PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Three distinct components of pragmatic language use: Social conventions, intonation, and world knowledge–based causal reasoning

Sammy Floyd[a,b,1,2], Olessia Jouravlev[c,1], Moshe Poliak[b], Zachary Mineroff[d], Edward Gibson[b,2,3], and Evelina Fedorenko[b,e,f,2,3]

Affiliations are included on p. 10.

Edited by Michael C. Frank, Stanford University, Stanford, CA; received November 30, 2024; accepted September 20, 2025 by Editorial Board Member Susan A. Gelman

Successful communication requires frequent inferences. Such inferences span a multitude of phenomena: from understanding metaphors, to detecting irony and getting jokes, to interpreting intonation patterns. Do all these inferences draw on a single underlying cognitive ability, or does our capacity for nonliteral language comprehension fractionate into dissociable components? Using an approach that has successfully uncovered structure in other domains of cognition, we examined covariation in behavioral performance on diverse nonliteral comprehension tasks across two large samples to search for shared and distinct components of pragmatic language use. In Experiment 1, n = 376 participants each completed an 8 h battery of 20 critical tasks. Controlling for general cognitive ability, an exploratory factor analysis revealed three clusters, which can be post hoc interpreted as corresponding to i) understanding social conventions (critical for phenomena such as indirect requests, conversational implicatures, and irony), ii) interpreting contrastive and emotional intonation patterns, and iii) making causal inferences based on world knowledge. This structure largely replicated in a new sample of n = 400 participants (Experiment 2, preregistered) and was robust to analytic choices. This research uncovers structure in the human communication toolkit and can inform our understanding of pragmatic difficulties in individuals with brain disorders. The hypotheses put forward here about the underlying cognitive abilities can now be evaluated in new behavioral studies, as well as using brain imaging and computational modeling, to continue deciphering the ontology of the component pieces of linguistic and nonverbal communication.

pragmatics | nonliteral language | communication | intonation | individual differences

Significance

Real-life language comprehension frequently requires nonliteral interpretation and inferences about speaker intent. What is the structure of these so-called pragmatic abilities? We applied a dimensionality reduction approach to a large behavioral dataset (776 participants, each completing an 8-hour battery of diverse nonliteral comprehension tasks). By examining covariation in performance across tasks, we identified three interpretable components of pragmatic language use: adherence to social conventions, extracting meaning from intonation, and causal reasoning based on world knowledge. Thus, pragmatic language use is relatively low-dimensional cognitively, and its distinct components may a) draw on dissociable neural substrates, b) exhibit distinct developmental trajectories and differential susceptibility to genetic brain disorders, and c) be variably challenging for artificial intelligence systems.

Language is a powerful cultural invention that enables us to share our knowledge, thoughts, and feelings with one another, and facilitates cooperative behaviors (1–3). Although effortless for most neurotypical healthy adults, communication through language requires a complex orchestration of perceptual, motor, linguistic, and cognitive processes, and the ontology of the component pieces of linguistic and nonverbal communication remains an active area of research (4–9). One important feature of linguistic exchanges is that they often require going beyond the literal meaning of the words. For instance, someone may describe a ruthless lawyer as a shark, or remark "Lovely weather!" when it is pouring rain. Across phenomena as diverse as metaphors, irony, indirect requests, and contrastive prosody, communicative success requires inferences about speaker intent, often based on contextual information (10–16). These kinds of inferences—pervasive in everyday language use—are often jointly referred to by the umbrella term "pragmatic inferences". However, the phenomenological diversity of nonliteral language raises the question of whether all pragmatic inferences are alike and draw on the same cognitive resources, or whether instead our capacity for nonliteral comprehension consists of dissociable skills.

Past theorizing has often emphasized shared features of different pragmatic phenomena and correspondingly similar cognitive requirements. For example, the Relevance Theory framework (12), building on Grice's earlier work (10), proposes a unified cognitive inferential operation, where inferences may draw on diverse knowledge sources, including general world knowledge. Other accounts focus more narrowly on social inferences, which require Theory of Mind, and which are argued to be inherent in many pragmatic phenomena (17–19). The same unifying flavor characterizes more recent, probabilistic reasoning accounts, including the Rational Speech Act framework (20–22), which is couched within the broader framework of rational inference approaches to cognition (23).

This article is a PNAS Direct Submission. M.C.F. is a guest editor invited by the Editorial Board.

[1]S.F. and O.J. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: sambfloyd@gmail.com, egibson@mit.edu, or evelina9@mit.edu.

[3]E.G. and E.F. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2424400122/-/DCSupplemental.

Published December 9, 2025.

These theoretical accounts make a testable prediction: Pragmatic phenomena that are hypothesized to rely on the same cognitive resources should show a positive relationship in behavioral performance (e.g., individuals who are good at understanding indirect requests should also be good at understanding irony), and should recruit similar neural mechanisms (e.g., the Theory of Mind brain areas if the critical shared operation is social inferencing; 24–26). However, these predictions remain largely untested. Instead, most studies in experimental pragmatics have focused on particular phenomena: for example, metaphors (27–29), scalar implicatures (30–32), conversational implicatures (33–35), or prosody comprehension (36–38). This trend also characterizes investigations of neurodivergent populations (39, 40) and neuroimaging investigations of nonliteral comprehension (41–43). As a result, to our knowledge, no experimental evidence exists to substantiate the hypothesized relationships among phenomena that are grouped under the umbrella of "pragmatic inference" (cf. a recent developmental study: 44) or to uncover dissociations in the space of nonliteral language comprehension.

To fill this gap in our understanding of the associations and dissociations among the mental computations related to pragmatic language use, we turn to the individual-differences approach (45, 46). In this approach, each participant performs diverse tasks, and then the correlational structure of task performance is examined across individuals. Strong positive relationships across tasks are taken as evidence that those tasks rely on the same cognitive and neural resources, and weak or no correlations are interpreted as evidence of distinct underlying resources. The approach was originally introduced by Spearman (47)[*] and led to the discovery of the "positive manifold"—the finding that performance on *all* tasks shows some degree of positive correlation. This shared component—termed "g" (for general ability or intelligence)—can be measured by standard IQ assessments (48, 49) and needs to be factored out when examining abilities in a particular domain. A key advantage of the individual-differences approach is that it allows for empirical, bottom–up discovery of which behaviors and cognitive abilities go together and which are distinct. This approach has been successful in uncovering structure across numerous domains—from vision (50), to audition (51), to numerical cognition (52), to executive functions (53), to intuitive physical reasoning (54). Inspired by these successes, we here apply this approach to probe the structure of pragmatic language use.

To foreshadow our critical finding, after factoring out general cognitive ability, our analyses of accuracy patterns revealed three clusters of phenomena, which replicated across experiments and analytic choices, and which, we hypothesize, correspond to three distinct underlying abilities: i) understanding social conventions, ii) interpreting intonation patterns, and iii) making causal inferences based on world knowledge.

Results

We carried out two large-scale web-based individual-differences experiments, where each participant completed diverse nonliteral language comprehension tasks, along with some control tasks, across four testing sessions. Experiment 1 (n = 376 participants) included 20 critical tasks spanning eleven phenomena. For Indirect Requests, Conversational Implicatures (sometimes referred to in past studies as "Gricean implicatures"), Irony, Polite Deceits (sometimes referred to as "white lies"), Metaphor, Humor, World Knowledge–Based Inferences, Contrastive Prosody, and Emotional Prosody, we included two tasks per phenomenon (usually a text-based and a picture-based version; Fig. 1), which allowed us to evaluate phenomenological construct validity (46). The remaining two phenomena—Scalar Implicatures and a particular Quantifier Scope Ambiguity ("every…not")—only had one task each. When possible, we used stimuli from past studies, often in adapted form (*SI Appendix*, section 1A). The control tasks included Kaufman Brief Intelligence Test (KBIT; 55), to factor out variability due to general cognitive ability, catch question trials at the end of each session to ensure that participants were attentive throughout the battery, and several other tasks and questionnaires for exploratory analyses (*SI Appendix*, section 1B; see *SI Appendix*, section 1C for the structure of Experiments 1 and 2).

Experiment 2 (n = 400 participants) served two goals. The first was to replicate the findings from Experiment 1 (preregistration: https://osf.io/dcvb8), so the tasks were largely the same, except for the exclusion of the Scalar Implicatures and Quantifier Scope Ambiguity tasks because they did not yield meaningful individual-level performance estimates for many participants (*Results*; *SI Appendix*, section 2A). The second goal was to test whether interindividual differences in auditory perceptual ability could partially explain the task grouping observed in Experiment 1. To this end, we added a low-level auditory task (tone discrimination; 56) and another critical task (an auditory version of an Irony comprehension task).

For the critical analyses, our main approach was Exploratory Factor Analysis (EFA) (47, 57), where underlying latent variables (factors) are found based on intertask correlations, and the influence of each factor on each task is quantified as a factor loading. Tasks that load on the same factor at 0.3 or higher are interpreted as tapping the same ability, which results in them consistently covarying across participants. A key advantage of EFA is that it models performance on each task as a combination of common and unique factors, which means that variance is not "forced" to be attributed to the common factors (cf. approaches such as Principal Component Analysis; 57). In this way, EFA allows for the discovery of possibilities as distinct as i) all phenomena load on the same factor (which would be the case if a single shared ability underlay performance on all nonliteral comprehension tasks), and ii) each phenomenon loads on its own factor (which would be the case if the ability to understand nonliteral language consisted of many distinct computations, each supporting a particular kind of inference depending on which nonliteral device is used). Whatever grouping of tasks is discovered can then give rise to hypotheses about the shared and distinct cognitive abilities underlying pragmatic language use.

The Critical Nonliteral Comprehension Tasks Exhibit Good Psychometric Properties. Good psychometric properties are essential for obtaining meaningful results in individual-differences investigations (46, 58). With the exception of two tasks in Experiment 1 (discussed below), all tasks across both experiments exhibited good psychometric properties. First, performance was variable across participants, falling in a dynamic range (*SI Appendix*, section 2A, i). Second, interitem consistency was high, as evidenced by split-half correlations across participants (*Methods*): average correlation was 0.45 in Experiment 1 (range across tasks: 0.22 to 0.92), and 0.61 in Experiment 2 (range: 0.36 to 0.94) (Fig. 2*A*; see *SI Appendix*, section 2A, iii for additional evaluation using Item Response Theory (IRT; 59, 60)). And third, the nine phenomena for which two tasks were included showed good construct validity, as assessed via Cronbach's alpha (61) (*Methods*; Fig. 2*B*). Furthermore, within-phenomenon correlations (e.g., a correlation between the text-based and picture-based versions of Indirect Requests) were

[*]This early work from Spearman does not discuss the relationship between intelligence and race, but it is important to acknowledge this work's role in subsequent pseudoscientific racist discourse.

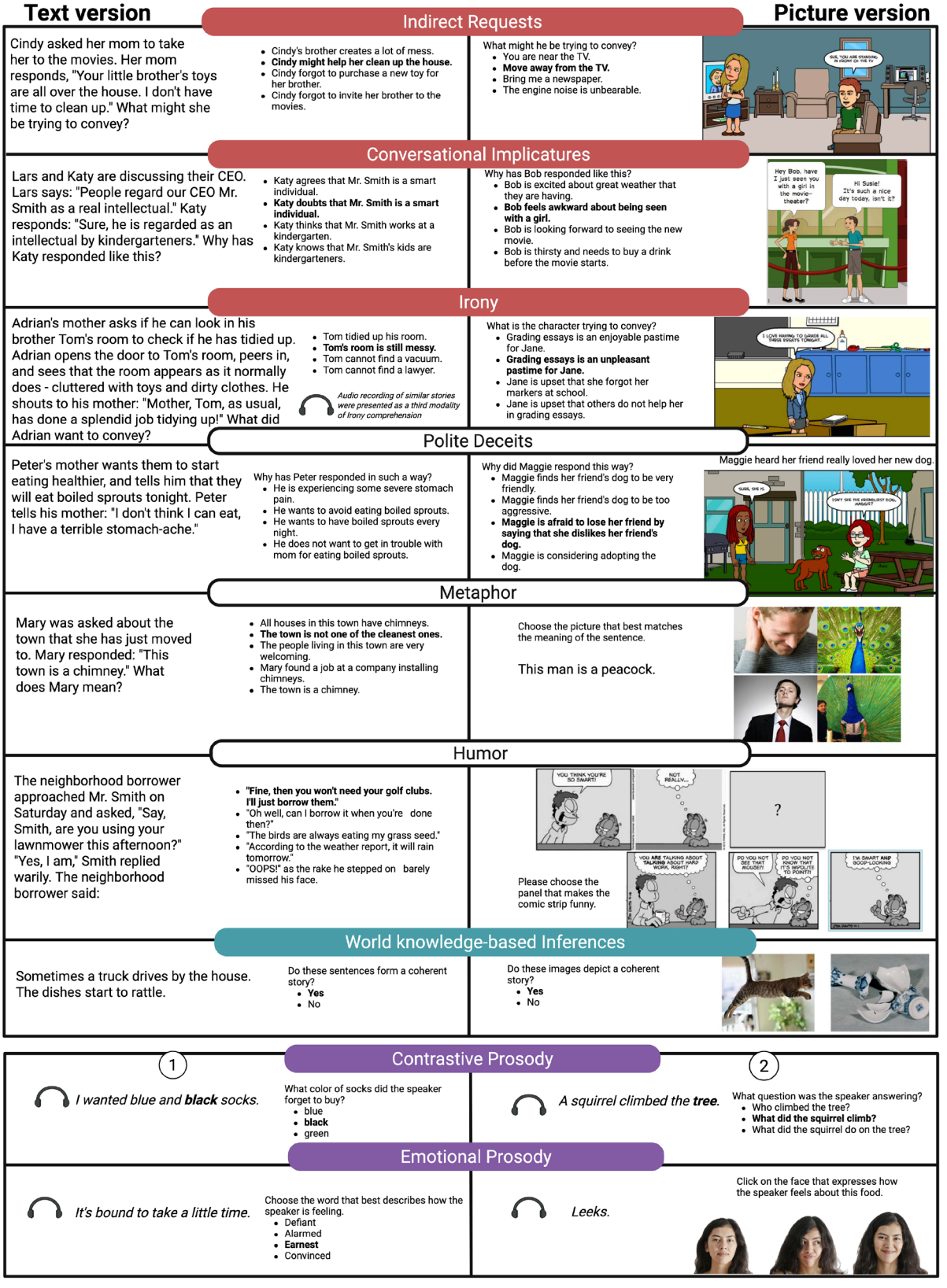

Fig. 1. Sample items from the 18 nonliteral comprehension tasks used across both experiments. For each phenomenon (row), two tasks were created from scratch or adapted from past work (*SI Appendix*, section 1A); for all but one phenomenon (Contrastive Prosody), one task used text-based stimuli and the other—picture-based stimuli; for Contrastive Prosody, the auditory stimuli were accompanied by text stimuli in both tasks. Additionally, for Experiment 2, we created a third task for Irony, which used auditory stimuli. The labels for the phenomena are colored according to the clustering that emerged in the current study (for ease of cross-reference with Figs. 3 and 4); phenomena that are not colored did not show consistent clustering patterns.

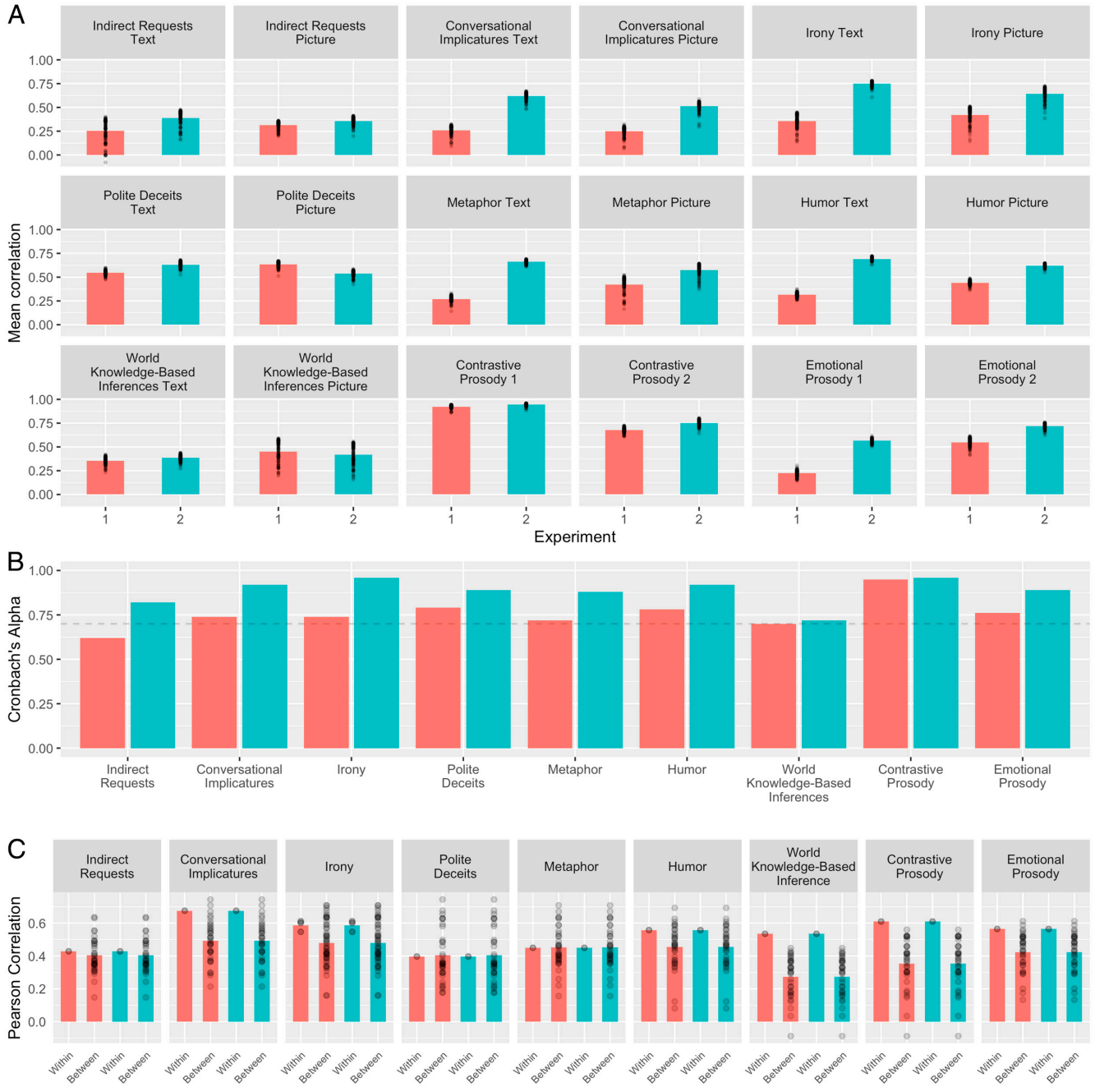

Fig. 2. Psychometric properties of the nonliteral comprehension tasks (red bars = Experiment 1; teal bars = Experiment 2). (*A*) Interitem consistency for the 18 nonliteral comprehension tasks used across both experiments. The bars show an average correlation across 100 random split-halves of items in each task, and the dots show individual-split correlation values (see *SI Appendix*, section 2A, i for scatterplot visualizations of a sample random split). (*B* and *C*) Construct validity for the nine phenomena for which two tasks were included. In *B*, the bars show Cronbach's alpha (61) computed across all items within each phenomenon (*Methods*). All phenomena except for Indirect Requests in Experiment 1 fall at or above the 0.7 threshold (marked by the dotted line), which is considered to index high reliability (62). The bars in C show within-phenomenon vs. between-phenomena correlations in Experiments 1 and 2. The "within" bars show the correlation (across participants) between the two tasks within a phenomenon, and the "between" bars show the average of pairwise correlations between each task for the phenomenon of interest and each task outside the phenomenon of interest (the dots represent correlations for individual task pairs). For 15 of the 18 pairs of bars, the within-phenomenon correlation is higher than the average of the between-phenomena ones.

generally higher than between-phenomena correlations (*Methods*; Fig. 2*C*), adding to the validity of these phenomenological constructs. (The phenomenon where this was not the case—Polite Deceits—did not show a consistent pattern of factor loadings in the critical analyses, as discussed below.)

Two tasks in Experiment 1 (the Scalar Implicatures and Quantifier Scope Ambiguity tasks) were excluded from the main analyses (*SI Appendix*, section 2A, iv; but see analyses in the last section of the Results). We had included these tasks because these phenomena—especially scalar implicatures—are often considered paradigmatic examples of nonliteral comprehension (10) and had been investigated extensively in past studies (30–32, 63); however, we could not find any tasks tapping these phenomena that would be suitable for use in individual-differences investigations. As a result, we optimistically used versions that had been developed in the Gibson lab and worked well in standard, group-averaging designs (64), but these tasks did not yield meaningful individual-level performance estimates for a substantial fraction of the participants, and their inclusion did not improve the model fit when a Confirmatory Factor Analysis (CFA; 65) was applied to Experiment 1 data (*SI Appendix*, section 2A, iv).

EFA Discovers a Replicable Task Grouping Across Experiments. To investigate our critical question—the structure of cognitive abilities underlying pragmatic language use—we submitted participants' scores on each task in Experiment 1 (residualized against our measure of IQ, to factor out variation due to general cognitive ability; 55) to an exploratory factor analysis (EFA), with promax rotation and principal axis factoring (66–67) (*SI Appendix*, section 2B). Horn's parallel analysis (68) determined that four factors were appropriate. To directly compare the results between Experiments 1 and 2, we performed the same analysis with four factors on Experiment 2 data, using the 18 intersecting tasks. The factor loadings are shown in Fig. 3. This 4-factor solution accounted for 32 and 55% of the cumulative total variance in Experiments 1 and 2, respectively.

In line with our analysis of phenomenological construct validity (Fig. 2 *B* and *C*), the tasks tapping the same phenomenon always loaded on the same factor, with the exception of the Polite Deceits tasks in Experiment 1. Critically, the grouping of the 18 core tasks was remarkably similar between the two experiments (Fig. 3; see *SI Appendix*, section 2C for evidence that the clustering results are not driven by task order).

The first factor [Factor 1 (F1) in Experiment 1, F4 in Experiment 2; the factor numbers are not meaningful for our purposes—see Fig. 3 caption] had high loadings from both versions of the tasks that tap comprehension of indirect requests, conversational implicatures, and irony. In addition, in Experiment 1, the text-based version of Polite Deceits loaded on this factor. Finally, both versions of Metaphor loaded on this factor in Experiment 2 (in Experiment 1, they showed positive but below-threshold loadings). The second factor (F2 in Experiment 1, F1 in Experiment 2) had high loadings from all four tasks that tap comprehension of intonation patterns: two versions each of Contrastive and Emotional Prosody. In Experiment 2, two tasks additionally loaded on this factor above threshold: picture-based versions of Humor and Metaphor. The third factor (F3 in Experiment 1, F2 in Experiment 2) had high loadings from the two tasks that tap world knowledge–based inferences. In Experiment 2, the text-based version of Indirect Requests additionally loaded on this factor (this task also showed a positive but below-threshold loading in Experiment 1). Finally, the fourth factor (F4 in Experiment 1, F3 in Experiment 2) did not show a consistent pattern of loadings across experiments. In Experiment 1, this factor had the two Humor tasks loading on it, but in Experiment 2, only the text-based version of Humor loaded reliably on this factor, along with a few tasks from other phenomena.

To evaluate the robustness of our results to the choice of analytic method, we additionally performed a Confirmatory Factor Analysis (CFA) on Experiment 2 data. A model which specified the groupings discovered in Experiment 1 (as described in the preregistration; https://osf.io/dcvb8) provided a good fit to the data, and a reliably better fit than a one-factor model (*SI Appendix*, section 2D).

Hypotheses About Three Interpretable Latent Factors Underlying Pragmatic Language Use. EFA revealed a consistent grouping of nonliteral comprehension tasks across experiments (see also *SI Appendix*, section 2D). The first three factors appear to be interpretable (i.e., it is possible to make a guess about a shared ability required for all high-loading tasks), but we remind the reader that these interpretations are necessarily post hoc given the data-driven structure-discovery nature of the approach. We hypothesize that the

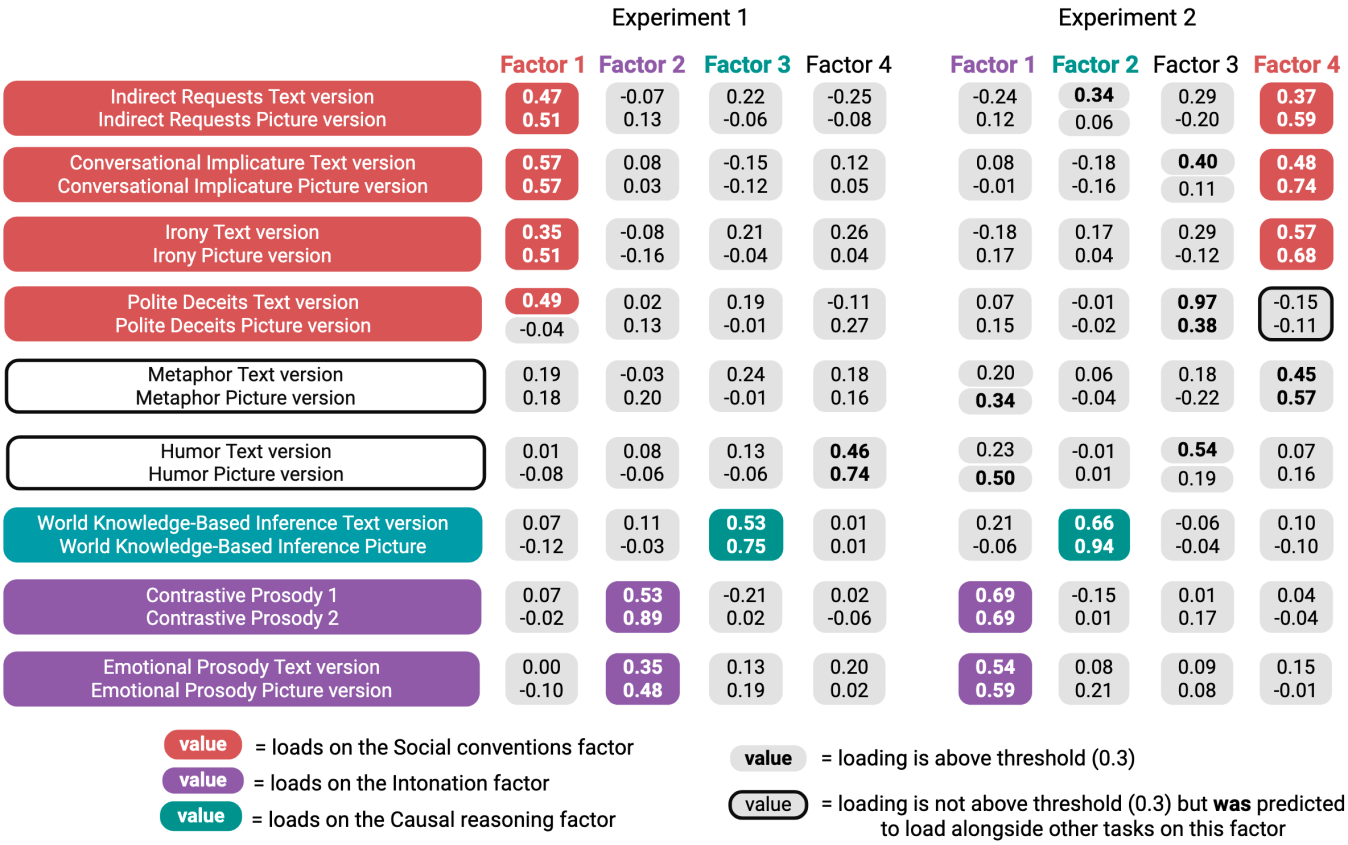

Fig. 3. Factor loadings for the 18 nonliteral comprehension tasks used across both experiments, obtained from an EFA performed on each experiment separately. Bolded values indicate loadings 0.3 or higher and are considered significant (66, 69). Within each experiment, factors (columns) are ordered according to relative variance they account for; as a result, factor order is not theoretically significant for establishing correspondences between the factors in Experiments 1 vs. 2; instead, we treat factors as "corresponding" based on the similarity of the task loading patterns between them (corresponding factors are marked in the same color in the *Top* row: e.g., F1 in Experiment 1 and F4 in Experiment 2).

first factor—which consistently underlies comprehension of indirect requests, conversational implicatures, and irony—reflects an ability to follow conventional social practices, such as phrasing requests, criticisms, or feedback in an indirect or untruthful way to maintain a sense of positive rapport between communicators (see ref. 13 for a related idea). We tentatively label this factor "Social conventions". The second factor is hypothesized to reflect an ability to infer a speaker's intended meanings and emotional states from the patterns of acoustic variation (pitch, tempo, pauses, etc.). We tentatively label this factor "Intonation". Finally, we hypothesize that the third factor reflects an ability to reason about causal relationships in the world, and we tentatively label this factor "Causal reasoning". The loadings on the fourth factor were not consistent across experiments (or even across analytic choices in Experiment 1, which we treated as a discovery sample and where we tried several dimensionality reduction approaches), leading to the exclusion of this factor from the preregistration; https://osf.io/dcvb8). We return to the interpretation of the first three factors in the Discussion.

The Grouping of Intonation Tasks Is Not Due to Low-Level Auditory-Processing Demands. One concern about the loading of the four intonation tasks on the same factor is that this grouping is due not to a shared ability to extract meaning from intonation (as we hypothesize above) but is instead explained by the fact that these tasks all require auditory processing, and participants may differ in auditory perceptual ability, or in the quality of audio hardware. To test this possibility, we submitted participants' scores from Experiment 2 to another EFA (preregistered: https://osf.io/dcvb8), where the scores for the critical auditory tasks (the four intonation tasks and the auditory version of Irony) were residualized on performance on a low-level auditory control task (56) in addition to IQ. Here, Horn's parallel analysis determined that a 5-factor solution was appropriate. This solution accounted for 57% of the variance, and the factor loadings are shown in *SI Appendix*, Fig. S2*E*.

The task grouping is remarkably similar to the original EFA on the data from Experiment 2 (Fig. 3, *Right*). A factor similar to the "Social conventions" factor emerged as F4. As in the 4-factor solution, six of the nine predicted tasks loaded on this factor; both versions of Metaphor also loaded on this factor. A factor similar to the "Causal reasoning" factor emerged as F2. Both of the predicted tasks loaded on this factor, along with the text-based version of Indirect Requests, as in the 4-factor solution. Critically, a factor similar to the "Intonation" factor emerged as F5. All four predicted tasks loaded on this factor, although the text-based version of Emotional Prosody loaded just below the 0.3 threshold (at 0.26). The fact that this grouping remains after factoring out low-level auditory ability, along with the fact that the auditory version of Irony did not load on this factor, rules out the low-level auditory explanation of the intonation task grouping.

The Main Results Are Robust, Emerging in a Bayesian EFA on the Combined Dataset and Another EFA Variant. To leverage the power of the large total number of participants in this study (n = 776) and ensure the robustness of the results to different analytic approaches, we conducted a Bayesian exploratory factor analysis (BEFA; 70, 71) that combined the data across the two experiments. Unlike the classic EFA, which assigns each task a loading for every factor, the BEFA implementation that we used essentially solves a discrete allocation problem, wherein each task can only load on at most one factor. BEFA also provides 95% credible intervals, which quantify the certainty of the loadings. Following the reviewers' suggestions, we a) included the Scalar Implicatures and Quantifier Scope Ambiguity tasks from Experiment 1, and b) divided each World Knowledge–Based Inference task into two based on condition ("Coherent" or "Incoherent") to address a potential concern that different inferences may be required for these conditions (*SI Appendix*, section 2A, ii). For comparability with our main analysis (Fig. 3), we set the number of factors to 4.

The factor loadings (Fig. 4) align in critical ways with the EFA of the two experiments separately (Fig. 3). A factor similar to the "Social conventions" factor emerged as F1, with seven of the nine predicted tasks loading on this factor, along with both versions of Metaphor. A factor similar to the "Intonation" factor emerged as F3, with three of the four predicted tasks loading on this factor. The Quantifier Scope task also loaded on the "Intonation" factor, perhaps because simulating a context-appropriate contrastive intonation contour is an important contributor to this task. And a factor similar to the "Causal reasoning" factor emerged as F4. Three of the four predicted tasks (and no other tasks) loaded on this factor. The re-emergence of a similar three-factor structure in the BEFA suggests that the factor structure identified in the EFA is robust.

Finally, following a reviewer's suggestion, we conducted another exploratory analysis, where we first analyzed each task using Item Response Theory (IRT; 60; *SI Appendix*, section 2A, iii), and then subjected the resulting thetas (the conceptual equivalent of the sum scores, used in the main analysis) to an EFA. Although the IRT analysis revealed that a few of our tasks may not be unidimensional and that for a couple of tasks, the thetas do not correlate strongly with the sum scores, the theta-based EFA recovered almost the same task grouping, as our main analyses (*SI Appendix*, section 2F). The main differences were that i) in addition to the Indirect Requests, Conversational Implicatures, and Irony tasks, the Metaphor tasks (similarly to BEFA; Fig. 4) and Humor tasks also loaded reliably on the "Social conventions" factor, and ii) one of the Contrastive Prosody tasks loaded on its own factor, so the "Intonation" factor got split into two.

Discussion

The present study asked a question that has not been previously investigated at scale: What is the cognitive structure of pragmatic language use? Historically, diverse phenomena have been grouped under the umbrella of "pragmatic inference" or "non-literal language comprehension", and many have explicitly argued that different phenomena require similar mental computations (10–16, 20–22). Here, we used an individual-differences approach with two large independent samples (n = ~400 participants each) and a diverse array of pragmatic tasks (21 total tasks across 11 phenomena) to evaluate this claim empirically. Across experiments and analytic choices, we found that some phenomena consistently group together, but that pragmatic language use encompasses at least three distinct components. We hypothesize that these components correspond to an ability to understand social conventions, an ability to extract meaning from intonation, and an ability to make causal inferences based on world knowledge. Below, we contextualize these findings in the broader landscape of the field and discuss their implications.

The Social Conventions Factor. Tasks across three phenomena—indirect requests, conversational implicatures, and irony—consistently loaded on the same factor. These phenomena share adherence to conventional social practices, which maintain positive rapport between communicators. The use of an indirect request can be thought of as maintaining the interlocutor's sense of freedom of action (13). Conversational implicatures have been

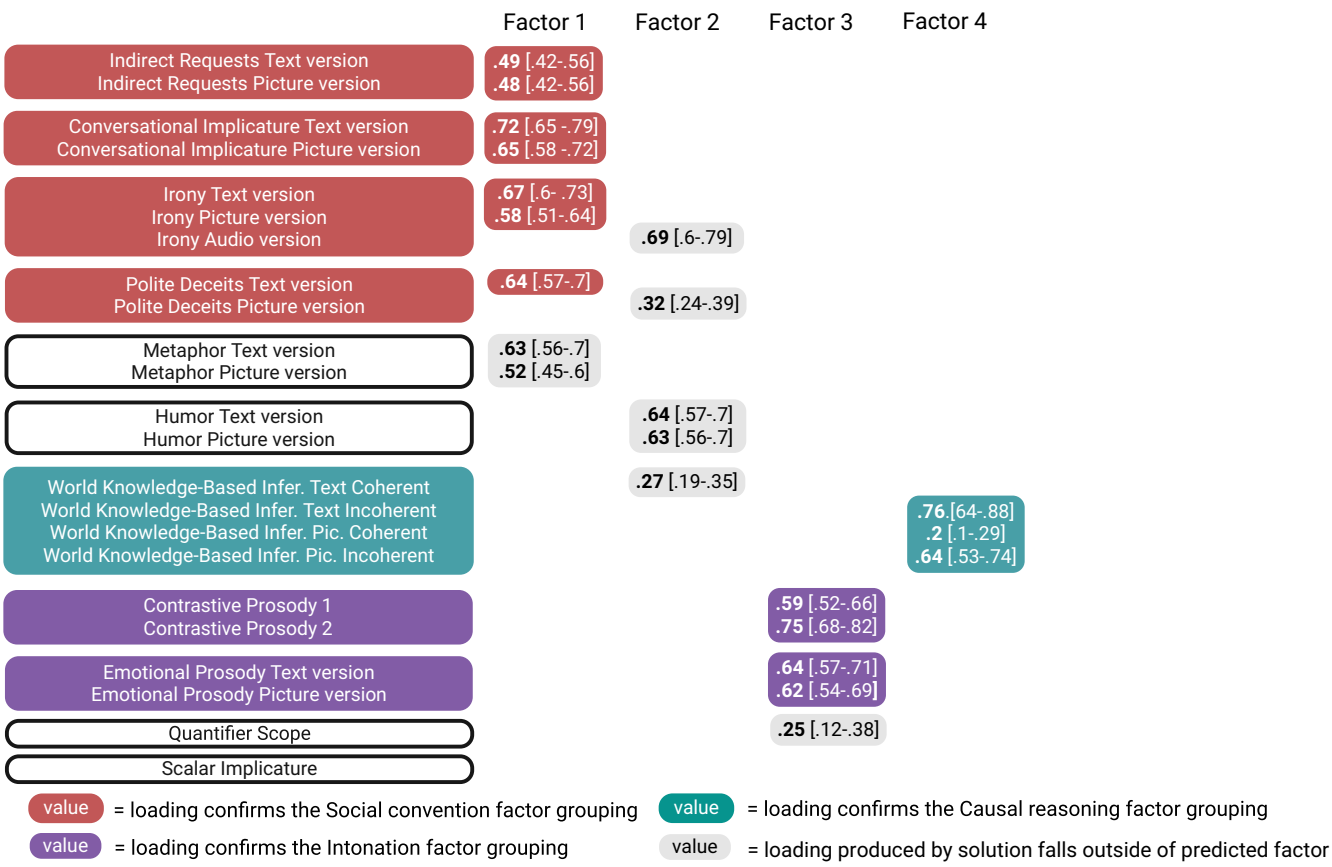

Fig. 4. Factor loadings for the critical tasks, obtained from a BEFA performed on the combined dataset, accompanied by 95% credible intervals (shown in square brackets for all loadings; unlike frequentist 95% CI, the interpretation of a 95% credible interval is intuitive: given the data and the model, there is a 95% chance that the true loading falls within the interval).

discussed as "face saving" acts: Changing the topic or remarking on some irrelevant details allows the speaker to convey (and the addressee to infer) something negative without directly expressing the negativity (13). And irony has also been argued to provide a way for the speaker to convey a negative evaluation in a way that imposes less on the addressee's viewpoint (72).

All of these phenomena appear to require consideration of the communication partner's mental state, or Theory of Mind (17–19). To correctly interpret the speaker's utterance, the listener is likely to consider the speaker's beliefs, attitudes, desires, and intentions given that the speaker is saying something that is obviously not truthful (e.g., in the case of irony) or does not directly pertain to the question or topic of conversation (e.g., in the case of conversational implicatures). Indeed, brain imaging investigations of these phenomena have reported putative responses in the social-reasoning brain areas (73–75; reviews/meta-analyses: 41–43), but often without including a Theory of Mind task, as needed to unambiguously establish overlap. In our exploratory analyses of correlations between the "Social conventions" factor score and measures of social cognition (the Autism Spectrum Quotient questionnaire; 76; and the Reading the Mind in the Eyes task: 77), we did not see a consistent relationship (*SI Appendix*, section 2G), but this may be due to the difficulty of measuring social abilities (78, 79). As a result, more research is needed to understand whether the shared computation indeed has to do with Theory of Mind or social competence more broadly.

Tasks for three additional phenomena—polite deceits, metaphors, humor—loaded on the "Social conventions" factor in some versions of the analyses. Polite deceits (or white lies) is a prime example of a face-saving social practice (13). However, this phenomenon showed poor construct validity, with the two tasks targeting it loading on distinct factors in most analyses. The Metaphor tasks loaded on the "Social conventions" factor in Experiment 2 and positively but below threshold in Experiment 1. And both the Metaphor tasks and the Humor tasks loaded on this factor in the EFA performed on IRT thetas (*SI Appendix*, section 2F), which may be less noisy than the sum scores. Better tasks will be needed to conclusively determine whether these additional phenomena share computations with the three consistently loading ones. For metaphors and humor, this effort may also require devising tasks that target particular types of metaphors/jokes, given that both are highly heterogeneous phenomena (80–81).

The Intonation Factor. Four tasks tapped the interpretation of intonational/prosodic cues: two focused on contrastive prosody, which conveys meaning differences (typically related to information structure; 82), and the other two—on emotional prosody, which conveys different emotional states (38, 83). The grouping of these tasks was consistent across experiments and analytic choices, and in Experiment 2, we ruled out a low-level explanation of this grouping in terms of reliance on the auditory modality.

The fact that the prosody tasks all load on a common factor suggests that some computations may be shared between the extraction of semantic and emotional information from acoustic speech patterns. This finding is consistent with some past claims of broad neural overlap between the processing of these two kinds of prosodic cues (84–85). This said, some versions of the analyses hint at additional complexity in the processing of intonation. For example, in the 5-factor solution of Experiment 2 (*SI Appendix*, section 2E), the two Emotional Prosody tasks—in addition to loading on the common "Intonation" factor—also loaded on a separate factor. This partial segregation aligns with a putative dissociation based on neural data, whereby the right hemisphere has been argued to play a greater role in emotional prosody, and the left hemisphere—in linguistic prosody (86–88). And in the EFA performed on IRT thetas (*SI Appendix*, section 2F), one of the Contrastive Prosody tasks loaded on a separate factor. Therefore, whether intonation processing draws on a single shared ability or whether additional fractionation will be uncovered with greater coverage of the task space and finer-grained contrasts remains to be discovered.

The Causal Reasoning Factor. Both versions of the task that requires reasoning about causal relationships between events based on world knowledge consistently loaded on the same factor. However, when splitting these tasks into coherent vs. incoherent trials (for the combined-data BEFA), only three of the four tasks loaded on the same factor; similarly, when splitting each task into two dimensions, as suggested by the IRT (*SI Appendix*, section 2A, iii), only three of the dimensions loaded on the same factor (*SI Appendix*, section 2F). We therefore place less confidence in this factor compared to the other two.

In general, in our battery, the tasks that target the understanding of event contingencies may seem the least directly related to pragmatic ability. We had included these tasks because many researchers consider discourse coherence a pragmatic inference process (89–91), and some have argued for a broad inferential ability underlying nonliteral comprehension, which subsumes inferences based on world knowledge (12). To the extent that the mental processes tapped by our World Knowledge–Based Inference tasks belong to the space of pragmatics-related processes, our results suggest that they are distinct from those that underlie phenomena such as indirect requests, conversational implicatures, and irony. Although some inferences about event contingencies may require thinking about other people (e.g., understanding that feeling stressed can make one's hands sweaty), many do not (e.g., understanding that a cat running around can result in a broken vase). In this way, these inferences stand in contrast to the inferences in the tasks that load on the "Social conventions" factor, which ubiquitously require reasoning about others' mental states based on their indirect utterances, in light of our knowledge of social rules.

An in-depth investigation of semantic knowledge and reasoning may reveal that the "Causal reasoning" factor discovered in this study is a part of a broader semantic or logical reasoning ability, or it may find that different domains of world knowledge and reasoning (e.g., physical vs. social) load on distinct factors. Future work should also aim to cover the space of discourse processing more broadly, beyond examining discourse relations that require reasoning about causation (92–93).

Scalar Implicatures Appear to Be Cognitively Distinct From the Phenomena Loading On the Social Conventions Factor. In the analyses where the Scalar Implicatures (SI) task was included, it did not load on the "Social conventions" factor, and—when allowed—loaded strongly on its own factor (*SI Appendix*, section 2F). Although this result should be replicated with additional SI tasks, this task's separation from other pragmatic tasks can potentially inform the debate on whether SIs should be considered a pragmatic phenomenon. In particular, some have argued that SIs, and quantity-based implicatures more generally, instead fall within the domain of grammatical competence (94; cf. 95). To the extent that the "Social conventions" factor has strong loadings from phenomena such as indirect requests and irony—which are standardly considered to be pragmatic in nature (10–11, 14)—it appears that SIs are distinct from those phenomena. However, to argue for a particular cognitive basis of SI processing, additional studies will be needed: E.g., to argue that SIs are a grammatical phenomenon, one would need to establish that SI tasks cluster with other grammar-taxing tasks in individual-differences investigations or that they recruit brain areas that have been linked to grammatical processing (cf. the Theory of Mind areas).

Shared and Distinct Components of Pragmatic Language Use. The current results have implications for how pragmatic language phenomena should be approached in research and in practice. On the one hand, our findings suggest the existence of a broad shared component that has to do with the ability to adhere to social conventions, including phenomena such as indirect requests, conversational implicatures, irony, and perhaps also polite deceits, metaphors, and humor. The existence of this "Social conventions" factor aligns with prior theoretical claims of diverse pragmatic phenomena sharing underlying cognitive demands (10–16, 20–22), and justifies, to some degree, the use of terms, such as "pragmatic abilities" or "pragmatic deficits", common not only in basic language research but also in work on communication disorders (96–98). However, our results call for more precision in using such terms given that not all nonliteral comprehension phenomena cluster onto a single factor, with both intonation processing and general world knowledge–based causal reasoning separating out as distinct components. This basic three-factor structure—with indirect requests, conversational implicatures, and irony tasks loading on the same, shared factor, contrastive and emotional intonation loading on a separate shared factor, and world knowledge–based reasoning loading on a third factor—should be replicable in future studies given its robustness across two large independent samples and analytic choices. Although pragmatics researchers would likely not be surprised by the fact that pragmatic language use is not a cognitive monolith, we provide an empirical foundation for a particular grouping of common comprehension phenomena and offer hypotheses for the underlying cognitive abilities.

Limitations, Open Questions, and Future Directions.

Interpretation of the observed groupings. Given the data-driven structure-discovery approach, our interpretation of the observed groupings should be construed as hypotheses about the shared cognitive abilities. We have attempted to rule out several trivial explanations for the observed groupings, including task order effects (*SI Appendix*, section 2C) and the contributions of low-level auditory ability to the "Intonation" factor (*SI Appendix*, section 2E), but other, theoretically interesting interpretations remain possible. Besides, factors other than pragmatics-relevant cognitive traits may play a role. For example, a reviewer suggested that personality traits may contribute (99). Future studies can now rigorously evaluate the hypotheses we put forward, as well as alternative possibilities, by testing additional tasks and phenomena in new individual-differences studies, and using other approaches discussed below.

Expanding the space of nonliteral comprehension tasks. The current study covers diverse phenomena, but it is obviously not comprehensive. Besides, for some phenomena, our tasks were imperfect. As a result, future studies should develop better tasks for those phenomena, expand the range of tasks for heterogeneous domains (including, as noted above, metaphors, humor, discourse comprehension, and world knowledge), and investigate other phenomena related to pragmatic language use (e.g., understanding mutual exclusivity, reasoning about alternatives outside the realm of scalar implicatures, partner-specific adaptation and collaborative planning in interactive settings, using nonverbal communicative cues, such as facial expressions and gestures, and so on; 5, 6, 8, 100–102). Doing so will be important for both evaluating the hypotheses we have put forward, and refining—and almost certainly expanding—the ontology of cognitive skills that jointly enable efficient communication.

Intercultural differences in pragmatics. Results from this study must be interpreted within the relevant cultural and linguistic context: Our participants were proficient English speakers, and the tasks were constructed in English and assumed a Western, industrialized cultural background. This limitation is not to be taken lightly (103–104). For pragmatic communication specifically, substantial differences exist across cultures (105–106). Therefore, investigating pragmatic abilities in other languages and cultures may challenge and enrich our findings.

The neural basis of pragmatic abilities. Our findings may help guide future work on the brain basis of nonliteral comprehension. As noted above, past fMRI studies had reported putative engagement of the Theory of Mind (ToM) brain areas in comprehension of several nonliteral phenomena (73–75). However, to conclude that ToM mechanisms are engaged requires the inclusion of a validated ToM paradigm within the same participants, which most studies have not done. Besides, many studies of nonliteral comprehension additionally report the engagement of the core left-hemisphere language areas, which support word retrieval and syntactic/semantic composition and are distinct from the ToM areas (7), their right-hemisphere homotopic areas, and additional cortical and subcortical areas (41–43). As a result, more work is needed to understand the division of labor among different brain systems during pragmatic language use. The dissociations discovered here make testable predictions about which phenomena should recruit the same vs. distinct brain mechanisms and can thus guide future studies.

Pragmatic development. Children's ability to understand nonliteral language shows some early competency, but also protracted development for some phenomena (107–108). The latent structure identified here predicts that phenomena loading on the same factor should emerge together during development, and phenomena that load on different factors may show distinct trajectories. Bohn et al. (44) have established that stable variation in pragmatic abilities is possible to measure in children aged 5 to 7 y and that variation related to pragmatic tasks can be dissociated from tasks that tap nonpragmatic cognitive demands. Similar investigations with diverse tasks will be important for understanding how the structure of adult pragmatic abilities emerges in children and how it may be affected by developmental disorders like autism.

Pragmatic abilities in artificial neural network language models. Artificial neural network language models (LMs) have achieved human-level performance on diverse language tasks, and their internal representations resemble those of humans (109), which makes these models promising as candidate models of human language processing. With respect to nonliteral comprehension, LMs have shown some successes (110), but also some limitations (111–113; review: 114). Using the text-based tasks from the current battery, Hu et al. (111) found that smaller LMs trained only on next-word prediction show chance-level performance, but larger LMs fine-tuned on diverse tasks succeed on several phenomena, including showing human-like error patterns. Future minimal-pair model comparisons can help illuminate which pragmatic phenomena can be solved by tracking linguistic statistics alone, and which may require additional data (e.g., visual grounding), additional training objectives, or neurosymbolic approaches (115). Furthermore, using model interpretability tools, we can begin to gain mechanistic-level insights into how different nonliteral phenomena may be solved, and whether the classes of phenomena that are dissociable in humans draw on distinct mechanisms in the models.

Conclusions

For decades, philosophers, linguists, and cognitive scientists have developed and evaluated theories of nonliteral linguistic communication, and much progress has been made in understanding particular phenomena. However, the relationship among pragmatic phenomena has not been empirically evaluated. By heavily sampling participants across diverse nonliteral comprehension tasks, we gained insight into the structure of pragmatic language use and offered hypotheses about the underlying latent abilities. After accounting for differences in general cognitive ability and auditory processing, we found three latent factors—replicable across experiments and analytic choices—which we interpreted as reflecting comprehension of social conventions, interpretation of intonation patterns, and the ability to make causal inferences based on world knowledge.

Methods

Participants. Participants were recruited through Amazon.com's online survey platform Mechanical Turk, under IRB protocol 403000040 approved by MIT's Committee on the Use of Humans as Experimental Subjects. All participants provided informed consent and were paid for participation. For Experiment 1 (conducted in 2017), 405 participants were recruited, and 29 were excluded (5 for technical issues during data collection, 16 for failure to complete the study, and 8 for poor catch trial performance), leaving 376 for analysis (177 females, M_{age} = 36.3 y, SD = 9.8). Participants were paid $45 for completing the task, and an additional $25 if they scored at or above 60% on each of the quality check sections (*SI Appendix*, section 1B). For Experiment 2 (conducted in 2022), we collected data until 400 participants (prescreened using a sentence completion task due to data quality decrease over the years (116, 117; *SI Appendix*, section 1B) met our inclusion criteria, as specified in the preregistration (https://osf.io/dcvb8) (183 females, M_{age} = 38.8 y, SD = 11.0). A total of 466 participants were recruited, and 66 were excluded (3 for technical issues, 5 for failure to complete the study, 57 for poor catch trial performance, and 1 for experimenter error of collecting data beyond the planned sample size). Participants were paid $60 and an additional $40 if they scored at or above 60% on the quality check sections.

Critical Tasks. In selecting the tasks, we tried to be as inclusive as possible with respect to covering the space of nonliteral comprehension phenomena (*SI Appendix*, section 1A, i). Each critical task consisted of 25–50 items (20–36 critical items that required pragmatic inference, and the rest served as control items; *SI Appendix*, section 1A, iii and iv). Prior to the study, most tasks underwent several rounds of piloting on independent participants with adjustments to the items made until the tasks showed good psychometric properties (*SI Appendix*, section 2A). After Experiment 1, we also made minor edits to the picture-based version of Polite Deceits to further improve psychometric properties. (For the details of all the noncritical tasks, see *SI Appendix*, section 1B.)

Procedure. Each experiment was divided into 4 testing sessions 1.5 to 2 h each. Participants had to complete all 4 sessions within 72 h. Because our analysis methods use variation across participants to infer relationships among pragmatic phenomena, task and item order were kept constant across participants (*SI Appendix*, Fig. S1*C*: Task order), as is common in individual-differences investigations (118–119).

Task evaluation.

Evaluation of interitem consistency. For each task, the items were divided in half (randomly), and a correlation was computed across participants (using IQ-residualized scores). This procedure was performed 100 times to derive an average correlation value for each task. This analysis effectively asks if a participant performs well on one half of the items in a task, do they also perform well on the other half of the items? (see *SI Appendix*, section 2A, iii for additional evaluation using IRT; 60).

Evaluation of construct validity. For the phenomena for which two tasks were included, we calculated Cronbach's alpha (61) using the full set of critical items within each phenomenon. For example, Polite Deceits had 20 text-based critical items and 20 picture-based items, so Cronbach's alpha was calculated across 40 items. To additionally determine how distinct different phenomena are from one another, we measured the strength of the relationships between pairs of tasks within a phenomenon vs. between phenomena using pairwise Pearson correlations (here, we again used participants' IQ-residualized scores).

Data, Materials, and Software Availability. Study data and materials are publicly available here (120).

ACKNOWLEDGMENTS. This research was supported by a grant from the Simons Foundation to the Simons Center for the Social Brain (SCSB) at MIT and by NIH award R01-DC016607 to EF. SF was supported by NSF award #2105136. EF was additionally supported by NIH awards R01-DC016950 and U01-NS121471. We thank the reviewers, the editor, as well as Nancy Kanwisher, Rebecca Saxe, Laura Schulz, Josh Tenenbaum, Leon Bergen, and Alexander Paunov for helpful comments on the manuscript and discussions of the project; Patrick Mair for input on the statistical analyses; Avital Baral and Frances Mulligan for help with stimulus creation and review; Saima Malik-Moraleda, Maya Taliaferro, and Emma Landry for creating the auditory version of the Irony task; and Malinda McPherson for sharing the auditory control task materials.

Author affiliations: [a]Sarah Lawrence College, Bronxville, NY 10708; [b]Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Institute of Cognitive Science, Carleton University, Ottawa, ON K1S 5B6, Canada; [d]Eberly Center at Carnegie Mellon University, Pittsburgh, PA 15213; [e]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [f]Program in Speech and Hearing in Bioscience and Technology, Harvard University, Cambridge, MA 02138

Author contributions: O.J., E.G., and E.F. designed research; S.F., O.J., Z.M., E.G., and E.F. performed research; S.F., O.J., and M.P. analyzed data; and S.F., O.J., E.G., and E.F. wrote the paper.

The authors declare no competing interest.

1. R. Jackendoff, S. Pinker, The faculty of language: What's special about it? *Cognition* **95**, 201–236 (2005).
2. R. M. Seyfarth, D. L. Cheney, *The Social Origins of Language* (Princeton Univ. Press, 2018).
3. E. Fedorenko, S. T. Piantadosi, E. A. F. Gibson, Language is primarily a tool for communication rather than thought. *Nature* **630**, 575–586 (2024a).
4. R. Jackendoff, *Foundations of Language: Brain, Meaning, Grammar, Evolution* (Oxford, 2002).
5. G. Vigliocco, P. Perniss, D. Vinson, Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **369**, 20130292 (2014), 10.1098/rstb.2013.0292.
6. L. V. Hadley, G. Naylor, A. F. d. C. Hamilton, A review of theories and methods in the science of face-to-face social interaction. *Nat. Rev. Psychol.* **1**, 42–54 (2022), 10.1038/s44159-021-00008-w.
7. E. Fedorenko, A. A. Ivanova, T. I. Regev, The language network as a natural kind within the broader landscape of the human brain. *Nat. Rev. Neurosci.* **25**, 289–312 (2024), 10.1038/s41583-024-00802-4.
8. P. Hagoort, A. Özyürek, Extending the architecture of language from a multimodal perspective. *Top. Cogn. Sci.* **17**, 12728 (2024), 10.1111/tops.12728.
9. Edward A. Gibson, *Syntax: A Cognitive Approach* (MIT Press, 2025).
10. H. P. Grice, *Logic and Conversation in Readings in Language and Mind* (Blackwell, 1975).
11. S. Levinson, *Pragmatics* (Cambridge University Press, 1983).
12. D. Sperber, D. Wilson, *Relevance: Communication and Cognition* (Harvard University Press, Cambridge, MA, 1986), vol. 142.
13. P. Brown, S. C. Levinson, *Politeness: Some Universals of Language Usage* (Cambridge University Press, 1987).
14. H. H. Clark, *Using Language* (Cambridge University Press, 1996).
15. R. Gibbs, *Figurative language in MIT Encyclopedia of the Cognitive Sciences*, R. A. Wilson, F. C. Keil Eds. (MIT Press. 1999), pp. 314–315.
16. S. Glucksberg, *Understanding Figurative Language: From Metaphors to Idioms* (Oxford University Press, 2001).
17. U. Frith, F. Happé, Language and communication in autistic disorders. *Phil. Trans. R. Soc. Lond., B* **346**, 97–104 (1994).
18. D. Sperber, D. Wilson, Pragmatics, modularity, and mindreading. *Mind Lang.* **17**, 3–23 (2002).
19. D. O'Neill, Components of Pragmatic Ability and Children's Pragmatic Language Development. Cognitive Pragmatics, H. J. Schmid Ed. (De Gruyter Mouton, 2012), pp. 261–287.
20. M. C. Frank, N. D. Goodman, Predicting pragmatic reasoning in language games. *Science* **336**, 998–998 (2012).
21. M. Franke, G. Jager, Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Z. Sprachwiss.* **35**, 3–44 (2016).
22. J. Degen, The rational speech act framework. *Annu. Rev. Linguist.* **9**, 519–540 (2023).
23. J. B. Tenenbaum, T. L. Griffiths, C. Kemp, Theory-based bayesian models of inductive learning and reasoning. *Trends Cogn. Sci.* **10**, 309–318 (2006).
24. P. C. Fletcher *et al.*, Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition* **57**, 109–128 (1995), 10.1016/0010-0277(95)00692-r.
25. K. Vogeley *et al.*, Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage* **14**, 170–181 (2001), 10.1006/nimg.2001.0789.
26. R. Saxe, N. Kanwisher, People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *Neuroimage* **19**, 1835–1842 (2003).
27. D. W. Allbritton, G. McKoon, R. J. Gerrig, Metaphor-based schemas and text representations: Making connections through conceptual metaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* **21**, 612–625 (1995), 10.1037/0278-7393.21.3.612.
28. D. Gentner, P. Wolff, Alignment in the processing of metaphor. *J. Mem. Lang.* **37**, 331–355 (1997).
29. R. Carston, X. Yan, Metaphor processing: Referring and predicating. *Cognition* **238**, 105534 (2023), 10.1016/j.cognition.2023.105534.
30. I. A. Noveck, When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* **78**, 165–188 (2001).
31. A. Papafragou, J. Musolino, Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition* **86**, 253–282 (2003).
32. J. Degen, M. K. Tanenhaus, Processing scalar implicature: A constraint-based approach. *Cogn. Sci.* **39**, 667–710 (2015), 10.1111/cogs.12171.
33. R. Carston, S. C. Levinson, Presumptive meanings: The theory of generalized conversational implicature. *J. Linguist.* **40**, 181–186 (2004).
34. J. C. Sedivy, Implicature during real time conversation: A view from language processing research. *Philos. Compass* **2**, 475–496 (2007).
35. R. Breheny, H. J. Ferguson, N. Katsos, Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition* **126**, 423–440 (2013).
36. M. Breen, E. Fedorenko, M. Wagner, E. Gibson, Acoustic correlates of information structure. *Lang. Cogn. Process.* **25**, 1044–1098 (2010).
37. T. B. Roettger, T. Mahrt, J. Cole, Mapping prosody onto meaning–the case of information structure in American English. *Lang. Cogn. Neurosci.* **34**, 841–860 (2019).
38. P. Larrouy-Maestri, D. Poeppel, M. D. Pell, The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspect. Psychol. Sci.* **20**, 17456916231217722 (2024).
39. H. Tager-Flusberg, R. Paul, C. Lord, "Language and communication in autism" in *Handbook of Autism and Pervasive Developmental Disorders* (2005), vol. 1, pp. 335–364.
40. G. Andreou, V. Lymperopoulou, V. Aslanoglou, Developmental language disorder (DLD) and autism spectrum disorder (ASD): Similarities in pragmatic language abilities. A systematic review. *Int. J. Dev. Disabil.* **70**, 777–791 (2022), 10.1080/20473869.2022.2132669.
41. P. Hagoort, S. C. Levinson, "Neuropragmatics" in *The cognitive neurosciences*, M. S. Gazzaniga, G. R. Mangun Eds. (Boston Review, ed. 5, 2014), pp. 667–674.
42. A. Reyes-Aguilar, E. Valles-Capetillo, M. Giordano, A quantitative meta-analysis of neuroimaging studies of pragmatic language comprehension: In search of a universal neural substrate. *Neuroscience* **395**, 60–88 (2018).
43. M. Hauptman, I. Blank, E. Fedorenko, Non-literal language processing is jointly supported by the language and theory of mind networks: Evidence from a novel meta-analytic fmri approach. *Cortex* **162**, 96–114 (2023).
44. M. Bohn, M. H. Tessler, C. Kordt, T. Hausmann, M. C. Frank, An individual differences perspective on pragmatic abilities in the preschool years. *Dev. Sci.* **26**, e13401 (2023).
45. J. B. Carroll, S. E. Maxwell, Individual differences in cognitive abilities. *Annu. Rev. Psychol.* **30**, 603–640 (1979).
46. N. J. Boogert, J. R. Madden, J. Morand-Ferron, A. Thornton, Measuring and understanding individual differences in cognition. *Phil. Trans. R. Soc.* **373**, B37320170280 (2018).
47. C. Spearman, "General intelligence", objectively determined and measured. *Am. J. Psychol.* **15**, 201–292 (1904).
48. R. B. Cattell, The measurement of adult intelligence. *Psychol. Bull.* **40**, 153 (1943).
49. J. Raven, The raven's progressive matrices: Change and stability over culture and time. *Cogn. Psychol.* **41**, 1–48 (2000).
50. J. D. Mollon, J. M. Bosten, D. H. Peterzell, M. A. Webster, Individual differences in visual science: What can be learned and what is good experimental practice?. *Vis. Res.* **141**, 4–15 (2017).
51. G. R. Kidd, C. S. Watson, B. Gygi, Individual differences in auditory abilities. *J. Acoust. Soc. Am.* **122**, 418–435 (2007), 10.1121/1.2743154.
52. J. L. Booth, R. S. Siegler, Developmental and individual differences in pure numerical estimation. *Dev. Psychol.* **42**, 189 (2006).
53. A. Miyake *et al.*, The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cogn. Psychol.* **41**, 49–100 (2000).
54. A. Mitko, J. Fischer, When it all falls down: The relationship between intuitive physics and spatial cognition. *Cogn. Res.* **5**, 24 (2020), 10.1186/s41235-020-00224-7.
55. A. S. Kaufman, *Kaufman Brief Intelligence Test: KBIT* (AGS, American Guidance Service, Circle Pines, MN, 1990).
56. M. J. McPherson, J. McDermott, Diversity in pitch perception revealed by task dependence. *Nat. Hum. Behav.* **2**, 52–66 (2018).

57. R. B. Cattell, Factor analysis: An introduction and manual for the psychologist and social scientist (Harper,1952).
58. P. Kline, *A Handbook of Test Construction Psychology Revivals: Introduction to Psychometric Design* (Routledge, Abingdon, UK, 2015).
59. R. K. Hambleton, H. Swaminathan, H. J. Rogers, *Fundamentals of Item Response Theory* (Sage Press, Newbury Park, CA, 1991).
60. R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford, ed. 2, 2022).
61. L. J. Cronbach, Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334 (1951).
62. J. Nunnally, I. Bernstein, *Psychometric Theory* (McGraw-Hill, ed. 3, 1994).
63. H. S. Kurtzman, M. C. MacDonald, Resolution of quantifier scope ambiguities. *Cognition* **48**, 243–279 (1993).
64. Z. Zhang, L. Bergen, A. Paunov, R. Ryskin, E. Gibson, Scalar implicature is sensitive to contextual alternatives. *Cogn. Sci.* **47**, e13238 (2023).
65. T. A. Brown, *Confirmatory Factor Analysis for Applied Research* (Guilford, ed. 2, 2015).
66. R. Gorsuch, *Factor Analysis* (Lawrence Erlbaum Associates, Hillsdale, NJ, ed. 2, 1983).
67. L. R. Fabrigar, D. T. Wegener, *Exploratory Factor Analysis* (Oxford University Press, 2011).
68. J. L. Horn, A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185 (1965).
69. P. Kline, *An Easy Guide to Factor Analysis* (Routledge, London, 2002).
70. G. Conti, A. Fruhwirth-Schnatter, J. J. Heckman, R. Piatek, Bayesian exploratory factor analysis. *J. Econom.* **183**, 31–57 (2014).
71. R. Piatek, BayesFM: An R Package for Bayesian Factor Modeling. Computer programme, CRAN - The Comprehensive R Archive Network. (2016).
72. D. Wilson, D. Sperber, Explaining irony. *Meaning Relevance*, 123–146 (2012).
73. N. Spotorno, E. Koun, J. Prado, J. B. Van Der Henst, I. A. Noveck, Neural evidence that utterance-processing entails mentalizing: The case of irony. *Neuroimage* **63**, 25–39 (2012), 10.1016/j.neuroimage.2012.06.046.
74. M. J. van Ackeren, D. Casasanto, H. Bekkering, P. Hagoort, S. A. Rueschemeyer, Pragmatics in action: Indirect requests engage theory of mind areas and the cortical motor network. *J. Cogn. Neurosci.* **24**, 2237–2247 (2012), 10.1162/jocn_a_00274.
75. G. Jang *et al.*, Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *Neuroimage* **81**, 61–72 (2013), 10.1016/j.neuroimage.2013.05.027.
76. S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, E. Clubley, The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J. Autism Dev. Disord.* **31**, 5–17 (2001), 10.1023/a:1005653411471.
77. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, I. Plumb, The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry Allied Discip.* **42**, 241–251 (2001).
78. F. Quesque, Y. Rossetti, What do theory-of-mind tasks actually measure? Theory and practice. *Perspect. Psychol. Sci.* **15**, 384–396 (2020), 10.1177/1745691619896607.
79. O. Jouravlev, K. Mahowald, A. Paunov, E. Gibson, E. Fedorenko, Evaluation of Psychometric Properties and Inter-test Associations for Three Popular Measures of Social Competence (PsyArXiv, 2023), 10.31234/osf.io/es9wt (Accessed 21 August 2025).
80. D. L. Long, A. C. Graesser, Wit and humor in discourse processing. *Discourse Process.* **11**, 35–60 (1988), 10.1080/01638538809544690.
81. G. Lakoff, M. Johnson, *Metaphors We Live By* (University of Chicago Press, 2008).
82. J. Cole, Prosody in context: A review. *Lang. Cogn. Neurosci.* **30**, 1–31 (2014), 10.1080/23273798.2014.963130.
83. D. Grandjean, T. Bänziger, K. R. Scherer, Intonation as an interface between language and affect. *Prog. Brain Res.* **156**, 235–247 (2006), 10.1016/S0079-6123(06)56012-1.
84. S. A. Kotz, M. Meyer, S. Paulmann, Lateralization of emotional prosody in the brain: An overview and synopsis on the impact of study design. *Prog. Brain Res.* **156**, 285–294 (2006), 10.1016/S0079-6123(06)56015-7.
85. J. Witteman, M. H. van Ijzendoorn, D. van de Velde, V. J. van Heuven, N. O. Schiller, The nature of hemispheric specialization for linguistic and emotional prosodic perception: A meta-analysis of the lesion literature. *Neuropsychologia* **49**, 3722–3738 (2011), 10.1016/j.neuropsychologia.2011.09.028.
86. K. M. Heilman, R. Scholes, R. T. Watson, Auditory affective agnosia. Disturbed comprehension of affective speech. *J. Neurol. Neurosurg. Psychiatry* **38**, 69–72 (1975), 10.1136/jnnp.38.1.69.
87. J. J. Schmitt, W. Hartje, K. Willmes, Hemispheric asymmetry in the recognition of emotional attitude conveyed by facial expression, prosody and propositional speech. *Cortex* **33**, 65–81 (1997), 10.1016/s0010-9452(97)80005-6.
88. A. Seydell-Greenwald, C. E. Chambers, K. Ferrara, E. L. Newport, What you say versus how you say it: Comparing sentence comprehension and emotional prosody processing using fMRI. *Neuroimage* **209**, 116509 (2020), 10.1016/j.neuroimage.2019.116509.
89. C. Roberts, Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semant. Pragmat.* **5**, 1–69 (1996).
90. S. Larsson, "Questions under discussion and dialogue moves." In Proceedings of the thirteenth Twente Workshop on Language Technology/TWENDIAL'98: Formal semantics and pragmatics of dialogue (1998).
91. A. Kehler, "Discourse coherence" in *The Handbook of Pragmatics* (2006), pp. 241–265.
92. J. R. Hobbs, *On the Coherence and Structure of Discourse. Technical Report 85-37* (Center for the Study of Language and Information (CSLI), Stanford, CA, 1985).
93. F. Wolf, E. Gibson, Representing discourse coherence: A corpus-based study. *Comput. Linguist.* **31**, 249–287 (2005).
94. G. Chierchia, Scalar implicatures and their interface with grammar. *Annu. Rev. Linguist.* **3**, 245–264 (2017).
95. B. Geurts, *Quantity Implicatures* (Cambridge University Press, 2010).
96. M. R. Perkins, "Pragmatic impairment" in *The Handbook of Language and Speech Disorders* (2010), pp. 227–246.
97. V. Bambini *et al.*, The communicative impairment as a core feature of schizophrenia: Frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Compr. Psychiatry* **71**, 106–120 (2016).
98. A. Parola *et al.*, Assessment of pragmatic impairment in right hemisphere damage. *J. Neurolinguistics* **39**, 10–25 (2016).
99. L. R. Dougherty, L. M. Guillette, Linking personality and cognition: A meta-analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, 20170282 (2018).
100. S. E. Brennan, H. H. Clark, Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 1482–1493 (1996), 10.1037//0278-7393.22.6.1482.
101. S. E. Brennan, J. E. Hanna, Partner-specific adaptation in dialog. *Top. Cogn. Sci.* **1**, 274–291 (2009).
102. S. Goldin-Meadow, Widening the lens: What the manual modality reveals about language, learning and cognition. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130295 (2014).
103. J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world?. *Behav. Brain Sci.* **33**, 61–83 (2010).
104. D. E. Blasi, J. Henrich, E. Adamou, D. Kemmerer, A. Majid, Over-reliance on English hinders cognitive science. *Trends Cogn. Sci.* **26**, 1153–1170 (2022).
105. S. Blum-Kulka, J. House, G. Kasper, "Cross-cultural pragmatics: Requests and apologies" in Advances in Discourse Processes, XXXI (Ablex, 1989).
106. S. Floyd, Conversation and culture. *Annu. Rev. Anthropol.* **50**, 219–240 (2021).
107. A. Ninio, C. Snow, *Pragmatic Development* (Routledge, 2018).
108. E. V. Clark, "Pragmatics and language acquisition" in *The Handbook of Pragmatics* (2006), pp. 562–577.
109. G. Tuckute, N. Kanwisher, E. Fedorenko, Language in brains, minds, and machines. *Annu. Rev. Neurosci.* **47**, 277–301 (2024).
110. P. Jeretic, A. Warstadt, S. Bhooshan, A. Williams, Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. arXiv [Preprint] (2020), https://doi.org/10.48550/arXiv.2004.03066 (Accessed 21 August 2025).
111. J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, E. Gibson, "A fine-grained comparison of pragmatic language understanding in humans and language models" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023), pp. 4194–4213.
112. L. Ruis *et al.*, The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. *Adv. Neural Inform. Process. Syst.* **36**, 20827–20905 (2023).
113. Y. Cong, Manner implicatures in large language models. *Sci. Rep.* **14**, 29113 (2024).
114. B. Ma *et al.*, Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. arXiv [Preprint] (2025), https://doi.org/10.48550/arXiv.2502.12378 (Accessed 21 August 2025).
115. B. Lipkin, L. Wong, G. Grand, J. B. Tenenbaum, Evaluating statistical language models as pragmatic reasoners. arXiv [Preprint] (2023), https://doi.org/10.48550/arXiv.2305.01020 (Accessed 21 August 2025).
116. M. Chmielewski, S. C. Kucker, An MTurk crisis? Shifts in data quality and the impact on study results. *Soc. Psychol. Personal. Sci.* **11**, 464–473 (2020).
117. S. A. Dennis, B. M. Goodson, C. A. Pearson, Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behav. Res. Account.* **32**, 119–134 (2020).
118. C. Hedge, G. Powell, P. Sumner, The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).
119. S. C. Goodhew, M. Edwards, Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Conscious. Cogn.* **69**, 14–25 (2019).
120. S. Floyd, M. Poliak, E. Gibson, E. Fedorenko, PragMega. Open Science Framework. https://doi.org/10.17605/OSF.IO/DPGE6. Deposited 27 June 2022.