.” While there are theoretical issues here (relating to what ‘conventional dictionary meaning’ is), this is a working definition. Particularised conversational implicatures are settled on as the ideal domain for testing pragmatics given their dependence on communicative context. The goal was to use a p.c. implicature test to find out how far pragmatic skill is separable from other language abilities.

1. It is argued that a focus on p.c. implicatures would be beneficial since conceptually it is the most distinctively pragmatic skill (whereas other skills that are nominally pragmatic have more potential overlap with, e.g., semantics). Does this mean that the results (finding a 2 factor solution) are not necessarily what would be found if we looked at other skills traditionally considered as pragmatic? I wondered to what extent picking the most distinctively pragmatic skill (the one at the far end of the continuum, so to speak) makes it difficult to assess the main question of whether core language abilities are separable from pragmatic abilities (in general).
2. Likewise, does use of CTT and IRT (where during test development items are excluded for poor fit) bias things in favour of finding separable factors? Essentially, the steps here are to devise a test, e.g., with a set of items that theoretically at least all require p.c. implicature. Then any items that do not pattern with most of the others are left out. This is desirable for test development. But I did wonder whether doing this at the same time as factor analysis essentially means that you have chosen items on the basis that they pattern together and then shown in factor analysis that they do. I doubt given the number of items excluded that this is really a problem but it might be worth clarification (whether this is an issue in principle or not, whether it is an issue in practice or not).
3. Given the goal of testing whether formal language skills group separately to pragmatic language skills, to what extent are the four ‘conversation comprehension’ tests cleanly tests of pragmatics? I can see the implicature test is a test of pragmatics (at least as far as it is possible to have a clean test, there is a convincing rationale here). But why are the Frith-Happe’ Animations – a test of mental state attribution intended to be as non-verbal as possible - taken as a measure of conversation comprehension? The discussion recognised this as a somewhat separate test. It seemed odd to go to great lengths to attempt to identify a cleanly pragmatic test (to the exclusion of tests of other nominally pragmatic skills) and then to include a test traditionally seen as tapping ToM and not involving conversation comprehension at all.  
Likewise, for the Awkward Dialogues, p.6 reads “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.” Could you clarify why this would be helpful for addressing the question at hand?  
It seems ‘pragmatics’ – already a broad tent - is being stretched further than necessary here. To find out whether that is the case perhaps exploratory factor analysis would help but a broader range of tests would be needed. For this paper, perhaps some further rationale could be given along with a revision to the introduction. (I had followed the introduction – largely focused on the implicature test – and then was surprised when I came to Table 3 at the choice of conversation comprehension measures).
4. Provision of test items. It was great to be able to read this paper, look at the pre-reg document, data and R code. What would have made it really great, and what is perhaps more important for conceptualising the findings, would have been able to see the full list of items from the tests. Without this, I found myself struggling to imagine what was going on exactly. The example p.c. implicatures given seem fairly straightforward such that I guess most of the adults who participated

answered correctly. However, there was variance for this measure, so some of the inferences must have been harder to make. I'd have been interested to get a sense of these items, not least because it is a challenge to make a difficult inference test item that has a correct answer. It is reported that some items on this test were more difficult than others and that scores for a given difficult item correlated well with total scores excluding that item, indicating that they reliably tapped a particular skill. I'd like to better understand what the skill is exactly. I guess it could be distracting to give one example that a reader could take issue with for reasons that are beside the point, but given the problems particular to this sub-field, being able to concretely point to what items had in common seems important. What would be really valuable would be to see the full set of stimuli with the difficult items highlighted so that we could try to imagine what those items have in common.

**Suggested revision:** Ideally, the full stimuli sets/list of items would be given so that we can understand the tasks. It is noted in a final section of the paper ('Extended data') that stimuli are not provided as test development is underway and future participants might access materials. I would be interested to see the full list but, for publication, would it be possible to give a few example items from the harder end of each test, with information about what % of participants got them 'right'? This is needed for the implicature, awkwardness, fillers and narrative tests.

More generally – perhaps for discussion – does it always make sense to say there is a correct answer? Or just one that most people alight on? For example, with the filler/give the floor test, there are probabilistic cues to who will speak next. Is the correct answer just the answer that most people would give? Are we looking at norms of behaviour (a participant is correct by virtue of doing what most other people would do) or is there a clear inferential process that should optimally be engaged in?

5. P.9 "Any individual with at least one outlying score on any test was excluded from the dataset". How many were excluded? Could this be noted in the participants section also?
6. In the results section, it wasn't entirely clear what question the exploratory analysis was intended to address. Could this be spelt out somewhere?
7. **Conclusions drawn.** I was interested in the discussion on p. 12 about what the second factor (non-grammar factor) represents. I wondered if it might be possible to interpret the CFA a little for people who don't use it regularly. Does this analysis show there is a second factor that holds together or is it just that a two factor solution is better than having just one factor, one sensible factor is grammar (with vocab) and another dimension is the inferencing test (with other tests also grouping with this but not really strongly related to the second factor). That is, this seems to show that p.c. implicature and grammar are quite distinct. But does it tell us that p.c. implicature and the other tests that are nominally tapping "pragmatic/communication" should be conceptualised as a single factor 'inferring meaning through integrating information in context'? Looking at the correlation matrix and the loadings, these variables seem quite loosely related. In answer to the main question of "whether core language abilities and sensitivity to social aspects of language are separable sets of skills in the general adult population", it seems the answer is yes but apparently only in so far as grammar is distinct from p.c. inferencing (which is reflected in the title and final conclusions). What everything else is doing, and what elements of the other nominally conversation measures are pragmatic (or social aspects of language use) seemed less clear to me. Some additional text explaining what this factor analysis allows us to conclude here (and how future studies could further test whether there is a pragmatics factor and if so what it reflects) would be very helpful.

In sum, the title and main conclusion about inferencing and grammar are supported but the question set up in the introduction had a broader scope and the conclusion wrt to that broader question could be clarified.

8. I was interested in the implications of this study for development. Would you predict separation of vocabulary/grammar and pragmatics earlier in development? Given that children start off without a

'dictionary/code/lexicon', perhaps some consideration of the co-development of these skills would be helpful for understanding developmental disorders.

9. Typo: well-correlated twith >> with

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

No

**If applicable, is the statistical analysis and its interpretation appropriate?**

I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Pragmatic development.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 11 Jul 2019

**Alexander Wilson**, University of Oxford, Oxford, UK

Many thanks for your comments on the paper. Please see below for our responses.

- Regarding the focus on p.c. implicatures rather than other pragmatic skills: "I wondered to what extent picking the most distinctively pragmatic skill (the one at the far end of the continuum, so to speak) makes it difficult to assess the main question of whether core language abilities are separable from pragmatic abilities (in general)."

This is an interesting point, and I think comes down to the issue of what we mean by pragmatic skills. Our argument is that pragmatic comprehension involves context-dependent inference (as per Relevance Theory), and so testing the question of whether pragmatics is distinct from core language requires a "pure" measure of context-dependent inference. We observe that many existing tests of pragmatics conflate pragmatics with other domains, so testing the relationship between performance on these tests and core language skills will ultimately only tell us that tests involving a significant demand on core language skills interrelate. It is also interesting that you use the metaphor of a continuum for pragmatics (vs. semantics, I assume). I think the language model that we are testing would look more like a Venn diagram, with overlapping domains for core language and pragmatic processing. Most tests would tap the central overlapping section, but we are trying as much as possible to design a test that puts most of its cognitive demand on the

pragmatic domain. We have included more detail in our Introduction about our model of language processing (which is ultimately heavily influenced by Relevance Theory).

- "Does use of CTT and IRT (where during test development items are excluded for poor fit) bias things in favour of finding separable factors?"

The purpose of using reliability analysis was to minimise the amount of measurement error in each test. This would mean that scores on our tests of implicature and grammar, for instance, represented as far as possible participants' "true scores" in these domains and were less subject to random error caused by poor items that did not reflect the domain well (perhaps because they did not have a "right" answer or were ambiguous in some way). If we're looking to test the relationship between, say, implicature and grammar, it makes sense to ensure that the tests measure these constructs with as little error as possible. Indeed, if implicature and grammar are closely related, then we would expect that removing random error would increase, not decrease, correlations between domains, as error attenuates true correlations. As such, we're not sure that conducting reliability analysis leads to the bias described.

- "Given the goal of testing whether formal language skills group separately to pragmatic language skills, to what extent are the four 'conversation comprehension' tests cleanly tests of pragmatics?"

The tests (besides the implicature comprehension test) were included not so much as "clean" tests of pragmatics, but instead to establish convergent validity between the implicature test and other tests that we might expect to tap pragmatics (as well as other skills), based on the idea that pragmatics involves sensitivity to communication contexts and is thought to be a subdomain of "theory-of-mind" under Relevance Theory. We include some further clarification on this in our Introduction. Please note that in our pilot study we found that the implicature test was unrelated to anything, so we wanted to establish whether it did relate to other tests with a pragmatic demand, as a means of exploring the nature of the construct. We have also renamed this factor as "social understanding", as we feel that more appropriately reflects the tests, and also is more in line with the theoretical basis.

- "Ideally, the full stimuli sets/list of items would be given so that we can understand the tasks. It is noted in a final section of the paper ('Extended data') that stimuli are not provided as test development is underway and future participants might access materials. I would be interested to see the full list but, for publication, would it be possible to give a few example items from the harder end of each test, with information about what % of participants got them 'right'? This is needed for the implicature, awkwardness, fillers and narrative tests."

In writing up this paper, we were conscious of needing to restrict access to our stimuli as we have ongoing work, and wanted to reduce the possibility of participants seeing our tests in advance of testing. You're right, however, that we have not shared our materials sufficiently for the results to be fully interpreted. We have taken your advice and included as extended data five implicature items with item-level statistics. All five of the awkward dialogues were of approximately the same difficulty; we share one dialogue in extended data with the associated questions. We share a few videos for the fillers task, with item-level statistics. The inferencing test is less important to our follow-up work. The whole test is now available as extended data.

- "More generally – perhaps for discussion - does it always make sense to say there is a correct answer? Or just one that most people alight on?"

This is a valid point. To some extent we are looking at consensus answers with some of the tests. With the fillers task, this is particularly true – although, in many ways the consensus answer can also be seen as the "correct answer", as individuals who alight on the non-consensus answer are likely to find themselves disadvantaged during communication with individuals who are sensitive to normative cues. We have provided some speculation in the Discussion about the role of norms and formal reasoning in our implicature test.



- "How many [participants] were excluded? Could this be noted in the participants section also?"

We have included all excluded scores in our results table, and note the number of participants included in our analysis in Results. For completeness, we now include the number of excluded participants in the participants section too.

- "In the results section, it wasn't entirely clear what question the exploratory analysis was intended to address. Could this be spelt out somewhere?"

The exploratory analysis was intended to quantify the amount of overlap between the implicature test and the other tests (especially when accounting for core language skills). We felt this was important given that our pilot study had shown no relationships between the implicature test and other measures, so we wanted to establish the extent to which it was tapping a general ability that might be reflected in other tests of communication/social understanding. Essentially, we were asking the question as to how specific/general the test was. We have clarified this in the Results.

- In relation to the Discussion: "Does this analysis [...] tell us that p.c. implicature and the other tests that are nominally tapping "pragmatic/communication" should be conceptualised as a single factor 'inferring meaning through integrating information in context'? Looking at the correlation matrix and the loadings, these variables seem quite loosely related. [...] Some additional text explaining what this factor analysis allows us to conclude here (and how future studies could further test whether there is a pragmatics factor and if so what it reflects) would be very helpful. [...] In sum, the title and main conclusion about inferencing and grammar are supported but the question set up in the introduction had a broader scope and the conclusion to that broader question could be clarified."

We think it is fair to question the extent to which the tests grouped under the communication factor really represent one general ability in this domain. As you point out, the correlations between the tests were relatively low, indicating that task-specific skills were important in determining performance on individual tasks, with a more general inferential ability accounting for a small amount of the variance in scores. We have clarified this in the conclusion, and have more distinctly addressed the questions of (a) implicature dissociating from grammar/vocab and (b) core language dissociating from pragmatics.

- "I was interested in the implications of this study for development. Would you predict separation of vocabulary/grammar and pragmatics earlier in development? Given that children start off without a 'dictionary/code/lexicon', perhaps some consideration of the co-development of these skills would be helpful for understanding developmental disorders."

This is an interesting point, and is relevant to some of our follow-up work. We provide some exploration of this in our Discussion.

- "Typo: well-correlated twith >> with"

Corrected.

**Competing Interests:** No competing interests were disclosed.