An individual differences perspective on the development of pragmatic abilities in the preschool years

Manuel Bohn[1], Michael Henry Tessler[2,3], Clara Kordt[4], Tom Hausmann[5], & Michael C. Frank[6]

[1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2] DeepMind, London, UK

[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA

[4] Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

[5] Brandenburg Medical School Theodor Fontane, Neuruppin, Germany

[6] Department of Psychology, Stanford University, Stanford, USA

Author Note

We are very thankful to Stella Christie for sharing the material for the relational match-to-sample task with us. M. Bohn received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 749229. M. H. Tessler was funded by the National Science Foundation SBE Postdoctoral Research Fellowship Grant No. 1911790. M. C. Frank was supported by a Jacobs Foundation Advanced Research Fellowship and the Zhou Fund for Language and Cognition. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors made the following contributions. Manuel Bohn: Conceptualization, Methodology, Formal Analysis, Visualization, Writing – original draft, Writing – review & editing; Michael Henry Tessler: Formal Analysis, Writing – review & editing; Clara Kordt: Investigation, Writing – review & editing; Tom Hausmann: Investigation, Writing – review & editing; Michael C. Frank: Conceptualization, Writing – review & editing.

Correspondence concerning this article should be addressed to Manuel Bohn, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: manuel_bohn@eva.mpg.de

Abstract

Pragmatic abilities are fundamental to successful language use and learning. Individual differences studies contribute to understanding the psychological processes involved in pragmatic reasoning. Small sample sizes, insufficient measurement tools, and a lack of theoretical precision have hindered progress, however. Three studies addressed these challenges in three- to five-year-old German-speaking children (N = 228, 121 female). Studies 1 and 2 assessed the psychometric properties of six pragmatics tasks. Study 3 investigated relations among pragmatics tasks and between pragmatics and other cognitive abilities. The tasks were found to measure stable variation between individuals. Via a computational cognitive model, individual differences were traced back to a latent pragmatics construct. This presents the basis for understanding the relations between pragmatics and other cognitive abilities.

Keywords: Pragmatics, language development, individual differences, cognitive modeling

Word count: 8558

An individual differences perspective on the development of pragmatic abilities in the preschool years

Introduction

Communication predates language. Before children produce their first words, they communicate with the world around them using vocalizations and gestures (Bates, Benigni, Bretherton, Camaioni, & Volterra, 1979; Bruner, 1974). The process of language learning recruits many of the social-cognitive processes that underlie pre-verbal communication (Bohn & Frank, 2019; E. V. Clark, 2009; Tomasello, 2009). Even for proficient language users, communication is not reducible to the words being exchanged. The common thread running through the different aspects of human communication is its inferential nature: what a speaker means – verbal or otherwise – is underdetermined by the parts that make up the utterance. It takes contextual social inferences, often referred to as *pragmatic* inferences, to recover the intended meaning (Bohn & Köymen, 2018; H. H. Clark, 1996; Grice, 1991; Levinson, 2000; Sperber & Wilson, 2001).

The development of pragmatics has been widely studied in recent years (for a recent review see Bohn & Frank, 2019). This research covers a range of different phenomena ranging from so-called *pure pragmatics* (Matthews, 2014) in non-verbal communication in infancy to sophisticated linguistic inferences developing much later (Huang & Snedeker, 2009; Papafragou & Skordos, 2016). A growing portion of this work is devoted to studying individual differences (Matthews, Biney, & Abbot-Smith, 2018; E. Wilson & Katsos, 2021). The motivation behind the move to study individual variation is twofold: first, individual differences offer insights into the underlying psychological processes. If two phenomena (e.g. pragmatic reasoning and executive functions) vary together this is consistent with shared cognitive processes (Kidd, Donnelly, & Christiansen, 2018; Matthews et al., 2018; A. Wilson & Bishop, 2022), though it is not definitive evidence for such a claim. Second, deficits in pragmatic abilities have been linked to maladaptive behavioral patterns and forms of language impairment (Helland, Lundervold, Heimann, & Posserud, 2014).

In their recent review, Matthews et al. (2018) identified three issues that significantly limit what we can learn from individual differences research on pragmatic abilities. First, most studies have insufficient sample sizes so that small and medium sized correlations among pragmatics tasks and between pragmatics tasks and measures for other cognitive abilities cannot be reliably detected (mirroring issues in estimating correlations across other fields, Schönbrodt & Perugini, 2013). Second, the tasks used to assess pragmatic abilities often have poor or unknown psychometric properties. For example, many tasks only have a single trial and are therefore unable to capture variation between children (see also Enkavi et al., 2019). Furthermore, reliability is not assessed, making it unclear if the task captures stable characteristics (Flake & Fried, 2020; Russell & Grizzle, 2008). Third, the cognitive processes underlying pragmatic inferences in a particular task are underspecified. As a consequence, there is often no clear rationale for why a particular target task should correlate with another cognitive measure.

In search of a better understanding of individual variation in pragmatic ability, the studies presented here directly address these issues. We identified six pragmatic reasoning tasks in children between three and five years of age and investigated their psychometric properties, in particular their re-test reliability. Reliable tasks are a necessary precondition for meaningful individual differences research (Fried & Flake, 2018; Hedge, Powell, & Sumner, 2018). Next, we investigated the relations among different pragmatic reasoning tasks as well as between pragmatic reasoning and other cognitive abilities in a sample large enough to detect small to medium sized correlations. For this purpose, we introduced computational cognitive models of pragmatic reasoning to the study of individual differences. Computational cognitive models formalize hypotheses about cognitive processes that could underlie pragmatic reasoning; thus, the use of these models provides a substantive theoretical account of why certain pragmatic reasoning tasks should be related to one another. Here, we use the formalism introduced by the Rational Speech Act (RSA) framework (Frank &

Goodman, 2012; Goodman & Frank, 2016). RSA models see pragmatic inferences as a special case of (Bayesian) social reasoning. A pragmatic listener interprets an utterance by assuming it was produced by a cooperative speaker. The speaker tries to be informative, that is, they provide messages that would increase the probability that the listener will recover their intended meaning. The informativeness of an utterance arises from a contrastive inference in which the effects of multiple – plausible – utterances are compared. We assume that this inference process is shared by some of the pragmatics tasks involved in this study and can thus be used to account for individual differences (see below).

The six tasks we selected were developmental adaptations of referential communication games inspired by research in experimental pragmatics (Noveck & Reboul, 2008; Noveck & Sperber, 2004). They all share a common trial-by-trial structure in which the test event always involved an agent producing an ambiguous utterance that the child had to resolve using pragmatic reasoning. This structure allowed us to run multiple trials per task, increasing reliability. We grouped the tasks into two broad categories (Figure 1). *Utterance-based tasks* asked children to derive inferences from the words and gestures the speaker produced in context. *Common ground/discourse-based tasks* asked children to derive inferences from the social interaction that preceded the utterance.

For the utterance-based category, we selected mutual exclusivity, informativeness inference, and ad-hoc implicature tasks. "Mutual exclusivity" describes the phenomenon that children tend to map a novel word to an unknown object (Bion, Borovsky, & Fernald, 2013; E. V. Clark, 1988; Halberda, 2003; Lewis, Cristiano, Lake, Kwan, & Frank, 2020; Markman & Wachtel, 1988; Merriman, Bowman, & MacWhinney, 1989). Following Lewis et al. (2020), we use the term "mutual exclusivity" as a convenient term to denote a specific task. This term is also related to a particular theoretical account of the phenomenon (Markman, 1990), but we do not presuppose that specific account. Informativeness inferences describe situations in which children identify the referent of a novel word by assuming that the speaker is trying to be informative. Being informative translates to using words that reduce ambiguity and help the listener to recover the intended meaning (Frank & Goodman, 2014). Ad-hoc implicature describes inferences that ask the child to contrast an utterance with alternatives that the speaker could have used but did not (Katsos & Bishop, 2011; Stiller, Goodman, & Frank, 2015; Yoon & Frank, 2019).

For the discourse-based category, we selected speaker preference, discourse novelty and discourse continuity tasks. In the speaker preference task, the child had to track the preference of a speaker in order to identify the referent of a novel word (Saylor, Sabbagh, Fortuna, & Troseth, 2009). Discourse novelty refers to a situation in which the child tracks the temporal appearance of objects and expects the speaker to refer to objects that are new in context (Akhtar, Carpenter, & Tomasello, 1996; Diesendruck, Markson, Akhtar, & Reudor, 2004). In the discourse continuity task, the child had to infer and track the topic of an ongoing conversation to resolve ambiguity (Akhtar, 2002; Bohn, Le, Peloquin, Köymen, & Frank, 2021).

In addition to the pragmatics tasks, we also included two additional cognitive tasks: one measuring executive functions (Zelazo, 2006) and the other analogical reasoning (Christie & Gentner, 2014). Executive functions refer to a family of top-down mental processes that enable us to inhibit automatic or intuitive responses and allow us to concentrate and focus attention on particulars (Diamond, 2013). A substantial body of research has investigated the link between executive functions and pragmatics – with mixed results (Matthews et al., 2018; Nilsen & Graham, 2009). Analogical reasoning refers to the ability to reason about abstract relations between stimuli (Carstensen & Frank, 2021) – an ability that, to our knowledge, has not been specifically linked to pragmatics – at least not to the same extent as executive functions.

Study 1 and 2 explored the re-test reliability of the pragmatics tasks and found it to be relatively good. Study 3 tested a larger sample of children to investigate relations between the three utterance-based tasks. We focused on these tasks for theoretical reasons: as noted above, we assume that – computationally – they share a common contrastive inference process. We formalize these assumptions in a computational cognitive model which we then use to study individual differences in this alleged process. Study 3 also included tasks for executive functions and analogical reasoning. Across analytical approaches, we found systematic relations among the pragmatics tasks as well as between pragmatics and executive functions, but not analogical reasoning. In the discussion, we use the structure of the cognitive model to speculate about the psychological processes shared between pragmatics and executive functions.

Taken together, this study introduces a set of tasks that reliably measure individual differences in pragmatic abilities in the preschool years. In addition, it introduces a new (formal) theoretical framework that help us understand individual differences on a process level and, with that, suggests answers to why pragmatic abilities relate to other cognitive abilities.

Study 1

Study 1 focused on the psychometric properties of four pragmatics tasks, in particular, their re-test reliability. We chose our sample size so that we would detect medium to high re-test correlations with sufficient power. Two of the tasks were from the utterance-based group and two from the common ground/discourse-based group. This design allowed us to explore whether tasks within one group are more related to one another than between groups. As a fifth task, we included a measure of executive functions. Methods and sample size were pre-registered at https://osf.io/6a723. All analysis scripts and data files can be found in the following repository: https://github.com/manuelbohn/pragBat. The same repository also contains the code to run the experiments.

Participants

For Study 1, we collected data from 48 children ($m_{age} = 3.99$, range$_{age}$: 3.10 - 4.99, 23 girls), of whom 41 were tested twice. For most children, the two test sessions were two days apart; the longest time difference was six days. Children came from an ethnically homogeneous, mid-size German city (~550,000 inhabitants, median income €1,974 per month as of 2020); were mostly monolingual and had mixed socioeconomic backgrounds. The study was approved by an internal ethics committee at the Max Planck Institute for Evolutionary Anthropology. Data was collected between November 2019 and January 2020.

Material and Methods

The study was presented as an interactive picture book on a tablet computer (Frank, Sugarman, Horowitz, Lewis, & Yurovsky, 2016). The tasks were programmed in `HTML/JavaScript` and run in a web browser. Pre-recorded sound files were used to address the child (one native German speaker per animal). Children responded by touching objects on the screen. Children were tested in a quiet room in their daycare or in a separate room in a child laboratory. An experimenter guided the child through the study, selecting the different tasks and advancing within each task. In the beginning of the study, children completed a touch training to familiarize themselves with selecting objects. After a short introduction to the different animal characters, children completed the following six tasks. Figure 1 shows screenshots for each task and the order in which they were presented.

Training. An animal was standing on a pile between two tables. On each table, a familiar object was located. The animal asked the child to give them one of the objects (e.g., "Can you give me the car"). The objects were chosen so that children of the youngest age group would easily understand them (car and ball). This procedure familiarized the child with the general logic of the animals making requests and the child touching objects. There were two training trials.

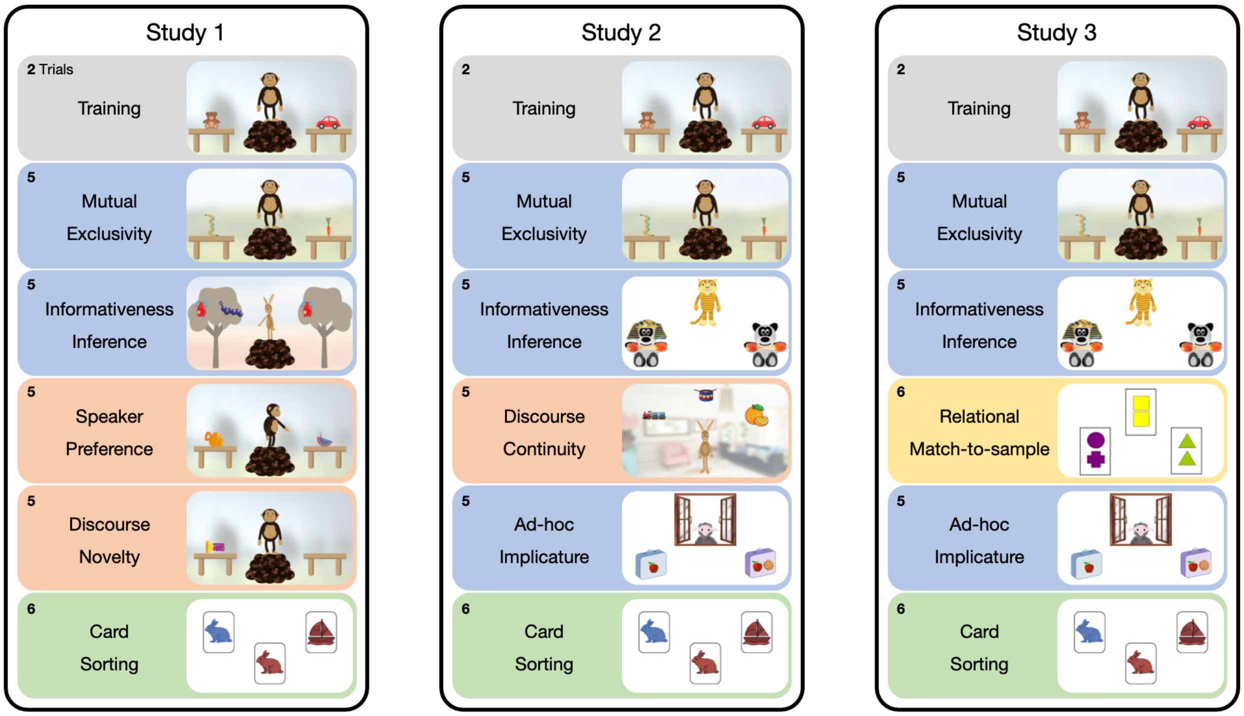

Figure 1. Overview of the tasks used in Study 1 to 3. Pictures show screenshots from each task. The vertical order corresponds to the order of presentation in each study. The colors group the tasks along the (assumed) cognitive processes involved. Blue: utterance-based inferences. Red common ground/discourse-based inferences. Green: Executive functions. Yellow: Analogical reasoning. Bold numbers show the number of trials per task.

Mutual exclusivity. This task was adapted from Bohn, Tessler, Merrick, and Frank (2021). The task layout and the procedure was the same as in the training. In each trial, one object was a novel object (drawn for the purpose of this study) while the other one was likely to be familiar to children. Both object types changed from trial to trial. Following Bohn, Tessler, et al. (2021), the familiar objects varied in terms of the likelihood that they would be familiar to children in the age range (carrot, duck, eggplant, garlic, horseshoe). For example, we assumed that most 3-year-olds would recognize a carrot, whereas fewer children would recognize a horseshoe. The animal always used a novel non-word (e.g., gepsa) in their request. We reasoned that children would identify the novel object as the referent of the novel word because they assumed the animal would have used the familiar word if they wanted to request the familiar object. Children's response was thus coded as correct if they selected the novel object. There were five trials, with the side on which the novel object appeared pseudo-randomized.

Informativeness inference. The task was adapted from Bohn, Tessler, Merrick, and Frank (2022). The animal was standing between two trees with objects hanging in them. In one tree, there were two objects (type A and B) and in the other tree there was only one (type B). The animal turned to the tree with the two objects and labeled one of the objects. It was unclear from the animal's utterance, which of the two objects they were referring to. We assumed that children would map the novel word onto the object of type A because they expected the animal to turn to the tree with only the object of type B if their intention was to provide a label for an object of type B. Next, the trees were replaced by new ones, one of which carried an object of type A and the other of type B. The animal then said that one of the trees had the same object as they labeled previously (using the same label) and asked the child to touch the tree. We coded as correct if the child selected the tree with the object of type A. The first two trials were training trials, in which there was only one object in each tree. There were five test trials. The location of the tree with the two objects in the beginning of each trial was pseudo-randomized and so was the location of the objects when the new trees appeared.

Speaker preference. This task was also adapted from Bohn et al. (2022). The animal was standing between the two tables, each of which had a novel object (drawn for the purpose of the study) on it. The animal turned to one table, pointed at the object and said that they very much liked this object (using a pronoun instead of a label). Next, the animal turned to the other table and said that they really did not like the object (again, using a pronoun and no label). Then the animal turned towards the participant and used a novel label to request an object in an excited tone. We assumed that children would track the animal's preference and identify the previously liked object as the referent. Thus, we coded as correct if the child selected the object the animal expressed preference for. There were five test trials. The location of the preferred object as well as whether the animal first expressed liking or disliking was pseudo-randomized across trials

Discourse novelty. This task was adapted from Bohn, Tessler, et al. (2021). Once again, the animal was standing between the two tables. One table was empty whereas there was a novel object on the other table. The animal turned towards the empty table and commented on its emptiness. Next, the animal turned to the other table and commented (in a neutral tone) on the presence of the object (not using a label). The animal then briefly disappeared. In the absence of the animal a second novel object appeared on the previously empty table. Then the animal returned and, facing the participant, asked for an object in an excited tone. We assumed that children would track which object was new to the ongoing interaction and identify the object that was new in context as the referent. We coded as correct when children selected the object that appeared later. There were five test trials. The location of the empty table and whether the animal first commented on the presence or absence of an object was pseudo-randomized across trials

Card sorting. This task was modeled after Zelazo (2006). The child saw two cards, a blue rabbit on the left and a red boat on the right. The experimenter introduced the child to the color game they would be playing next. In this game, all blue cards (irrespective of objects depicted) would go to the left card and all red cards to the right. Next, a third card appeared in the middle of the screen (red rabbit or blue boat) and the experimenter demonstrated the color sorting by moving the card to the one with the same color. After a second demonstration trial, the child started to do the color sorting by themselves. After six trials, the experimenter said that they were now going to play a different game, the shape game, according to which all rabbits would go to the card with the rabbit (left) and all boats to the card with the boat (right). The experimenter repeated these instructions once and without any demonstration the child continued with the sorting according to the new rule. There were six test trials. The shape on the card was pseudo-randomized across trials. We only coded the trials after the rule change and coded as correct when the child sorted according to shape.

Each child received exactly the same version of each task and completed the tasks in the same order, with the same order on the two days. This ensured comparability of performance across children.

Analysis

We analyzed the data in three steps. First we investigated developmental effects in each task, then we assessed re-test reliability, and finally, we looked at relations between the tasks. All analyses were run in `R` (R Core Team, 2018) version 4.1.2. Regression models were fit as Bayesian generalized linear mixed models (GLMM) using the function `brm` from the package `brms` (Bürkner, 2017). We used default priors for all analysis.

To estimate developmental effects in each task, we fit a GLMM predicting correct responses (0/1) by age (in years, centered at the mean) and trial number (also centered). The model included random intercepts for each participant and random slopes for trial within participants (model notation in `R`: `correct ~ age + trial + (trial|id)`). We pre-registered the inclusion of random intercepts for item. We deviate from this here because the order of items was fixed and the same for all participants so that trial and item were confounded for each task. For each task, we inspected and visualized the posterior distribution (mean and 95% Credible Interval (CrI)) for the age estimate.

We assessed re-test reliability in two ways. First, for each task we computed the proportion of correct trials for each individual in the two test sessions and then used Pearson correlations to quantify re-test reliability. Second, we used a GLMM based approach suggested by Rouder and Haaf (2019). Here, a GLMM was fitted to the trial-by-trial data for each task with a fixed effect of age (in years, centered at the mean), a random intercept for each participant and a random slope for test day (`correct ~ age + (0+test_day|id)`). The notation `0+test_day` yields a separate intercept estimate for each test day and subject instead of an intercept estimate for day 1 and a slope for the difference between day 1 and day 2. As a consequence, the model estimates the correlation between the two test days instead of a correlation between an intercept and the slope for test day. The correlation between test days can be interpreted as the re-test reliability. This approach has several advantages. First, it uses trial-by-trial data and avoids information loss that comes with data aggregation. Second, it uses hierarchical shrinkage to obtain better participant-specific estimates. Finally, it allows us to get an age-independent estimate for reliability. One worry when assessing re-test reliability in developmental studies is that re-test correlations can be high because of domain general cognitive gains and not because of task-specific individual differences. By including age as a fixed effect in the model, the estimates for each participant are independent of age and so is the correlation between estimates for the two test days – the re-test reliability.

Finally, we used aggregated data from both test days for each participant and task to compute Pearson correlations between the different tasks. Given the small sample size in Study 1, this part of the analysis was mostly exploratory.

Results

We found developmental effects in most of the tasks. Figure 2 shows the data and visualizes the developmental trajectories based on the model. Figure 3 shows the model estimates for age. In the mutual exclusivity task, performance was reliably above chance level and increased with age. For informativeness inference, the pattern was quite different: Performance was at chance level with only minor developmental gains. In the speaker preference task, performance was again clearly above chance with developmental gains resulting in a ceiling effect for older children. In the discourse novelty task, performance was also above chance with no clear developmental effects. The card sorting task showed the strongest developmental effects with younger children performing largely below chance and older children performing above chance.

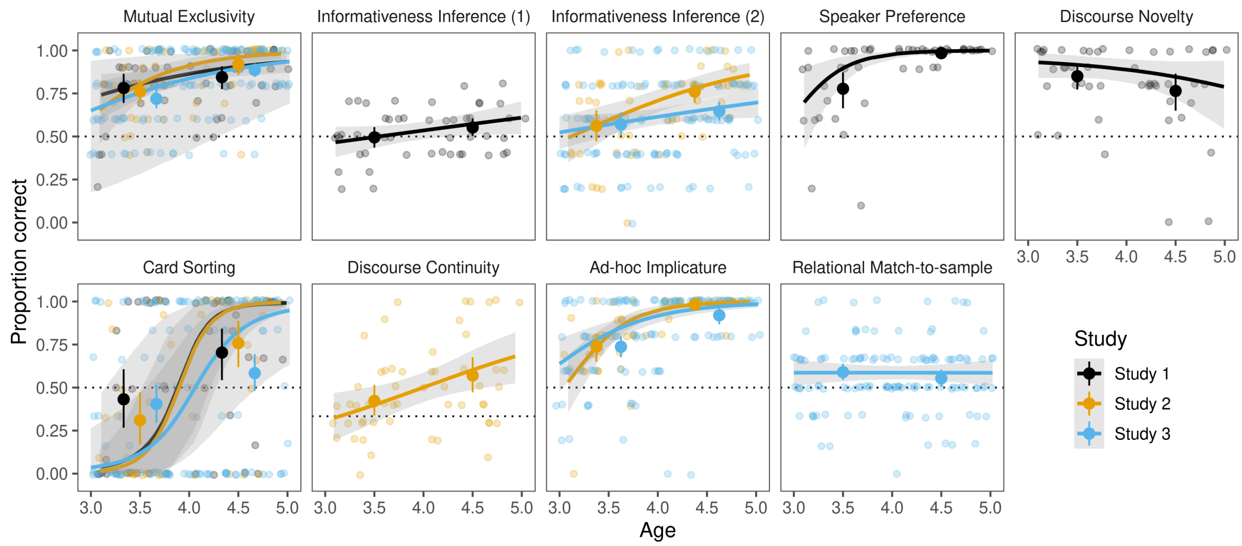

Figure 2. Results by task for studies 1 to 3. Each panel shows the results for one task. Regression lines show the predicted developmental trajectories (with 95% CrI) based on by-task GLMMs, with the line type indicating the study. Colored points show age group means (with 95% CI based on non-parametric bootstrap) with the different shapes corresponding to the different studies. Light shapes show the mean performance for each subject by study. Dotted line shows the level of performance expected by chance.

Re-test reliability was high for most tasks (see Figure 4). Raw correlation between the two test sessions was above .7 for mutual exclusivity, speaker preference and discourse novelty, though it was slightly lower for card sorting (.62). The model based – age independent – reliability estimates yielded similar results suggesting that the tasks did capture task specific individual differences. A notable exception was the informativeness inference task, which was not reliable according to any of the methods of computing re-test reliability (Figure 4). We suspected the overall low variation in performance to be responsible for this.

Most correlations between the tasks were low and ranged between r = -0.2 and 0.2 (see Figure 2). A notable exception was the correlation between mutual exclusivity and card sorting (r = 0.31, 95% CI[0.03 - 0.55]).

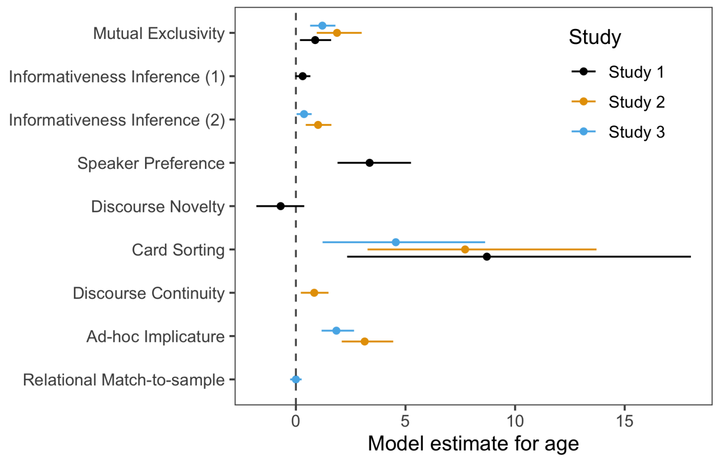

Figure 3. Model estimates (with 95% CrI) for age (in years, centered at the mean) based on GLMMs for each task and study.

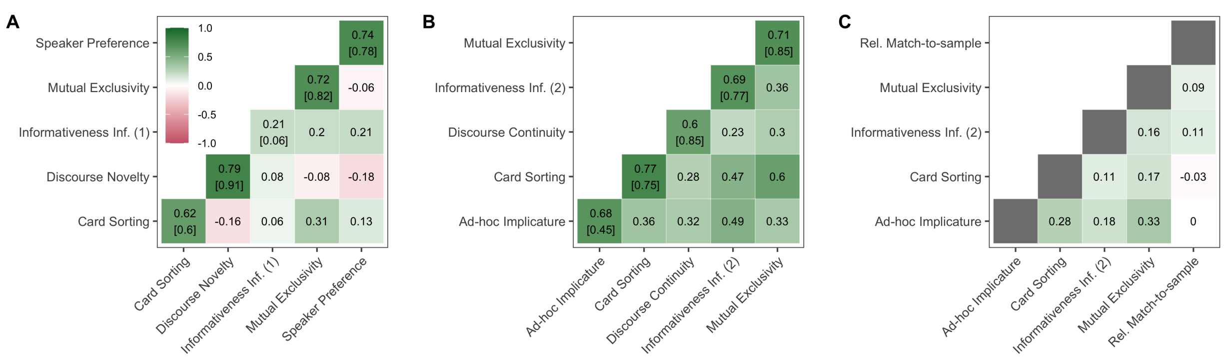

Figure 4. Re-test and task correlations for Study 1 (A), 2 (B) and 3 (C). The diagonal in A and B shows the re-test reliability based on aggregated raw test scores (top row) and based on a GLMM that accounted for participant age (see main text for details).

Discussion

Study 1 showed that the different tasks were – for the most part – age appropriate and reliable. A notable exception was the informativeness inference task which generated no systematic variation in the age range we studied here. Correlations between the tasks were generally low, with the exception of the relation between mutual exclusivity and card sorting. Given the small sample size, we avoid overly strong claims, however, it was interesting to see that the relation between the two tasks tapping into discourse-based inferences (speaker preference and discourse novelty) were – if anything – negatively correlated.

Study 2

The goal of Study 2 was to assess the re-test reliability in a new set of tasks. We retained the mutual exclusivity and card sorting tasks because of the interesting relation between the two found in Study 2. We simplified the informativeness inference task to be more age appropriate with the hope of inducing more variation in performance. We removed the speaker preference and discourse novelty tasks – despite their excellent re-test reliability – because they seemed to be unrelated to one another and also unrelated to the other tasks. We also added new tasks focused on ad-hoc implicature and discourse continuity. As noted in the introduction, we had theoretical reasons to expect the ad-hoc implicature task to be related to the mutual exclusivity and informativeness inference tasks. We had no such strong predictions for the discourse continuity task.

Methods and sample size were pre-registered at https://osf.io/hp9f7. Data, analysis scripts and experiment code can be found in the associated online repository.

Participants

Participants for Study 2 were recruited from the same general population. We collected data from 54 children (m_{age} = 3.97, range$_{age}$: 3.09 - 4.93, 24 girls), of whom 40 were tested twice. The two test sessions were again two days apart; the longest time difference was 14 days. Data was collected between March and October 2020.

Material and Methods

The general setup and mode of presentation was the same as in Study 1. We added two new tasks and modified the informativeness inferences task, which we will describe in detail below. The training, mutual exclusivity and card sorting tasks were the same as in Study 1.

Informativeness inference. The general structure of the task was the same as in Study 1, however, we replaced the stimuli from Bohn et al. (2022) with those used originally by Frank and Goodman (2014). We suspected that many children did not treat the novel objects hanging in the tree as properties of the tree but rather viewed the tree as a "container" for the novel objects. The alleged inference, however, relies on seeing the objects as properties of the referent. With the new stimuli, we emphasized that the novel objects were mere properties of the referent by making the referent more salient and more different across trials. The animal was located between two identical objects, which had different properties (see Figure 1). For example, the child saw two bears, one with a Pharaoh-style crown and a match in its hand, the other only with the match. The animal then turned to the object with the two properties and described it by referring to one of the properties (e.g., a bear with a [non-word]). Next, the objects disappeared, and the same objects re-appeared but this time, each of them had only one property (e.g., one bear with a crown, the other with the match). The animal then asked which of these objects had the aforementioned property (e.g., which bear has a [non-word]). We coded responses as correct if the child selected the object with the property that was unique to the object during labeling. The first two trials were training trials, in which each object only had one property. There were five test trials. The location of the object with the two properties in the beginning of each trial was pseudo-randomized and so was the location of the properties when the new objects appeared.

Discourse continuity. This task was adapted from Bohn, Le, et al. (2021). Children were told that they were going to visit the animals in their home. An animal greeted the child and told them that they would show them their things. During exposure trials, the child saw three objects from three different categories (e.g., train (vehicle), drum (instrument), orange (fruit); see Figure 1). The animal named one of the objects and asked the child to touch it. On the next exposure trial, the child saw three new objects but from the same categories (e.g., bus (vehicle), flute (instrument), apple (fruit)). The animal asked the child to touch the object from the same category as previously (only naming the object, not the category). There were 5 such exposure trials. On the following test trial, the animal used a pronoun to refer to one of the objects (i.e., can you touch *it*). We assumed that children would use the exposure trials to infer that the animal was talking about a certain category and would use this knowledge to identify the referent of the pronoun. Children received five test trials, each with a different category as the target. The position of the objects in exposure trials as well as test trials was pseudo-randomized.

Ad-hoc implicature. This task used the general procedure and stimuli developed in Yoon and Frank (2019). The animal was located in a window, looking out over two objects (see Figure 1). Both objects were of the same kind, but had different properties. As properties we chose objects that were well known to children of that age range. One object had one property (A), while the other had two (A and B). For example, objects were lunch boxes, one with an orange and the other with an orange and an apple. The animal then asked the child to hand them their object which was the one with the property that both objects shared (A). We assumed that children would pick the object with only property A because they expected the animal to name property B if they had wanted to refer to the object with both properties. There were five test trials, preceded by two training trials in which the objects did not share a common property. The positioning of the objects (left and right) was pseudo-randomized.

Analysis

We used the same methods to analyze the data as in Study 1.

Results

We found substantial developmental gains in all five tasks (Figure 2 and 3). For mutual exclusivity and ad-hoc implicature performance was above chance across the entire age range. For the informativeness inference and discourse continuity tasks, performance was close to chance for younger children and reliably above it for older children. Like in Study 1, we found the strongest developmental effect for card sorting, with performance below chance for 3-year-olds and above chance for 4-year-olds.

Re-test reliability based on aggregated data was good for all tasks with most estimates around 0.7. The model-based reliability estimates were similar, with lower values for ad-hoc implicature and higher ones for discourse continuity. Notably, the revised informativeness inference task showed a much-improved re-test reliability compared with the estimate from Study 1.

Correlations between tasks were generally higher compared to Study 1. In fact, confidence intervals for correlation coefficients were not overlapping with 0 except for the correlation between the discourse continuity and informativeness inference tasks (Figures 4. Once again, we found the strongest relation between card sorting and mutual exclusivity (r = 0.60, 95% CI[0.40 - 0.75]). Other notable relations were those between card sorting and informativeness inference (r = 0.47, 95% CI[0.23 - 0.65]) as well as between ad-hoc implicature and informativeness inference (r = 0.49, 95% CI[0.25 - 0.67]).

Discussion

In Study 2 we found good results from a measurement perspective: all tasks had acceptable re-test reliability. This result extended to the informativeness inference task, which had very low reliability in Study 1. Higher average performance and increased variability both suggest that our changes to the stimuli made the task easier for children.

As in Study 1, we found a relatively strong correlation between the mutual exclusivity and card sorting tasks. This finding supports the idea that these tasks share common processes. We also found substantial relations between the three utterance-based inference tasks (mutual exclusivity, ad-hoc implicature, informativeness inference). The correlations between these tasks and the discourse continuity task were numerically lower.

Study 3

In Study 3, we focused explicitly on the relations between the different tasks. In particular, we explored the idea that the three utterance-based inference tasks share common cognitive processes. Once again, we also included the card sorting task and added a new task of analogical reasoning as a control for which we did not expect strong relations with the other tasks. To be able to test predictions about cross-task variation, we collected data from a comparatively larger sample of children.

The reliability estimates from Study 1 and 2 helped us plan the sample size for Study 3. The focal tasks had a re-test reliability around 0.7. Because the highest plausible correlation between two tasks is the product of their reliabilities (higher correlations would mean that the task is more strongly related to a different task than to itself), the highest we could expect were correlations between two tasks around $0.7 * 0.7 = 0.49$. We planned our sample so that we could detect correlations between two tasks of 0.3 with 95% power. The first author drafted a pre-registration and shared it with the last author but forgot to register it at OSF. Thus, the study was not officially pre-registered. Data, analysis scripts and experiment code can be found in the associated online repository.

Participants

For Study 3, we collected data from 126 children ($m_{age} = 4.00$, range$_{age}$: 3.00 - 5.02, 74 girls) from the same general population. Data was collected between June and November 2021. Children were tested only once.

Materials and Methods

From Study 2, we used the mutual exclusivity, ad-hoc implicature, informativeness inference and card sorting tasks. We added the relational match-to-sample task, which we now describe in more detail.

Relational match-to-sample. The task was modeled after (and used the original stimuli from) Christie and Gentner (2014). The child saw three cards, one on top (the sample) and two at the bottom (the potential matches; see Figure 1). The experimenter guided the child through the study and read out the instructions. The child was instructed to match the sample card to one of the lower ones based on similarity, that is, they were instructed to pick the card that was "like" the sample. All cards had two geometrical shapes of the same color on them. The sample card showed two identical shapes and so did one of the potential matches. The other card showed two different shapes. We assumed that children would match the sample to the match that showed the same relation between shapes (sameness). Children received six test trials, preceded by two training trials in which one of the potential matches was identical to the sample. The position of the same-match was pseudo randomized.

Analysis

Study 3 had only one test session. Therefore, we did not investigate re-test reliability. We estimated age effects and raw correlations between tasks in the same way as in Studies 1 and 2. We used two additional methods to investigate the structure of individual differences between tasks.

First, we used Confirmatory Factor Analysis (CFA). Models were fit in a Bayesian framework using the `R` package `blavaan` (Merkle & Rosseel, 2018) using default priors. As outlined above, our focal model assumed that mutual exclusivity, ad-hoc implicature and informativeness inference load on a common pragmatics factor. The card sorting and relational match-to-sample tasks were included as separate factors. We used Posterior Predictive P-Values (PPP) to evaluate model fit (Lee & Song, 2012). A good model fit is indicated by a PPP close to 0.5 and should not be smaller than 0.1 (Cain & Zhang, 2019). We also fit two alternative models: one including only a single factor on which all tasks loaded and a second with a separate factor for each task. We compared models using WAIC (widely applicable information criterion) scores and weights (McElreath, 2018). WAIC is an indicator of out-of-sample predictive accuracy with lower values indicating better fit. WAIC weights transform WAIC values to give the probability that a particular model (out of the models considered) provides the best out-of-sample predictions. Within the focal model, we inspected the posterior estimates (with 95%CrI) for the factor loadings and the variance in the task explained by the factor for the three pragmatics tasks. In addition, we evaluated the correlations between the pragmatics factor and the other two tasks.

Second, we used computational cognitive models from the Rational Speech Act (RSA) framework to relate the three pragmatics tasks to one another (Frank & Goodman, 2012; Goodman & Frank, 2016). In contrast to the CFA model above, the RSA models are models of the tasks, and not of the data. That is, they include a schematic representation of the experimental tasks and provide a computational account of how participants make inferences in this context. RSA models see pragmatic inferences as a form of Bayesian social reasoning where the listener tries to infer the speaker's meaning (here: the intended referent) by assuming that the speaker is helpful and informative. Being helpful and informative means that the speaker chooses a message based on the probability that it would help the listener to recover the speaker's intended meaning (i.e., select the intended referent). Thus, RSA models have a recursive structure in which the the listener reasons about a speaker who is reasoning about the listener. To avoid an infinite regress, the speaker is assumed to reason about a literal listener, who interprets utterances according to their literal semantics.

The studies from which we took the mutual exclusivity and informativeness inference tasks also formalized these tasks in an RSA-style model (Bohn, Tessler, et al., 2021; Bohn et al., 2022). We refer to this earlier work for a more detailed description of the models. For the present study, we formalized the ad-hoc implicature task within the same RSA framework. The common model structure is formally defined as:

$$P_{L_1}(r|u) \propto P_{S_1}(u|r) \cdot P(r)$$

In the above equation, the listener (P_{L_1}) is trying to infer the speaker's (P_{S_1}) intended referent r by imagining what a rational speaker would say, given the referent they are trying to communicate and the listener's prior expectations about the referent $P(r)$ (which we assumed to be uniform over potential referents). The speaker is an approximately rational Bayesian actor (with degree of rationality α) who produces utterances as a function of their informativity.

$$P_{S_1}(u|r) \propto Informativity(u;r)^{\alpha}$$

The informativity of an utterance for a referent is taken to be the probability with which a naive listener (P_{L_0}), who only interprets utterances according to their literal semantics, would select a particular referent given an utterance.

$$Informativity(u;r) = P_{L_0}(r|u)$$

The three models differ in the types of utterances that are being produced, however, they share the same contrastive inference process according to which the listener (P_{L_1})

compares the speaker's (P_{S_1}) utterance to a set of alternative, possible utterances. As noted above, the listener expects the speaker to be informative (with degree α) that is, choose the utterance that best communicates the intended message. In the mutual exclusivity task, the speaker produced an unfamiliar word; thus, the alternative utterance for the speaker would have been to use a familiar word. In the case of the informative inference task, the speaker pointed to the object with two properties; thus, the alternative would have been to point to the object with only one property. For the ad-hoc implicature task the speaker referred to the property shared by the two objects, which contrasts with referring to the property that was unique to one of the objects. In all cases, these alternative utterances would be better suited to communicate about the respective other referent.

As noted above, models for the different tasks shared one common parameter: the speaker informativeness parameter α. This commonality offers a way of relating performance in the three tasks to one another by constraining the three models to use the same value for α. We then used Bayesian inference to estimate the posterior distribution for α that best explained performance in the three tasks. To adapt this framework to the study of individual differences, we allowed a separate parameter for each participant (α_i). We estimated α_i in a hierarchical model as a deviation from a hyper parameter: $\alpha_i \sim \mathcal{N}(\alpha_j, \sigma^{\alpha})$. Given the developmental nature of our data, we defined α_j via a linear regression as a function of the child's age (age_i): $\alpha_j = \beta_0^{\alpha} + age_i \cdot \beta_1^{\alpha}$. Thus, the participant-specific value for α was not only constrained by the performance in the three tasks but also by the child's age.

To account for differences in difficulty between the tasks due to other factors, we added a scale parameter to the model that adjusted α for each task in comparison to a reference task (ad-hoc implicature).

To validate this approach, we first applied this model to the data from Study 2 separately for each test session. This allowed us to compute the re-test reliability of α and see if it captures individual differences equally well compared to the raw test scores. After finding excellent re-test reliability, we applied it to the data from Study 3 and correlated the results with the card sorting and relational match-to sample tasks. For this correlational analysis, we converted the posterior distribution for each participant into a single value by taking the mode (and 95% highest density interval – HDI). The cognitive models were implemented in `WebPPL` (Goodman & Stuhlmüller, 2014) and the corresponding code, including information on prior distributions (which we omit here for space), can be found in the associated online repository.

Results

The age effects in Study 3 largely replicate those of Study 2 for the four overlapping tasks (see Figure 2 and 3). There were no substantial developmental gains in the newly added relational match-to-sample task and performance was close to chance for both age groups. Thus – in the absence of information on re-test reliability – it is unclear if the variation in performance reflects systematic individual differences in analogical reasoning or not.

Overall, the correlations between the tasks were lower compared to Study 2. This was to some extent expected given that there were only half the number of trials per task in Study 3 and, hence less "signal" (systematic, non-error variability) for capturing individual differences. Nevertheless, the overall pattern resembles that found in Study 2 (Figure 4). We saw the strongest bi-variate relation between the mutual exclusivity and the ad-hoc implicature task (r = 0.33, 95% CI[0.16 - 0.48]) followed by ad-hoc implicature and card sorting (r = 0.28, 95% CI[0.11 - 0.44]). The relational match-to-sample task showed no substantial correlations with any of the other tasks.

Next, we turn to the results of the confirmatory factor analysis. Our focal model – including a latent factor for pragmatic reasoning – fit the data well (PPP = 0.50) and with a WAIC of 1,753.45 (se = 32.02, weight = 0.74) better compared to the two alternative models (individual factors model: PPP = 0.51, WAIC = 1,756.48, se = 32.51, weight = 0.16; one

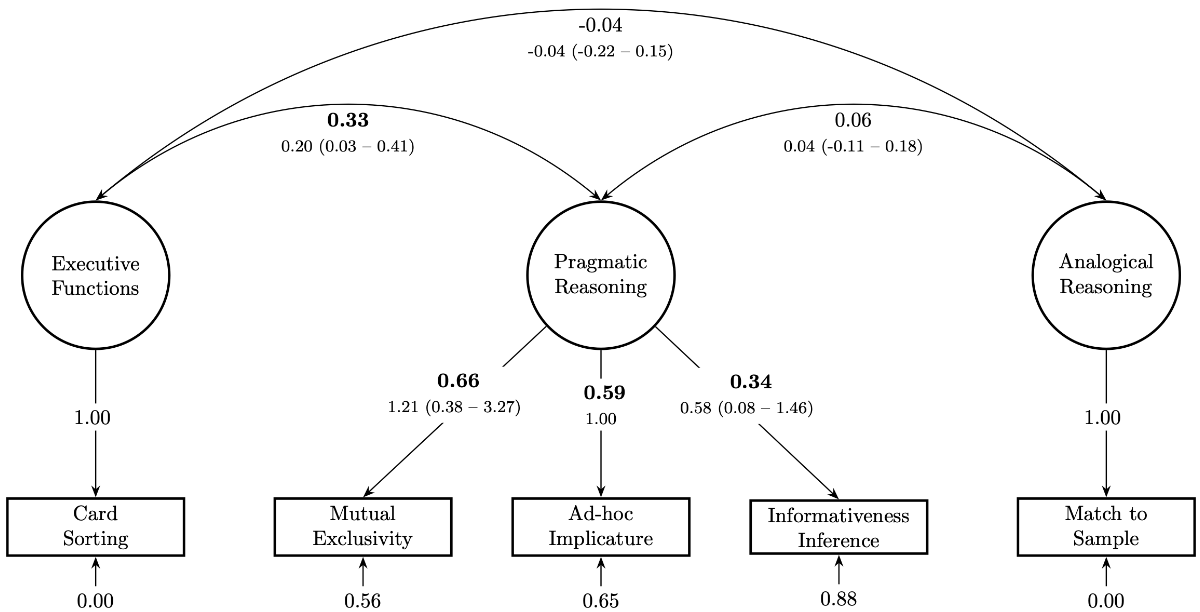

Figure 5. Graphical overview of CFA model for Study 3. Arrows from latent variable (circles) to observed variable (rectangles) show factor loadings. Bottom arrows to observed variables give the residual variance not explained by the factor. Bent arrows between latent variables show correlations. Bottom rows show model estimates with 95% CrI. Top rows show standardized estimates (bold if 95 % CrI does not include 0).

factor model: PPP = 0.36, WAIC = 1,758.10, se = 32.37, weight = 0.07).

Figure 5 shows factor loadings for the individual tasks as well as their residual variance. The latent pragmatic reasoning factor best explained the mutual exclusivity task, followed by the ad-hoc implicature and the informativeness inference task. The correlation between pragmatic reasoning and executive functions (indicated by the card sorting task) was estimated to be reliably different from zero (r = 0.33; model estimate = 0.20, 95% CrI [0.02 - 0.39]). There was no systematic relation between pragmatic reasoning and analogical reasoning (as indicated by the relational match-to-sample task): r = 0.06; model estimate = 0.04, 95% CrI [-0.11 - 0.18]. However, the latter result should be taken with a grain of salt given the unknown psychometric properties of the relational match-to-sample task.

Finally, we present the results of the cognitive modeling analysis. Using the data from Study 2, we saw that participant specific speaker informativeness parameters (α) were highly reliable (Figure 6B). The scale parameter suggested that the mutual exclusivity task was easier and the informativeness inference task was harder compared to the ad-hoc implicature task (Figure 6C). When correlating α with performance in the other two tasks, the cognitive modeling approach yielded similar conclusions compared to the confirmatory factor analysis (Figure 6A): There was a substantial correlation with the card sorting ($r = 0.31$, 95% CI[0.15 - 0.47]) but not the relational match-to-sample task ($r = 0.03$, 95% CI[-0.15 - 0.20]). The same limitations apply to the latter result as for the confirmatory factor analysis.

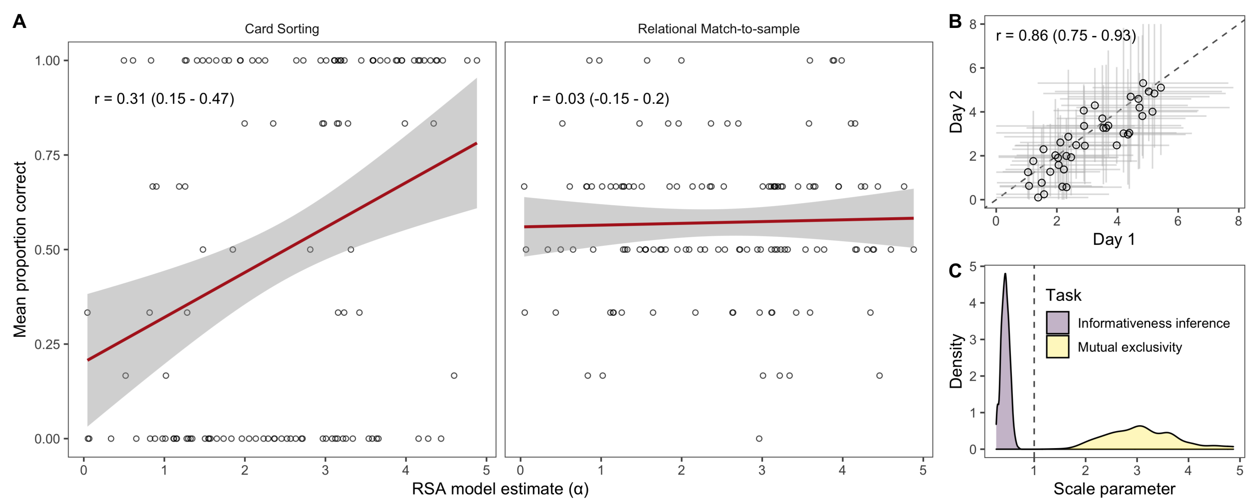

Figure 6. Results of cognitive model analyses. A: Correlation between the speaker informativeness parameter α and the performance in the card sorting and relational match to sample tasks. Regression line (with 95% CI) is based on a linear model. B: Re-test reliability for α based on the data from Study 2. C: Scale parameter for α in relation to the ad-hoc implicature task. Values below 1 indicate a more difficult task, values above 1 an easier task. Correlation coefficients show Pearson correlation with 95% CI.

Discussion

Using a diversity of analytical tools, we found that performance in the three utterance-based pragmatic inference tasks was related in a way that points to shared cognitive processes. In the confirmatory factor analysis, we found that a model including a latent pragmatic reasoning factor fit the data well and better compared to alternative models. The latent factor explained substantial portions of the variance in each of the three tasks. The cognitive modeling approach provides an explicit theory of what the shared cognitive processes may look like: according to the model, the pragmatic inference in each task was driven by contrasting the utterance the speaker produced and alternative utterances. Individual differences were thought to arise from differential expectations about how informative the speaker is.

Both analytic strategies point to systematic relations between pragmatic reasoning and executive functions as indicated by the card sorting test. We found no such relations with analogical reasoning as indicated by the relational match-to-sample task. However, given the unknown psychometric properties of the latter task, this result should be interpreted with caution.

General Discussion

In this paper, we explored the development of pragmatic inferences in the preschool years. We identified six tasks covering a broad range of pragmatic phenomena. We found them to have generally good re-test reliability. We then selected three utterance-based inference tasks for a well-powered study of relations among different types of pragmatic abilities and between pragmatics and other cognitive abilities. The results showed systematic relations between the utterance-based tasks, consistent with a latent cognitive construct. We used a computational cognitive model of pragmatic reasoning to formalize the cognitive processes we believed the tasks to share. Finally, we found pragmatic abilities to be related to a task of executive functions

One of the main contributions of this paper is that it presents six pragmatic inference tasks that are highly robust and reliable. Whenever we used a task in two studies (mutual exclusivity, informativeness inference, ad-hoc implicature), we found developmental results that replicated previous findings. In Study 1 and 2, all tasks showed good re-test reliability – even when corrected for age. A notable exception was the informativeness inference task in Study 1. However, after making some procedural changes, it turned out to be robust and reliable as well. Taken together, the tasks are suitable for individual differences research, advancing the agenda of Matthews et al. (2018). These materials are freely available via the associated online repository.

We grouped our pragmatics tasks into utterance-based and common ground/discourse based. This grouping broadly captured the kind of information that we assumed to be relevant to compute the inference. For Study 3, we focused on the three utterance-based tasks. The main reason was theoretical. We were able to build on earlier work (Bohn, Tessler, et al., 2021; Bohn et al., 2022) and formalize the inferences involved in these tasks in a common computational framework. We specified the structural overlap between the tasks and identified a parameter in the model that we used to capture individual differences. The shared structural features involve a recursive social inference process according to which the listener expects the speaker to select the most informative of a set of possible utterances. The individual difference parameter captured how informative the listener expected the speaker to be. Previous accounts would not have predicted such an overlap. In particular, theoretical accounts of mutual exclusivity as arising from heuristics or principles unconnected with pragmatic reasoning (reviewed in Lewis et al., 2020) do not make the prediction of correlations with other pragmatic tasks.

Our formal model also allowed us to speculate about why we saw a systematic relation across the three studies between pragmatic inference and the card sorting task as a measure of executive functions. Before we do so, we want to emphasize that the model is first and foremost a computational description of the tasks and not a model of a psychological process (cf. Goodman & Frank, 2016). Here we speculate, assuming a bit more psychological realism in our interpretation of the RSA model than previous authors have. The card sorting task asks the child to switch between rules after having practiced the first rule over the course of several trials. This switch requires inhibiting a pre-potent response and attending to different features of the cards. Similarly, pragmatic inference in the RSA model involves contrasting the observed utterance with alternative plausible utterances. This process, too, could be described as requiring inhibiting available, plausible interpretations and contrasting different interpretations before making a response. To pursue this connection further, the next step should be to model card sorting and the pragmatics tasks jointly to substantiate such a verbal analysis.

Limitations

The studies we presented here have important limitations. Our focus on the utterance-based pragmatic inference tasks meant that we did not study or analyze the common ground/discourse-based tasks with the same level of detail. That is, we did not formalize them in a cognitive model and did not study relations between them in a larger sample. Future research should address these shortcomings. Nevertheless, the work presented here is an important first step because it showed that the common ground/discourse tasks themselves have good psychometric properties and are therefore suitable for individual differences research.

We presented the tasks as interactive picture books on a tablet computer with animal characters as agents. This methodological step improved the quality of our measurement because it allowed us to experimentally isolate the different inferences and run multiple trials in each task. However, it also means that – like most experimental work – our tasks lack ecological validity. In real-world conversations, multiple information sources are available to listeners to draw inferences from (Bohn, Tessler, et al., 2021; Bohn et al., 2022). Furthermore, by design our experimental paradigm prevented the use of strategies that are an integral part of real-world conversations, like asking questions or seeking clarification (Arkel, Woensdregt, Dingemanse, & Blokpoel, 2020; H. H. Clark & Brennan, 1991). However, we want to highlight that the results from our tasks replicated many findings from interactive versions of these tasks (Akhtar et al., 1996; Frank & Goodman, 2014; Markman & Wachtel, 1988; Saylor et al., 2009).

Finally, we only studied one sample of children from a Western, affluent setting. Thus, it is unclear if and how the results would transfer to other settings (Nielsen, Haun, Kärtner, & Legare, 2017). The tasks used here were largely developed and tested with English-speaking children in the US. The fact that they transferred well to the German setting of the current studies is at least a small hint that they might also be suitable to study pragmatic inference in other cultural and linguistic settings. Future research will hopefully test whether that is the case.

Conclusion

The studies reported here addressed some fundamental challenges in the study of individual differences in pragmatic abilities (Matthews et al., 2018). We developed and validated new methodological and theoretical tools that helped to study the relations between different types of pragmatic inferences as well as between pragmatics and other cognitive abilities in a more reliable and valid way. This approach emphasizes the interdependent nature of theoretical and methodological progress and provides a roadmap for future work.

References

Akhtar, N. (2002). Relevance and early word learning. *Journal of Child Language*, *29*(3), 677–686.

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development*, *67*(2), 635–645.

Arkel, J. van, Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020). *A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis.*

Bates, E., Benigni, L., Bretherton, I., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Cognition and communication in infancy.* New York: Academic Press.

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39–53.

Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, *1*(1), 223–249.

Bohn, M., & Köymen, B. (2018). Common ground and development. *Child Development Perspectives*, *12*(2), 104–108.

Bohn, M., Le, K. N., Peloquin, B., Köymen, B., & Frank, M. C. (2021). Children's interpretation of ambiguous pronouns based on prior discourse. *Developmental Science*, *24*(3), e13049.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour*, *5*(8), 1046–1054.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2022). Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings. *Journal of Experimental Psychology: General.*

Bruner, J. S. (1974). From communication to language—a psychological perspective. *Cognition*, *3*(3), 255–287.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Cain, M. K., & Zhang, Z. (2019). Fit for a bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 39–50.

Carstensen, A., & Frank, M. C. (2021). Do graded representations support abstract thought? *Current Opinion in Behavioral Sciences*, *37*, 90–97.

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science*, *38*(2), 383–397.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(2), 317–335.

Clark, E. V. (2009). *First language acquisition*. Cambridge: Cambridge University Press.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition.* (pp. 127–149). American Psychological Association.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

Diesendruck, G., Markson, L., Akhtar, N., & Reudor, A. (2004). Two-year-olds' sensitivity to speakers' intent: An alternative account of samuelson and smith. *Developmental Science*, *7*(1), 33–41.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development*, *17*(1), 1–17.

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, *31*(3).

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2014). *The design and implementation of probabilistic programming languages.* http://dippl.org.

Grice, H. P. (1991). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

Helland, W. A., Lundervold, A. J., Heimann, M., & Posserud, M.-B. (2014). Stable associations between behavioral problems and language impairments across childhood–the importance of pragmatic language problems. *Research in Developmental Disabilities*, *35*(5), 943–951.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers:

Insight into the semantics–pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169.

Lee, S.-Y., & Song, X.-Y. (2012). *Basic and advanced bayesian structural equation modeling: With applications in the medical and behavioral sciences*. John Wiley & Sons.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT press.

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Matthews, D. (2014). *Pragmatic development in first language acquisition* (Vol. 10). John Benjamins Publishing Company.

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

McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*, 1–30.

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, i–129.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38.

Nilsen, E. S., & Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology*, *58*(2), 220–249.

Noveck, I. A., & Reboul, A. (2008). Experimental pragmatics: A gricean turn in the study of language. *Trends in Cognitive Sciences*, *12*(11), 425–431.

Noveck, I. A., & Sperber, D. (2004). *Experimental pragmatics*. Springer.

Papafragou, A., & Skordos, D. (2016). Scalar implicature. In J. Lidz, W. Snyder, & J. Pater (Eds.), *The Oxford Hanbook of Developmental Linguistics* (pp. 611–632). Oxford University Press.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467.

Russell, R. L., & Grizzle, K. L. (2008). Assessing child and adolescent pragmatic language competencies: Toward evidence-based assessments. *Clinical Child and Family Psychology Review*, *11*(1), 59–73.

Saylor, M. M., Sabbagh, M. A., Fortuna, A., & Troseth, G. (2009). Preschoolers use speakers' preferences to learn words. *Cognitive Development*, *24*(2), 125–132.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612.

Sperber, D., & Wilson, D. (2001). *Relevance: Communication and cognition* (2nd ed.). Cambridge, MA: Blackwell Publishers.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176–190.

Tomasello, M. (2009). *Constructing a language*. Cambridge, MA: Harvard University Press.

Wilson, A., & Bishop, D. V. (2022). A novel online assessment of pragmatic and core language skills: An attempt to tease apart language domains in children. *Journal of Child Language*, *49*(1), 38–59.

Wilson, E., & Katsos, N. (2021). Pragmatic, linguistic and cognitive factors in young children's development of quantity, relevance and word learning inferences. *Journal of Child Language*, 1–28.

Yoon, E. J., & Frank, M. C. (2019). The role of salience in young children's processing of ad hoc implicatures. *Journal of Experimental Child Psychology*, *186*, 99–116.

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*(1), 297–301.