The Dynamics of Pragmatics: Individual Differences and Adaptation in Pragmatic Processing

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von
Alexandra Mayn
aus Kasan, Russland

Saarbrücken, 2026

Dekanin der Fakultät P: Prof. Dr. Nine Miedema
Erstberichterstatterin: Prof. Dr. Vera Demberg
Zweitberichterstatter: Prof. Dr. Matthew W. Crocker
Drittberichterstatter: Prof. Dr. Ira Noveck
Tag der letzten Prüfungsleistung: 3 Juni 2026

Abstract

When people communicate with each other, they often do not say exactly what they mean, whether it is because they are being polite or trying not to hurt someone's feelings, making a joke, or not mentioning something that they expect the interlocutor to know or be able to infer from context. Still, people are able to understand each other. In pragmatics, this is commonly explained by the *cooperative principle* (Grice, 1975), which holds that interlocutors behave cooperatively and follow shared communicative norms, such as being sufficiently informative and clear. This account has received substantial experimental support. However, real-world communication is more complex than such idealized theories suggest: individuals vary widely in their pragmatic abilities (Matthews et al., 2018; Kidd et al., 2018), and these abilities may be further influenced by factors such as inattention or fatigue.

This thesis aims to advance pragmatic theories by investigating the variation in people's pragmatic abilities. Through experimental work and computational modeling, it investigates the cognitive factors which underlie individual differences in pragmatic reasoning and examines how comprehenders accommodate differences in speakers' pragmatic competence by adjusting their interpretations.

This dissertation bridges two strands of research which had previously mostly been separate – on individual differences in pragmatic processing and on modeling pragmatic inferences mathematically in the Rational Speech Act (RSA) framework (Frank & Goodman, 2012) – by investigating cognitive factors which account for variability in performance on a pragmatic reference game (Frank et al., 2016), which is theoretically motivated by the predictions of RSA models of different complexity. We show that people's pragmatic reasoning in reference games is predicted by logical reasoning ability and Theory of Mind – the ability to reason about others' mental states – but not by working memory capacity. By identifying cognitive factors which predict individual variability in pragmatic reasoning in reference games, we lay the groundwork for the study of underlying cognitive mechanisms and for processing models of this variability.

A further contribution of this dissertation is an account of how comprehenders accommodate variation in speakers' pragmatic abilities. Previous work has shown that comprehenders' interpretations are sensitive to certain speaker characteristics, such as speaker persona (Beltrama, 2018; Beltrama & Schwarz, 2021, 2024), language proficiency and accent (Ip & Papafragou, 2021, 2023; Puhacheuskaya & Järvikivi, 2025, 2022). The findings of this dissertation extend this work by showing that comprehenders' beliefs about the speaker's pragmatic competence affect interpretation. Specifically, the same ambiguous utterance may be interpreted literally or as an implicature depending on the speaker's perceived competence and cooperativity. Moreover, our findings suggest that the perceived pragmatic competence of the speaker affects the depth of the comprehender's reasoning: when the speaker is not believed to be sufficiently pragmatically competent or cooperative, comprehenders may be less

willing to expend the cognitive effort required for deeper pragmatic inference. This finding aligns with the core claim of relevance theory (Sperber & Wilson, 1986) that nonliteral meanings are derived only when the required cognitive effort is justified.

Across two experimental paradigms, we show that people rapidly adapt their interpretations to specific speakers over the course of an interaction, dynamically constructing and updating a model of the speaker's pragmatic competence. We also find evidence for individual differences in adaptation behavior which can be explained by differences in prior beliefs and tendency to generalize across speakers.

Finally, this dissertation makes a methodological contribution by proposing a simple yet effective method for eliciting strategies that people use in pragmatic reference games which can also be applied in other paradigms. We demonstrate the effectiveness of this method for uncovering and reducing bias in experimental stimuli and gaining additional insight into the used reasoning strategies which would not be apparent by examining performance.

In sum, the research presented in this dissertation contributes to more comprehensive theories of pragmatics by offering new insights into variability and adaptation in pragmatic processing. The findings of this dissertation support a nuanced view of pragmatic processing, whereby people's pragmatic reasoning is influenced by both cognitive and contextual factors, and whereby interlocutors anticipate and accommodate these differences in one another.

Kurzzusammenfassung

Wenn Menschen miteinander kommunizieren, sagen sie oft nicht genau das, was sie meinen und zwar aus verschiedenen Gründen: Sie versuchen vielleicht, höflich zu sein oder die Gefühle anderer nicht zu verletzen, machen einen Witz oder erwähnen Informationen nicht explizit, von denen sie annehmen, dass ihre Gesprächspartner:innen sie kennen oder aus dem Kontext ableiten können. Dennoch sind Menschen in der Lage, einander zu verstehen. In der Pragmatik wird dies üblicherweise mit dem *Kooperationsprinzip* (Grice, 1975) erklärt, welches besagt, dass Gesprächspartner:innen kooperativ handeln und gemeinsame Kommunikationsnormen befolgen, wie z. B. so viele Informationen wie nötig zu liefern und klar zu kommunizieren. Diese Erklärung wurde durch zahlreiche empirische Studien belegt. Die Realität der Kommunikation ist jedoch viel komplexer, als solche idealisierten Theorien vermuten lassen: Die pragmatischen Fähigkeiten der Menschen variieren stark (Matthews et al., 2018; Kidd et al., 2018) und können zusätzlich durch Faktoren wie Unaufmerksamkeit oder Müdigkeit beeinträchtigt werden.

Diese Dissertation zielt darauf ab, pragmatische Theorien durch die Untersuchung individueller Variabilität und der Anpassung von Zuhörenden an ihre Gesprächspartner:innen voranzubringen. Sie verbindet experimentelle Forschung mit rechnergestützter Modellierung, um kognitive Faktoren zu identifizieren, die individuellen Unterschieden im pragmatischen Schlussfolgern zugrunde liegen. Darüber hinaus wird untersucht, wie Zuhörende Unterschiede in der pragmatischen Kompetenz von Sprecher:innen berücksichtigen, indem sie ihre Interpretationen entsprechend anpassen.

Diese Dissertation verbindet zwei Forschungsstränge, welche bisher weitgehend getrennt waren – individuelle Unterschiede in der pragmatischen Verarbeitung und die mathematische Modellierung pragmatischer Schlussfolgerungen im Rational Speech Act (RSA)-Framework (Frank & Goodman, 2012) – indem sie kognitive Faktoren untersucht, die der Variabilität der Leistung in pragmatischen Referenzspielen zugrunde liegen (Frank et al., 2016), welche durch die Vorhersagen von RSA-Modellen unterschiedlicher Komplexität theoretisch motiviert ist. Dabei wird gezeigt, dass pragmatisches Schlussfolgern in Referenzspielen durch logisches Denkvermögen und *Theory of Mind* – die Fähigkeit, die Gedanken und Gefühle anderer zu verstehen und korrekt zu interpretieren – vorhergesagt wird, nicht jedoch durch das Arbeitsgedächtnis. Durch die Identifizierung kognitiver Faktoren, die individuelle Variabilität in pragmatischem Schlussfolgern in Referenzspielen vorhersagen, legen wir die Grundlage für die Untersuchung der zugrunde liegenden kognitiven Mechanismen sowie für die Entwicklung kognitiver Modelle des pragmatischen Schlussfolgerns auf der algorithmischen Ebene.

Ein weiterer Beitrag dieser Dissertation besteht darin, zu untersuchen, wie Zuhörende Variationen in den pragmatischen Fähigkeiten von Sprecher:innen berücksichtigen. Frühere Arbeiten haben gezeigt, dass Interpretationen der Zuhörenden durch bestimmte Eigenschaften der sprechenden Person beeinflusst werden können, wie z. B. deren Persönlichkeit (Beltrama, 2018; Beltrama & Schwarz, 2021, 2024), Sprachken-

ntnisse oder Akzent (Puhacheuskaya & Järvikivi, 2022, 2025; Ip & Papafragou, 2021, 2023; Fairchild & Papafragou, 2018). Die Ergebnisse dieser Dissertation erweitern diese Forschung, indem sie zeigen, dass Überzeugungen der Zuhörenden über die pragmatische Kompetenz der sprechenden Person einen maßgeblichen Einfluss auf die Interpretation haben. Dieselbe mehrdeutige Äußerung kann je nach der wahrgenommenen Kompetenz und Kooperationsbereitschaft des Sprechers wörtlich oder als Implikatur interpretiert werden.

Darüber hinaus deuten unsere Ergebnisse darauf hin, dass die wahrgenommene pragmatische Kompetenz des Sprechers einen Einfluss darauf hat, wie tiefgehend Zuhörende über eine Äußerung nachdenken: Wenn sie die sprechende Person nicht für hinreichend pragmatisch kompetent oder kooperativ halten, sind sie offenbar weniger geneigt, den zusätzlichen kognitiven Aufwand aufzubringen, der für eine nicht-wörtliche Interpretation einer Äußerung erforderlich wäre. Diese Erkenntnis steht im Einklang mit der zentralen Annahme der Relevanztheorie (Sperber & Wilson, 1986), wonach eine Äußerung nur dann als Implikatur interpretiert wird, wenn der dafür notwendige kognitive Aufwand durch die Relevanz der gewonnenen Informationen für die zuhörende Person gerechtfertigt ist.

Über zwei experimentellen Paradigmen hinweg wird gezeigt, dass Menschen ihre Interpretationen im Laufe einer Interaktion schnell an eine bestimmte Sprecherin oder einen bestimmten Sprecher anpassen und dabei dynamisch ein mentales Modell der pragmatischen Kompetenz dieser Sprecher:innen aufbauen und fortlaufend aktualisieren. Zudem finden wir Hinweise auf individuelle Unterschiede im Anpassungsverhalten, die sich sowohl durch unterschiedliche Vorannahmen der Zuhörenden über die pragmatische Kompetenz von Sprecher:innen als auch durch Unterschiede in der Tendenz zur Verallgemeinerung über verschiedene Sprecher:innen hinweg erklären lassen.

Darüber hinaus leistet diese Dissertation einen methodologischen Beitrag, indem sie eine einfache, aber effektive Methode zur Identifikation von Strategien vorschlägt, die Menschen in pragmatischen Referenzspielen anwenden. Diese Methode ist auch auf andere experimentelle Paradigmen übertragbar. In mehreren Experimenten dieser Dissertation hat sie sich als nützlich erwiesen, um Verzerrungen in experimentellen Stimuli aufzudecken und zu reduzieren sowie zusätzliche Einblicke in die Strategien zu gewinnen, die Menschen anwenden, welche durch eine reine Untersuchung der Leistung nicht sichtbar wären.

Zusammenfassend leistet die vorliegende Dissertation einen Beitrag zu umfassenderen Theorien der Pragmatik, indem sie neue Einblicke in die individuellen Unterschiede in pragmatischen Fähigkeiten und Anpassung der Zuhörenden an die pragmatische Kompetenz der Sprecher:innen gewährt. Die Ergebnisse sprechen für eine differenziertere Auffassung pragmatischer Verarbeitung, nach der pragmatisches Schlussfolgern sowohl von kognitiven als auch von kontextuellen Faktoren beeinflusst wird und Gesprächspartner:innen solche Unterschiede in der pragmatischen Kompetenz ihres Gegenübers antizipieren und ihre Intepretationen entsprechend anpassen.

Ausführliche Zusammenfassung

Wenn wir mit anderen Menschen kommunizieren, sagen wir oft nicht genau das, was wir meinen, und zwar aus verschiedenen Gründen: Wir versuchen vielleicht, höflich zu sein oder die Gefühle anderer nicht zu verletzen, machen einen Witz, verwenden eine Metapher oder erwähnen Informationen nicht explizit, von denen wir annehmen, dass unser:e Gesprächspartner:in sie kennt oder aus dem Kontext ableiten kann. Die Disziplin der Pragmatik befasst sich damit, wie Menschen die Lücken zwischen der wörtlichen Bedeutung einer Äußerung und der Bedeutung, welche die Sprecherin oder der Sprecher vermitteln wollte, mithilfe von Kontexthinweisen füllen.

Wie schaffen Menschen es, sich gegenseitig zu verstehen, wenn oft ein großer Teil der beabsichtigten Bedeutung nicht explizit gesagt wird? Nach dem einflussreichen *Kooperationsprinzip*, das vom Sprachphilosophen Paul Grice formuliert wurde, verhalten wir uns bei der Kommunikation kooperativ und befolgen eine Reihe von Regeln, die Grice als *Konversationsmaximen* bezeichnet hat, wie z. B. so viele Informationen wie nötig und nicht mehr als nötig zu liefern, ehrlich zu sein und klar zu kommunizieren. Da von allen Gesprächspartner:innen erwartet wird, dass sie diese Regeln kennen, sind sie in der Lage, Rückschlüsse auf die beabsichtigte Bedeutung des anderen zu ziehen. Sprecher:innen können absichtlich gegen die Konversationsmaxime verstoßen, um zusätzliche Bedeutungen zu vermitteln und etwas zu implizieren, ohne es ausdrücklich zu sagen. Solche Äußerungen, die eine implizierte Bedeutung enthalten, welche über das wörtlich Gesagte hinausgeht, werden in der Pragmatik als *Implikaturen* bezeichnet.

Nehmen wir das folgende Beispiel:

> A: Wien liegt in Deutschland, oder?
> B: Ja, und Paris liegt in Portugal!

In diesem Beispiel verstößt B absichtlich gegen eine der Konversationsmaximen, die Maxime der Qualität, indem sie etwas sagt, das nicht der Wahrheit entspricht. Sie geht jedoch davon aus, dass A weiß, wo Paris liegt, und sagt etwas, das nicht wahr ist, um einen humoristischen Effekt zu erzielen und A zu signalisieren, dass seine Annahme über die Lage von Wien offensichtlich falsch ist.

Grice' einflussreiche Theorie der konversationellen Implikatur bildet eine Grundlage der Pragmatik. Sie ist jedoch allgemein gehalten und trifft keine spezifischen Aussagen darüber, unter welchen Bedingungen die zuhörende Person eine implizite, nicht wörtliche Bedeutung einer Äußerung ableiten kann. Verschiedene Ansätze bauen auf Grice' Theorie auf und konkretisieren, in welchen Situationen eine zuhörende Person eine mehrdeutige Äußerung nicht wörtlich interpretieren wird (Horn, 1984; Levinson, 2000; Chierchia et al., 2004; Sperber & Wilson, 1986). Ein Ansatz, den wir hier hervorheben möchten, ist die Relevanztheorie (Sperber & Wilson, 1986). Diese Theorie geht davon aus, dass eine nicht-wörtliche Interpretation nicht immer von

der zuhörenden Person abgeleitet wird, sondern nur dann, wenn diese Interpretation ausreichend relevant ist und durch den Kontext unterstützt wird. Im Unterschied zu anderen grundlegenden pragmatischen Theorien berücksichtigt die Relevanztheorie die kognitiven Anstrengungen der Gesprächspartner:innen und stellt diesen Faktor in den Vordergrund, anstatt lediglich zu spezifizieren, was unter idealen Bedingungen geschehen sollte.

Die Theorien, welche die Grundlage der Pragmatik bilden, wurden größtenteils von Philosoph:innen entwickelt, die ihre eigenen Intuitionen nutzten, um ihre Behauptungen zu untermauern. Das relativ junge interdisziplinäre Feld der experimentellen Pragmatik (Noveck, 2001; Noveck & Reboul, 2008; Noveck, 2018) setzt Methoden aus der experimentellen Psychologie und Psycholinguistik ein, um empirische Belege für die Vorhersagen dieser Theorien zu sammeln, sie weiter zu präzisieren und die kognitiven Mechanismen zu untersuchen, die dem Prozess des pragmatischen Schlussfolgerns zugrunde liegen.

Die meisten grundlegenden Theorien der Pragmatik legen fest, wie Kommunikation unter idealen Bedingungen ablaufen sollte. Dabei wird davon ausgegangen, dass sich die Gesprächspartner:innen immer rational verhalten und schlussfolgern. Individuelle Unterschiede zwischen Menschen und kontextbezogene Faktoren, welche die optimale Kommunikation erschweren könnten, werden nicht berücksichtigt. Die Relevanztheorie bildet insofern eine Ausnahme, als sie sich auf kognitive Anstrengungen bezieht und nicht von unbegrenzten kognitiven Ressourcen ausgeht, aber auch sie trifft keine klaren Vorhersagen hinsichtlich individueller Unterschiede oder der Anpassung an Kontextfaktoren. Die Realität der Kommunikation ist jedoch viel komplexer, als es idealisierte Theorien der pragmatischen Kompetenz vermuten lassen: Es wurde gezeigt, dass Menschen sich in ihren pragmatischen Fähigkeiten – d. h. in ihrer Fähigkeit, kooperativ zu kommunizieren und Bedeutungen zu erschließen, die gemeint, aber nicht ausdrücklich gesagt werden – stark unterscheiden (Matthews et al., 2018; Kidd et al., 2018). Diese Fähigkeiten können zudem durch situative Faktoren wie Müdigkeit und Unaufmerksamkeit beeinträchtigt werden.

Lange Zeit konzentrierten sich experimentelle Studien in der Kognitionswissenschaft, darunter auch solche im Bereich der experimentellen Pragmatik, auf Effekte auf der Populationsebene, d. h. darauf, wie sich eine Gruppe von Menschen im Durchschnitt verhält. Individuelle Unterschiede zwischen Menschen wurden als statistisches Rauschen behandelt (Kidd et al., 2018). In den letzten Jahren begann jedoch das Interesse an der Untersuchung individueller Unterschiede zu wachsen. Es setzt sich zunehmend die Erkenntnis durch, dass Unterschiede zwischen Menschen allgegenwärtig sind und es für umfassende Theorien des menschlichen Verhaltens und Sprachverstehens nicht nur sinnvoll, sondern sogar unerlässlich ist, diese Unterschiede zu untersuchen und zu berücksichtigen.

Das Ziel dieser Dissertation ist es, pragmatische Theorien durch die Untersuchung individueller Variabilität und der Anpassung der Zuhörenden an ihre Gesprächspartner:innen voranzubringen. Sie wendet eine Kombination aus experimen-

teller Arbeit und computergestützter kognitiver Modellierung an, um die folgenden Forschungsfragen zu beantworten: **Welche kognitiven Faktoren liegen individuellen Unterschieden in der Fähigkeit zum pragmatischen Schlussfolgern zugrunde? Und wie berücksichtigen Zuhörende diese Unterschiede – passen sie ihre Interpretationen an die Sprecher:innen an?**

Kapitel 2 stellt den für die folgenden Kapitel relevanten theoretischen Hintergrund der Arbeit dar. Es führt in zentrale Theorien der Pragmatik ein und gibt einen Überblick über einschlägige Forschung zu kognitiven Faktoren, die individuellen Unterschieden in der Fähigkeit zu pragmatischem Schlussfolgern zugrunde liegen. Darüber hinaus werden Forschungsarbeiten vorgestellt, die sich damit befassen, wie die Eigenschaften der sprechenden Person die Interpretation ihrer Äußerungen beeinflussen und wie Zuhörende ihre Interpretationen an unterschiedliche Sprecher:innen anpassen. Außerdem werden die zentralen Ziele und Methoden der experimentellen Pragmatik vorgestellt, welche darauf abzielt, pragmatische Theorien empirisch zu testen. Zudem werden gängige Experimentalparadigmen der experimentellen Pragmatik erörtert, darunter pragmatische Referenzspiele (Frank et al., 2016; Franke & Degen, 2016), ein Paradigma, das in dieser Dissertation für die Mehrheit der Experimente verwendet wird.

Referenzspiele sind einfache Kommunikationsspiele, an denen zwei Personen beteiligt sind: eine sprechende Person und eine zuhörende Person. Die Spielenden kommunizieren mithilfe einer begrenzten Auswahl an bildlichen Symbolen. Die Aufgabe der sprechenden Person besteht darin, ein Symbol auszuwählen, um dadurch die zuhörende Person auf eines der Objekte im gemeinsamen Sichtfeld hinzuweisen. Die Aufgabe der zuhörenden Person besteht darin, das von der sprechenden Person übermittelte Signal zu interpretieren und das beabsichtigte Objekt zu identifizieren. Wir verwenden Referenzspiele aufgrund ihrer Flexibilität und weil sie sich auf natürliche Weise in Modellen des Rational Speech Act-Frameworks (Frank & Goodman, 2012) formulieren lassen.

Das Rational Speech Act-Framework (RSA) ist ein probabilistisches, Bayes'sches Framework, das Grices kooperatives Prinzip als rekursiven Schlussfolgerungsprozess zwischen einer zuhörenden und einer sprechenden Person formalisiert. Laut diesem Framework erreichen die Gesprächspartner:innen erfolgreiche Kommunikation, indem sie über die Kommunikationsziele und Überzeugungen der jeweils anderen Person schlussfolgern und dabei alternative Äußerungen sowie Interpretationen berücksichtigen. RSA wurde erfolgreich verwendet, um eine Vielzahl pragmatischer Sprachphänomene, wie z. B. skalare Implikaturen (Goodman & Stuhlmüller, 2013), Höflichkeit (Yoon et al., 2016) und Metaphern (Kao et al., 2014a) als Computermodelle zu formulieren. In dieser Dissertation wird RSA verwendet, um zu modellieren, wie Zuhörende Unterschiede in der Fähigkeit zu pragmatischem Schlussfolgern bei Sprechenden berücksichtigen und in die Interpretation einfließen lassen.

Da es unser Ziel ist, individuelle Unterschiede in der Fähigkeit zum pragmatischen Schlussfolgern anhand von Referenzspielen zu untersuchen, beginnen wir den experi-

mentellen Teil dieser Arbeit mit der Frage, in welchem Maß Referenzspiele tatsächlich diese Fähigkeit messen. Oft wird angenommen, dass eine gute Leistung der Teilnehmenden in pragmatischen Aufgaben auf erfolgreiche Schlussfolgerungen hinweist. Es können jedoch auch andere Faktoren eine Rolle spielen, wie etwa Verzerrungen im Versuchsaufbau, die dazu führen, dass die Teilnehmenden die Aufgaben korrekt lösen, ohne tatsächlich pragmatisch schlussfolgern zu müssen.

In Kapitel 3 wird eine einfache, aber effektive Methode vorgeschlagen, um einen besseren Einblick in die Strategien zu erhalten, die Menschen in pragmatischen Referenzspielen anwenden. Diese Methode kann auch in anderen experimentellen Paradigmen verwendet werden. Die Methode besteht darin, die Teilnehmenden nach der letzten Aufgabe zu bitten, ihre Strategie zu erklären. Anschließend werden die Antworten der Teilnehmenden anhand eines von uns entwickelten Annotationsschemas annotiert und mit deren Leistung im Referenzspiel verglichen. In Kapitel 3 wenden wir diese Methode in Kombination mit gezielten Manipulationen der Eigenschaften der experimentellen Stimuli an, um die in mehreren früheren Studien (Franke & Degen, 2016; Degen et al., 2012, 2013) verwendeten Stimuli auf mögliche Verzerrungen zu untersuchen. Unsere Analyse zeigte, dass bestimmte visuelle Eigenschaften der Stimuli dazu führten, dass die Teilnehmenden Aufgaben korrekt lösen konnten, ohne tatsächlich pragmatisch schlussfolgern zu müssen.

Daraufhin überprüften wir mit derselben Methode eine andere Version des Referenzspiels, die abstrakte Stimuli wie geometrische Formen und Farben verwendete. Diese Version erwies sich als weniger anfällig für Verzerrungen und spiegelte die pragmatischen Fähigkeiten der Teilnehmenden genauer wider. Diese weniger verzerrte Version der Stimuli wurde in den weiteren Experimenten der Dissertation verwendet.

Somit leistet Kapitel 3 einen methodologischen Beitrag, indem es am Beispiel der Referenzspiele verdeutlicht, wie wichtig es ist, nicht einfach davon auszugehen, dass eine experimentelle Aufgabe die Fähigkeit misst, die sie zu messen vorgibt, sondern den Versuchsaufbau zu hinterfragen, um mögliche Verzerrungen zu identifizieren und zu beheben. Es wird eine einfache, aber effektive Methode vorgeschlagen, die dafür verwendet werden kann. Darüber hinaus erwies sich die vorgeschlagene Strategieerfragungsmethode in den folgenden Kapiteln der Dissertation in mehrfacher Hinsicht als nützlich: Sie wurde eingesetzt, um zusätzliche Einblicke in die von den Teilnehmenden verwendeten Strategien zu gewinnen, Teilnehmende zu identifizieren, die die Aufgabe missverstanden haben und daher ausgeschlossen werden mussten, sowie andere von Teilnehmenden angewendete Strategien zu identifizieren, die anhand der bloßen Leistung im Referenzspiel nicht erkennbar gewesen wären, wie z. B. Meta-Schlussfolgerungen über die pragmatische Kompetenz der sprechenden Person.

Eines der zentralen Ziele dieser Dissertation ist die Untersuchung individueller Unterschiede in pragmatischen Fähigkeiten sowie der kognitiven Faktoren, die diesen Unterschieden zugrunde liegen. In der Forschung existiert mittlerweile eine wachsende Zahl von Arbeiten zu individueller Variation in der Pragmatik sowie eine Reihe von Studien, die pragmatische Effekte probabilistisch modellieren, u. a. im oben genann-

ten Rational Speech Act-Framework. Diese beiden Forschungsstränge blieben jedoch bisher weitestgehend getrennt, mit wenigen Ausnahmen in jüngerer Zeit (Bohn et al., 2022; van Tiel et al., 2022).

Einer der Beiträge dieser Dissertation besteht darin, diese beiden Literaturstränge miteinander zu verbinden, indem kognitive Faktoren, die der individuellen Variabilität in pragmatischen Referenzspielen zugrunde liegen, untersucht und mit der Modellierung der Komplexität des pragmatischen Schlussfolgerns im RSA-Framework in Verbindung gebracht werden. Franke & Degen (2016) zeigte, dass die Leistung von Menschen in pragmatischen Referenzspielen erheblich variiert und dass sich diese Variation in drei Leistungsklassen einteilen lässt, die mit den Vorhersagen von drei RSA-Modellen unterschiedlicher Komplexität übereinstimmen. Aufbauend auf diesen Befunden untersucht Kapitel 4, *warum* sich Menschen in ihrer Leistung bei pragmatischen Referenzspielen unterscheiden.

Frühere Studien haben widersprüchliche Ergebnisse hinsichtlich der Rolle des logischen Denkens, der Theory of Mind – der Fähigkeit, die Gedanken und Gefühle anderer zu verstehen und korrekt zu interpretieren – sowie der Kapazität des Arbeitsgedächtnisses im pragmatischen Schlussfolgern geliefert. In Kapitel 4 wird die von Franke & Degen (2016) beobachtete individuelle Variabilität in der Leistung in Referenzspielen an einer deutlich größeren Stichprobe mit weniger verzerrten Stimuli repliziert. Dabei wird gezeigt, dass pragmatisches Schlussfolgern in Referenzspielen durch logisches Denkvermögen und Theory of Mind vorhergesagt werden kann, während sich keine Hinweise auf einen Einfluss des Arbeitsgedächtnisses finden. Durch die Identifizierung kognitiver Faktoren, die individuelle Variabilität im pragmatischen Denken in Referenzspielen erklären, legen wir die Grundlage für eine genauere Untersuchung der zugrunde liegenden kognitiven Mechanismen sowie für die Entwicklung kognitiver Modelle des pragmatischen Schlussfolgerns auf der algorithmischen Ebene.

Während die Unterschiede in der Leistung von Menschen im Referenzspiel, die wir in Kapitel 4 untersucht haben, theoretisch durch die Vorhersagen von drei RSA-Modellen unterschiedlicher Komplexität motiviert sind, erweisen sich diese Unterschiede in unserer Studie als eher graduell und es bilden sich keine drei klar abgegrenzten Cluster. Dies deutet darauf hin, dass es zwar drei grobe Kategorien von Personen gibt, die sich in der Tiefe ihres pragmatischen Schlussfolgerns unterscheiden, innerhalb dieser Gruppen jedoch erhebliche Variabilität besteht und die Grenzen zwischen ihnen flexibel sind.

Ein weiteres zentrales Ziel dieser Dissertation ist es, zu untersuchen, wie die individuellen Unterschiede in der pragmatischen Kompetenz der Sprechenden von Zuhörenden berücksichtigt wird. Frühere Arbeiten haben gezeigt, dass Interpretationen der Zuhörenden von bestimmten Eigenschaften der sprechenden Person beeinflusst werden können, wie z. B. von der Persönlichkeit des Sprechers (Beltrama, 2018, 2020; Beltrama & Schwarz, 2024), Sprachkenntnissen und Akzent (Puhacheuskaya & Järvikivi, 2022, 2025; Ip & Papafragou, 2023; Fairchild & Papafragou, 2018). Diese Dissertation leistet einen Beitrag, indem sie untersucht, wie die Wahrnehmung der pragmatischen

Kompetenz der sprechenden Person durch die zuhörende Person ihre Interpretation beeinflusst.

In Kapitel 5 wird untersucht, wie die geschätzte pragmatische Kompetenz bestimmter Sprecher:innentypen die Interpretationen der zuhörenden Person beeinflusst. Zu den von uns untersuchten Typen von Sprechenden gehörten Erwachsene, 4-jährige Kinder, der sprachmodellbasierte Chatbot ChatGPT und Sprechende, die Sprache wörtlich verwenden. Wir stellten fest: Wenn Menschen glaubten, dass eine mehrdeutige Äußerung von einem Kind kam, neigten sie eher dazu, sie wörtlich zu interpretieren. Wenn dieselbe Äußerung allerdings von einer erwachsenen Person kam, wurde sie eher als Implikatur – eine Äußerung mit einer implizierten nicht-wörtlichen Bedeutung – interpretiert.

Der Effekt der Identität der sprechenden Person auf die Interpretation der Zuhörenden hatte eine kategoriale sowie eine graduelle Komponente: Bei einer ambigen Äußerung, von der die Teilnehmenden glaubten, dass sie von einem Kind kam, haben mehr Menschen diese Äußerung wörtlich interpretiert als wenn dieselbe Äußerung von einem Erwachsenen stammte. Diejenigen, die diese Äußerung als Implikatur interpretiert haben, waren sich tendenziell weniger sicher in ihrer Interpretation. Der graduelle Teil des Effekts spiegelt die Unsicherheit der Menschen hinsichtlich der pragmatischen Kompetenz von Kindern wider. Die kategoriale Komponente dieses Effekts – d. h. die Tatsache, dass mehr Menschen die Äußerung wörtlich interpretierten, wenn sie dachten, dass sie von einem Kind kam – deutet darauf hin, dass die wahrgenommene pragmatische Kompetenz der sprechenden Person beeinflusst, wie tiefgehend Zuhörende über eine Äußerung nachdenken. Wenn sie die sprechende Person nicht für hinreichend pragmatisch kompetent oder kooperativ halten, sind sie offenbar weniger geneigt, den zusätzlichen kognitiven Aufwand aufzubringen, der für eine nicht-wörtliche Interpretation einer Äußerung erforderlich wäre. Diese Erkenntnis steht im Einklang mit der zentralen Annahme der Relevanztheorie (Sperber & Wilson, 1986), wonach eine Äußerung nur dann als eine Implikatur interpretiert wird, wenn der kognitive Aufwand, der dafür erforderlich ist, um zu dieser Interpretation zu gelangen, durch Relevanz dieser Informationen für die zuhörende Person gerechtfertigt ist.

Ein weiteres Experiment, das in Kapitel 5 vorgestellt wird, untersucht die Wahrnehmung der pragmatischen Fähigkeiten des LLM-basierten Chatbots ChatGPT. Wie bereits erwähnt, besagt das Grice'sche *Kooperationsprinzip*, dass Menschen sich in der Kommunikation kooperativ verhalten und sich an gemeinsame Konversationsregeln halten. Es ist allerdings deutlich weniger klar, ob dieses Prinzip in gleicher Weise auch für künstliche Agenten gilt. Unsere Ergebnisse zeigen eine auffällige Diskrepanz: In expliziten Bewertungen schrieben die Teilnehmenden ChatGPT sehr hohe pragmatische Kompetenz zu, während ihr Verhalten im Referenzspiel darauf hindeutete, dass sie ChatGPT implizit als weniger pragmatisch kompetent behandelten als einen erwachsenen Menschen und daher häufiger zu wörtlichen Interpretationen gelangten. Wie bereits erläutert, erfordert pragmatisches Schlussfolgern oft einen erheblichen ko-

gnitiven Aufwand und setzt eine entsprechende Motivation voraus. Diese Motivation scheint in der Interaktion mit anderen Menschen eher gegeben zu sein als mit einem künstlichen Agenten, da es Menschen leichter fällt, sich auf andere Menschen zu beziehen und deren kommunikative Intentionen nachzuvollziehen. Insgesamt tragen diese Befunde zu einem besseren Verständnis künstlicher Agenten als Kommunikationspartner:innen bei.

Im letzten Experiment von Kapitel 5 wurde untersucht, wie Zuhörende mehrdeutige Äußerungen einer sprechenden Person benutzen, von der sie wissen, dass sie die Sprache immer wörtlich benutzt. Das Rational Speech Act-Framework sagt voraus, dass pragmatisch kompetente Zuhörende selbst dann zu nicht-wörtlichen Interpretationen gelangen können, wenn die sprechende Person strikt wörtlich kommuniziert. Dies geschieht durch Schlussfolgern über alternative Äußerungen, die der sprechenden Person zur Verfügung gestanden hätten, sowie über deren jeweilige Wahrscheinlichkeiten. In unserem Experiment interpretierten die Teilnehmenden Äußerungen, von denen sie annahmen, dass sie von einem sehr einfachen Computerprogramm stammten, das zufällig eine beliebige wahre Beschreibung auswählte. Entgegen den Vorhersagen des RSA-Modells des pragmatischen Zuhörenden, zeigten die Ergebnisse, dass die Teilnehmenden die Äußerungen überwiegend wörtlich interpretierten und offenbar nicht die alternativen Äußerungen oder deren Wahrscheinlichkeiten in ihrer Interpretation berücksichtigten. Diese Befunde deuten darauf hin, dass Schlussfolgern über kommunikative Intentionen der Gesprächspartner:innen für Menschen zwar naheliegend ist, Nachdenken über mögliche Alternativen und deren Wahrscheinlichkeiten jedoch weniger selbstverständlich zu sein scheint und offenbar keine standardmäßige Interpretationsstrategie darstellt.

Die Befunde von Kapitel 5 deuten darauf hin, dass die Interpretationen der Zuhörenden von ihrer Wahrnehmung der pragmatischen Kompetenz der sprechenden Person beeinflusst werden. In Kapitel 5 wurden den Teilnehmenden die Eigenschaften der Sprecher:innen explizit mitgeteilt: Sie wurden darüber informiert, dass sie Äußerungen eines bestimmten Sprechertyps interpretieren würden, etwa die einer anderen erwachsenen Person, eines Kindes oder eines künstlichen Agenten. In vielen alltäglichen Kommunikationssituationen sind Faktoren, welche die pragmatischen Fähigkeiten einer sprechenden Person beeinflussen, jedoch nicht direkt beobachtbar. So kann man beispielsweise mit einer Person interagieren, die man gerade erst kennengelernt hat und über die man nur sehr wenig weiß, oder mit einer Person, die müde oder unaufmerksam ist. In solchen Fällen müssen Zuhörende die pragmatische Kompetenz ihres Gegenübers im Laufe der Interaktion erschließen.

Frühere Studien zu anderen pragmatischen Phänomenen legen nahe, dass Menschen ihre Interpretationen auch ohne explizite Hinweise auf relevante Eigenschaften der sprechenden Person dynamisch an ihre Gesprächspartner:innen anpassen. So konnte beispielsweise gezeigt werden, dass Menschen Unterschiede darin wahrnehmen, wie verschiedene Sprecher:innen bestimmte Wörter verwenden, etwa Ausdrücke der Unsicherheit wie “möglicherweise” oder “wahrscheinlich” (Schuster & Degen, 2020)

oder auch wie stark Sprecher:innen dazu neigen, Humor zu verwenden (Loy & Demberg, 2024).

In den Kapiteln 6 und 7 wurde untersucht, ob Menschen ihre Interpretationen im Laufe der Interaktion an die pragmatische Kompetenz der sprechenden Person anpassen. Kapitel 6 befasste sich mit der Anpassung an zwei Sprecher:innen in einem Referenzspiel, von denen sich eine Person pragmatisch kompetenter verhielt als die andere. In Kapitel 7 weiteten wir die Untersuchung der dynamischen Anpassung an die pragmatische Kompetenz der Sprecher:innen auf ein anderes Paradigma aus, nämlich auf ein Bauspiel, in dem die Teilnehmenden dreidimensionale Strukturen aus Bausteinen auf Grundlage schriftlicher Anweisungen von zwei Sprecher:innen bauten. Eine dieser Personen ließ ausschließlich Informationen aus, die aus dem Kontext erschlossen werden konnten, während die andere Person manchmal auch Informationen ausließ, die nicht zuverlässig inferiert werden konnten.

Über beide Paradigmen hinweg fanden wir Hinweise auf eine schnelle Anpassung der Teilnehmenden an die pragmatische Kompetenz beider Sprecher:innen: Die Teilnehmenden wurden sich ihrer Interpretationen im Laufe der Interaktion weniger sicher, wenn sie mit der weniger pragmatisch kompetenten Person interagierten. Mit der pragmatisch kompetenteren Person wurden sie hingegen mit der Zeit sicherer. Diese Befunde sprechen dafür, dass Menschen sensibel für die pragmatischen Fähigkeiten ihrer Gesprächspartner:innen sind und ihre Interpretationen entsprechend anpassen. Wichtig ist jedoch, dass wir zwar auf der Populationsebene starke Evidenz für solche Anpassungsprozesse fanden, zugleich aber unterschiedliche Anpassungsmuster zwischen den Menschen beobachteten. Einige passten ihre Interpretationen überhaupt nicht an und unter denjenigen, die eine Anpassung zeigten, bestanden Unterschiede. Diese lassen sich gemäß dem im Folgenden diskutierten rechnergestützten kognitiven Modell durch unterschiedliche Vorannahmen über die pragmatische Kompetenz der Sprecher:innen sowie durch Unterschiede im Ausmaß der Generalisierung zwischen Sprecher:innen erklären. Darüber hinaus deuten unsere Ergebnisse darauf hin, dass erfolgreiche pragmatische Anpassung ein hinreichendes Maß an pragmatischer Kompetenz bei der zuhörenden Person voraussetzt.

Das letzte Ziel dieser Dissertation war es, ein rechnergestütztes kognitives Modell zu entwickeln, das zeigt, wie die geschätze pragmatische Kompetenz der sprechenden Person die Interpretation der zuhörenden Person beeinflusst. In Kapitel 5 wurde ein rechnergestütztes kognitives Modell im Rahmen des Rational Speech Act-Frameworks vorgestellt, in dem die Annahmen der Zuhörenden über die pragmatischen Fähigkeiten der sprechenden Person als Unsicherheit über den Typ der sprechenden Person formalisiert wurden. Konkret wird angenommen, dass Zuhörende nicht sicher sind, ob sie es mit einem *literal speaker* zu tun haben – also einer sprechenden Person, die alle wörtlich passenden Äußerungen mit gleicher Wahrscheinlichkeit verwendet – oder mit einem *pragmatic speaker*, der pragmatisch kompetent ist und seine Äußerungen gezielt so auswählt, dass der kommunikative Erfolg maximiert wird. In Kapitel 6 wurde dieses Modell erweitert, um die Anpassung der zuhörenden Per-

son an die pragmatische Kompetenz der sprechenden Person im Laufe der Interaktion einzubeziehen.

Die Modellierung liefert wichtige Einsichten in die kognitiven Mechanismen pragmatischer Anpassung. Erstens legen die Ergebnisse nahe, dass Zuhörende ihre Überzeugungen über die pragmatische Kompetenz einer sprechenden Person nicht immer auf Grundlage deren tatsächlichen Verhaltens aktualisieren. Insbesondere dann, wenn eine Perspektivübernahme kognitiv aufwendig ist und sich die Perspektive des Gegenübers deutlich von der eigenen unterscheidet, scheinen Zuhörende stattdessen von ihrer eigenen Interpretation auszugehen. Dies zeigt sich darin, dass Lern- bzw. Anpassungseffekte gerade in solchen Durchgängen auftraten, die aus Sicht der Zuhörenden mehrdeutig waren, obwohl sie aus der Perspektive der sprechenden Person eindeutig waren. Zweitens bietet das vorgeschlagene rechnergestützte Modell eine Erklärung für die beobachteten individuellen Unterschiede in den Mustern der pragmatischen Anpassung. Diese Unterschiede lassen sich auf unterschiedliche Vorannahmen über die pragmatische Kompetenz von Gesprächspartner:innen sowie auf Unterschiede darin zurückführen, in welchem Ausmaß Erfahrungen mit einer bestimmten Sprecherin oder einem bestimmten Sprecher die Erwartungen gegenüber anderen Sprecher:innen beeinflussen.

Zusammenfassend leistet die vorliegende Dissertation einen Beitrag zu umfassenderen Theorien der Pragmatik, indem sie wichtige Einblicke in individuelle Unterschiede in der Fähigkeit zum pragmatischen Schlussfolgern und die Anpassung der Zuhörenden an die pragmatische Kompetenz der Sprecher:innen gewährt.

Die dargelegten Ergebnisse sprechen für eine differenziertere Auffassung pragmatischer Verarbeitung, nach der pragmatisches Schlussfolgern sowohl von kognitiven als auch von kontextuellen Faktoren beeinflusst wird und Gesprächspartner:innen solche Unterschiede in der pragmatischen Kompetenz ihres Gegenübers antizipieren und ihre Interpretationen entsprechend anpassen.

Darüber hinaus wirft diese Arbeit wichtige Fragen für die zukünftige Forschung auf, darunter den Bedarf einer Untersuchung konkreter kontextueller Bedingungen, die tiefgehenderes oder eher oberflächliches pragmatisches Schlussfolgern begünstigen, sowie die Untersuchung, ob sich spezifische kognitive Merkmale identifizieren lassen, welche die individuellen Unterschiede in pragmatischen Anpassungsmustern, die in dieser Arbeit festgestellt wurden, erklären können.

Acknowledgements

I am grateful to so many people who have touched my life in the past five years and from whom I could learn to be a better scientist and a kinder person.

First of all, I would like to thank my supervisor Vera Demberg, whose ability to grasp a complex idea within seconds and offer new insights when I'd been puzzling over it for weeks will never fail to impress me. Vera, thank you for the perfect balance of guidance and freedom to explore, and for all your support through these years.

I would like to thank my collaborators – John Duff, Emilia Ellsiepen, Jia E. Loy, Kata Naszádi, Margarita Ryzhova and Sebastian Schuster – for the many interesting discussions and for allowing me to learn with you and from you. Many thanks to the amazing research assistants Nataliia Bila and Yuan Zhang for their help with annotation, experiment implementation and stimuli creation.

I'm very fortunate to have done my PhD in such a warm and welcoming lab, where we celebrated each other's successes and encouraged and supported each other in times of struggle. Special thanks to Marian and Ana who were a delight to share an office with. Thanks to Mayank who has helped me countless times and who is always willing to share tea from his extensive collection. Many thanks to Khanda Adams for her enormous help with navigating bureaucratic labyrinths, her patience and her contagious joie de vivre.

I would like to say many thanks to the organizers and attendees of XPRAG. I have felt very welcome in this community and have really enjoyed exchanging ideas with its members and looked forward to seeing them at conferences. It has felt like everyone is an equal and everyone's voice is heard and valued, whether they are a senior researcher or someone who, like me, was doing their PhD.

Outside of academia, too, there are many people whom I want to thank.

I am incredibly fortunate to have found friends who have made Saarbrücken feel like home, who have been there to share both joys and struggles with, and who have supported and encouraged me through these years. I am very grateful to them and look forward to many more wonderful years of friendship. Danke, Anna, für deine Unterstützung und dafür, dass du es ermutigst und feierst, wenn ich auf meine Grenzen achte, und dass ich gemeinsam mit dir Muster reflektieren und umlernen darf. Danke, Dany, für dein offenes Ohr, deine Kochkünste und häufiges (wunder)leckeres Bekochen in den Endspurtmonaten und für gemütliche Abende. Danke, Raphaela, dass ich bei dir immer das Gefühl habe, so okay zu sein, wie ich bin, dass du immer für mich da bist und einen (Premium-)Platz auf deiner Couch für mich hast. Danke, Tabea, für deine warme, fürsorgliche Art, die mich immer aufmuntert und geborgen fühlen lässt, und deine unglaubliche Empathie. Danke an Sarah, die beste Co-Pilotin, fürs Mitfliegen und Mitfiebern, An-mich-Denken und Aufmuntern. Many thanks to my dear Laura for the wonderful hikes, dinners and advice.

I am very grateful to my partner Judith, who has stood by my side through the

ups and downs of the past months and has offered comfort and lightness in difficult times. Danke für die zahllosen Flaschen Granatapfelmate und Tassen Ingwertee, mit denen du mich an meinen Schreibtagen und -nächten versorgt hast, für spätabendliche Ausflüge zu Peace Kebab und McDonald's, für all die Coworking-Samstage und das Korrekturlesen. Danke fürs Mitfiebern, Mitfeiern und Mitleiden, fürs Zuhören und Trösten, und für dein Geduldig-mit-mir-(und-meiner-Neigung-zur-Substantivierung)-Sein – aber vor allem dafür, dass du so eine Lichtgestalt in meinem Leben bist und mir immer das Gefühl gibst, dass ich bei dir so sein darf, wie ich bin. Das ist ein unglaubliches Geschenk.

I am very thankful to my dear friend Lia who has been in my life for over 10 years and is like a sister to me. I know that she is always there for me, even when we don't see each other for a while and are many kilometers apart. Her strength and her ability to approach even the most difficult of times with the world's best sense of humor are a source of hope and inspiration for me. Лиечка, спасибо, что ты есть.

I am grateful to my long-distance friends who are very dear to my heart – Jia, Marjolein and Alish. Thank you for being better at keeping in touch than I am.

I am very grateful to my parents, Galina and Eugene, for their love and unconditional support and for always believing in me.

I would like to say thanks to the choir of the htw saar, Half-Tone Waves, for making Monday a day to look forward to.

Ich möchte ganz herzlich dem Team der FrauenGenderBibliothek Saar danken, die zu einem Wohlfühlort für mich geworden ist. Es bedeutet mir sehr viel, Teil einer Community mit solcher Solidarität und so viel Frauen*-Power zu sein.

I started my PhD in May 2021. Less than a year later, in February 2022, Russia – my home country – invaded Ukraine, starting a devastating war that cast a dark shadow over very many lives, including mine. I am very honored to have met so many incredibly brave people who were forced to flee their homes in Ukraine and rebuild their lives here from scratch; their resilience has been both inspiring and deeply humbling. In response to these events, I have tried to contribute in whatever small ways I could to counteract the devastating and criminal actions of the Russian government. I thank Kultur- und Lesetreff Malstatt and Sebastian Matthes, FrauenGenderBibliothek Saar and Dr. Annette Keinhorst and Dr. Eva Nita, IHK Sprachmittlerschulung and Kultur- und Sprachmittler e.V. for helping me find ways to do something useful. I am also very grateful to my supervisor Vera Demberg for her patience and understanding during a time when my attention was necessarily divided between my academic work and my engagement with refugee support. I hope for peace, freedom and independence for Ukraine. Нет войне. Слава Україні.

Contents

Chapter 1

Introduction

When we communicate, what we say often does not exactly match the meaning we intend to convey, whether it is because we are trying to be polite, making a joke, or not stating something that we expect our interlocutor to be able to infer from context. The discipline of pragmatics is concerned with the gap between what is literally said and what is meant. It investigates how interlocutors use contextual clues to bridge this gap and achieve understanding.

One of the key figures in the field of pragmatics, philosopher of language Paul Grice, introduced the influential *cooperative principle*. This principle states that in communication, people generally behave cooperatively and follow a set of communicative rules – such as being informative, staying on topic and communicating clearly – which allows them to draw inferences about the interlocutor's intended meaning. This theory has been widely supported by experimental research, which has shown that, to varying degrees, these principles hold in many settings. However, the reality of communication is more complex than such idealized theories of pragmatic competence suggest: individuals have been shown to differ vastly in their pragmatic abilities (i.e., ability to communicate cooperatively and infer things which are not explicitly said; Matthews et al. (2018)), and their performance may be affected by factors such as inattention or fatigue.

For a long time, experimental studies in cognitive science, including those in the field of experimental pragmatics, have focused on population-level effects, i.e., how a group of people behaves on average, treating any variation between individuals as noise (Kidd et al., 2018). However, recently there has been a growing interest in this variability and an increased understanding that given the pervasive differences in individuals' pragmatic abilities, they are not only meaningful but also essential for comprehensive theories of Pragmatics.

This thesis supports the view that, given that individual variation exists, theories of pragmatic processing need to be developed to account for it. Through a combination of experimental work and computational modeling, we aim to answer the following

research questions: **Which cognitive factors underlie individual differences in pragmatic inference? And how do comprehenders accommodate these differences – do they tailor their interpretations to specific speakers?**

By answering these questions, we aim to contribute to more comprehensive theories of variability in pragmatics. The research goals of this dissertation are presented below, followed by the contributions of the dissertation. The chapter concludes with an overview of the dissertation.

1.1 Research goals

This thesis aims to investigate the sources of individual differences in pragmatic reasoning and comprehenders' adaptation of their interpretations based on the characteristics of the speaker. More concretely, this thesis pursues four research goals, which are outlined below.

1. The first goal of the thesis is to **develop a method for eliciting strategies which people use when deriving pragmatic inferences**, i.e., interpretations which go beyond the literal meaning of the speaker's words. In Chapter 3, we propose a method for eliciting participants' strategies on a pragmatic reasoning task, along with an annotation scheme for categorizing the reported explanations. This method is useful for determining the extent to which performance reflects the underlying reasoning. A paper I co-authored with Margarita Ryzhova during the dissertation period which is not part of this thesis (Ryzhova et al., 2023) also contributes to this goal: in the context of inferences about redundant mentions of predictable events, we show that a similar strategy elicitation technique allows us to identify people who acknowledge the possibility of an inference but reject it due to implausibility.

2. The second goal of this thesis is to **identify cognitive factors which underlie individual differences in pragmatic reasoning ability**. Previous work has shown that people differ in their ability to derive non-literal meanings (e.g., Heyman & Schaeken (2015); Fairchild & Papafragou (2021); Yang et al. (2018). In formal probabilistic models of pragmatics, such as the widely used Rational Speech Act framework (Frank & Goodman, 2012), differences in pragmatic abilities can be expressed as depth of recursive reasoning of the listener and the speaker about each other: the greater the recursion depth, the more complex the nonliteral meanings a communicator is expected to be able to derive. Franke & Degen (2016) showed that their participants' performance aligned with predictions of three models of different recursion depth, i.e., three reasoning types. In this thesis, we aim to investigate whether cognitive factors can be identified which underlie these differences in the complexity of people's pragmatic reasoning. In Chapter 4, we examine whether working memory capacity, logical reasoning ability and Theory of Mind – ability to understand and

reason about other people's mental states – predict performance in a pragmatic reference game. Two further co-authored papers not included in this dissertation contribute to this goal. In a paper with Margarita Ryzhova (Ryzhova et al., 2023), we investigated individual differences in the processing of redundant utterances about predictable events, finding evidence of association of pragmatic reasoning with logical ability. In a paper with Sebastian Schuster (Schuster et al., 2023), we showed that people's ability to associate distinct uses of uncertainty expressions, such as "might" and "probably", with different speakers is predicted by their working memory updating ability.

3. The third goal of this thesis is to **investigate how comprehenders adapt their interpretations to the pragmatic competence of specific speakers**. Given that individuals differ in their pragmatic abilities, the same utterances may have different meanings when coming from different speakers. In this thesis, we investigate whether people derive speaker-specific interpretations and sometimes do not access the more complex nonliteral meaning if they have reasons to believe the speaker to be insufficiently pragmatically competent or cooperative to have intended to communicate it. In Chapter 5, we investigate whether people interpretations' are affected by their beliefs about various speaker types, such as adults, children, and artificial agents. In Chapters 6 and 7, we ask whether people adjust their interpretation of a speaker's utterances during interaction based on solely on the evidence of the speaker's behavior, in the absence of explicit clues to the speaker's pragmatic competence.

4. The final goal of this thesis is to **propose a computational cognitive model of how reasoning about the speaker is incorporated into utterance interpretation by comprehenders**. Cognitive modeling is useful since it allows us to make the theory maximally specific, making assumptions explicit and generating quantitative predictions. In Chapter 5, we propose a computational model in the Rational Speech Act framework of how participants incorporate their beliefs about the speaker's pragmatic sophistication into their interpretations. In Chapter 6, we extend this model to incorporate belief updating through interaction.

1.2 Research contributions

By addressing the above research goals, the research in this dissertation makes the following contributions.

- **Methodology: Technique for eliciting participants' strategies**. We propose a simple but effective technique for eliciting strategies that people use in a pragmatic reference game, which can also be applied in other paradigms. It consists of asking participants to provide a post-hoc explanation of the strategy

they used on one critical item, annotating the answers based on the proposed annotation scheme, and comparing participants' performance to the annotated explanations. In Chapter 3, we show that, combined with careful manipulation of stimuli properties, this technique can be useful for identifying biases in the experimental design: participants explanations revealed that certain characteristics of the stimuli inflated performance for reasons unrelated to pragmatic reasoning. In other chapters of the thesis, this technique is applied for gaining additional insight into participants' strategies, identifying participants who misunderstood the instructions or applied other kinds of strategies of interest, such as meta-reasoning about the speaker's pragmatic competence.

- **Individual differences in ad-hoc pragmatic inferencing**. The research presented in Chapter 4 extends the research on cognitive individual differences in pragmatic processing. Previous work, which is reviewed in Section 2.4, yielded conflicting findings regarding the role of working memory capacity, logical reasoning and Theory of Mind in pragmatic processing, which may be related to the fact that different pragmatic phenomena may tax cognitive resources to different extent. We build on findings by Franke & Degen (2016), who showed that participants' performance on a pragmatic reference game align with predictions of three formal probabilistic models of different complexity, by replicating this finding on a much large sample with less biased stimuli and investigating the relationship of performance on this task to cognitive individual differences. We show that pragmatic reasoning in the reference game is predicted by logical reasoning and Theory of Mind but do not find any evidence of an effect of working memory.

- **Speaker effects in pragmatic processing**. Previous work has found that certain properties of the interlocutor, such as their perceived persona (Beltrama, 2018, 2020), language fluency and nativess (Puhacheuskaya & Järvikivi, 2022, 2025; Ip & Papafragou, 2021, 2023), may affect the comprehender's interpretation. In Chapter 5, we extend previous research by showing that comprehenders incorporate the perceived pragmatic competence of the speaker into their inference: when the speaker is perceived to be incapable of performing complex pragmatic reasoning, e.g., when the speaker is a 4-year-old child, people are more likely to interpret their utterance literally as opposed to as a pragmatic implicature. We show that there is both a gradient and a categorical component to this effect: when the speaker is believed to be less pragmatically competent or cooperative, people tend to derive less strong pragmatic inferences, and fewer people derive an inference.

- **Perception of artificial agents as pragmatic speakers**. Previous research has yielded mixed findings on people's perception of the pragmatic abilities of artificial agents, with some studies suggesting greater adaptation to artificial

agents than to humans (Fischer, 2006; Branigan et al., 2011), and others finding the opposite pattern (Loy & Demberg, 2023). This discrepancy may be related to the shift of people's perception of the competence of artificial agents over time. In Chapter 5, we extend this line of work by examining people's perception of the pragmatic competence of a large language model–based chatbot, ChatGPT. While participants rated ChatGPT's pragmatic abilities very highly when asked explicitly, on the pragmatic task itself they were less likely to draw pragmatic inferences when ChatGPT was the speaker than when the speaker was another adult human. This discrepancy between explicit judgments and behavior may reflect the fact that participants engage in deeper pragmatic reasoning with human speakers than with artificial agents, possibly because they are more uncertain about the cooperativity of artificial agents.

- **Dynamic adaptation to the speaker's pragmatic competence**. Previous work shows that people may adjust their interpetations dynamically based on observation or interaction with the speaker (Ryskin et al., 2019; Gardner et al., 2021; Loy & Demberg, 2023; Schuster & Degen, 2020). In Chapters 6 and 7, we extend this line of research by providing empirical evidence that people dynamically update their beliefs about the speaker's pragmatic competence during interaction. Additionally, we show evidence for individual differences in adaptation patterns, reflecting variation in prior expectations and in the extent to which people generalize from their experience with one speaker to other speakers.

- **Computational cognitive modeling of reasoning about the speaker's pragmatic competence**. In Chapter 5 of this dissertation, we propose a computational model in the Rational Speech Act framework of how listeners incorporate their beliefs about the speaker into their inferences. It formalizes the listener's reasoning about the speaker's pragmatic competence as uncertainty about the speaker type – literal or pragmatic – that they are is interacting with. In Chapter 6, we extend this model to incorporate updating of listener's beliefs about the speaker during interaction. Our modeling work yielded insights into the dynamics of pragmatic adaptation. First, it suggested that in settings where perspective-taking is effortful, comprehenders may be updating their beliefs not through direct observation of the speaker but by reasoning about their own interpretations. Second, it provided evidence of systematic differences in participants' adaptation patterns, which can be explained by differing prior expectations about speakers' pragmatic competence and by varying degrees of generalization across speakers.

1.3 Overview of the dissertation

The research goals outlined above are addressed in this dissertation through a combination of experiments and computational modeling. Below, we provide a brief overview of each chapter of the dissertation.

Chapter 2 provides the relevant theoretical background for the following chapters of the thesis. It introduces the foundational theories of Pragmatics and discusses the connection between pragmatic reasoning and (bounded) rationality. It also introduces reference games, a paradigm, variations of which are used for the majority of the experiments in this thesis, and discusses their strengths and limitations. Next, it provides an overview of relevant existing work on cognitive individual differences in pragmatic processing, as well as of the effects of speaker identity in Pragmatics and adaptation to specific speakers. Finally, it introduces the Rational Speech Act (RSA) framework (Frank & Goodman, 2012), a probabilistic Bayesian modeling framework which will be used for the computational models proposed in this thesis, and discusses how the effects of interest – bounded rationality, speaker effects and adaptation to specific speakers – may be modeled.

In **Chapter 3**, we address Research Goal 1 by proposing a method for examining people's reasoning strategies and assessing the extent to which they align with performance on the pragmatic task we use, the reference game paradigm. We elicit participants' strategies and classify them using the annotation scheme we propose. This allows us to gain insights into the reasoning strategies used by participants, as well as to uncover underlying biases. Through four experiments, in which we carefully manipulate the stimuli and elicit participants' strategies using the proposed method, we show that the original version of the stimuli used by several previous studies contains several biases, which lead to inflated performance due to factors unrelated to reasoning. We also validate a less biased version of the stimuli for use in the subsequent chapters. We argue for the importance of validating one's chosen measures and ensuring that they indeed reflect the construct of interest, and demonstrate the effectiveness of two complementary techniques – systematic manipulation and comparison of stimuli and elicitation of explanations – for doing so.

Chapter 4 addresses Research Goal 2 by investigating cognitive individual differences which underlie variation in pragmatic reasoning. We demonstrate substantial variation in performance on the pragmatic reference game and relate it to cognitive factors. Performance in the pragmatic reference game was positively associated with logical reasoning ability and Theory of Mind, but not with working memory capacity. By identifying cognitive factors underlying pragmatic reasoning in the pragmatic reference game, Chapter 4 lays the groundwork for processing-level models of pragmatic reasoning. In a paper with John Duff, which is not included in this dissertation (Duff et al., 2025), we build on these findings and propose a processing-level model of the pragmatic reference game in the ACT-R framework, which associates different idealized pragmatic reasoning types with distinct interpretation strategies and makes

predictions about the time course of inferences.

Chapter 5 focuses on how differences in speakers' pragmatic competence are accommodated by comprehenders, addressing Research Goal 3. We examine how people's pragmatic interpretations are influenced by their beliefs about the speaker's pragmatic abilities. We conduct four experiments with different speaker types – adults, 4-year-old children, artificial agents, and explicitly literal speakers – and show that pragmatic interpretation is influenced by the identity of the speaker: when the speaker was believed to be less pragmatically competent or cooperative, participants were less likely to derive a pragmatic inference, instead interpreting the utterance literally, or derive a less strong inference. Our findings also suggest that people may reason less deeply when they expect their interlocutor to be less pragmatically competent, indicating that speaker identity affects not only comprehenders' inferences but also their own depth of pragmatic reasoning. We also propose a computational model of how the comprehender's beliefs about the speaker's pragmatic abilities affect interpretation in the Rational Speech Act framework, addressing Research Goal 4.

The experimental part of **Chapter 6** addresses Research Goal 3. Chapter 6 builds on the findings of Chapter 5 and extends the investigation of comprehenders' accommodation of variation in speakers' pragmatic competence to cases where it is not apparent from context and needs to be inferred from interaction. We investigate whether people dynamically adjust their interpretations to two different communication partners based on interaction feedback and find evidence of such adaptation. In Chapter 6, we also extend the computational model of comprehenders' reasoning about the speaker's pragmatic competence from Chapter 5 to incorporate learning through interaction, in line with Research Goal 4. Our modeling results yield valuable insights about the nature of this adaptation: they suggest that in settings where perspective-taking is effortful, people may update their beliefs by reasoning about their own interpretations of the speaker's utterances, as opposed to by observing the speaker's behavior directly. We also find evidence of individual differences in adaptation strategies, which can be explained by prior beliefs about the speaker's pragmatic competence and different degrees of generalization from the first speaker to the next one.

Chapter 7 extends the speaker adaptation experiment from Chapter 6 to a more complex and naturalistic paradigm, a collaborative block building game, where participants follow ambiguous instructions by two simulated partners and build simple 3D block structures. We find that the results are consistent with those of the previous experiment: participants dynamically adjust their interpretations to the two speakers. Chapter 7 addresses Research Goal 3 and provides further evidence for the generalizability of our findings.

Chapter 8 reviews the dissertation's findings in the context of related work, outlines possible directions for future work, and concludes the dissertation.

1.4 Dissemination

The research in this dissertation is partially based on the following peer-reviewed publications:

- **Mayn, A.**, & Demberg, V. (2023). High performance on a pragmatic task may not be the result of successful reasoning: On the importance of eliciting participants' reasoning strategies. *Open Mind*, 7, 156-178. (Chapter 3)
- **Mayn, A.**, & Demberg, V. Sources of individual variability in a pragmatic reference game: Effects of reasoning and Theory of Mind. To appear in *PLOS One*. (Chapter 4)
- **Mayn, A.**, & Demberg, V. (2022). Individual differences in a pragmatic reference game. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44). (Chapter 4)
- **Mayn, A.**, Loy, J. E., & Demberg, V. (2025). Beliefs about the speaker's reasoning ability influence pragmatic interpretation: Children and adults as speakers. *Open Mind*, 9, 89-120. (Chapter 5)
- **Mayn, A.**, Loy, J., & Demberg, V. (2024). Capable but not cooperative? Perceptions of ChatGPT as a pragmatic speaker. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46). (Chapter 5)
- **Mayn, A.**, Duff, J., Bila, N., & Demberg, V. (2025). People do not engage in ad-hoc reasoning about alternative messages when interacting with a literal speaker. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 47). (Chapter 5)
- Naszádi, K.*, **Mayn, A.***, Duff, J., Monz, C., & Demberg, V. Listeners Adapt to Speakers' Pragmatic Competence. (Chapters 6 & 7) (in prep.) *Equal contribution

While the above papers are all co-authored, I was primarily responsible for the conceptualization, methodology, investigation, formal analysis, and the writing and revision of the original drafts. Co-authors contributed through discussion during the conceptualization phase and by providing feedback and edits on manuscript drafts. Specific deviations from this contribution pattern are detailed below:

- Natalia Bila wrote the original draft of the experimental design section of Mayn, A., Duff, J., Bila, N., & Demberg, V. (2025).
- For Naszádi et al. (in prep), conceptualization, methodology, investigation and formal analysis were conducted jointly with Kata Naszádi and John Duff, building on Kata Naszádi's initial conceptual idea. The RSA modeling was proposed

and implemented by me, with input from co-authors regarding model fitting. I wrote the original draft of the experiments, modeling and discussion sections of the manuscript. Kata Naszádi authored the original draft of the introduction and theoretical background of the article; however, only material written by me is included in this dissertation.

Research conducted during the dissertation period but not included in this thesis is reported in the following publications:

- Ryzhova, M., **Mayn, A.**, & Demberg, V. (2023). What inferences do people actually make upon encountering informationally redundant utterances? an individual differences study. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).
- Schuster, S., **Mayn, A.**, & Demberg, V. (2023). Working memory updating modulates adaptation to speaker-specific use of uncertainty expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).
- Duff, J., **Mayn, A.**, & Demberg, V. (2025). An ACT-R model of resource-rational performance in a pragmatic reference game. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47).
- Duff, J., **Mayn, A.**, & Demberg, V. (2026). The role of reinforcement learning in pragmatic reasoning tasks: Modeling and validating the sources of individual differences. To appear in *Open Mind*.

Chapter 2

Theoretical background

In this chapter, we provide the theoretical background relevant to the experimental studies and computational models described in the later chapters of this thesis. The aims of this chapter are to provide the theoretical foundations that serve as the basis the research presented in this dissertation, to give an overview of the current state of the experimental and modeling literature, and to motivate the research questions of the dissertation.

Section 2.1 introduces the foundational pragmatic theories, with a focus on Grice's influential account, and discusses the connection between (bounded) rationality and pragmatics. Section 2.2. motivates the use of experimental methods for studying pragmatic phenomena and gives an overview of paradigms commonly used in experimental pragmatics. It also introduces reference games, a paradigm which will be used for the majority of experiments in this thesis. Section 2.3. provides a review of the literature on the role of cognitive individual differences in pragmatic processing and motivates the choice of individual differences for our experiments. Section 2.4. reviews the literature on effects of speaker identity and adaptation to the speaker in pragmatic processing and language processing more generally. Section 2.5. discusses probabilistic modeling approaches to pragmatics. It introduces the Rational Speech Act (RSA) framework, which will be used for the computational models proposed in this thesis, and discusses related modeling work and components of RSA relevant for modeling the effects of interest to us: individual differences, bounded rationality, and adaptation to the speaker. Section 2.6. presents the specific variant of pragmatic reference games which will be used in this thesis and discusses the predictions of various RSA models for the performance on this task. Section 2.7. concludes with a brief summary and outlook for the remainder of this dissertation.

2.1 Theoretical foundations of pragmatics

2.1.1 What is pragmatics?

Before delving into the theoretical foundations of pragmatics, we begin by defining it and what falls under the umbrella of pragmatics. Informally put for a start, pragmatics concerns itself with what an influential pragmaticist and the founding father of experimental pragmatics Ira Noveck called *gappiness* of language (Noveck, 2018): the gap between the literal meaning of the sentence and the meaning intended by the speaker that is filled using contextual information. The intended meaning is often richer and more complex than the literal meaning of the uttered words. Speakers may, for instance, use irony or figurative language, aim to be polite and not hurt someone's feelings, or omit information they expect the interlocutor to already know or be able to infer from the context. Pragmatics investigates how people bridge this gap between the literal meaning of the utterance and the context-dependent meaning intended by the speaker. Phenomena which fall under the umbrella of pragmatics include implicature (which will be defined more precisely later in this chapter, but for now it can be described as implied, not literally stated meaning of an utterance), figurative language, irony, under- and overspecification when referring to objects, indirect requests and politeness, and so on.

In this section, we first provide a brief overview of the foundational theoretical ideas in pragmatics, with a particular focus on the influential work of Paul Grice and subsequent approaches that build on his theory. We then discuss the relationship between pragmatics and (bounded) rationality.

Connection to philosophy of language

Several theoretical ideas that have been central to pragmatics originate in the philosophy of Ordinary Language, a tradition concerned with studying language as it is used in everyday contexts rather than expressions in isolation (Leech & Thomas, 2002). A key figure in this tradition is J. L. Austin, who in his lecture series *How to Do Things with Words* (1975) argued against the narrow focus of philosophy on descriptive propositions alone. Austin argued that a different approach is needed in order to capture the wide variety of functions language serves, including asking questions, making requests, issuing commands, and expressing wishes.

To account for the different uses of language, Austin developed Speech Act Theory, which distinguishes between three levels of action involved in an utterance: the locutionary act (the utterance itself), the illocutionary act (the speaker's intended communicative action), and the perlocutionary act (the effect the utterance has on the world). For instance, the utterance "Can you pass the salt?" constitutes a locutionary act; its illocutionary force is that of making a request, and its perlocutionary effect is that the interlocutor passes the salt to the speaker. John Searle, a student of

Austin, further developed Speech Act Theory by proposing a more fine-grained taxonomy of speech acts and by introducing the distinction between direct and indirect speech acts (1979).

Ordinary Language philosophy and Speech Act Theory were instrumental in drawing attention to aspects of meaning of language that arise from context and use and that are distinct from the semantic meaning of the utterance itself. Next, we will describe the influential ideas of Paul Grice, philosopher of language who is associated with Ordinary Language philosophy but gradually distanced himself from it (Lüthi, 2006; Neale, 1992; Chapman, 2010). Grice's theory of cooperative communication is widely regarded as foundational for pragmatics and has served as the basis for many contemporary pragmatic theories.

2.1.2 Gricean pragmatics

In his seminal work *Logic and conversation* (1975), Paul Grice made two contributions which are foundational to pragmatics: (1) a new definition of meaning, which distinguishes between utterance meaning and the speaker's intended meaning, and (2) proposing the cooperative principle and defining a set of rules of communication which communicators are expected to adhere to.

According to Grice, *sentence meaning*, which is the meaning of the words alone derivable in the absence of context, is distinct from *speaker meaning*, that is, the meaning that the speaker intended to communicate. His proposal was that the intended meaning is recovered by enriching the sentence's literal meaning though inference.

He further formalized a set of principles, so-called maxims of conversation, which interlocutors are expected to adhere to in order to communicate successfully. These maxims together he termed *the cooperative principle*. These maxims are:

- Maxim of Quantity: Be as informative as required and no more informative than that.
- Maxim of Quality: Be truthful.
- Maxim of Relation: Make your contribution relevant.
- Maxim of Manner: Be clear and brief (Or, in Grice's arguably less clear words: "avoid unnecessary prolixity").

Grice argues that interlocutors are aware of these principles and are sensitive to deviations from them. His account makes a distinction between violating a maxim and flouting it. *Violating* a maxim constitutes uncooperative behavior and should be avoided. Here's an example of a violation of a Gricean maxim[1]:

[1]The examples in this section are adapted from Detmar Beurer's Handout 4 (April 12, 2004), Linguistics 201, University of Tübingen: `https://sifnos.iwm-tuebingen.de/dm/04/spring/201/04-pragmatics-grice.pdf`

A: I saw John and Mary the other day. They have two children now.
B: Are they planning on having a third?
A: Well, actually, they already have a third child.

From the logic perspective, A is strictly speaking not being untruthful: the statement that someone has two children is logically compatible with the situation where someone has three. The issue here is the informativity maxim: A is not being as informative as is required by the situation. This is likely to result in unsuccessful communication and confusion and perhaps even frustration on B's part.

In contrast, a cooperative speaker may intentionally not observe one or more of the conversational maxims to convey nonliteral meaning. That is called *flouting* a maxim. In that case, the listener is expected to realize that the speaker's behavior is intentional and derive the intended meaning accordingly. Consider the following example:

A: Vienna is in Germany, isn't it?
B: Yeah, and Paris is in Portugal!

In this example, B intentionally violates the Maxim of Quality: she is not being truthful, but she assumes that the location of Paris is in the common ground – known to both interlocutors – and flouts the maxim, creating a humorous effect and signaling to A that his assumption about the location of Vienna is obviously incorrect. This kind of nonliteral meaning implied by B is known as an *implicature*.

In *Logic and conversation*, Grice distinguishes between the *particularized* conversational implicature and the *generalized* conversational implicature (p. 56), also known as an ad-hoc implicature. Particularized implicatures are highly context-dependent and take on the specific meaning only in that specific context. The example we just discussed falls in that category. By contrast, generalized implicatures are assumed to have a conventional meaning that they normally express. We will expand on this distinction in the next section, where we discuss approaches which build on Grice's account.

2.1.3 Post-Gricean theories

Grice's communication maxims are influential and build the foundation of contemporary pragmatics. Several approaches built on that foundation, consolidating the maxims into a smaller set of principles and making more specific predictions, for instance, about the contexts in which implicatures are expected to be derived and whether this derivation is associated with cognitive effort.

So-called Neo-Gricean approaches of Laurence Horn (1972; 1984) and Stephen Levinson (Levinson, 2000) aimed to simplify and refine Gricean principles and integrate them with linguistic theory (Noveck, 2018).

An important contribution by Horn is proposing that the derivation of generalized conversational implicatures relies on pre-existing scales and that the use of a weaker item on a scale implicitly negates the stronger item (Horn, 1972). For example, the statement "Some students passed the exam" is logically compatible with the interpretation that *all* students passed the exam, i.e., the literal, logical meaning of *some* is *some and possibly all*. However, upon hearing this statement, the listener may infer that some but not all students passed the exam, since if the speaker had wanted to say that all students had, they should have been more informative and said exactly that. <some,all> is thought to constitute a scale, and by uttering *some*, the speaker is thought to negate the stronger alternative *all*, resulting in the enriched meaning *some but not all*. <some,all> is one of many sets of lexical items thought to exist on a scale (see van Tiel et al. (2016) for an investigation of different scales and their properties); <warm, hot> is another example. However, the <some, all> scale has historically received a lot of attention. Derivation of an implicature through negation of a stronger scalemate is widely known as *scalar implicature*, which is a kind of generalized implicature.

Horn also proposed to consolidate Grice's maxims in two opposing principles: the Q-principle (Quantity) and the R-Principle (Relation) (1984). The Q-principle is based on the Gricean maxim of Quantity and states that one should make their contribution *sufficient* and say as much as one *can*. The R-principle builds on the Gricean maxims of Relation, Quantity and Manner and states that, at the same time, one needs to make their contribution *necessary* and say no more than one *must*.

Levinson's theory of generalized conversational implicature (GCI) (Levinson, 2000) is similar to Horn's account in that it also consolidates the Gricean maxims into a smaller set of principles, but it is less concerned with the speaker's intention. Instead, it holds that the meaning of generalized conversational implicatures, such as scalar implicatures, arises by default, licensed by a set of heuristic principles without effortful computations related to the speaker's intended meaning. Each of the three principles he proposed has a speaker's maxim associated with it, stating what the speaker should say, and a listener's maxim, stating what the comprehender should infer. The Q-principle states that what is not said is not the case. The I-principle states that what is said in an unmarked, i.e., usual, way represents a stereotypical situation. Conversely, the M-principle states that a marked utterance – i.e., utterance said in an abnormal way – signals an atypical situation. Thus, Levinson's claim is that generalized implicatures are generated by default but can be overridden by contextual information.

Gennaro Chierchia's grammatical account specifies contexts in which scalar implicatures are not derived – those are assumed to be the downward entailing contexts (2004). It contrasts with Gricean and neo-Gricean accounts in that it is not concerned with the speaker's intention at all but rather with grammatical environments that enable implicature derivation (Noveck, 2018).

In contrast to Levinson's and Chierchia's approaches, which view the pragmatic

meaning of generalized implicatures as the default, Sperber and Wilson's *relevance theory* (1986) assumes that narrowing of the meaning from the logical to the pragmatic is context-dependent and only happens when it is sufficiently relevant. Thus, it can be viewed as a corollary of relevance theory that the derivation of the pragmatic meaning may be associated with additional effort. Relevance theory is also distinct from the accounts discussed above in that it is a cognitive account: instead of specifying what should happen in ideal conditions, it includes cognitive effort as a factor: inferences should be derived in a way which is as economical as possible in terms of cognitive costs while maximizing *cognitive effects*, i.e., benefits for the listener depending on their goal. Relevance theory is also different from the other accounts described above in that it allows for variation by speaker and situation. It does not assume that the pragmatic meaning of implicatures is derived by default. Instead, it assumes that it is only derived if supported by the context and if it is relevant enough to justify the cognitive costs associated with the derivation in that it leads to additional cognitive effects.

So far, we mostly discussed generalized conversational implicatures, that is, implicatures which are assumed to have a conventional meaning. What of particularized implicatures? Particularized implicatures have received less attention than generalized ones (Kravtchenko, 2022). Generally, the Default accounts, including those of Horn, Levinson and Chierchia, assume that generalized implicatures are computed by default if they are licensed, whereas particularized implicatures are computed at a later stage, where contextual information is taken into account (Breheny et al., 2006). Thus, derivation of generalized implicatures is assumed by these accounts to be relatively effortless and automatic, whereas derivation of particularized implicatures is expected to be associated with more effort. In contrast, Sperber and Wilson's relevance theory does not put an emphasis on the distinction between generalized and particularized implicatures and instead argues that both types of implicature arise through the same process and require substantial contextual support. It therefore follows from relevance theory that the derivation of both kinds of implicature is expected to be effortful. Empirical evidence on this is mixed, with increasing evidence that even generalized implicatures are often associated with some cognitive costs (see Khorsheed et al., 2022 for a review).

As we see, these theories make distinct predictions about automaticity and potential effortfulness of implicature derivation. It is not possible to tease apart these theories on the basis of intuition and theorizing alone. The discipline of experimental pragmatics emerged in the past three decades, applying experimental methods to quantitatively test predictions of different pragmatic theories. The aims and methods of experimental pragmatics are discussed in Section 2.2.

2.1.4 Pragmatics and rationality

Grice famously suggested that language is "a special case of purposive, indeed rational, behavior" (Grice, 1975, p. 47). Rationality plays an important role in pragmatics. The assumption that pragmatic behavior is rational allows researchers to use formal tools developed to study other rational processes, such as game theoretic approaches and probabilistic Bayesian models (Goodman & Frank, 2016), to study pragmatic phenomena. Moreover, if pragmatic inference is a form of rational reasoning, then insights from theories of rationality, including those of bounded rationality, should have implications for how pragmatic inferences are drawn under resource constraints.

In this section, we first define rationality and briefly review its theoretical foundations. We then discuss approaches that focus on bounded rationality and individual differences in rational thought and consider their implications for studying variation in pragmatic processing, which is the central goal of this thesis.

Defining rationality

Multiple disciplines have been concerned with the study of rationality, including philosophy, psychology, economics, and game and decision theory (see Chapter 1 of Mele & Rawling (2004) for a detailed overview). These fields approach rationality from different perspectives, but a common distinction is drawn between weak and strong definitions of rationality. The weak definition, which can be traced back to Aristotle, defines rationality as basing one's actions on reason. This notion is binary (rational vs. arational) and relatively unrestrictive (Stanovich, 2012). By contrast, the strong definition of rationality is normative (Chater & Oaksford, 2012; Stanovich, 2012): it specifies standards of thinking and action and evaluates rationality in terms of deviations from those standards.

In behavioral economics and decision and game theory, rationality is typically formalized in terms of utility maximization for the agent's goal fulfillment (Von Neumann & Morgenstern, 2007; Weber & Dawes, 2005). This definition of rationality has been adopted by the probabilistic models of pragmatics, which we introduce in Section 2.5.

The other perspective on rationality relevant to this thesis is that of psychology and cognitive science. The question of how rational people actually are has motivated extensive debate, often referred to as the *(Great) Rationality Debate* (Stanovich, 2020; Stein, 1996). Numerous studies have shown that people frequently violate principles of logic and probability, draw conclusions based on irrelevant contextual information, and misapply Bayesian reasoning (Tversky & Kahneman, 1974; Stengård et al., 2022; Fox & Levav, 2004a; Evans, 2016). However, does that necessarily imply that humans are fundamentally irrational? There are different views on this issue.

Within the heuristics-and-biases tradition – most prominently associated with Tversky and Kahneman – human judgment is argued to rely on error-prone heuris-

tics and to systematically fall short of normative standards of rationality (Tversky & Kahneman, 1973, 1974). This position, sometimes termed *meliorist*, posits that such systematic deviations from the normative standard can in principle be reduced or corrected (Stanovich, 1999). In contrast, the Bayesian rationality tradition rejects the view of humans as "failed logicians" (Oaksford & Chater, 2009) and instead reconceptualizes rationality as reasoning under uncertainty. On this view, rationality means maintaining well-calibrated beliefs and acting consistently on the basis of those beliefs. Reanalyses of classic reasoning tasks, such as the Wason selection task, suggest that human performance appears substantially more rational when people's reasoning is viewed as probabilistic rather than purely logical (Oaksford & Chater, 2007). Proponents of Bayesian rationality argue that human reasoning is inherently context-sensitive, whereas in formal logic context is irrelevant and conclusions follow solely from argument structure, which rarely holds in everyday cognition, making logic inappropriate as a normative standard of rationality (Oaksford & Chater, 2009).

Related arguments have been put forward by proponents of ecological rationality, such as Gerd Gigerenzer (Gigerenzer & Goldstein, 1996; Gigerenzer & Selten, 2002), and by researchers working on cognitive architectures, notably John R. Anderson (Anderson, 1991). These approaches challenge the assumption that traditional logical or probabilistic norms should serve as the benchmark for rationality. Instead, they argue that human cognition is adapted to the structure of the environment. From this perspective, heuristics, which within the heuristics-and-biases framework are normally taken to be evidence of irrationality, can instead be understood as instances of rational optimization under constraints (Gigerenzer & Goldstein, 1996). Such heuristics have been argued to be well-suited for use in real-world environments, enabling fast and efficient problem solving while making effective use of limited cognitive resources (Gigerenzer & Selten, 2008).

Resource constraints and individual differences

As we mentioned above, a recurring theme in debates about rationality is that human reasoning is embedded in context and subject to constraints imposed by both the environment and limited cognitive resources. The term *bounded rationality* was coined by Herbert Simon (1955), whose aim was to develop a realistic notion of rationality. He argued that, rather than optimizing globally, humans *satisfice*: they make choices that are good enough given their resource limitations. Work on fast and frugal heuristics builds on this approach and identifies simple decision strategies that often yield good outcomes at a fraction of the time and cognitive cost required by more complex, information-intensive methods (Gigerenzer & Selten, 2002; Gigerenzer et al., 2000; Love et al., 2023).

In the domain of language processing, the notion of bounded rationality is reflected in Ferreira's *good enough processing* account (Ferreira et al., 2009; Ferreira & Patson, 2007), which holds that real-world language comprehension is often shallow

and partial. Instead of computing complete syntactic and semantic representations, comprehenders frequently rely on partial parses that are sufficient for current communicative goals.

A framework closely related to bounded rationality is resource-rational analysis (Griffiths et al., 2015; Lieder & Griffiths, 2020). Building on Anderson's rational analysis (Anderson, 1991), this approach explicitly incorporates the cost of cognitive operations and evaluates reasoning strategies in terms of their efficiency under resource constraints. The aim of resource-rational analysis is to identify the "best biologically feasible mind out of the infinite set of bounded-rational minds" (Lieder & Griffiths, 2020) by identifying the optimal trade-off between solution optimality and available cognitive resources.

Since a central aim of this thesis is to investigate individual variability in pragmatics, research on individual differences in rational thinking is particularly relevant. It has been repeatedly shown that individuals differ substantially in their performance on reasoning tasks: while many participants rely on heuristics or exhibit systematic biases discussed above, there is consistently a subset of individuals who give the normative response, whether the norm is assumed to be logical or probabilistic (Stanovich & West, 1998; Stanovich, 2012). Performance on reasoning tasks, such as syllogistic reasoning and Bayesian probability problems, has been found to correlate with measures of fluid intelligence, suggesting that limitations in cognitive resources contribute to variability in rational thinking. However, these correlations are typically moderate, and in some cases absent or even reversed, indicating that intelligence and rationality are related but distinct constructs (Stanovich & West, 1998; Lopes & Oden, 1991).

To account for this dissociation, Stanovich (2012) proposed an extension of the popular dual-process theories of cognition (Evans & Stanovich, 2013; Kahneman, 2011). Dual-process theories distinguish between automatic, fast Type 1 processing which relies on heuristics and controlled, slow Type 2 processing which is responsible for effortful analytical thinking. Stanovich further subdivides Type 2 processing into the *algorithmic mind* and the *reflective mind*. According to his account, automatic Type 1 processing can be overridden by the algorithmic mind, but it is the reflective mind that initiates that override. The algorithmic mind is responsible for carrying out complex computations and is closely linked to fluid intelligence, whereas the reflective mind governs the initiation of analytic processing and is associated with rational thinking dispositions. This framework helps explain why measures of fluid intelligence, such as Raven's Progressive Matrices (Raven & Court, 1938), are consistently but moderately correlated with measures like Need for Cognition (Cacioppo & Petty, 1982), which is intended to measure intrinsic motivation to engage in effortful cognitive activities, and Actively Openminded Thinking (Stanovich & West, 1997), which aims to measure a general tendency towards openminded thinking, including the willingness to consider other perspectives and openness to changing one's mind (Stanovich & Toplak, 2023).

Implications for pragmatic processing

What implications do these accounts of rationality have for pragmatic processing?

First, as mentioned above, treating communication as a form of rational behavior allows us to apply tools developed for the study of rational behavior to pragmatics. The notion of an agent maximizing their utility in order to achieve a goal while minimizing costs has been adapted by probabilistic Bayesian modeling approaches, such as the Iterated Best Response model (IBR) (Jäger, 2012; Franke, 2009) and the Rational Speech Act (RSA) model (Frank & Goodman, 2012). The RSA model, which we introduce in Section 2.5, is used in this thesis to address Research Goal 4 defined in Introduction, which is to model how comprehenders' beliefs about speakers' reasoning abilities influence interpretation.

Second, accounts of bounded and resource-rational cognition hold that the agent's goal is to find sufficiently good solutions given their resource limitations (Simon, 1955, 1990; Lieder & Griffiths, 2020). Since individuals differ in cognitive resources, as reflected in individual differences in working memory capacity and fluid intelligence (Jarrold & Towse, 2006; Conway & Kovacs, 2013), it follows that different pragmatic strategies may be resource-rational for different individuals. Under the assumption that pragmatic communication is a form of rational behavior, systematic individual differences in pragmatic ability related to cognitive resource limitations may therefore be expected.

Finally, the observation that there are individual differences in rationality, which are predicted by measures of intelligence and rational thinking dispositions associated with Type 2 processing, suggests that, if pragmatic inference is effortful, these constructs may also modulate pragmatic performance. This raises a central question: is pragmatic inference primarily automatic (Type 1), or does it rely on controlled (Type 2) processing? Stanovich (1999) makes a distinction between interactional and analytic intelligence, whereby interactional intelligence, including application of Gricean conversational principles, is assumed to be generally effortless and automatic, whereas analytic intelligence is conscious and effortful. However, a large body of experimental work has provided evidence that implicature derivation can be effortful and highly context-dependent (see Khorsheed et al. (2022) for a review). This suggests that controlled processing is involved, at least in some cases. The study presented in Chapter 4 addresses this issue by examining the relationship between performance on an ad-hoc implicature derivation task – a pragmatic reference game – and individual differences in Raven's matrices (Raven & Court, 1938) and the Cognitive Reflection Test (Frederick, 2005).

2.2 Experimental pragmatics

The experiments described in this dissertation use the methods of experimental pragmatics, an interdisciplinary science which investigates pragmatic phenomena using

psycholinguistic methods. In this section, we discuss the main questions that experimental pragmatics aims to answer and give a brief overview of commonly used methods. Finally, we introduce a behavioral paradigm which will be used for the majority of experiments in this dissertation - pragmatic reference games.

2.2.1 From theory to experimentation

In Section 2.1, we described the foundational theories of pragmatics, which were developed primarily by philosophers of language, who mostly used their own intuitions to find evidence for substantiating their claims (Noveck & Sperber, 2004, Chapter 1). While that has proven to be fruitful – after all, these theories were developed through theorizing – it has been argued that relying on one's own intuitions is insufficient: since pragmatics is all about the specific context in which an utterance is produced and interpreted by a specific listener, it is important to elicit interpretations of utterances in that context as opposed to merely hypothesizing what the interpretation would be (Noveck, 2018). Experimental pragmatics aims to go from idealized theories of communication to real-life communication by people whose cognitive resources are limited.

These are some central aims of experimental Pragmatics:

- **Confirming or disconfirming hypotheses** Pragmatic theories give rise to hypotheses which can then be tested experimentally. For instance, Levinson (2000)'s default account suggests that the derivation of scalar implicature happens automatically, and therefore the pragmatic meaning should generally be more easily accessible than the literal meaning. Relevance theory (Sperber & Wilson, 1986), on the other hand, predicts that accessibility of the literal or pragmatic meaning should be determined by its relevance in a given context. These competing hypotheses can then be tested empirically.

- **Refining pragmatic theories** The possibility of testing pragmatic theories experimentally puts pressure on these theories to be sufficiently specific to be testable (Noveck & Sperber, 2004, Chapter 1). Theories can be continuously refined on the basis of experimental evidence. Below, we will discuss several examples where that has been the case for different pragmatic phenomena.

- **Investigating processing mechanisms** Gricean and neo-Gricean theories predict behavior under idealized conditions. However, there is ample experimental evidence that real-time language processing is often heuristic-based and error-prone (e.g., Ferreira et al. (2002)'s good-enough processing account, which states that people often use heuristics and do not derive full sentence interpretations unless required to by context). How are these accounts of pragmatic competence implemented in the view of cognitive constraints? What is the timing and cognitive costs of inferences?

- **Assessing & explaining variability** Theories of idealized communication do not provide an explanation of experimental variability. It has been observed that individuals vary a lot in their tendency to access nonliteral meanings (e.g., Heyman & Schaeken (2015) showed that a subgroup of their participants regularly derived scalar inferences while the other subgroup consistently responded literally). What cognitive and personality traits drive those differences?

Applying experimental techniques to the study of various domains of pragmatics has resulted in a deepened understanding of those pragmatic phenomena and more accurate theories. For example, experimental investigation has shown that in scalar implicature, the pragmatic meaning appears to be derived from the literal meaning and not accessible by default as evidenced by processing delays and by the fact that people generally show more literal responding under cognitive load (Breheny et al., 2006; Dieussaert et al., 2011; Khorsheed et al., 2022), contrary to what default accounts predict. However, the same does not appear to hold for figurative language: for conventionalized metaphors, the figurative meaning has been shown to be accessed very fast and sometimes before the literal meaning if presented in realistic contexts (Gibbs Jr & Colston, 2012). Other domains where experimental methods helped develop and refine theories include Clark's experimental work showing that indirect speech acts are interpreted by participants as having both the direct meaning and the speaker's intended meaning (Clark, 1979) and Matsui's work explaining bridging inferences using relevance theory (Matsui, 2000).

We will now present some approaches commonly used in experimental pragmatics. We will then introduce a paradigm which will be used extensively in this dissertation – pragmatic reference games.

2.2.2 Commonly used methods

There is a growing number of methods and paradigms in experimental Pragmatics, and the brief overview we give below is by no means an exhaustive list. The purpose of this overview is to give the reader the sense of the breadth of paradigms that exist. Ultimately, the choice of method depends on the question that the study sets out to answer.

The experiments in this dissertation use outcome-based tasks – i.e., where task performance, as opposed to e.g. its timecourse, is used as the measure – and correlation with cognitive individual difference measures. Our data is collected using crowdsourcing. The reason for this is that our hypotheses relate to the mechanisms underlying processing but not its timecourse, and also that the paradigms we use are generally too coarse-grained for e.g. reaction time information to be meaningful. However, for our adaptation experiments described in Chapters 6 and 7, we collected the time it took participants to complete each trial, so in case future work develops predictions concerning the timing of pragmatic adaptation in the reference game or block building game, they can be tested with our data.

Behavioral methods: Outcome-focused measures

Outcome-based behavioral measures focus on participants' responses on a given task. In other words, they are centered around *what* a participant infers (as opposed to, for instance, *how fast*).

Examples of such measures include truth-value judgment tasks, where participants are asked to evaluate whether a sentence is true or false, forced-choice selection tasks and gamified designs.

It is a common method for studying scalar implicatures to ask participants for acceptability ratings of an underinformative sentence ("Some giraffes have long necks"), either as a binary true-or-false response or as a response on a scale. This approach helped reveal that adults reject underinformative sentences more often than children (Noveck, 2001) and that lexical scales differ in their derivation rate (van Tiel et al., 2016).

Many studies in pragmatics use more gamified designs, where participants are asked to either describe a picture of an object in context or to pick out an object based on a description. These types of designs are common for studies whose results are then mathematically modeled in the Rational Speech Act framework. I will describe one such paradigm, reference game, in more detail later in this chapter since it will be used in many of the experiments of this dissertation. An example of a comprehension experiment in this kind of paradigm is a study on ironic language comprehension by Kao & Goodman (2015), where participants are shown pictures of weather contexts and sentences (e.g. "Wow, this weather is amazing!") and asked to rate how likely the sentence is to be literal vs. ironic. Since it has been shown that sufficient motivation is needed for participants to invest effort in perspective-taking (e.g., Gibbs Jr et al., 1991), often these designs involve either an actual partner who the participant is interacting with in the lab (e.g., Hanna & Tanenhaus, 2004; Loy et al., 2020) or over the internet (e.g., Hawkins et al., 2017; Graf et al., 2016), or the partner is either simulated or it is said that the messages are coming from the participant of an earlier study (e.g. Franke & Degen, 2016; an approach we use in our experiments).

Online measures: Reaction times, eye-tracking, and neuroscientific methods

Timecourse-focused paradigms are used to test hypotheses related to the *timing* of an inference. Reaction time (RT) studies often collect by-word processing of a sentence, and longer RTs are often interpreted as indicating processing difficulty. In self-paced reading (SPR) tasks, participants read sentences word by word and progress through the sentence using a space bar. Bergen & Grodner (2012) used SPR to investigate the effect of speaker knowledge on scalar inferences and observed a delay on the words following the quantifier in underinformative some-sentences when the context suggested full speaker knowledge, and a speedup on the continuation when it was compatible with the recently computed interpretation. While generally longer RTs are associated

with processing difficulty, Nicenboim et al. (2016) argued that sometimes high RTs may be indicative of less careful reading and not committing to an interpretation.

Eye tracking, i.e., recording participants' eye movements, is typically used in experimental pragmatics for investigating reading – in that case, reading times are usually of interest, but additional reading measures can be derived compared to SPR, e.g., first-pass reading time, total reading time, and number of fixations on a given region – or in the so-called visual world paradigm, where participants hear sentences while looking at an array of objects. Turcan & Filik (2016) used eye-tracking to investigate the processing of written sarcasm, revealing that sarcastic comments generally resulted in longer reading times compared to literal comments. An example of a visual world study in pragmatics is the investigation of scalar implicature processing by Degen & Tanenhaus (2016).

In addition to reaction time (RT) and eye-tracking measures, neuroscientific methods are used to investigate online pragmatic processing, most notably functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). fMRI makes it possible to localize brain regions involved in processing by detecting changes in blood flow and oxygenation; however, it has relatively low temporal resolution (Constable, 2023). EEG, by contrast, allows for better time-locking of neural activity in response to stimuli, but provides less precision regarding where in the brain this activity occurs (Burle et al., 2015).

An example of the use of fMRI pragmatics is the study by Feng et al. (2017), who studied the effects of contextual relevance in implicature derivation and demonstrated distinct brain activation patterns in response to contextually relevant vs. irrelevant replies.

In EEG research, two components that are particularly prominent in language research are the N400 – a negative deflection occurring approximately 400 milliseconds after stimulus onset – and the P600 – a positive deflection occurring around 600 milliseconds after stimulus onset. The N400 has been linked to lexical retrieval and semantic incongruency, whereas the P600 has been associated with syntactic violations as well as more general processes of repair and integration. However, the mapping between these components and specific cognitive processes is nuanced and subject to ongoing debate (Noveck, 2018; Delogu et al., 2019; Seyednozadi et al., 2021), and a detailed discussion lies outside the scope of this thesis. Both N400 and P600 effects have been observed when a speaker's utterance does not align with stereotypes associated with their identity or communicative style (N400: van Berkum et al. (2008); P600: Regel et al. (2010); Lattner & Friederici (2003); both N400 and P600: Wu & Cai (2025)).

Effortfulness: Dual-task setups

While longer response times are often interpreted as an indication of effortul processing, another commonly used technique for studying the amount of cognitive effort

involved in processing of a given pragmatic phenomenon is dual-task setups, where participants read or listen to sentences while performing another task at the same time. Since cognitive resources are assumed to be limited and shared between the two tasks, the second task is said to increase the cognitive load. If participants' performance on a task diminishes under cognitive load, that is taken to mean that processing is effortful and hinges on availability of cognitive resources. De Neys & Schaeken (2007) and Dieussaert et al. (2011) had participants memorize dot patterns concurrently with interpreting underinformative statements and found that the number of logical responses grows with increased cognitive load.

Studying individual variability: Categorizing responders and correlation with cognitive measures

Significant individual variability has been observed in various pragmatic tasks. Two common approaches for studying it have included categorizing responders into groups and studying the correlation of pragmatic task performance with cognitive measures.

To make sense of variability between participants, it can be useful to categorize them into groups based on their response patterns. For instance, it has been observed that while some participants consistently derive implicatures, others consistently prefer the literal meaning, and some participants are inconsistent in their responding. Approaches which have been used for that include categorization based on a threshold (Degen & Tanenhaus, 2016; Ryzhova et al., 2023), Latent Profile Analysis (Heyman & Schaeken, 2015) and hierarchical Bayesian modeling (Franke & Degen, 2016). The performance of each group can then be examined separately.

A common approach to studying the sources of individual variability is examining which cognitive individual differences measures (e.g., working memory capacity, Theory of Mind) correlate with participants' performance on a pragmatic task. Often, a battery of individual difference measures is collected and used as predictors of pragmatic performance in a linear mixed-effects model or a structural equation model. Like with any correlational study, caution is called for when interpreting the results, since it may be a third variable which underlies the correlation between a cognitive variable and performance on a pragmatic task. Since contributing to the understanding of individual variation in pragmatic processing is one of the central aims of this thesis, Section 2.3 of this chapter includes a review of literature on cognitive individual differences in pragmatics.

Crowdsourcing as a data collection approach

While it is not a method in experimental pragmatics per se, crowdsourcing is a data collection technique which has gained popularity in experimental pragmatics in recent years and which was used for collecting all data used in the experiments in this dissertation.

Crowdsourcing is a data collection method where, instead of inviting participants to come to the lab, they are recruited through an online platform. Compared to a traditional laboratory design, crowdsourcing involves more limited control over the environment in which participants take part in the study (Noveck, 2018) and higher risk of participants getting distracted or being insufficiently motivated (Gould et al., 2016; Shapiro et al., 2013), but it allows data collection at a much increased speed and lower cost and potentially gives access to a more naturalistic diverse sample (Paolacci & Chandler, 2014). Additionally, it gives researchers the option of one-shot or few-shot setups (e.g., Frank & Goodman, 2012), where each participant only completes one or a few trials, which is uncommon in the lab context.

While concerns have been raised about data quality and participant motivation, studies show that, provided that best practices are followed, crowdsourced data quality is generally quite high (Shapiro et al., 2013), and techniques such as using more engaging gamified experiment designs (Krause & Kizilcec, 2015) and framing the task as meaningful (Chandler & Kapelner, 2013) have been shown to result in improved data quality (see Daniel et al. (2018) for a review of different quality assurance strategies).

Two crowdsourcing platforms commonly used for the collection of psycholinguistic data are Amazon Mechanical Turk (MTurk) and Prolific Academic. Compared to MTurk, participants on Prolific have been shown to produce higher quality data, be more attentive, and provide more meaningful answers (Douglas et al., 2023; Albert & Smilek, 2023).

2.2.3 Reference games

The experiments presented in this dissertation, with the exception of Chapter 7, use the so-called reference game paradigm. In this section, I will discuss the paradigm and its variations and consider its limitations.

Reference games are simple communication games involving two players, a speaker and a listener. They are in part inspired by the concept of signaling games in game theory (Franke, 2009), where two players try to achieve an equilibrium. They are also related to Wittgenstein's concept of *language games*, where simple language is used to induce action (Wittgenstein, 2009).

In reference games, a speaker and a listener communicate using a limited language which often uses symbols or pictures. The speaker's task is to send the listener a message to refer to one of the objects in a shared view. The listener's task, in turn, is to identify the referent of the speaker's message.

Reference games yield themselves naturally to being described as computational models in the Rational Speech Act framework (Frank & Goodman, 2012), which I will describe in Section 2.5, since the set of possible alternative utterances and of possible referents are explicitly defined and shared between the speaker and the listener.

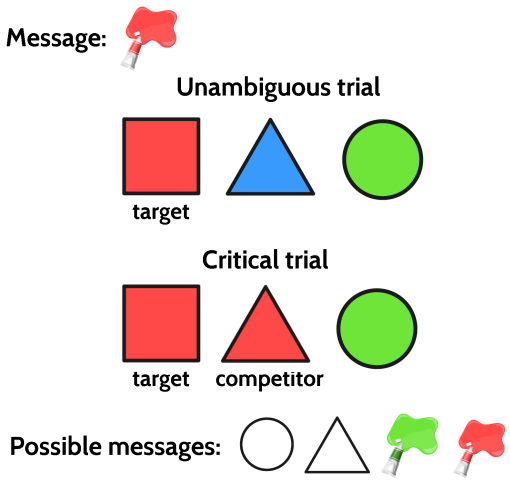


Figure 2.1: Examples of an unambiguous and an implicature reference game trials.

Let's look at an example of such a reference game. Figure 2.1 shows a set of referents which differ along two dimensions, color and shape. There is also a restricted set of possible messages which is known to both players. The speaker's task is to select one of the four possible messages to describe a given referent such that another participant may be able to recover the selected referent. The listener's task is the opposite: given a speaker's message, their task is to identify the object the speaker was trying to refer to.

Cases where the message is ambiguous are interesting. In the bottom panel of Figure 2.1, the message "red" sent by the speaker is on the surface ambiguous between the target (red square) and the competitor (red triangle). One could resolve this ambiguity by reasoning about other messages the speaker could have used: since "square" is not in the set of available messages, the message "red" is the only way to refer to the red square. If the speaker wanted to refer to the red triangle, the more optimal unambiguous message would have been "triangle". Since the speaker did not use it, one could infer that the speaker meant to refer to the red square.

We will call the red square the target and the red triangle the competitor. Picking the target means correctly solving the implicature, i.e., resolving the ambiguity by reasoning about the speaker's intended meaning. A literal interpretation of a message on an implicature trial is therefore ambiguous between the target and the competitor, whereas a pragmatic interpretation favors the target as a result of reasoning about alternatives the speaker could have used. That is the reasoning formalized by Rational Speech Act models.

Frank & Goodman (2012) showed that participants tended to select the pragmatic target even in a one-shot design despite the minimalistic and nonlinguistic nature of the task. However, Sikos et al. (2021a) later examined the stimuli used by Frank & Goodman (2012) more carefully and observed that only a small subset of the items used were ambiguous implicature trials, and that performance on those trials was far less optimal than what the authors of the original paper claimed. Still, in

multitrial designs, it has been shown that participants are generally able to solve such implicature trials correctly and that significant individual variation exists (Sikos et al., 2021b; Franke & Degen, 2016).

There are different options available for the dependent measure in these paradigms, including (1) forced choice, where participants select one of the objects (used e.g. in Franke & Degen, 2016 and Qing & Franke, 2015), (2) a betting paradigm, where participants are instructed to allocate $100 between the options (used in Frank & Goodman, 2012 and Goodman & Stuhlmüller, 2013), and (3) a likert scale (used in Kao & Goodman, 2015). Frank et al. (2016) compared these dependent measures and found that forced choice resulted in the highest proportion of target selections, while the scale and betting measures allowed for more hedging, which conservative participants took advantage of. Qing & Franke (2015) argue that these different measures tap into different constructs: the forced choice paradigm measures *action* – the likelihood of the agent selecting a given utterance or referent – whereas the graded measures are more likely to tap into the agent's *beliefs* and be associated with introspection. In this dissertation, we thus use different dependent measures depending on the goal of the experiment. Chapters 4 and 6 use a forced choice paradigm, whereas Chapter 5 uses a variant of the betting paradigm because we are interested in collecting more graded beliefs.

Advantages and limitations

We use reference games for most of the experiments in this thesis, with the exception of Chapter 7. Below, we motivate this paradigm choice and discuss the advantages and limitations of this paradigm.

To begin with limitations, since reference games are often non-linguistic and use a highly constrained set of possible messages, it can be argued that what we can learn from them has limited generalizability to real-life communication using language, where the space of alternative utterances is very large and potentially unbounded. That is a valid concern. However, we show in Chapter 7 that the adaptation effects observed in the reference game are very similar to those that we observe in a more naturalistic paradigm with linguistic stimuli, a collaborative building game. That, along with the fact that reference game formalizes implicatures similar to those used in language, suggests that valuable insights about pragmatic processing can be gained from this paradigm.

Reference games have a number of advantages which make them suitable to the aims of this dissertation, which are to study variability and adaptation in pragmatic reasoning. They are tightly controlled – the set of available alternatives is known to both interlocutors and therefore does not need to be assumed – leading to interpretable results. In addition, in reference games, implicatures of different complexity can be defined, and formal probabilistic pragmatic models make specific predictions about which implicatures are expected to be harder than others, which may be less

clear for linguistic implicatures.

We argue that, despite the very simple design, reference games allow us to learn useful theoretical facts about pragmatic processing that can be straightforwardly modeled in the Rational Speech Act framework. In Chapter 4, we show that pragmatic reasoning on reference games is associated with logical reasoning ability and Theory of Mind; these cognitive traits have been shown to modulate performance on other pragmatic phenomena, such as metaphor comprehension (Puhacheuskaya & Järvikivi, 2025), scalar implicatures (Fairchild & Papafragou, 2021), and indirect requests (Trott & Bergen, 2019). In Chapter 5, we use reference games to study the effect of speaker identity on comprehension and show that implicature derivation is sensitive to the speaker's perceived pragmatic competence: when the speaker was believed to be a child or an artificial agent, participants were less likely to derive a pragmatic interpretation than when it was believed to be another adult. In Chapter 6, we show that participants track their partners' perceived pragmatic competence during interaction and may cancel an implicature if they believe the speaker to be insufficiently pragmatically competent.

Reference games thus provide a strong foundation for follow-up studies in other, more naturalistic paradigms. Chapter 7 of this thesis builds on the findings of the previous chapters and uses a more complex paradigm, a collaborative building game with linguistic instructions. In Section 8.2 of the conclusion, we provide recommendations for extending the findings of this dissertation to other paradigms.

2.3 Cognitive individual differences as a source of variability in pragmatics

Bounded-rational accounts of cognition state that processing is constrained by the available cognitive resources. Thus, under the view of communication as a kind of rational behavior, it follows that variation in cognitive resources may give rise to variation in pragmatic processing. A growing body of work has attempted to relate the observed variability in pragmatic processing to cognitive individual differences.

The space of cognitive individual differences is broad. However, in this section we focus on a subset that we expect to be particularly relevant to the pragmatic phenomena investigated in this thesis. Specifically, working memory capacity, Theory of Mind, and logical reasoning ability (including reflective thinking and pattern recognition) are expected to modulate pragmatic processing in the reference game. In addition, working memory updating is expected to play a role in adaptation over time. This section is based on, and in parts identical to, Mayn & Demberg (2026).

Before reviewing related work on these individual differences in detail, we include a brief summary of other factors that have been shown to influence pragmatic processing.

Previous work has demonstrated that pragmatic and discourse processing can be modulated by linguistic abilities, such as vocabulary knowledge (Matthews et al., 2018) and exposure to print (Scholman et al., 2020). However, the variant reference game task used in this dissertation is not language-based and uses symbols as possible messages and referents. Thus, for this particular setting, we did not expect linguistic skills to have an effect.

Other studies in pragmatics have reported effects of personality traits, such as openness (Verhoeven & Vermeer, 2002), socio-pragmatic traits measured by the Autism Spectrum Quotient (AQ) (Yang et al., 2018), and self-reported thinking dispositions, such as Need for Cognition (NfC) (Cacioppo & Petty, 1982). These measures are not examined in this thesis for several reasons. First, in an early pilot study of individual differences in a reference game (Mayn & Demberg, 2022), we found no evidence for effects of NfC or AQ. Second, the focus of this dissertation is on cognitive traits that influence pragmatic processing. This is not to suggest that personality or neurodevelopmental factors do not play a role. However, their investigation is outside the scope of this thesis.

2.3.1 Working memory

Working memory capacity (WMC) is a core cognitive function, which is usually assumed to consist of a memory storage component and controlled attention, which allows for maintenance goal-related information in memory in the face of interference (Engle et al., 1999; Conway et al., 2003). Research on the role of WMC in pragmatic inference has yielded mixed results. While some studies have found a relationship between WMC and implicature derivation, suggesting that implicature derivation may be effortful and may depend on the availability of cognitive resources, other studies have reported no effect of working memory on pragmatic processing.

There are two main ways in which previous work has investigated whether working memory is involved in pragmatic processing. The first way is to use a measure of working memory capacity, such as Operation Span (Unsworth et al., 2005) or Reading Span (Daneman & Carpenter, 1980), as a predictor of performance on a pragmatic task. The second approach, commonly employed in studies of scalar implicature, relies on a dual-task setup, in which participants process underinformative sentences while simultaneously performing a secondary task designed to tax cognitive resources.

Studies which used the dual-task setup generally found that people tended to respond more logically under cognitive load, which has been interpreted as evidence that implicature derivation may engage cognitive resources (De Neys & Schaeken, 2007; Dieussaert et al., 2011; Marty & Chemla, 2013; Marty et al., 2013). Dieussaert et al. (2011) additionally found that the effect of cognitive load was more pronounced for individuals with low WMC.

In contrast, studies that used WMC measures as predictors of pragmatic performance have produced mixed findings. Antoniou et al. (2016) found that participants

with higher WMC, as measured by the Backward Digit Span Test and the Reading Span task, were more likely to reject underinformative statements. By contrast, Heyman & Schaeken (2015) found that WMC, as measured by a modified Operation Span task, was not a significant predictor of scalar implicature derivation. Antoniou et al. (2016) suggested that features of their experimental design may have placed additional demands on working memory. This raises the possibility that implicature derivation is cognitively effortful, but that in some cases—such as generalized implicatures with strong contextual support—the required effort is low enough that even individuals with lower WMC can derive them. Under such conditions, relationships between WMC and implicature derivation may only emerge when working memory is explicitly taxed, as in a dual-task paradigm.

Consistent with this view, in a study co-authored during the dissertation period but not included in the thesis (Ryzhova et al., 2023), we found no effect of verbal working memory, as measured by Reading Span, on the rate of inferences about event typicality triggered by redundant mentions of highly typical events (*atypicality inferences*; Kravtchenko & Demberg, 2022). However, using a dual-task design, Ryzhova & Demberg (2020) found evidence that the derivation of such inferences may be effortful.

Yang et al. (2018) showed that participants with higher WMC were better able to incorporate contextual information into scalar implicature derivation. Bambini et al. (2021) showed that WMC is a significant predictor of implicature derivation and figurative language comprehension in older adults.

In addition to having a direct influence on pragmatic processing, working memory may also play an auxiliary role, where it supports some other process which, in turn, is involved in pragmatic processing. Fairchild & Papafragou (2021) investigated whether individual differences in executive function (EF), a construct closely related to working memory capacity, predicted pragmatic responding in scalar implicatures, as well as indirect request and metaphor comprehension. They found that EF was a significant predictor of pragmatic responding on its own, but when a measure of Theory of Mind was added to the model, EF ceased be significant, suggesting that EF may play a supporting role in that it supports perspective taking.

Overall, the literature points to a complex role of WMC in pragmatics. Working memory may be differentially engaged across pragmatic phenomena, and even within a single phenomenon, such as scalar implicature, its involvement may depend on factors such as contextual support and task demands.

Working memory updating

An ability related but distinct from WMC is *updating*, an executive function which is assumed to be responsible for monitoring representations in working memory and updating outdated information (Miyake et al., 2000). It has been consistently shown to be strongly linked to WMC (Shipstead et al., 2016; Ecker et al., 2010), and some

have argued that updating tasks are also valid measures of WMC (Schmiedek et al., 2009). However, updating additionally emphasizes the temporal component of replacing outdated information (Miyake et al., 2000), and it has been shown in developmental research that high updating scores don't always co-occur with high WMC (Panesi et al., 2022).

Working memory updating, often measured using the Keep Track Task (Yntema, 1963), has been shown to predict pragmatic adaptation. Loy & Demberg (2023) found that updating ability but not WMC was related to participants' ability to adapt to two speakers with distinct communicative styles, suggesting that the modification of mental representations plays a key role in pragmatic adaptation. Similarly, in a paper co-authored during the dissertation period but not included in this thesis (Schuster et al., 2023), we found that participants with higher updating scores were more likely to adapt to two speakers' distinct uses of uncertainty expressions.

Although the research presented in this thesis does not include a measure of working memory updating, this ability is relevant for interpreting the findings of Chapter 6. In particular, individual differences in updating may help explain observed variation in patterns of dynamic adaptation to speakers' pragmatic competence.

2.3.2 Theory of Mind

Theory of Mind (ToM), also sometimes called mentalizing (Trott & Bergen, 2019) or mindreading, is the ability to attribute mental states, such as beliefs and desires, to oneself and others (see Quesque et al. (2024) for a discussion of subtle distinctions between these terms). It has been shown to play a role in pragmatic processing in numerous studies on children and clinical populations, as well as in a growing number of studies on neurotypical adults.

A large number of studies investigated the relationship between Theory of Mind and pragmatic processing in children and adolescents and how it changes with age. Longitudinal studies on the role of Theory of Mind in metaphor comprehension have shown that ToM supports metaphor comprehension in children of ages 8-9 but reliance on ToM diminishes with age (Bambini et al., 2025; Tonini et al., 2023; Del Sette et al., 2020). A positive relationship between children's ToM skills and comprehension of irony and deceit has been shown (Nilsen et al., 2011; Bosco & Gabbatore, 2017).

Studies in clinical populations, such as adults with schizophrenia (see Bosco et al. (2018) for a review) have observed a co-occurrence of ToM impairment with impairment of pragmatic skills. There is also a debate in the literature about ToM impairment, as well as its role in pragmatic processing in adults with autism, with some studies claiming that there is a ToM impairment in individuals with autism (Barnes & Baron-Cohen, 2011; Cummings, 2013), and other studies calling for caution due to the fact that many existing ToM tests are based on neurotypical conception of social norms and behavioral expressions and may unfairly disadvantage neurodivergent populations (Long et al., 2025; Milton, 2012).

A growing body of work suggests a link between Theory of Mind and pragmatic processing in neurotypical adults as well, which is our population of interest. Several studies observed that ToM facilitated processing of indirect requests. Marocchini et al. (2022) showed that participants with higher ToM, as measured by the Strange Stories task (Happé, 1994), showed facilitated processing of conventionalized indirect requests in a self-paced reading paradigm. Trott & Bergen (2019) and Fairchild & Papafragou (2021) also observed that ToM, measured by the Short Story Task (Dodell-Feder et al., 2013) and a composite score of the Strange Stories task and the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001) respectively, modulated indirect request comprehension. Higher Theory of Mind abilities has also been found to be predictive of humor comprehension: Bischetti et al. (2023) showed that older adults with higher ToM demonstrated better comprehension of mental but not phonological jokes, and Loy & Demberg (2024) showed that participants with better mentalizing abilities were better at tracking whether a speaker was likely to produce a humorous utterance. ToM abilities have also been shown to modulate scalar implicature derivation and metaphor comprehension in neurotypical adults (Fairchild & Papafragou, 2021).

2.3.3 Logical reasoning

Pragmatic reasoning, variation in which this thesis investigates, is a form of reasoning. Since deriving inferences, particularly in the case of particularized implicatures, requires reasoning about the context, the speaker's intentions, and alternative utterances the speaker could have produced, pragmatic reasoning may draw on general reasoning abilities. Here, we use the term *logical reasoning* broadly to distinguish it from pragmatic reasoning. We use this term synonymously with what has been referred to as *general reasoning* (Scholman et al., 2024) and *analytical reasoning* (Williams et al., 2019). Logical reasoning is associated with the ability to reason flexibly, reflect on new information, identify patterns, and solve novel problems. Within dual-process models of cognition, this ability is typically linked to System 2, which is responsible for effortful analytical processing and hypothetical reasoning (Stanovich, 2012; Kahneman, 2011; Evans & Stanovich, 2013).

Here, we review existing work which has used two related but distinct measures of reasoning – Raven's Progressive Matrices (RPM) (Raven & Court, 1938) and Cognitive Reflection Test (CRT) (Frederick, 2005) – as predictors of pragmatic and discourse processing. RPM is often used as a measure of fluid intelligence and is assumed to tap into pattern recognition and ability to reason in novel situations. Which constructs the CRT taps into and how closely it is linked with intelligence has been debated in the literature: some studies suggest that the CRT measures reflective thinking and the ability to override the intuitive incorrect response (Frederick, 2005; Toplak et al., 2014; Campitelli & Gerrans, 2014), while other, more recent studies suggest that it is primarily a cognitive measure which is strongly associated with intelligence (Otero et al., 2022; Welsh, 2022). It has also been shown that the CRT

is a distinct predictor on tasks of rational thinking after fluid intelligence has been accounted for (Toplak et al., 2011, 2014). The results of the experiment reported in Chapter 4, together with those of a co-authored study on individual differences in atypicality inferences not included in this dissertation (Ryzhova et al., 2023), support treating RPM and CRT as a combined measure of logical reasoning.

Heyman & Schaeken (2015) investigated the effect of the CRT on scalar implicature derivation. Contrary to their hypothesis, they did not find that individuals with higher CRT were more likely to derive an implicature. Instead, their results suggest that higher-CRT individuals were more consistent in their responses, whether they were logical or pragmatic.

Scholman et al. (2020) examined the effects of RPM and CRT on connective interpretation in discourse. They observed a positive effect of RPM, such that participants with higher RPM scores showed better comprehension of connectives. In contrast, higher-CRT participants did not perform better overall on connective comprehension, but instead provided lower coherence ratings across conditions. The authors interpreted this pattern as indicating that higher-CRT individuals are more critical of connective–relation mappings and more sensitive to the narrative structure of the materials.

In the co-authored study on atypicality inferences Ryzhova et al. (2023), we found that individuals with a higher composite reasoning score, consisting of the CRT and the RPM, were more likely to derive an inference. In a study on irony comprehension in written texts, Kyriacou & Köder (2024) found an effect of RPM.

Overall, although the available evidence is limited, existing work generally supports the view that logical reasoning plays a role in pragmatic processing. Studies using RPM generally indicate that higher RPM scores are associated with more pragmatic responding. In contrast, findings involving the CRT present a more complex picture, suggesting that the CRT may be related not only to pragmatic responding but also to response consistency and to critical evaluation of the structure of the materials.

2.4 Effects of speaker identity & adaptation to specific speakers

Communication does not exist in a vacuum, and inferences are drawn in context. Recent work has shown that the profile of the interlocutor is part of this context: the same utterance may be interpreted differently depending on who it is said by.

In this section, we review related work on the effects of speaker identity on interpretation. I will start with reviewing key findings from two related areas and explain how they can inform our understanding of speaker effects in pragmatics. Studies in speech recognition have shown that information about the speaker is processed very early on and is integrated with the semantic meaning of the sentence. In addition, we discuss a theoretical model of speaker effects on utterance processing, which is also applicable

for pragmatic processing and can used to place studies on pragmatic adaptation in this dissertation in relation to existing theoretical models. Then we review psycholinguistic work on social meaning and identity construction which informed pragmatic work on the influence of speaker persona on utterance interpretation. Finally, we review studies specifically in pragmatics which have suggested that comprehenders' inferences are affected by speaker characteristics and that comprehenders dynamically adapt their interpretations to specific speakers.

This dissertation focuses on pragmatic processing in comprehension as opposed to production. Thus, we review studies on the effects of speaker characteristics on the listener. There is, however, also the other side of the picture – the role of the addressee. It has been shown that speakers adjust the way they speak based on factors such as familiarity with the listener (Krauss & Fussell, 1991), the listener's knowledge (Fussell & Krauss, 1989), and whether the listener is a native or a nonnative speaker of the language (Hu, 2022). Tailoring utterances to be maximally helpful to a specific listener is known as *audience design* (Bell, 1984).

2.4.1 Speaker effects in spoken word processing

Processing studies have shown that speaker identity is integrated into interpretation very early on, as evidenced by N400 and P600 effects in EEG experiments. Listeners noticed when the content of the utterance did not match their expectations of the speaker. van Berkum et al. (2008) observed an N400 effect when the content of the utterance was in conflict with the persona stereotypically associated with the voice of the utterance (e.g., a child drinking wine, a person with an accent associated with the working class going to the opera, or a man engaging in a stereotypically female-coded activity such as getting a manicure), whereas Lattner & Friederici (2003) observed a P600 effect. Wu & Cai (2025) observed an N400 effect when the activity violated social stereotypes but was possible and a P600 effect when the literal interpretation was highly unlikely; the N400 effect was additionally modulated by openness.

It has also been found that information about the speaker that stems from acoustic signal influences meaning access. When hearing a polysemous word for which different meanings are more frequent in British vs. American English (e.g., *flat*), the meaning was more quickly accessed if spoken with an accent in which it is prevalent; and even when the accent was neutral, participants still exhibited the same pattern if the expectation of the speaker's dialect was established (Cai et al., 2017). Priming with a specific social type has also been shown to influence the perception of vowel boundaries (D'onofrio, 2018).

The work described above shows that traits associated with a demographic are integrated with the semantic content of the utterance early on. There is also research showing that listeners appear to build a speaker-specific model as well when exposed to a speaker's utterances. For instance, it has been shown that participants expected a speaker to use the same label to refer to the same referent and were slower to locate

it when the same speaker started using a new label than when a different speaker stated using a new label (Metzing & Brennan, 2003). Roettger & Rimland (2020) showed that listeners interpreted intonation clues in a speaker-specific manner: when a speaker's intonation patterns were consistent, listeners used them to disambiguate between referents, whereas when a speaker's intonation patterns were unreliable, intonation was not used for disambiguation. Demographic-based and speaker-specific expectations were also shown to interact: Wu et al. (2024) found that label switching was associated with an N400 and a P600 effect, and that the N400 was greater when a child switched labels than when an adult did, indicating an expectation that children are less flexible in their language use.

In response to the above findings, theoretical models have been proposed to explain the integration of speaker information with the semantic content of the utterance. Cai et al. (2017) proposed a dual-pathway model, where the lexical-semantic pathway is responsible for word identification and processing and the other pathway, *the speaker model*, uses information related to the characteristics of the speaker, such as their age, gender and dialect, inferred from the utterance to modulate the construction of lexical-semantic meaning. Wu & Cai (2024) further refined Cai et al. (2017)'s model, introducing a distinction and an interplay between an individual speaker model, reflecting experience interacting with an individual speaker, and a demographic speaker model, reflecting experience interacting with a population, both of which inform each other and jointly guide processing and meaning access. According to Wu & Cai (2024)'s account, when exposed to a new speaker, a comprehender initially uses a general demographic model. As they gain more experience with a specific speaker, they gradually develop an individual speaker model, which may also override or modify the demographic model. While this model was proposed for spoken word processing, similar principles can be expected to hold for interpretation of written utterances or of non-linguistic signals at a higher level.

In Chapters 6 and 7 of this thesis, we find that when interacting with two speakers of different pragmatic competence, participants exhibited a gradual shift in confidence which was distinct for the two speakers, consistent with the speaker-specific model, but also that their initial confidence for the speaker they saw second was influenced by the first speaker they were exposed to, suggesting that exposure to the first speaker modified their model of the whole population, which was then used to guide the interpretation of the signals of the second speaker.

2.4.2 Speaker identity and social meaning in sociolinguistics

Another perspective on the influence of speaker identity on meaning construction comes from sociolinguistics. While pragmatics is largely concerned with how non-literal meaning is shaped by the context and the communicative intent of the speaker, sociolinguistics focuses on the related but distinct question about how social identities, such as socioeconomic status, political orientation, age and gender co-occur with and

are indexed by linguistic cues (Eckert, 2019).

The concept of *social meaning*, which is distinct from the utterance's semantic meaning, is the combination of properties that the utterance, including word choice and pronunciation, conveys about the speaker's social identity (Beltrama, 2020). Social meaning is not fixed and is constantly open to being reevaluated. Unlike pragmatic meaning, which is usually associated with the speaker's intent, social meaning may or may not be intentional (Beltrama, 2020; Bucholtz & Hall, 2005).

Social meanings are flexibly instantiated in a given context to give rise to social identities (Eckert, 2008) and to signal relationships between interlocutors (Jeong, 2021). It has been argued, for instance, that the use of uncertainty words like *probably* is associated with politeness (Bonnefon & Villejoubert, 2006), as are rising declaratives (Jeong, 2021), and that demonstratives can carry the social meaning of solidarity (Acton & Potts, 2014).

In a series of studies at the intersection of sociolinguistics and pragmatics, Andrea Beltrama and Florian Schwarz's work has investigated the influence of social persona on the interpretation of numeric expressions. They found that participants had the expectation that "nerdy" characters would utter more precise expressions than "chill" characters (Beltrama & Schwarz, 2021, 2024). Going in the other direction, they found that degree of numeric precision also gives rise to inferences about the speaker's persona: more precise speakers are expected to be more intelligent and educated, but also more uptight and annoying (Beltrama, 2018). Such studies show that social meaning is intertwined with pragmatics and that speaker characteristics may influence the outcome of pragmatic reasoning (Beltrama & Schwarz, 2024).

2.4.3 Speaker effects in pragmatics

Above we discussed two theoretical strands which have formalized the role of speaker identity in utterance processing. Next, we'll review existing work specifically in pragmatics which suggests that comprehenders' interpretations are sensitive to the properties of their interlocutor.

Several studies have investigated the effect of the speaker's nativeness and language experience on interpretation. It has been shown that ironic utterances were rated as less ironic (Puhacheuskaya & Järvikivi, 2022) and metaphors as less sensical (Puhacheuskaya & Järvikivi, 2025) when uttered by a nonnative speaker. The metaphor effect was also modulated by individual differences: it was strongest for participants with less cognitive reflection and coming from less diverse backgrounds. Nativeness also has been shown to influence interpretations of informativeness: Ip & Papafragou (2021) showed that when underinformative utterances came from native speakers, that resulted in the speaker being perceived as untrustworthy, whereas when they came from nonnative speakers, their trustworthiness was less affected because underinformativeness was attributed to limited language proficiency. Similarly, Fairchild & Papafragou (2018) found that underinformative sentences ("Some people

have noses with two nostrils") were judged to be more acceptable when produced by nonnative speakers and that this effect was stronger for less proficient speakers.

Studies on communication with artificial agents versus humans show that perceived communicative competence of the interlocutor influences perspective-taking. Fischer (2005) found that participants gave simpler, more streamlined instructions to a nonverbal robot, whereas instructions to a verbal robot were more complex, varied, and included some egocentric references, suggesting that beliefs about the robot's limited communicative competence influenced their behavior. Similarly, Branigan et al. (2011) reported that participants exhibited greater lexical alignment – i.e., repeated their interlocutor's words – when interacting with a computer than with a human, with alignment increasing when the computer was presented as less capable. In a more recent visual perspective taking study, Loy & Demberg (2023) also found that participants descriptions changed when they were interacting with a computer partner, but in the opposite direction than in the earlier studies: participants behaved more egocentrically with a computer than with a human partner, and more egocentrically with a modern computer than with an older computer. The difference in the findings of these two studies may be explained by a shift in people's perception of competence of artificial agents.

Speaker identity has also been shown to have an effect on online pragmatic processing. In a reaction time study, Bergen & Grodner (2012) showed that speaker knowledge affects incremental implicature computation: when the context suggested full speaker knowledge, participants had difficulty integrating an underinformative *some*-sentence. In a visual world study, Grodner & Sedivy (2011) found that speaker reliability influenced the way scalar adjectives were interpreted. They showed that their participants interpreted scalar adjectives contrastively by default (i.e., "a tall glass" suggests that there is a short glass in the shared view), as evidenced by a speedup in visually locating the referent when there was a contrasting item present in the display, but that this contrastive interpretation did not arise when participants were told that the speaker had an impairment "that caused language and social problems".

Dynamic adaptation to specific speakers

The studies discussed above showed that listeners' pragmatic interpretation is affected by their expectations based on the speaker's properties. Like in spoken language processing, several studies in Pragmatics have investigated adaptation to speakers over time on the basis of bottom-up evidence during interaction.

Two follow-up studies to Grodner & Sedivy (2011) – Ryskin et al. (2019) and Gardner et al. (2021) – investigated whether the same effect on contrastive adjective interpretation can emerge based on bottom-up information alone, i.e., through observing the behavior of a speaker who uses scalar adjectives non-contrastively. They showed that, indeed, bottom-up information was sufficient to alter contrastive infer-

ences given enough exposure.

In the field of visual perspective taking, Hawkins et al. (2021) showed that listeners initially expected the speaker to be sufficiently informative, but when that expectation was violated, they quickly adapted their interpretations to the speaker's behavior.

It has been shown that participants learn speaker-specific use of uncertainty expressions such as *might* and *probably* for events of different likelihood (Schuster & Degen, 2020) and learn to speakers with a distinct communicative style (Loy & Demberg, 2024). In both of those cases, this adaptation ability was shown to be subject to individual variation: participants with higher memory updating scores were better adapters (Schuster et al., 2023; Loy & Demberg, 2024).

2.5 Modeling pragmatic communication

Grice's theory of language as rational action is a verbal theory, and, as is the case with verbal theories, it is not always clear what specific predictions it would make in many specific cases. This is where computational models come into play, as they can provide quantitative predictions that can be tested empirically.

In this section, we introduce probabilistic approaches to pragmatics and describe one of them, the Rational Speech Act (RSA) model (Goodman & Stuhlmüller, 2013), which we will be using for modeling in this dissertation, in detail.

2.5.1 Probabilistic pragmatics

As discussed in Section 2.1.4, construing language as a kind of rational behavior licenses application of tools used for modeling such behavior to modeling communication (Frank et al., 2016). Tools that have been applied to formalizing rational behavior in other domains include the game-theoretic notion of utility (Jäger, 2012) and Bayesian inference (Oaksford & Chater, 2009). Game theory has been widely applied to the study of behavior in economics (Samuelson, 2016); in the realm of linguistic communication, it was first applied by David Lewis in 1969 (Lewis, 2008), and, starting in the 1990s and 2000s, picked up again and applied to language more widely (e.g., Dekker & Van Rooy, 2000; Benz et al., 2005).

Studying pragmatics probabilistically has several advantages. First, as Franke & Jäger (2016) argue, probabilistic pragmatics operates at the level of reasons: it predicts certain behavior based on rational agents' motivation to maximize the probability of communicative success while expending as little effort as possible. Thus, it does not simply describe observed effects, as constraints and propositions do, but is also able to make predictions for novel scenarios. Secondly, a probability distribution is a natural way to describe uncertainty and graded beliefs, and communicators are likely to have some uncertainty about their interlocutor's goals, beliefs and alternatives available to them.

Widely used recent approaches which incorporated these probabilistic tools include the Iterated Best Response (IBR) model (Jäger, 2012) and the Rational Speech Act (RSA) model (Frank & Goodman, 2012). Both of these frameworks formalize recursive reasoning in the communication between two rational agents who select an action based on their probabilistic beliefs about the other player, using Bayesian reasoning to infer the other's beliefs based on observed evidence. In this dissertation, we will use the RSA framework (Frank & Goodman, 2012) for modeling speaker and adaptation effects. We opted for this framework due to its flexibility and ease of implementing it as a computer simulation which can generate specific predictions. We describe it in detail in the following section.

It is important to note that probabilistic approaches are not the only effective way to study Pragmatics. Franke & Jäger (2016) argue that they should be seen not as a replacement but as a complement to alternative, e.g., formal approaches. RSA and other probabilistic approaches are computational-level in Marr's (1982) terms: they leave a lot of considerations related to processing mechanisms and limitations unspecified, and they do not make predictions related to e.g. processing time or cognitive load. We will show below that there have been several attempts to extend the RSA model to incorporate resource-rational considerations (Hawkins et al., 2021) and to study variation in pragmatic reasoning ability (Franke & Degen, 2016); our approach is in inspired by those models.

Since we are using the RSA framework for modeling, the models are at the computational level. We believe that it is valuable to also specify pragmatic models at the processing level, however, this is beyond the scope of the current work. We hope that the insights obtained in the experiments presented in this dissertation, in particular, in Chapter 4, where cognitive individual differences in pragmatic reasoning are investigated, can be useful for formalizing processing models of pragmatic reasoning. In a co-authored paper not included in this dissertation, we proposed a processing-level model of reasoning in the reference game in the processing-level ACT-R framework and refer the interested reader to it for an example of how a processing model can be defined on the basis of an RSA-inspired computational-level approach (Duff et al., 2025).

2.5.2 The Rational Speech Act framework

The Rational Speech Act (RSA) framework (Frank & Goodman, 2012) is a probabilistic pragmatic modeling framework which builds on ideas from Bayesian social reasoning and game theoretic probabilistic approaches. It makes quantitative predictions about pragmatic behavior which can then be compared against experimental data.

RSA formalizes the Gricean cooperative principle and frames communication as a recursive reasoning process between a listener and a speaker who arrive at a shared meaning by taking into account alternative utterances and interpretations. It has been

successfully applied to modeling a variety of nonliteral language use phenomena, such as scalar implicatures (Goodman & Stuhlmüller, 2013), hyperbole (Kao et al., 2014b), and metaphor (Kao et al., 2014a).

In the RSA model, the listener L_N and speaker S_{N-1} are Bayesian agents who recursively reason about each other. The speaker's goal is to refer to one of the three characters by choosing an utterance. The listener then interprets the utterance probabilistically.

It is assumed that the set of possible referents as well as possible alternative utterances are shared between the speaker and the listener. This assumption is, of course, is a fairly strong one; in the next section, we will discuss a resource-rational extension to the RSA model by Hawkins et al. (2021) which allows for asymmetry of speaker and listener perspectives.

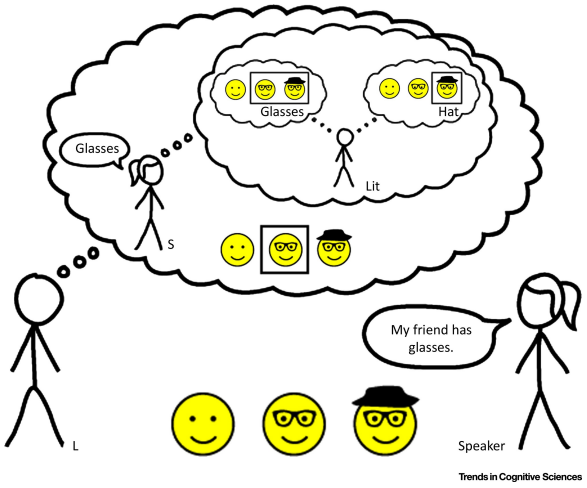


Figure 2.2: Visualisation of reasoning in the RSA framework from Goodman & Frank (2016).

We will illustrate how the model works using the example in Figure 2.2. There are three characters in the shared view: a man with no hat and no glasses, a man with a hat but no glasses, and a man with both hat and glasses. For simplicity, let's assume the following set of possible utterances: ["glasses","hat"]. Of course, in the real world many more utterances are possible, such as "glasses and no hat", but let's assume this minimal example for now. In fact, since the RSA is a rational model which maximizes communicative success while minimizing utterance cost, it is possible to set up the model in such a way that the alternative utterance "glasses and no hat" is available but almost never chosen because of high production cost, which would generate the same predictions as this minimal model.

We also assume that there are two possible referents – the character with glasses and no hat and the character with glasses and a hat – ignoring the character without glasses and hat.

Upon hearing the utterance "glasses", the pragmatic listener reasons about the alternative utterances the pragmatic speaker could have used. The pragmatic speaker

selected their utterance by reasoning about how the literal speaker would have interpreted these alternative utterances.

This recursion starts at the literal listener L_0. L_0 does not take the speaker's perspective into account and simply assigns equal probability to all referent objects o of which the utterance u is literally true. The probability distribution of the literal listener is defined as $L_0(o|u) \propto [[u]] \cdot P(o)$, where $[[u]]$ is the boolean function of the utterance's meaning, corresponding to whether it is literally true of the object o, and *P(o)* is the prior probability that the object will be referred to. *P(o)* is often defined as the object's salience which can be measured empirically (Frank & Goodman, 2012; Sikos et al., 2021a). In the literature, L_0 is often viewed not as an actual model of a conversational agent but as a dummy component of more complex RSA models that grounds the meaning of utterances and gets recursion off the ground (Degen et al., 2020). However, as we'll see later in this chapter as well as in our own experiment in Chapter 4, it has been observed that some participants make choices consistent with the predictions of L_0 and can therefore be considered literal listeners.

Assuming a uniform prior over objects $P(o)$, upon hearing the utterance "glasses", the literal listener L_0 assigns equal probability to the two referents that match the utterance:

	glasses	glasses_hat
"glasses"	0.5	0.5
"hat"	0	1

The pragmatic speaker S_1 chooses her utterances in a way that soft-maximizes the expected utility, which is defined as informativity of the utterance (probability that it will be correctly interpreted by the listener) subtracted by utterance cost:

$S_1(u|o) \propto exp(\alpha \cdot U(u; o))$, where $U = logL_0(o|u) - Cost(u)$.

The parameter α is the temperature parameter which controls the extent to which the speaker probability-matches or probability-maximizes when choosing an utterance. If α =1, the speaker's distribution mirrors the listener's probability distribution. As $\alpha \to \infty$, the speaker's probability of using the highest-utility utterance approaches 1 and she will never use any other utterance. Thus, $alpha = 1$ can be thought of as placing one's bets on utterances in proportion of how likely they are to be correctly understood, and $\alpha \to \infty$ corresponds to betting everything on the most likely option. When $\alpha = 0$, the speaker chooses any literally true utterance at random, essentially becoming a literal speaker. Often, α is fit to the data (e.g. Goodman & Stuhlmüller, 2013; Yoon et al., 2020), or a fixed value (often $\alpha = 1$) may be assumed for simplicity, for instance in simulation studies (e.g. Prasertsom et al., 2025; Qing et al., 2016).

The cost term represents the production cost of a given utterance. Utterance cost may be defined differently depending on the phenomenon, and may correspond to difficulty of retrieval, production, or other factors (Degen, 2023). Often, zero utterance costs are assumed for simplicity (e.g., Frank et al., 2016; Goodman & Stuhlmüller, 2013; Kao et al., 2014a; Qing et al., 2016), or utterance length is used as a proxy for cost (e.g. Lassiter & Goodman, 2017), sometimes with a scaling factor inferred from

the data (e.g., Yoon et al., 2020). Other factors, such as word frequency, can also be incorporated (e.g., Graf et al., 2016).

For our example, assuming $\alpha = 1$ and $Cost(u) = 0$ for both utterances for simplicity, the pragmatic speaker is twice as likely to use the utterance "hat" than "glasses" to refer to the man without the hat:

	"glasses"	"hat"
glasses	1	0
glasses_hat	0.33	0.66

Finally, the pragmatic listener L_2 reasons about the pragmatic speaker. The pragmatic listener is defined as $L_2(o|u) \propto S_1(u|o) \cdot P(o)$, where $P(o)$, like in L_0, is the prior probability of the object being referred to. It mirrors the definition of L_0, with the speaker probability distribution in place of the literal meaning. The pragmatic listener uses the asymmetry in the speaker probability distribution and concludes that "glasses" is three times more likely to refer to the character without a hat:

	glasses	glasses_hat
"glasses"	0.75	0.25
"hat"	0	1

Thus, while the literal meaning of the utterance "glasses" is compatible with multiple referents, the pragmatic listener derives an enriched meaning "glasses and no hat" by reasoning about alternatives which the speaker could have used.

It is, in fact, more common in the literature to refer to the pragmatic listener model which reasons about S_1 as L_1 and not L_2. However, we call it L_2 here to distinguish it from the listener model reasoning about the literal speaker S_0. This latter model we call L_1, following the notation from Franke & Degen (2016), who also make this distinction. We will introduce this model in the next section when we discuss modeling individual variation in RSA.

This is the basic RSA model. It has been extended to be applied to a variety of pragmatic phenomena. A common extension is the introduction of the Question Under Discussion, or QUD. The listener may have uncertainty about what it is that the speaker is trying to communicate, e.g. the literal meaning or, for instance, the speaker's affect (Kao et al., 2014b; Kao & Goodman, 2015).

While the classic RSA model includes only two levels of recursion, additional levels can in principle be defined. However, deeper recursive models are unlikely to be cognitively plausible given limitations in human working memory capacity (Yuan et al., 2018). Although Frank et al. (2016) found that increased recursion depth correlated with improved fits to human behavior, the gains beyond two levels of recursion were minimal. Moreover, Frank et al. (2016) showed that recursion depth trades off with the temperature parameter α, such that increasing either leads to higher performance, thereby limiting the identifiability of the specific contribution of recursion depth. Consistent with this view, Franke & Degen (2016) allowed recursion

depth to vary across individuals and found that most participants' behavior was best explained by a model assuming a recursion depth of 1, with only a small subset of participants exhibiting evidence for a recursion depth of 2.

In a simulation study, Yuan et al. (2018) showed that, although accuracy continues to improve with additional levels of recursion, the largest gains occur within the first few levels—especially in large utterance and meaning spaces. This suggests that the cognitively plausible shallow levels of recursion often suffice to achieve near-optimal performance.

Experimental evidence for the RSA and its limitations

The original influential study by Frank & Goodman, which introduced the RSA (Frank & Goodman, 2012), used a one-shot design such that every participant only saw one reference game item. They found that the RSA model predictions provided a very good fit to participants' judgements (r=0.99 with all trials and r=0.87 when ratings of 0 and 100 were removed), which was taken to suggest that participants engaged in ad-hoc pragmatic reasoning despite the artificial setting.

However, this observation later faced some criticism. Sikos et al. (2021a) observed that among the context used by Frank & Goodman (2012), only a small subset (about 17%) involved pragmatic reasoning. They asked whether pragmatic reasoning was necessary for solving the task, or whether a simpler model, relying just on the literal meaning of signals and their salience, could explain human responses. The authors conducted three follow-up one-shot experiments with a larger set of referential contexts and showed that a baseline literal salience-driven model predicted human performance equally well and sometimes even outperformed the full RSA model. While the literal listener model is often not viewed as a full cognitive model but instead as a dummy component to get recursion off the ground in RSA (Degen et al., 2020), Sikos et al. (2021a) suggested that it could be viewed as a cognitive model in its own right. They argued that the RSA is a computational-level model which defines an ideal communicator and that factors like engagement and resource bounds may influence what degree of reasoning is actually realized. We will return to this point in the next section where we discuss approaches to modeling resource rationality in RSA.

While the original Frank & Goodman study used a one-shot design, later studies which used the RSA to model pragmatic reasoning in reference games tended to use multi-shot designs (e.g., Franke & Degen, 2016; Achimova et al., 2023; Degen et al., 2013). It does not seem implausible that participants behave more pragmatically if they have more time to grasp this artificial task. A significant effect of trial number in multi-shot designs could be viewed as evidence of learning and becoming more pragmatic over time. Franke & Degen (2016) conducted longer experiments with two implicature types of different complexity. They did not observe a significant effect of trial number, which was crucial for their analysis where they classified their partici-

pants into three reasoner types of different complexity, which relied on a participant staying within the same type for the duration of the experiment. In our follow-up studies, however, we repeatedly find some evidence of a small learning effect (Ch 3-5), suggesting that some adaptation to the task does occur. One possible reason for the difference in findings is that the study by Franke & Degen (2016) used different stimuli.

The RSA model assumes a fixed agent role, i.e., each agent is either a speaker or a listener. In real-life communication, however, we play both roles, often at the same time. Sikos et al. (2021b) investigated whether participants' performance on the listener task improved after doing the speaker task first, and vice versa. They found that participants' performance on the listener task improved when they played the role of the speaker first, suggesting enhanced ability to reason about the other perspective after playing the interlocutor's role.

These aspects of the RSA framework are important to keep in mind as we aim to gain more understanding of its validity as a cognitive model and ways to extend it to improve its cognitive plausibility. Despite the limitations discussed above, a large body of work has shown that the RSA has provided a good fit to participants' performance in studies in various domains of pragmatics (e.g., politeness (Yoon et al., 2016, 2020), figurative language (Kao et al., 2014a,b), and projection (Qing et al., 2016)) at the population level.

In this dissertation, we am interested in investigating modeling variation and adaptation in pragmatic inference, as opposed to static population-level effects, which the RSA has been primarily used for in previous work. In the next section, we discuss which components of the RSA model can be used for this purpose and review extensions that have been proposed to model individual differences and imperfect rationality.

2.5.3 Beyond static population-level effects: Modeling individual differences, bounded rationality and learning

In previous work, RSA has mostly been used for modeling static population-level pragmatic effects. Previous studies often elicited population-level priors and modeled average effects (e.g., Yoon et al., 2020; Goodman & Stuhlmüller, 2013; Kao & Goodman, 2015). However, the focus of this thesis is on cognitive individual differences and adaptation to the speaker. In this section, we will review the limited existing work on extending the RSA for modeling those effects.

Modeling individual differences

Franke & Degen (2016) proposed that individual differences in performance on pragmatic reference games are linked to depth of reasoning, corresponding to *recursion depth* in RSA. While by default RSA assumes that comprehenders are L_2 listeners and

speakers are S_1 speakers. Franke & Degen (2016) investigated whether this assumption holds (1) at the population level and (2) at the level of individual participants. They had two separate groups of participants play reference games as listeners and speakers respectively. The reference games in this study used stimuli of two different difficulty levels, which Franke & Degen (2016) called simple and complex implicatures.

Crucially, probabilistic models of different recursion depth make distinct predictions for these two stimuli types:

	type		can solve?	
level	listener	speaker	simple	complex
0	literal	literal	✗	✗
1	exhaustive	**Gricean**	✓	✗
2	**Gricean**	hyper-pragmatic	✓	✓

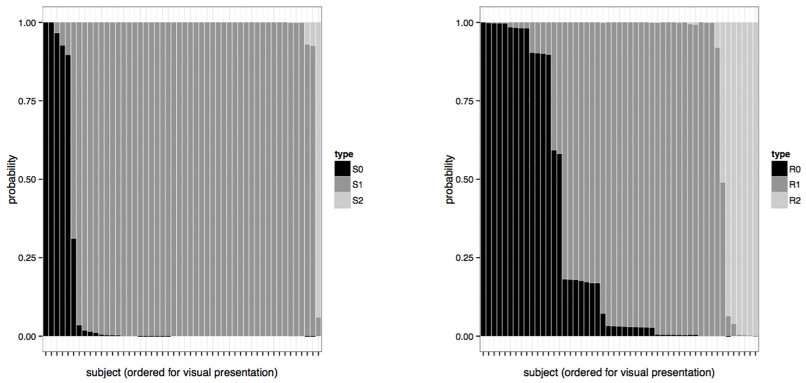
Table 2.1: Reasoning types in Franke & Degen (2016).

This way, participants can be categorized as being 0, 1 or 2-level listeners and speakers based on their performance. Franke & Degen (2016) fit a hierarchical Bayesian model to their participants' data. They found that while at the population level, an L_2 listener model and an S_1 speaker model provided a good fit, the assumption that all individual participants were L_2-level listeners did not hold: a model which allowed each participant to have one of the three recursion depths provided a better fit to the individual-level data. The same did not hold for the production experiment, however: most participants' data was best explained by the S_1 model, as is typically assumed in RSA. Figure 2.3 shows the marginal posterior of the hierarchical Bayesian model for individual participants for the production and comprehension experiments.

Franke & Degen (2016)'s approach shows that individual differences exist and can be captured by allowing reasoning depth to vary. However, it is agnostic to what causes this individual variation. In Chapter 4, we follow up on this question and, in addition to the listener reference game task, collect a battery of cognitive individual differences. We find that performance on the reference game is predicted by logical reasoning ability (pattern recognition and ability to override the first intuitive response) and Theory of Mind.

Modeling bounded rationality

The default RSA model is a computational-level model, and it is quite minimal in terms of machinery for modeling cognitive limitations: it has the temperature parameter α and utterance cost in the speaker utility term which could be defined as word frequency or length. By default, it assumes that interlocutors always consider all relevant alternatives during interaction.



(a) speaker **(b)** listener

Figure 2.3: Model fit to individual participants' performance for the production (left) and comprehension (right) experiments in Franke & Degen (2016).

There have been a couple of recent attempts of relating α to cognitive resources by inferring it from data for different populations. Bohn et al. (2022) fit α as a function of the participants' age in their model of novel word learning, and van Tiel et al. (2022) divided their participants into two groups based on their scores on the AQ questionnaire and fit a model to each group's data on uncertainty expressions and observed that the group with the higher AQ score, suggesting higher autistic traits, was associated with a lower α.

Since full Bayesian inference over all possible alternatives every time would be highly costly, implementations of RSA have been proposed where inference is amortized and direct optimization with an internal speaker model is performed without recursive reasoning (White et al., 2020) and where full inferences is replaced by a sampling-based approach (Scontras et al., 2021). In a similar vein, Yuan et al. (2018) showed in simulations that, when both the speaker and the listener sample only a small set of the most likely utterances and meanings, discarding the rest during computation, smaller samples can yield higher expected accuracy, as the speaker and the listener quickly converge on a shared mapping.

Hawkins et al. (2021) proposes a resource-rational extension to the RSA model in order to capture the intuition that perspective-taking is cognitively costly and that this cost is divided between the speaker and the listener. This is one of the very few attempts to extend RSA from a purely rational model to a model which captures some aspect of processing. The authors situate their model in the context of visual perspective-taking in a director-matcher paradigm, where the director gives the matcher instructions referring to an object but some of the objects are occluded. It has been observed in the literature that this kind of perspective-taking is not automatic: people often first reach for an object which is hidden from shared view (Lin et al., 2010).

Hawkins et al. (2021) posit that there are two perspective-taking extremes, the "ideal" perspective-taker who always takes their partner's perspective (i.e., pays per-

fect attention to which objects are occluded) and the fully egocentric agent who only pays attention to their own view. These two models are then combined in a probabilistic weighting model (Heller et al., 2016) with a mixture weight w which determines the degree of perspective-taking. For instance, for the speaker model:

$$U_{S_1}^{mix}(u; o, C, w_S) = w_s \cdot U_{S_1}^{asym}(u; o, C) + (1 - w_S) \cdot U_{S_1}^{ego}(u; o, C)$$

When $w_s = 0$, the speaker is completely egocentric and does not consider what is visible from the partner's perspective; when $w_s = 1$, the speaker completely takes the listener's perspective into account. The listener model is also equipped with a mixture weight w_L. The two agents are aware that the other is a mixture model and have uncertainty about which mixture weight their partner is using. Therefore, they will marginalize over possible mixture weights when choosing an action. As a result, the two models are inter-dependent and the actions of each agent are going to change depending on the mixing weight they expect their partner to be using.

For instance, the listener will marginalize over the listener's mixing weights:

$$P_{L_1}(o|u, C, w_L) \propto \int_{w_S} P(w_S) \cdot P_{L_1}^{mix}(o|u, C, w_L, w_S) dw_S$$

The authors conduct a resource-rational analysis (Lieder & Griffiths, 2020) of these models, where they find the optimal mixture weight. They assume that perspective-taking is costly and, for simplicity, that the cost is linear in the degree of perspective-taking. They derive the optimal weights w_L^* and w_S^* for which the optimal tradeoff between the benefit of perspective taking (i.e., communicative accuracy) and its cost are maximized.

Modeling adaptation

Adaptation through repeated interaction is traditionally modeled in terms of *Bayesian belief updating*. The agent updates their knowledge about the partner through repeated observations of their behavior:

In the general case, this look as follows:

$$P(\Theta|D) \propto P(D|\Theta) \cdot P(\Theta) = P(\Theta) \cdot \prod_i P_{S_1}(u_i|o_i, \Theta)$$

The listener has some prior beliefs about the speaker's property Θ (some examples: the speaker's willingness to take the listener's perspective, as in Hawkins et al. (2021), their lexicon, as in Hawkins et al. (2017), the speaker's probability thresholds for using different uncertainty expressions, as in Schuster & Degen (2020)). They then update their prior beliefs based on a sequence of observations, consisting of the speaker's utterance and the intended meaning.

The assumption in the literature (Hawkins et al., 2017, 2021; Schuster & Degen, 2020; Roettger & Franke, 2019; Kleinschmidt & Jaeger, 2015) so far has been that agents learn by reasoning about the other agent: i.e., that updating of Θ from the listener's point of view is based on the speaker model. However, we will show in Chapter 6 that in case of adaptation to speaker rationality in a reference game, not the speaker model but the listener model appears to be used for learning: our experimental findings suggest that, when perspective-taking is effortful, listeners may make inferences about the speaker's rationality based on trials which are ambiguous from the listener's perspective, not the speaker's perspective.

2.6 RSA predictions for the reference game

The majority of this thesis – the experiments described in Chapters 3-6 – uses variations of the reference game paradigm, which was introduced in Section 2.2.3. In these experiments, the reference game setup is based on the simple and complex implicatures from Franke & Degen (2016) (henceforth F&D), whose study was briefly described above in Section 2.5. Here, we will first describe the two implicature conditions in detail, and then discuss the predictions that probabilistic pragmatic models of different recursion depths make for them. This section is based on the Background section of Mayn & Demberg (2026).

2.6.1 Reference game setup: Simple and complex implicatures from Franke & Degen (2016)

The two implicature trial types are shown in Figure 2.4. F&D's study, as well as the earlier studies it is based on (Degen et al., 2012, 2013), use creatures (monsters and robots) and accessories (hats and scarves) as stimuli. However, we show in Chapter 3 that these stimuli are susceptible to biases which may inflate participants' performance. Therefore, in the experiments in Chapters 4-6 we use abstract stimuli, shapes and colors, instead. Therefore, in this section, we will use these abstract stimuli to describe the two implicature conditions; underlyingly, however, they are exactly the same as the two implicature conditions in the comprehension experiment in F&D.

In this setup, on each experimental trial, participants see three objects in the display. The objects differ along two dimensions – shapes (square, circle and triangle) and color (red, green and blue). Participants also see a message (a non-filled shape or a paint tube of a certain color) that they are told was sent by the previous participant in order to get them to pick out one of the three objects. Participants are also told that only two of the three shapes and two of the three colors were available to the previous participant to send as messages (square and blue were not available; we'll call them *inexpressible features*). The available messages are always displayed at the

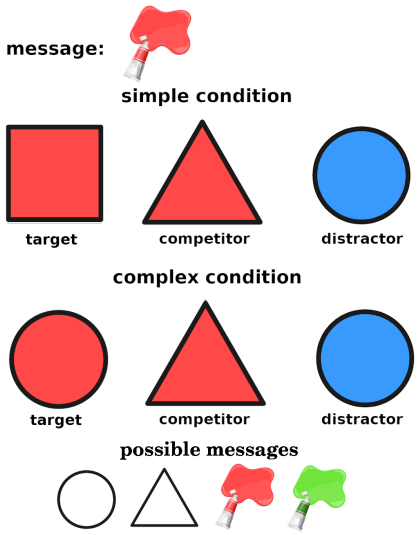


Figure 2.4: Simple and complex implicature conditions from Franke & Degen (2016).

bottom of the screen.

There are two types of implicature trials, simple and complex ones. On *simple* implicature trials (top panel of Figure 2.4), the target contains the message (which is one of the four available messages) and the inexpressible feature along the other dimension. For instance, if the message is the color red, the target would be a red square, since "square" is not an available message. The competitor is composed of the message and an expressible feature along the other dimension. Continuing with our example, the competitor could be a red triangle. Finally, the distractor is composed of any two features not present in the target or competitor. In our example, it could be a blue circle.

On *complex* implicature trials (bottom panel of Figure 2.4), the target contains the message along with an expressible feature along the other dimension. So if the message is the color red, the target could be a red circle. The competitor consists of the message and the other expressible feature along the other dimension. In this example, the competitor would be a red triangle. The distractor combines the other target feature not denoted by the sampled message with the inexpressible feature along the other dimension. In our example, the distractor would be a blue circle.

As we mentioned in Section 2.5 when we introduced F&D's study, the complex implicatures are predicted to be more difficult than the simple ones since a formal probabilistic model can be defined which can solve the simple but not the complex implicatures. We will explain this in more detail below. Both F&D's experimental results and the results of our experiments described in Chapters 3 and 4 with abstract stimuli show that participants' performance on the complex condition is indeed lower than on the simple condition.

In addition, all experiments feature completely unambiguous and completely am-

biguous filler trials. The exact configuration of fillers differs by experiment and is described in each chapter.

2.6.2 Models of reasoning complexity in the reference game

As mentioned above, F&D show that formal probabilistic models make distinct predictions with regards to simple and complex implicatures in the reference game.

As discussed in detail in Section 2.5, in the Rational Speech Act model (Frank & Goodman, 2012) and the closely related Iterated Best Response model (Franke, 2009), listener L_N and speaker S_{N-1} are Bayesian agents who recursively reason about each other and the alternative possible utterances by the speaker and interpretations by the listener. Models of different recursion depths make different predictions for performance on the reference game.

The simplest model is the *literal listener* L_0 who interprets the speaker's message m literally and assigns an equal probability to all objects o of which the message is literally true. The probability distribution of the literal listener is defined as $L_0(o|m) \propto [[m]] \cdot P(o)$, where $[[m]]$ is the boolean function of the message's meaning, corresponding to whether it is literally true of the object o, and *P(o)* is the prior probability that the object will be referred to. *P(o)* is often defined as the object's salience. L_0 will only be able to correctly solve unambiguous trials. On both simple and complex critical trials, L_0 will assign equal probability to the target and the competitor since they both match the message.

The next model in terms of complexity is L_1, which F&D termed *exhaustive listener* since its predictions closely align with an exhaustive interpretation in pragmatic inferences, which is a formalization of pragmatic inference from theoretical linguistics. L_1 reasons about a literal speaker S_0, who is equally likely select any true utterance to refer to an object. L_1 is defined as $L_1(o|m, \alpha_{L_1}) \propto exp(\alpha_{L_1} \cdot S_0(m|o; \alpha_{S_0} \rightarrow \infty)$, where α_{S_0} and α_{L_1} are the speaker's and the listener's temperature, or rationality, parameter respectively, which controls the degree to which the agent maximizes their utility. L_1 is able to solve simple but not complex trials. Let's show why that is the case. On simple trials, S_0 has only one way of referring to the target, which is the uttered message, because the other feature of the target is inexpressible, and two ways of referring to the competitor. Therefore, L_0 will reason that the message on simple trials is twice as likely to refer to the target, since S_0 will use it 100% of the time to refer to the target and only 50% of the time to refer to the competitor (the other 50% of the time, S_0 will use the other available message). On the complex condition, L_1 will be at chance since both the target and the competitor have two messages which can be used to refer to them.

Finally, the most complex listener model is the classic pragmatic listener model L_2, who reasons about the pragmatic speaker model S_1, which is most commonly used in RSA modeling literature. It is, in fact, more common in the literature to refer to the pragmatic listener model which reasons about S_1 as L_1 and not L_2. However,

we call it L_2 here, following F&D to distinguish it from L_1 listener model reasoning about a literal speaker S_0. L_2 reasons about the pragmatic speaker S_1, who, in turn, reasons about the literal listener L_0. S_1 is defined as $S_1(u|o, \alpha_{S_1}) \propto exp(\alpha_{S_1} \cdot (\log L_0(o|m) - Cost(m))$. The speaker seeks to maximize their informativity, that is, the probability that the literal listener L_0 will correctly identify the intended referent, while minimizing utterance cost. L_2 is then defined as $L_2(o|m) \propto S_1(u|m; \alpha_{S_1 \to \infty}) \cdot P(o)$, where $P(o)$, like in L_0, is the prior probability of the object being referred to. It mirrors the definition of L_0, with the speaker probability distribution instead of the literal meaning. L_2 is powerful enough to solve both implicature types. It will solve the complex condition by reasoning that S_1 will prefer to use the uttered message to refer to the target since the competitor has an unambiguous message ("triangle" in the example in the bottom panel of Figure 2.4) that could be used to refer to it.

As discussed in Section 2.5.3, F&D fit a hierarchical Bayesian model to their experimental data from the reference game, whereby one of the three listener types (L_0, L_1 or L_2) was assigned to each individual participant, and showed that this individual-level model provided a better fit to the data than a model that assumed that all participants had the same reasoning type. In Chapter 4, we will investigate which cognitive factors underlie this variability.

2.7 Summary and Outlook

In this chapter, we provided an overview of the foundational pragmatic theories, as well as of experimental work indicating that cognitive and situational speaker-related factors play a role in pragmatic inference. This section will discuss how the findings described in the previous sections motivate the research in this dissertation.

While classic theories of pragmatics – with the exception of Relevance Theory (Sperber & Wilson, 1986) – assume perfect rationality, experimental work has demonstrated vast individual variability in pragmatic performance and a connection of that variability to cognitive individual differences such as working memory capacity and Theory of Mind, in line with acounts of bounded rationality. The individual differences perspective on pragmatics is relatively new and growing, with mixed evidence of the effects of individual differences for various phenomena. This thesis aims to extend our understanding by investigating whether differences in logical reasoning, working memory and Theory of Mind can explain the observed variability in pragmatic reasoning in reference games (Chapter 4).

Another emerging line of research highlights the role of speaker identity — including factors such as language proficiency and the speaker's persona — in shaping the listener's inferences. In this thesis, we build on this work by investigating comprehenders' sensitivity to the *speaker's perceived pragmatic competence* and examine how it affects the comprehenders' inferences. Chapter 5 investigates how comprehenders' beliefs about the pragmatic competence of differen speaker types – adults, 4-year-old

children, and artificial agents – influences the derived inferences. Chapters 6 and 7 investigate adaptation to the speaker's pragmatic competence over the course of the interaction across two paradigms. We propose models within the RSA framework of how comprehenders' beliefs about a speaker's pragmatic competence influence interpretation and drive pragmatic adaptation.

To sum up, this thesis contributes to theories of pragmatics by investigating its dynamic aspects, including cognitive individual differences in pragmatic reasoning and comprehenders' adaptation to variation in speakers' pragmatic competence. We argue that variability in pragmatics is not noise or deviation from an ideal standard, but a central feature that comprehensive theories should explain. Using a combination of experimental and computational approaches, this thesis advances our understanding of the factors underlying this variability and of the ways in which interlocutors accommodate it in communication.

Chapter 3

Do reference games measure pragmatic reasoning? Uncovering biases through stimulus manipulation and strategy elicitation

In this dissertation, we investigate variability in pragmatic reasoning caused by cognitive as well as situational factors. For all experiments in this dissertation, with the exception of the final chapter, we use the reference game paradigm, since it is a simple yet flexible paradigm in which the space of alternative meanings and utterances is clearly defined. Besides, the current chapter and the following chapter of this dissertation are based on the work by Franke & Degen (2016) (henceforth F&D), who find evidence of substantial individual variability in the reference game, as discussed in the theoretical background (Section 2.5.3).

However, before we can proceed to use the reference game paradigm for investigating variability of pragmatic reasoning, the question of the validity of this paradigm for measuring pragmatic reasoning presents itself. As discussed in the theoretical background (Section 2.5), formal probabilistic models, such as the Rational Speech Act model, are widely used for formalizing the reasoning involved in various pragmatic phenomena, and when a model achieves good fit to experimental data, that is interpreted as evidence that it successfully captures some of the underlying processes. However, it is possible that participants' performance reflects not only successful pragmatic reasoning but also artifacts of the experimental design. Sikos et al. (2021a) revisited the influential RSA study by Frank & Goodman (2012) and found only modest evidence of people engaging in the assumed pragmatic reasoning. Thus, the authors argued that the good fit of Frank & Goodman (2012)'s model to empirical data was largely due to a combination of non-pragmatic factors. To ensure the validity of drawn conclusions, it is important that the experimental design is as free

from bias as possible.

This chapter investigates the validity of the reference game as a measure of pragmatic reasoning through careful manipulation of stimuli and elicitation of participants' strategies. We find that the original stimuli used by F&D resulted in inflated performance due to biases in experimental stimuli unrelated to pragmatic reasoning, but that these biases are greatly reduced when using abstract shapes. We thus find that the reference game paradigm with abstract stimuli is suitable for investigating variability in pragmatic reasoning and use it for experiments in subsequent chapters of this thesis.

Also, this chapter addresses Research Goal 1 formulated in the introduction by proposing a method for eliciting strategies that participants use for solving this task, which can be used to gain additional insights into the reasoning participants actually perform. We develop an annotation scheme for the elicited strategies and show that, despite warranted concerns about self-reports about one's own cognition which have been raised in previous work, strategy elicitation can help identify participants who misunderstood the task as well as those who used alternative reasoning strategies that are not identifiable from performance alone.

This chapter is based on, and in parts identical to, the following publication:

- **Mayn, A.**, & Demberg, V. (2023). High performance on a pragmatic task may not be the result of successful reasoning: On the importance of eliciting participants' reasoning strategies. *Open Mind*, 7, 156-178.

Data availability

All data and analysis scripts for the experiments presented in this chapter are available at `https://github.com/sashamayn/refgame_stimuli_methods`.

3.1 Introduction

Formal probabilistic models, such as the Rational Speech Act model (RSA, Frank & Goodman (2012)), are widely used to formalize the reasoning involved in various pragmatic phenomena, such as scalar implicatures (Goodman & Stuhlmüller, 2013), hyperbole (Kao et al., 2014b), and irony (Kao & Goodman, 2015). RSA assumes that the speaker and the listener are cooperative and reason recursively about each other to arrive at a shared interpretation. These models are then evaluated by being fit to experimental data, and if they achieve a close fit, that is interpreted as evidence that the model successfully captures some of the underlying processes.

Yet how can we be sure that participants' high performance on the task is indeed the result of successful reasoning, and not some other factor related to the experimental setup? Sikos et al. (2021a) revisited the influential RSA study by Frank &

Goodman (2012) and found only modest evidence of people engaging in the assumed pragmatic reasoning. They showed that a simple literal listener model, which is governed by the salience prior over objects, provided an equally good fit to the data as the full pragmatic listener model. The authors therefore argued that the good fit of the RSA model to empirical data was largely due to a combination of non-pragmatic factors. Neglecting to carefully investigate whether the experimental design contains biases hence may lead to making unwarranted conclusions about the phenomena we study.

In a study in the reference game paradigm, F&D asked their participants to identify the referent of an ambiguous message and observed that there is quite a lot of individual variability in performance, and that the data is better captured by a model which assumes that each participant has their own reasoning type, corresponding to predictions of three probabilistic models of different degrees of complexity, than by a population-level model that assumes that all participants belong to the same reasoning type. Therefore, better performance on the task is interpreted as evidence of participants successfully employing higher complexity of reasoning.

In this chapter, we conduct a series of experiments based on the task used by F&D (originally introduced in Degen et al., 2012) where we carefully manipulate the properties of the stimuli and also elicit participants' reasoning strategies. We show that certain biases are present in the task which allow participants to arrive at the correct answer without engaging in the assumed pragmatic reasoning. We then design a version of stimuli aimed at mitigating the identified biases and repeat the experiment, obtaining a somewhat smaller effect size, and, importantly, more reliable individual-level results. We argue that probing an experimental design for biases is crucial for ensuring that we can draw meaningful conclusions, and strategy elicitation is a simple and efficient way of doing so.

3.2 Background

3.2.1 Reference game

The task is the reference game, which is the comprehension experiment in F&D and the Experiment 1 in Degen et al. (2012). The setup of simple and complex implicatures is described in detail in the theoretical background (Section 2.6.2), with the only difference that F&D use creatures and accessories as opposed to abstract stimuli. I summarize F&D's setup here given that their stimuli differ from the ones used in subsequent chapters, but please refer to Section 2.6.2 for a detailed description of the two reference game conditions and the corresponding predictions of Bayesian pragmatic models.

On each trial, participants see three objects, each of which is a creature wearing an accessory. There are three possible creatures (green monster, purple monster, and

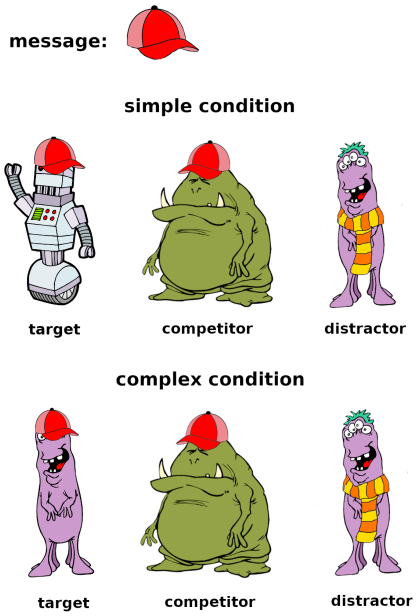


Figure 3.1: Example of a simple and a complex critical trial in Experiment 1.

robot) and three possible accessories (red hat, blue hat, and scarf). We can think of the objects on the screen, therefore, as varying across two feature dimensions – the creature feature and the accessory feature. Participants also see a message that they are told was sent by the previous participant. The message is always either a creature (without an accessory) or an accessory. Importantly, participants are told that not all creatures and accessories are available as messages: there are no messages *scarf* or *robot*, hence these are so-called *inexpressible* features. Participants' task is then to pick the creature they believe the previous participant was referring to.

The task consists of 66 experimental trials, of which 24 are critical and 42 are fillers. Each trial display consists of the target (correct answer), competitor and distractor, presented in random order.

On the critical trials, the message is ambiguous. Half of the critical trials are *simple implicature trials*. An example of a simple implicature trial is presented in the top panel of Figure 3.1. On those trials, the target contained the sampled message and the inexpressible feature along the other dimension. In our example, the sampled message was the red hat, so the target is a robot with a red hat. The competitor was composed of the message and an expressible feature along the other dimension. In our example, it is a green monster with a red hat. Finally, the distractor was composed of any two features not present in the target or competitor. In this example, it is a purple monster with a scarf. To draw the implicature, one can reason that if the speaker had meant to refer to the competitor (green monster), they could have used the unambiguous message *green monster*, whereas there is no way to refer to the target (robot) unambiguously since *robot* is not an available message, hence the red hat must be referring to the robot.

On complex implicature trials (bottom panel of Figure 3.1), the target contained the message along with an expressible feature along the other dimension. So if the sampled message was the red hat, the target could be a purple monster with a red hat. The competitor contained the message and the other expressible feature along the other dimension. In this example, the competitor would be a green monster with a red hat. The distractor combined the other target feature not denoted by the sampled message with the inexpressible feature along the other dimension. In our example, the distractor would be a purple monster with a scarf. To draw the implicature, one needs to reason that if the speaker had wanted to refer to the competitor (green monster), they could have used the unambiguous *green monster*. In contrast, there is no unambiguous way of referring to the target since its other feature (purple monster) is shared with the distractor, hence the red hat must be referring to the purple monster with the red hat.

Complex implicatures are predicted to be more difficult than simple ones since a probabilistic Bayesian model can be defined which can solve the simple but not the complex implicatures. The literal listener model L_0 can solve neither implicature type; an exhaustive listener L_1 who reasons about a literal speaker S_0 is able to solve simple but not complex implicatures, and the Gricean listener L_2 is able to solve both implicature types. Please refer to the theoretical background (Section 2.6.2) for a detailed explanation of these predictions.

Of the 42 fillers, 33 are completely unambiguous and 9 completely ambiguous. On completely unambiguous trials, only the target has the feature expressed by the message. For example, the message is a purple monster and there's only one purple monster on the screen. Those trials are used as an attention check. On completely ambiguous trials, the target and the competitor are identical. For example, the message is a purple monster and there are two identical purple monsters wearing a blue hat and a robot with a red hat on the screen. Since there is no way of knowing which of the two identical creatures is the target, performance on the ambiguous trials constitutes a random baseline.

3.2.2 Annotation of reasoning strategies

Since we were interested in how the participants solve the reference game, in addition to collecting participants responses to the experimental stimuli, in each of our experiments, we also elicited their reasoning strategies. After completing the main experiment, participants saw one simple and one complex item again, randomly selected and presented in randomized order. When they clicked on one of the creatures, a red box appeared around it, along with the question "Why did you make that choice?" and a textbox. This was done as a probe into participants' reasoning.

All responses were annotated by two annotators, one of whom was blind to the purpose of the experiment. All disagreements were resolved jointly: the two annotators met and each made a case for their choice of tag. If one of the annotators

successfully convinced the other, the tag was changed accordingly; otherwise, the tag *unclear* was assigned to the item and it was removed from future analysis.

Annotation scheme

Participants' responses were assigned one of five tags. Example responses for all annotation tags are presented in Table 3.1. The complete annotation scheme with more examples for each tag can be found in Appendix A.

Annotation tag	Subtag	Example explanation
correct_reasoning		The speaker could have used the message *green monster* if they meant to refer to the other creature with the red hat, but they didn't, which makes me think that they wanted me to choose the robot.
guess		It's a 50/50 guess between the two monsters.
other_reason	visual_resemblance	The robot's hat is pointing to the right, like the message.
other_reason	odd_one_out	It is the only creature not wearing a hat.
other_reason	salience	It is the biggest picture with a red cap.
other_reason	preference	Robots are cool.
other_reason	other	Non-sound reasoning: If the speaker had meant to communicate one of the other creatures, they would have used the *purple monster* message.
misunderstood_instr		The speaker could not choose the robot so it must be the green monster.
unclear		The robot

Table 3.1: Examples of each annotation tag.

The category *correct_reasoning* was assigned to hypothetical reasoning about alternatives of the kind described by the RSA model. An example for the top panel of Figure 3.1 would be "The speaker could have used the message *green monster* if they meant to refer to the other creature with the red hat, but they didn't, which makes me think that they wanted me to choose the robot".

Random responses were assigned to the category *guess*. Some participants indicated screen location as an explanation (often choosing the middle option) or referred to their responses on previous trials (e.g. "I have chosen a purple monster a lot in the past so now I will choose the robot"). Those were categorized as guessing since participants indicated that they had no way of differentiating between the target and the competitor and then used some other superficial criterion to break the tie.

Cases where participants described a reason for their choice which was something other than hypothetical reasoning about alternatives were labeled *other_ reason*. Within that category, we further assigned one of the following five subcategories to the explanations. The tag *visual_ resemblance* was assigned when the participant stated that they selected the creature which they found to be "visually the most similar" to the message. Some of the explanations included a justification revealing the nature of the similarity the participant picked up on. For instance, a common reason for selecting the target in the top panel of Figure 3.1 was that the robot's hat is facing in the same direction as the message, whereas the competitor's hat is facing the other way. *odd_ one_ out* was assigned when the participant reported selecting a creature that stands out because it is the only one that has a certain feature. For instance, a participant might select the distractor in the top panel of Figure 3.1 as it is the only creature with a scarf and without a hat. The tag *salience* was assigned when the participant reported selecting a certain creature because it stood out to them. An example response for the top panel of Figure 3.1 which was assigned the *salience* tag is selecting the competitor "Because it is the biggest picture with a red cap". The subcategory *salience* is quite closely related to *odd_ one_ out* since both of those correspond to selecting a response because it stands out. The difference is that in the *odd_ one_ out* case, a selection is made based on a creature being different from the other two based on a feature (exact match or mismatch), as opposed to being bigger or brighter, which is a relative difference. Often *odd_ one_ out* corresponded to adapting an inverse interpretation of the speaker's message, e.g., using the red hat message to refer to the only creature that *is not* wearing a red hat. The tag *preference* was assigned when the explanation constituted a personal preference like "Robots are cool".

One could argue that *preference* should be categorized as guessing since the decision did not involve reasoning about why the speaker sent a given message but instead broke the tie using the participant's own preference. Our motivation was the following: guessing is by its nature not consistent, so we assigned an explanation to the *guess* category if, when presented with the same trial again, possibly with different randomization, the participant could have made a different selection. In the case of selecting a creature because it's in the middle, since the screen location is randomized, a different creature could have been in the middle. In the case of a personal preference for robots, however, we assume that the participant would have consistently selected the robot in this situation if presented with this trial again.

Responses that involved a strategy that did not fall into one of the aforementioned

categories were assigned to the subcategory *other*. An example would be a participant attempting to reason but their reasoning not being sound. For the bottom panel of Figure 3.1, incorrect reasoning could be selecting the competitor since "if [the speaker] had meant to communicate one of the other creatures, they would have used the purple monster message".

Sometimes participants' answers revealed that they misunderstood the instructions and took the fact that a feature is inexpressible (i.e., that it cannot be referred to directly) to mean that a creature that has the inexpressible feature could not be referred to at all. An example of an answer in that category for the top panel of Figure 3.1 would be "The speaker could not choose the robot so it must be the green monster". Such responses were labeled *misunderstood_instructions*.

Answers where it was unclear what the participant meant, e.g., very brief answers just stating their selection ("The robot") were labeled *unclear* and excluded from further analysis. We also excluded items where in their explanation the participant changed their mind as to their answer since that indicates that their explanation does not reflect their original reasoning strategy.

We investigated how internally consistent participants were when they saw one of the experimental items again during the strategy elicitation part of the experiment. We'd expect people who used *correct_reasoning* to be consistent, while *guess*ers may have picked the target the first time and the competitor the second time, or the other way around. That is indeed what we find. In the simple condition, collapsing across experiments since the pattern is very similar for all experiments, 78 (88.6%) participants who used the *correct_reasoning* strategy are consistent, selecting the target both times, and the remaining 10 (11.4%) selected the competitor the first time and the target the second time around. As predicted, *guess*ers are more mixed: 17 (32%) participants selected the target both times (presumably by chance), 15 (28.3%) participants selected the competitor both times, and the remaining 21 (39.6%) were inconsistent. Those breakdowns of responses were significantly different, with *correct_reason*ers consistently selecting the target both times more often ($\chi^2(2, N = 141) = 52.63, p < 0.0001$).

Among *other_reason* responses, it is notable that 33 (86.8%) participants who relied on *visual_resemblance* picked the target, 3 (7.9%) picked the competitor both times, and 2 (5.3%) were inconsistent. Other *other_reason* subcategories were pretty mixed.

In the complex condition, 44 (84.6%) *correct_reason*ers picked the target both times and 8 (15.4%) were inconsistent and picked the target the second time. *guess*ers, on the other hand, are pretty evenly split, with 34 participants (37.8%) selecting the target both times, 29 (32.2%) of participants selecting the competitor both times, and 27 (30%) inconsistent participants. The difference between *correct_reason*ers and *guess*ers is again significant ($\chi^2(2, N = 142) = 32, 77, p < 0.0001$).

It is worth asking how reliable strategy explanations obtained through introspection are. Post-hoc explanations have been criticized for not always accurately re-

flecting the reasoning in the moment (Cushman, 2020). Nisbett & Wilson (1977) argued that humans lack access to introspective processes in the moment and therefore post-hoc explanations are rationalizations made after the fact. As an example, in the Wason selection task, people appear to be influenced by what has been termed *matching bias*: when they are asked to indicate which cards need to be turned over to verify a logical rule (e.g., "If a card has a D on one side, it has a 3 on the other"), they are more like to select a card if it was mentioned before (a D or a 3 in this example). However, subjects appear to not be conscious of this bias, as it never comes up in post-hoc explanations (Evans, 2019).

We argue that, despite these limitations, post-hoc explanations have the potential to yield important insights. In our experiments below, we show that strategy elicitation can help reveal certain biases in the task, which some subjects are conscious of and which may be influencing other subjects subconsciously. We also see that people whose explanation featured *correct_reasoning* tend to be quite consistent in that they select the target both times when they see the same trial twice, and in fact a lot more consistent than those who reported guessing. Later in the chapter we also show that there is considerable alignment between reported strategies and performance once biases present in the stimuli are accounted for. So whether or not the *correct_reasoning* explanation reflects correct hypothetical reasoning about alternatives about in the moment, it does seem to be a good predictor of solving the task correctly. Also, in these experiments we combine analysis of strategy explanations with careful manipulation of the stimuli and comparison of effect sizes. The fact that these two measures together yield consistent results gives more confidence in the validity of the provided explanations.

3.3 Experiment 1: Replication of F&D with reasoning elicitation

In this experiment, we replicated the comprehension experiment by Franke & Degen (2016), additionally eliciting participants' reasoning strategies in order to get an insight into how participants solve the task.

3.3.1 Participants

60 native speakers of English with an approval rating of at least 95% were recruited through the crowdsourcing platform Prolific.

3.3.2 Materials and procedure

Participants completed the reference game described in Reference game. After completing the main experiment, participants saw one simple and one complex item again,

presented in randomized order. When they made their selection, a red box appeared around it, along with the question "Why did you make that choice?" and a textbox. Participants' strategies were then annotated as described in Annotation scheme.

3.3.3 Results

One participant's data was not saved on the server. Two further participants were excluded from analysis for not having paid enough attention because their performance on the unambiguous filled trials was below 80%. The data from the remaining 57 participants entered the analysis.

F&D fit a logistic mixed-effect regression model to verify that participants perform significantly better on the simple condition than they do on the complex, and that their performance on the complex condition is above the chance baseline (the ambiguous filler condition). The model in F&D included the maximal effect structure that allowed it to converge, which, in addition to per-participant random intercepts, included random slopes for message type (accessory or species) and trial number. It did not include a random slope for condition, or a per-item random intercept, presumably for convergence reasons. We believe, however, that it is important to include a random slope for condition in the model since the motivation for individual-level modeling is that individual participants may perform differently in the two conditions.

We also faced convergence issues when attempting to fit generalized linear mixed-effects models for our data despite using an optimizer (*bobyqa*, Powell (2009)): the only random effect structure with which the models for all 4 of our experiments converged was one that included only per-participant random intercepts. Therefore, in order to be able to keep the random effect structure maximal, we fit all our models using Bayesian regression using the *brms* package in R (Bürkner, 2017).

We ran 4 chains of 4000 iterations each, with the first 2000 iterations of each chain discarded as warm-up. To assess the evidence for the effects, we examined whether the credible intervals for the parameter estimates included zero. If the credible interval excluded zero, we concluded that there was a meaningful relationship between the predictor and the dependent variable.

Like in F&D, we exclude from analysis trials on which distractors were selected (for Experiment 1, that corresponds to 2.1% of trials) and regress the binary correctness variable (whether target or competitor was selected) onto condition (simple, complex, or ambiguous, dummy-coded, with complex as the reference level), trial number, the interaction between trial number and condition, message type (creature or accessory), and position of the target creature on the screen (left, center, or right, dummy-coded with left as the reference level). The random effect structure was maximal and included per-participant random intercepts and random slopes for condition, message type and trial number, and per-item random intercepts. The results are reported in Table 3.2.

	F & D (51)	replication (57)	remapped (55)	all messages (56)	shapes (60)
Intercept	-0.15 (0.11), p=0.18	-0.29 (0.27), [-0.83, 0.24]	0.33 (0.17), [-0.00, 0.66]	-0.21 (0.21), [-0.63, 0.21]	0.20 (0.20), [-0.20, 0.61]
condition (simple vs. complex)	1.28 (0.12), p<0.0001	2.16 (0.37), [1.47, 2.94]	0.46 (0.33), [-0.17, 1.15]	1.15 (0.26), [0.66, 1.68]	1.24 (0.30), [0.67, 1.86]
condition (ambig vs. complex)	-0.44 (0.13), p<0.001	-0.19 (0.27), [-0.74, 0.34]	-0.33 (0.17), [-0.68, 0.01]	-0.06 (0.24), [-0.54, 0.41]	-0.42 (0.21), [-0.85, -0.03]
trial number	0.00 (0.00), p<0.3	0.00 (0.01), [-0.01, 0.01]	-0.00 (0.00), [-0.01, 0.01]	-0.00 (0.00), [-0.01, 0.01]	0.01 (0.00), [0.00, 0.02]
simple vs. complex : trial	0.00 (0.01), p<0.9	0.00 (0.01), [-0.01, 0.02]	0.02 (0.01), [0.01, 0.04]	-0.01 (0.01), [-0.02, 0.01]	-0.02 (0.01), [-0.03, -0.00]
ambig vs. complex : trial	0.01 (0.01), p<0.33	0.00 (0.01), [-0.01, 0.02]	-0.01 (0.01), [-0.02, 0.01]	0.01 (0.01), [-0.01, 0.02]	-0.02 (0.01), [-0.03, -0.00]
target pos (middle vs. left)	1.28 (0.14), p<0.0001	0.52 (0.15), [0.23, 0.80]	0.20 (0.14), [-0.07, 0.47]	0.30 (0.13), [0.04, 0.56]	0.38 (0.13), [0.11, 0.64]
target pos (right vs. left)	0.74 (0.13), p<0.0001	0.54 (0.15), [0.25, 0.83]	0.09 (0.13), [-0.17, 0.34]	-0.20 (0.13), [-0.46, 0.06]	-0.04 (0.14), [-0.31, 0.23]
msg type (accessory vs. species)	-0.02 (0.12), p<0.85	0.25 (0.21), [-0.15, 0.68]	-0.19 (0.14), [-0.48, 0.09]	0.48 (0.20), [0.08, 0.89]	0.22 (0.13), [-0.03, 0.47]

Table 3.2: Effect size estimates, standard errors, and 95% credible intervals for the 4 experiments, as well as effect size estimates, standard errors, and p-values for the original study by F&D. Participants who did not reach 80% accuracy threshold on unambiguous trials were excluded from analysis.

As in the original study by F&D, in the replication we find strong evidence that participants performed better on the simple trials than on the complex ones ($\hat{\beta}$ =2.16 (0.37), 95% CrI = [1.47, 2.94]); the effect size estimate is larger than in the original study (2.16 vs. 1.28). We do not find evidence for an effect of the difference between the population-level performance on the complex trials and the chance baseline ($\hat{\beta}$ =-0.19 (0.27), 95% CrI = [-0.74, 0.34], whereas in the original study, that difference was significant (β =-0.44 (0.13), p<0.001). We think that this difference may in part be due to a different participant sample, but also due to the original study using a different strategy to code the ambiguous condition. Since in the ambiguous condition, two of the three creatures are identical, it needs to be decided somehow which one is target and which one is competitor. Since F&D report an uneven split in the ambiguous condition (46% of target choices vs. 51% of competitor choices), it appears that their implementation was to randomly decide on each trial which one of the two identical creatures was target and which one was competitor (let's call this method *coin-flipping*). We, on the other hand, ensure an even split at the population level by having exactly half of the ambiguous non-distractor responses be correct. We chose this randomization method since we believe that it is a better approximation

of a chance baseline and has smaller variance depending on the initialization[1]. Our simulation results suggest that the difference between the ambiguous and the complex conditions in the original study was amplified by the fact that F&D's random initialization happened to result in more competitor responses.

Like in the original study, we find evidence of participants being more likely to select the target if it was in the center ($\hat{\beta}$ =0.52 (0.15), 95% CrI = [0.23, 0.80]]) or on the right ($\hat{\beta}$ =0.54 (0.15), 95% CrI = [0.25, 0.83]). We find no evidence of an effect of trial or of the interaction between trial and condition, indicating absence of learning effects, nor an effect of message type (creature or accessory).

We now take a look at the annotations. Inter-annotator agreement was substantial, Cohen's κ=0.65, 95% CI [0.56,0.75]. In the simple condition, 1 response labeled *exclude* and 8 responses labeled *unclear* out of 57 annotations were excluded. 48 of the 57 annotations (84%) entered the analysis. In the complex condition, 2 *exclude* and 5 *unclear* responses were excluded, and 50 of the 57 annotations (88%) entered the analysis.

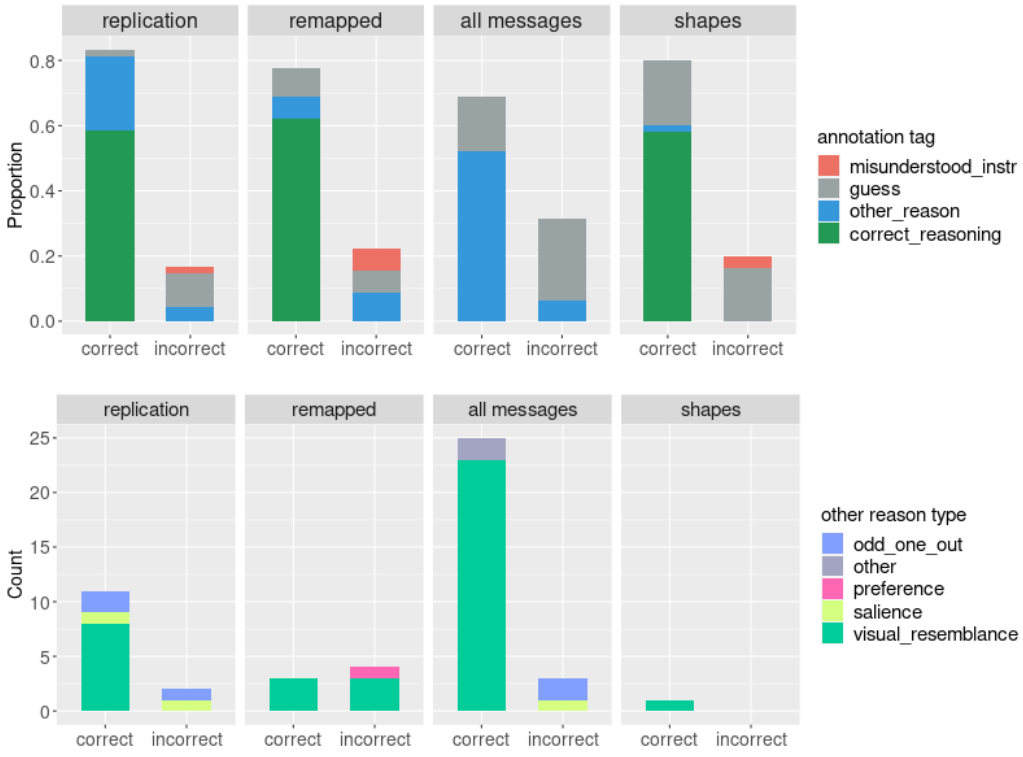


Figure 3.2: Participants' explanations by tag in the simple condition (top – all explanations, bottom – *other_reason* explanations).

As can be seen in the top panel of Figure 3.2, 27.5% of correct responses (and

[1]When we ran each randomization method 100 times on our replication data, the standard deviations of effect size estimates were larger for the coin-flipping method. Related to that, there was a significant difference between the complex and the ambiguous conditions 10 out of the 100 times when the coin-flipping method was used and none of the times when we ensured a 50-50 split.

27.1% of total responses) in the simple condition fell into the *other_ reason* category, suggesting that some participants arrived at the correct answer not via the assumed reasoning. When we examine the correct *other_ reason* responses (bottom panel of Figure 3.2), we see that the majority of them used the *visual_ resemblance* strategy. For the remainder of this chapter, we will be discussing the simple condition, which turned out to be more susceptible to bias and more helpful for revealing it, but corresponding graphs for the complex condition can be found in Appendix B.

In Figure 3.3, we plot participants' average performance on the simple and complex trials. We see that the majority of participants who gave an *other_ reason* explanation in the simple condition exhibit near-ceiling performance, so based on performance alone they would be indistinguishable from correct reasoners, although their answers are correct due to factors unrelated to RSA-style reasoning about alternatives at least some of the time.

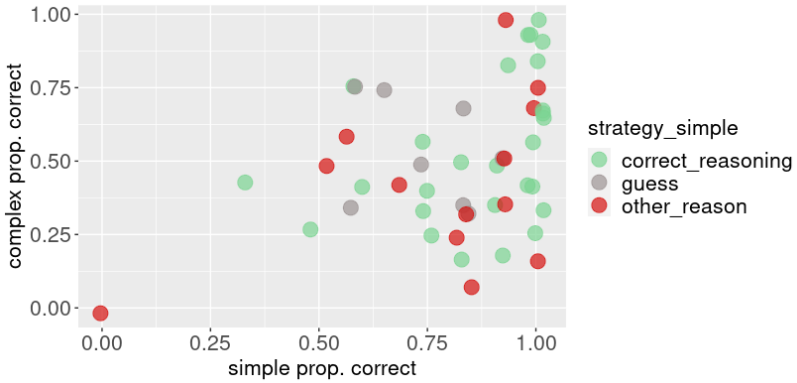


Figure 3.3: Each participant's average performance in Experiment 1, with the color corresponding to the strategy label on the simple condition. The red dots correspond to participants who applied a strategy other than *guess* or *correct_ reasoning*. Quite a few of the red dots are pretty far on the right, indicating that these participants got a trial correct for the wrong reason at least some of the time; in other words, these participants' performance is likely inflated. Based on the performance alone, however, they are indistinguishable from correct reasoners.

3.4 Experiment 2: Remapped version

We observed that in the original stimuli, the pairs of expressible features (that is, the features for which there is a message available to the speaker) constitute a kind of conceptual grouping – the two monsters are expressible as messages but the robot is not, the two hats are expressible as messages but the scarf is not.

The inexpressible features, therefore, because they are the odd-ones-out – robot is the only non-monster creature and the scarf is the only non-hat accessory – may

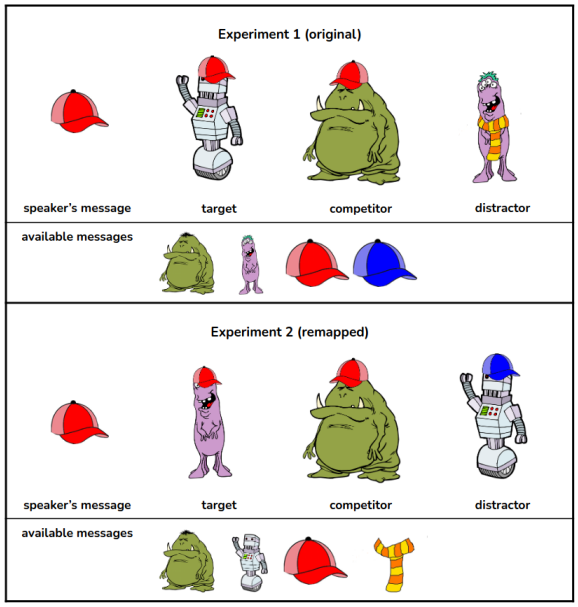


Figure 3.4: An example of a remapped trial (the simple condition from Figure 3.1). Because the robot gets swapped with the purple monster, the target becomes purple monster with the red hat, the competitor stays the same, and the distractor becomes a robot (swapped with the purple monster) with a blue hat (swapped with a scarf).

stand out a lot more. To further probe the effect of stimuli on this task, we, therefore, swapped the expressible and inexpressible features around, breaking this conceptual grouping: the robot and the scarf were swapped for the purple monster and the blue hat respectively. The swapping is illustrated in Figure 3.5: along the creature dimension, we made the robot expressible and the purple monster inexpressible instead, and along the accessory dimension, we made the scarf expressible and the blue hat inexpressible instead.

So in this experiment, the simple trial in Figure 3.1 looks as follows (illustrated in Figure 3.4): the message stays the red hat since no swapping had been performed there; the target is now a purple monster with a red hat (since we swapped the robot and the purple monster around), the competitor remains the same since no swapping has been performed for either the green monster or the red hat, and the distractor is now a robot with a blue hat (since the purple monster is swapped with the robot and the scarf is swapped with the blue hat). Therefore, underlyingly, all trials remained the same as in Experiment 1, the only difference is that the inexpressible features are represented with different images. We used the images from the original study so no changes were performed e.g. to the hat orientation.

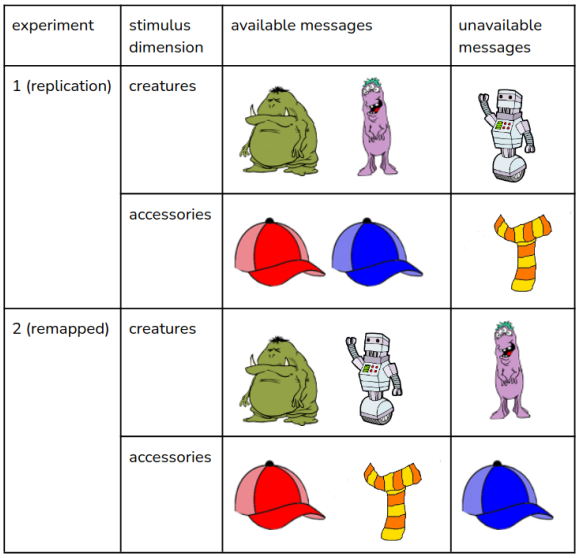


Figure 3.5: For Experiment 2, we changed which features were expressible as messages.

3.4.1 Participants

60 native speakers of English with an approval rating of at least 95%, were recruited through the crowdsourcing platform Prolific.

3.4.2 Materials and procedure

We changed which features were expressible as messages, swapping the robot and the scarf for the purple monster and the blue hat respectively. Otherwise the experiment was identical to the replication. An example of how the swapping was performed is shown in Figure 3.4.

3.4.3 Results

One participant's data was not saved on the server. Four further participants were excluded from analysis for not having paid enough attention because their performance on the unambiguous filler trials was below 80%. The data from the remaining 55 participants entered the analysis.

As can be seen in Table 3.2, we observe a stark change in the effect for simple vs. complex condition: the effect size is a lot smaller than in the replication, and the 95% credible interval crosses 0 ($\hat{\beta}$=0.46 (0.33), 95% CrI = [-0.17,1.15]). Like in the replication, we don't find evidence for difference in performance on the complex condition and the chance baseline ($\hat{\beta}$ =-0.33 (0.17), 95% CrI = [-0.68, 0.01]).

Strategy responses were again annotated by two annotators, one of whom was blind to the purpose of the experiment. Inter-annotator agreement was again substantial,

Cohen's κ=0.77, 95% CI [0.68,0.86]. In the simple condition, 4 responses labeled *exclude* and 6 responses labeled *unclear* out of 55 annotations were excluded. 45 of the 55 annotations (82%) entered the analysis. In the complex condition, 3 *exclude* and 10 *unclear* responses were excluded, and 42 of the 55 annotations (76%) entered the analysis.

When we examine the breakdown of annotation tags (Figure 3.2), we see that a smaller proportion of correct answers and a larger proportion of incorrect answers is comprised of *other_reason* responses, and while in the original study, *visual_resemblance* led participants to always select the correct answer, in the remapped version, it was now incorrect some of the time; salience, preference, and *odd_one_out* reasoning now pointed participants to the competitor instead of the target. For instance, in the simple trial depicted in Figure 3.4, in the original version, the creature that is the odd-one-out and therefore a more salient one is the robot, which happens to be the target, so if a participant selects it based on *salience* or *odd_one_out* reasoning, they will happen to be correct. In the remapped version, on the other hand, the target is less visually salient because it is one of the two monsters, therefore, the same *salience* or *odd_one_out* reasoning may lead the participant to select the distractor, the robot, which would be incorrect. This suggests that in the remapped version of the task, some of the existing biases in the stimuli were now having the opposite effect from the original study, that is, "deflating" participant performance.

3.5 Experiment 3: Original experiment, all messages available

We wanted to know how often participants would make the right choice coincidentally in the original experiment, in a setting where correct hypothetical reasoning about alternatives is made impossible, thus isolating stimuli effects.

In order to isolate stimuli effects and make *correct_reasoning* impossible, we made all six features expressible. As a result, *correct_reasoning* is no longer possible in the simple condition, since there are now unambiguous messages available to refer to the target as well as the competitor. For example, in the top panel of Figure 3.1, both the target and the competitor can now be referred to unambiguously via the messages "robot" and "green monster" respectively, so the message "red hat" is now completely ambiguous.

Note that in the complex condition, *correct_reasoning* is still possible because the target still cannot be unambiguously identified by a message, while both the competitor and the distractor can be.

Thus, in this case, none of the formal models should be able to solve the simple implicature condition and, like in the original setup, only L_2 should be able to solve the complex one.

3.5.1 Participants

60 native speakers of English with an approval rating of at least 95%, were recruited through the crowdsourcing platform Prolific.

3.5.2 Materials and procedure

One participant's data was not saved on the server. Three further participants were excluded from analysis for not having paid enough attention because their performance on the unambiguous filled trials was below 80%. The data from the remaining 56 participants entered the analysis.

Apart from the feature expressibility manipulation, the experiment was identical to the replication.

3.5.3 Results

Table 5.2 shows that, with all messages available, the correct option is still selected for reasons that are not correct hypothetical reasoning quite often, as evidenced by the larger signficant effect of condition (complex vs. simple, $\hat{\beta}$=1.15 (0.26), 95% CrI = [0.66, 1.68]), notably, considerably more often than in the remapped experiment ($\hat{\beta}$=0.46 (0.33), 95% CrI [-0.17, 1.15]).

Inter-annotator agreement was again substantial, Cohen's κ=0.67, 95% CI [0.57,0.78]. In the simple condition, 1 response labeled *exclude* and 7 responses labeled *unclear* out of 56 annotations were excluded. 48 of the 56 annotations (86%) entered the analysis. In the complex condition, 6 *unclear* responses were excluded, and 50 of the 56 annotations (89%) entered the analysis.

When we take a look at the performance by tag (Top panel of Figure 3.2), we see that when participants used an *other_reason* strategy in the simple condition, the vast majority of the time (89.2% of cases) their strategy coincidentally led them to select the target. When we examine the *other_reason* tags in more detail, we see that the strategy that resulted in accidental correct responses is *visual_resemblance*, that is, similarity of the target to the message based on the head of a scarf-wearing creature being uncovered, similar to the message (that creature without an accessory) or the orientation of the hat message matching that of the referent.

This corroborates the claim that factors unrelated to reasoning, most importantly incidental visual resemblance of the message to the target, inflate performance in the original experiment.

3.6 Experiment 4: Mitigating bias by using abstract stimuli

Having shown that the visual resemblance bias accounts for the vast majority of coincidental target choices, we repeated the original experiment with a version of the stimuli that we hoped would be less susceptible to the *visual_resemblance* bias.

3.6.1 Participants

60 native speakers of English with an approval rating of at least 95%, were recruited through the crowdsourcing platform Prolific.

3.6.2 Materials and procedure

There were no exclusions based on accuracy in this experiment, as all 60 participants scored above 80% on the unambiguous filler trials.

Having identified the biases in the original stimuli, we attempted to mitigate them. We hypothesized that using more abstract stimuli, of the kind used in the original RSA study (Frank & Goodman, 2012), would not create opportunity for a visual resemblance bias between the message and the target.

We therefore ran a version of the experiment with abstract stimuli, using geometric shapes and colors instead of creatures and accessories: square, triangle and circle corresponding to the robot, green monster and purple monster respectively, and the colors blue, green and red corresponding to the scarf, red hat and blue hat respectively. When the message was a color, it was represented by a color tube, and when it was a shape, it was represented by a shape contour. An example of trial from Experiment 4, corresponding directly to Figure 3.1 from Experiment 1, is included in Figure 3.6.

Thus, underlyingly, the experiment remained the same as the original, the only difference being what images were used to represent the messages and the referents.

3.6.3 Results

Inter-annotator agreement was again substantial, Cohen's κ=0.80, 95% CI [0.71,0.89]. In the simple condition, 2 responses labeled *exclude* and 3 responses labeled *unclear* out of 60 annotations were excluded. 55 of the 60 annotations (92%) entered the analysis. In the complex condition, 1 *exclude* and 6 *unclear* responses were excluded, and 53 of the 60 annotations (88%) entered the analysis.

We see in Figure 3.2 that with abstract stimuli, barely any *other_reason* explanations are given, indicating that people do not seem to rely on clues like visual similarity but instead employ either guessing or correct hypothetical reasoning about alternatives, as assumed by formal probabilistic models. Only one *visual_resemblance*

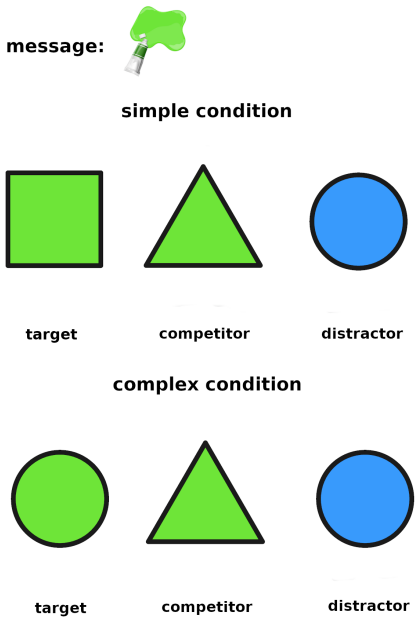


Figure 3.6: Example of a simple and a complex trial for Experiment 4. These trials directly correspond to those depicted in Figure 3.1 for Experiment 1.

response is given, and it does happen to accidentally lead to selecting the target: when the message is a circle, the blue circle is selected with the justification that the blue color is "nearest to clear", so a clear circle is most similar to the blue circle.

We conclude the abstract stimuli appear to be less susceptible to biases that inflated performance on the original experiment, and in particular, the visual resemblance bias. Therefore, we can have more confidence that the task results more accurately reflect participants' pragmatic reasoning ability. In terms of the effect sizes, this study's results closely resemble those of the original F&D study and our replication: $\hat{\beta}$ =1.24 (0.30), 95% CrI = [0.57, 1.86] for simple vs. complex condition in this study vs. β =2.16 in the replication; like in the original study and unlike in our replication, there is evidence for a difference between the complex condition and chance ($\hat{\beta}$ =-0.42 (0.21), 95% CrI = [-0.85,-0.03]). Thus, it appears that the original study, despite the biases, fairly accurately captures the effects at the population level. However, now we can be more certain that they accurately reflect individual-level performance as well.

3.7 General discussion

In the study described in this chapter, we elicited participants' strategies and manipulated which messages were expressible to probe the stimuli first introduced by Degen et al. (2012) and used in Degen et al. (2013) and F&D for biases. Effect sizes varied greatly between experiments and appear to be inflated in the original task

by factors unrelated to reasoning. The most prominent bias was visual resemblance, where participants selected the creature most similar to the message, which in most cases happened to be the target. We then replicated the experiment with a version of the stimuli that we showed to be less susceptible to the visual resemblance bias.

There remains the question of whether the effect sizes in Experiments 1 and 4 are even comparable: the stimuli in the original task have biases, inflating the effect of condition, but they are also visually more complex, potentially leading to higher task difficulty and lower effect of condition than the abstract stimuli in Experiment 4. In other words, there appears to be no way to know exactly how much smaller the "true" effect size of condition in the original experiment would have been if there were no stimuli-related biases, because when we use different stimuli that do not have the biases, those stimuli also have different perceptual properties which affect task complexity, and, correspondingly, effect sizes. We would expect the "true" effect sizes to lie somewhere between those in Experiment 1 and Experiment 2, since in the original experiment, biases inflate participants' performance, and in the remapped version (Experiment 2), biases still exist but they lead participants to choose incorrectly.

The final experiment, which used different stimuli, was shown to be much less susceptible to biases present in the original stimuli. That means that we can have more confidence that participants' performance on the task is a more reliable reflection of their reasoning ability, at the population level, and especially at the individual level.

To investigate individual-level performance in Experiments 1 and 4 in more detail, we ran Latent Profile Analysis on participants' average performance on simple and complex conditions using the *tidyLPA* package in R to obtain reasoner classes. We decided to use LPA and not the Bayesian model from F&D because the latter is too strict for our purposes: it presupposes the existence of only three classes (the theoretically motivated classes L_0, L_1, and L_2), and since we saw that many participants used other strategies, we wanted to have a data-driven way of identifying classes which may potentially include classes other than the ones corresponding to the three probabilistic listener models.[2] Recall that Experiments 1 and 4 are underlyingly the same and the only difference is what images are used as the representation of the messages and referents, hence L_0, L_1 and L_2 RSA models make the same predictions for each trial of the two experiments.[3]

For both experiments, the best fit, as measured by AIC and BIC, was obtained by models with 4 classes (model fit for each number of classes can be found in the Appendix B).[4] As can be seen in Figure 3.7, three of the four identified classes approximately correspond to the predictions of the three formally defined reasoning types

[2]Incidentally, applying F&D's class assignment model to our data yielded uninterpretable classes.

[3]That is also true for Experiment 2, but not for Experiment 3, where neither L_1 nor L_2 are predicted to be able to solve the simple condition since it is made completely ambiguous.

[4]For Experiment 4, the 5-class model has the best fit according to AIC while the 4-class model has the best fit according to BIC. The only difference between those two clusterings is that for the 5-class model, one participant whose accuracy is 0 for both simple and complex is in their own separate class. Therefore, for comparability with Experiment 1, we use the 4-class model.

from F&D, L_0 (at chance on both conditions), L_1 (can solve the simple but not the complex trials), and L_2 (can solve both conditions). Additionally, there's the smallest fourth class, consisting of 4 and 2 participants respectively for the two experiments, of participants who perform below chance on the simple condition and at or below chance on the complex. Strategy annotations also approximately match our expectations, but the pattern is more pronounced for Experiment 4: participants assigned to the class L_0 guess in the simple condition while participants assigned to classes corresponding to L_1 and L_2 apply correct reasoning in the simple condition. The corresponding scatterplot for the complex condition can be found in the Appendix B.

We see that in Experiment 1, of the participants whose strategy in the simple condition was labeled *other_reason*, 3 are assigned to the L_0 class, 5 to L_1, and 3 to L_2. As we saw in the above experiments, these participants' performance was likely inflated by factors unrelated to reasoning; therefore, they were possibly estimated to have a more advanced reasoning type than they actually have. In Experiment 1, only one participant used an *other_reason* strategy in the simple condition; that participant was assigned to L_0, although they are on the border with L_2 (bottom panel of Figure 3.7). As we saw in Experiment 4, this participant happened to choose the target for the wrong reason (a creative application of the *visual_resemblance* strategy); therefore, their performance is also inflated. However, the fact that this is only 1 participant is a large improvement compared to Experiment 1.

When we look at the breakdown by tag for the identified classes (Figure 3.8), we see that the responses in Experiment 4 much more cleanly map onto the corresponding formal classes than the responses from Experiment 1 (L_0 guesses for both types of trials, L_1 applies correct reasoning about alternatives for the simple condition and guesses on the complex, and L_2 applies correct reasoning in both conditions), again supporting the claim that the debiasing of the stimuli was successful and constitute a better proxy for participants' reasoning ability as defined by formal probabilistic models. In order to quantitatively corroborate this finding, we conducted cluster analysis. Each participant was assigned a class label based on performance. For that, we used the four classes identified by LPA: L_0, L_1, L_2, and *other* (people who ended up in the fourth, *below_chance* class). Each participant was also assigned a class label based on the annotated strategies: L_0 if they used the *guess* strategy in both conditions, L_1 if their strategy was labeled *correct_reasoning* in the simple condition and *guess* in the complex, and L_2 if their strategy in both condition was *correct_reasoning*. Finally, participants who used an *other_reason* strategy in either condition were given the label *other*. Participants whose strategy was labeled *unclear* or *exclude* were excluded from this analysis, resulting in 42 participants for Experiment 1 and 47 participants for Experiment 4 entering the analysis. The motivation for the *other* label is that we would ideally want people who use other reasoning to be distinguishable from the three reasoning types based on performance. In order to see how well the performance-based and annotation-based classes aligned, we computed cluster homogeneity and completeness for Experiment 1 and Experiment 4,

with performance class labels as reference. For Experiment 1, homogeneity is 0.13 and completeness is 0.13, and for Experiment 4, homogeneity is 0.56 and completeness is 0.51. This further supports the observation that more abstract stimuli lead to a better alignment between performance and the underlying strategy.

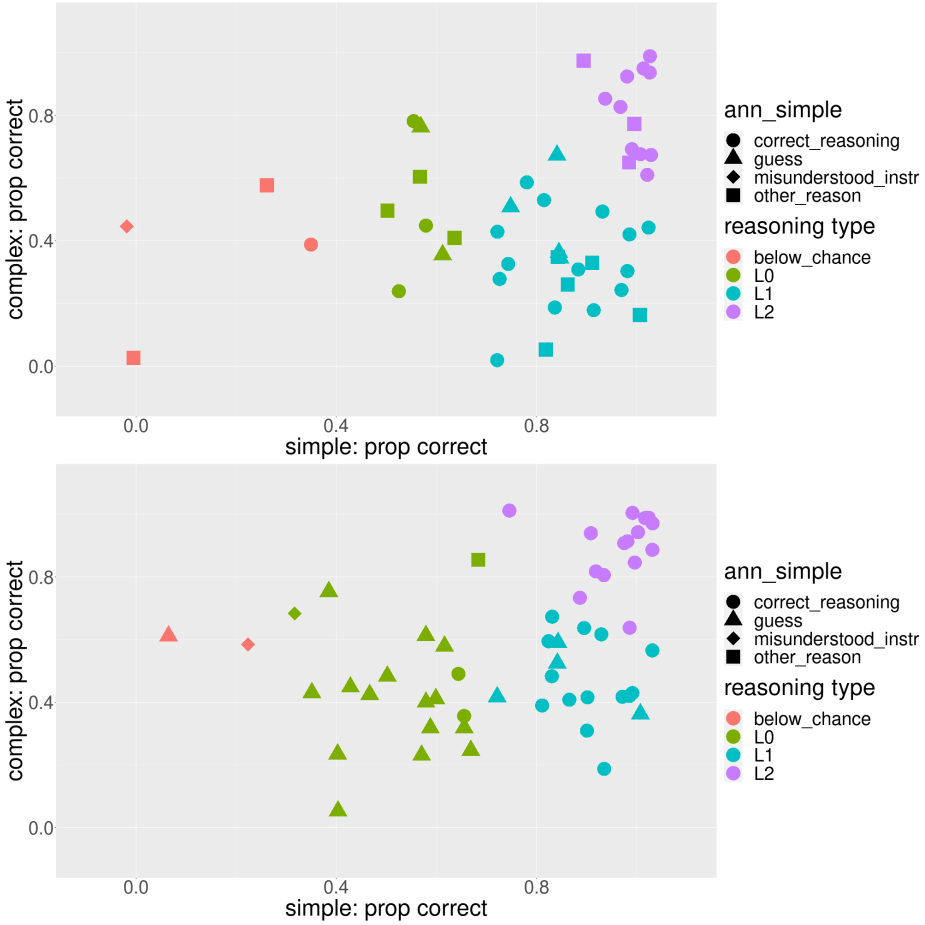


Figure 3.7: Classes of participants identified by LPA (top – Experiment 1 (replication), bottom – Experiment 4 (debiased stimuli)). Three of the identified classes approximately correspond to the theoretical predictions of listener models L_0, L_1 and L_2. The fourth LPA class, which is the smallest for both experiments, corresponds to participants who perform below chance level in the simple condition.

We used post-hoc strategy elicitation to probe the experimental design of the reference game for biases. Post-hoc explanations have been criticized for not to always accurately reflecting the reasoning in the moment (e.g., Cushman, 2020). If we look at the relationship between participants' performance and the elicited strategy explanation in Experiment 1 (Figure 3.3), we see that while most participants whose strategy explanation on the simple trial was labeled *correct_reasoning* performed near ceiling on average, there are a few participants who show chance-level average performance. One reason why this might be the case is that the fact that a participant applied a certain strategy, e.g. *correct_reasoning*, on a given trial does not mean that they

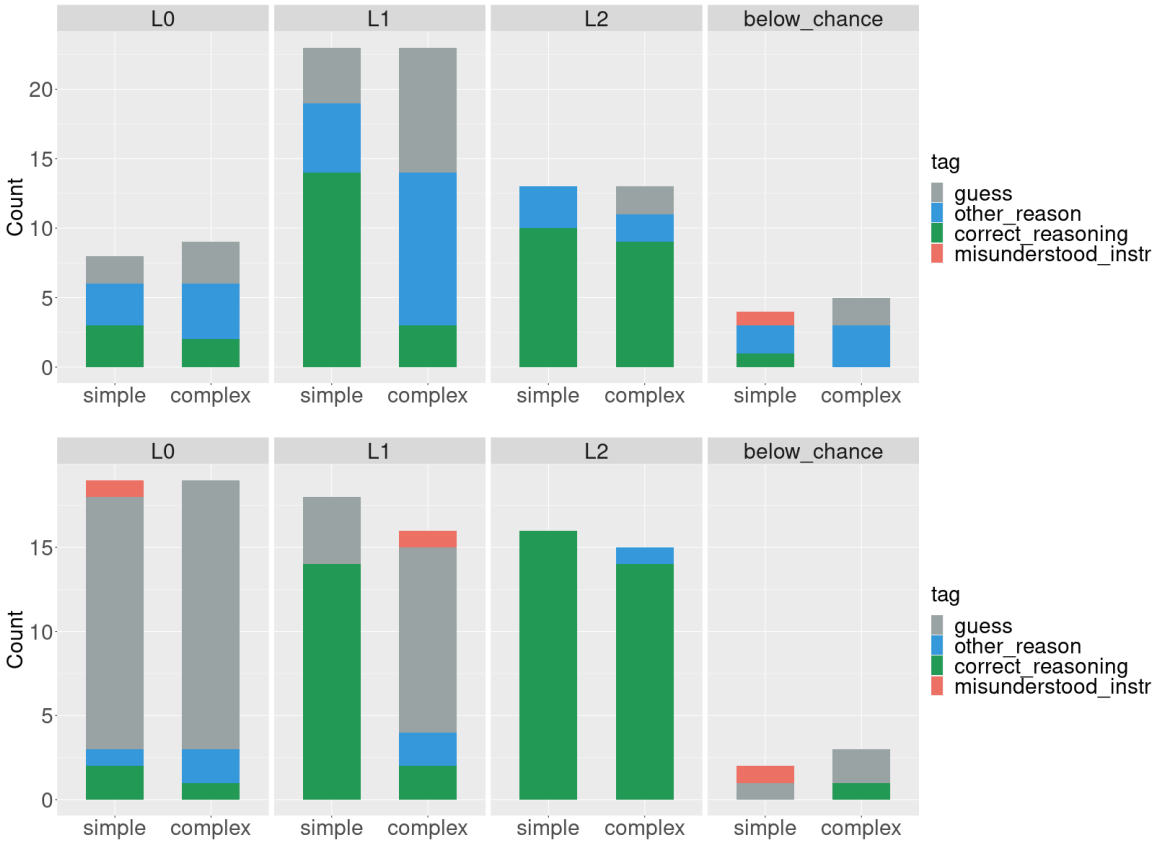


Figure 3.8: Tags in each condition for the classes identified by LPA (top - Experiment 1 (replication), bottom - Experiment 4 (debiased stimuli)).

applied it successfully on all trials. Another possibility is that these participants were actually guessing in the moment and came up with the correct reason for their selection after the fact. This suggests that these explanations on a single trial are not a perfect proxy for participants' reasoning strategies throughout the experiment. However, the fact that participants who provided a *correct_reasoning* reasoning on the simple trial have much better average performance than guessers (0.87 (0.17) vs. 0.66 (0.32)) suggests that, while imperfect, post-hoc explanations are a proxy for the strategies participants used during the experiment. While there are limitations to elicited post-hoc explanations, they can be a useful tool for revealing biases in an experimental design that could otherwise be missed, as was the case in this study.

The strategy elicitation method has been shown to be effective for another pragmatic phenomena, atypicality inferences. In a paper which I co-authored (Ryzhova et al., 2023), we applied a method quite similar to the strategy elicitation method used in the current study to a different pragmatic phenomenon, atypicality inferences, in order to gain more insight into how participants solved the task. When an event which is very predictable in a given context is overtly mentioned, it may be inferred that the event is actually atypical for the referent (e.g. "John went to the grocery store. He paid the cashier" may result in the atypicality inference that John does

not usually pay at the grocery store). In addition to collecting ratings of how typical participants expected the redundantly mentioned activity to be for the referent, we asked participants to justify the ratings they gave. Based on the provided justifications, we were able to make a distinction between two groups of participants which would have been unidentifiable based on ratings alone: those who did not make the atypicality inference and those who initially made it but then rejected it (e.g. "Not paying would be stealing, therefore it's unlikely that John does not pay") and gave a high typicality rating. Adding explanations to the experimental measure allowed us to identify different reasoning which results in the same performance, much like in the current work where we show that high task performance may be the result of correct reasoning but also of unintended biases in the task where people get away with applying simpler strategies. The fact that strategy elicitation has proved useful for two different pragmatic phenomena – the reference game in the study described in this chapter and in the case of atypicality inferences – suggests that it can be useful more broadly and may be worth applying when investigating other pragmatic phenomena and seeking to better understand the processes underlying pragmatic reasoning.

More broadly, we believe that, when designing experiments, it is very important to ask the question how to make sure that the measure that is being used is indeed tapping into the construct of interest. Might there be other explanations for why participants are behaving a certain way? Here, we showed two ways that can be used for probing an experimental design: careful manipulation of the stimuli and strategy elicitation. Of course, what methods are appropriate depends on the experimental setup: for instance, strategy elicitation seems to mostly apply when the studied phenomenon is expected to involve relatively conscious reasoning or reflection; manipulating stimuli in systematic ways and examining how that affects the results seems to be applicable more broadly. There may be biases which are not reported because subjects are not aware of them, just like participants don't report matching bias in Wason's selection task. Therefore, we believe that it is worthwhile to explore the application of on-line measures, such as eye gaze, to getting a better sense of participant strategies, and learn more about how those relate to performance and explanations obtained through introspection. How people actually solve a task is a difficult question as we cannot look directly into people's heads and have to rely on proxies like performance or explanations, but it is a very important one, and therefore we should make it a priority to search for possible ways to get closer to the answer.

3.8 Conclusion

In the experiments described in this chapter, we probed the reference game stimuli from Franke & Degen (2016) for biases by conducting four experiments where the stimuli properties were manipulated and by eliciting explanations of strategies used by participants. In doing so, the chapter made three contributions.

The first one is a methodological contribution relevant not only to research in Pragmatics but to experimental studies more broadly. We argue that it is important to validate that one's chosen measures reflect the construct of interest, as opposed to artifacts of the experimental design. We demonstrate the effectiveness of two complementary techniques which can be applied for probing an experimental design: (1) systematic manipulation and comparison of stimuli variants and (2) elicitation of explanations of participants' strategies.

Our second contribution was validating the reference game paradigm for use in subsequent experiments. We showed the reference game paradigm with abstract stimuli used in Experiment 4 was less susceptible to biases than the original stimuli used by Franke & Degen (2016) and thus more accurately captured the underlying pragmatic reasoning at the individual level. While it is unrealistic to expect no stimuli effects, it is important to know what biases are present and how strong they are because that influences the conclusions that are drawn. This is particularly important for individual difference studies, where spurious correlations of performance with individual measures may emerge or real ones may not be identified due to such stimulus effects. This is essential since studying individual variability is one of the core goals of this thesis and the main focus of Chapter 4.

The final contribution of this chapter is addressing Research Goal 1 by proposing a strategy elicitation method and demonstrating its effectiveness in yielding additional insights about the applied strategies. The strategy annotation scheme proposed in this chapter serves as the basis for analyses conducted in Chapters 4 and 5, with appropriate modifications for each experiment. For instance, in the experiments involving a child and an adult speaker described in Chapter 5, an additional *meta_reasoning* tag is introduced, which captures a participant's uncertainty about whether the speaker's reasoning ability is sufficient for selecting the optimal message.

In this chapter, we replicated variability in performance on the reference game with abstract stimuli, which supports F&D's finding that individuals differ widely in their pragmatic reasoning on this task. However, it remains an open question what gives rise to this variability: why do some participants perform better than others, giving responses that align with predictions of more complex probabilistic pragmatic models? The investigation of cognitive factors underlying the observed individual variability is the focus of the next chapter.

Chapter 4

Cognitive factors modulating individual variability in the pragmatic reference game: Effects of logical reasoning and Theory of Mind

As discussed in detail in Theoretical background, both cognitive and situational factors give rise to variability in pragmatic processing. Cognitive factors are the focus of the current chapter, while the remaining chapters examine a situational factor, namely speaker identity.

Chapter 3 replicated variability in a pragmatic reference game reported by Franke & Degen (2016) (F&D) with abstract stimuli which were shown to be less susceptible to bias. We replicated individual variability observed by F&D, where clusters emerged whose performance approximately corresponds to the predictions of three probabilistic Bayesian models of different reasoning depth, from literal listener L_0 to a second-level pragmatic listener L_2. However, the origins of these clusters remain unclear – what makes someone perform more or less pragmatically on the reference game, or, in RSA terms, what makes someone an literal vs. a pragmatic listener? It is also an open question whether reasoning types are stable and related to cognitive traits, or whether they are more indicative of one's attention or disposition on a given day. Answering these questions is essential for constructing plausible models of pragmatic processing. If performance on the reference game can be predicted by performance on tasks measuring cognitive traits, that would serve as evidence for stability of reasoning types and for variability being driven at least in part by cognitive factors.

In this chapter, we replicate the comprehension experiment from F&D on a much larger sample (N=292 vs. 51) using abstract stimuli, which were introduced and validated in Chapter 3, and also collect measures of three individual differences – working memory capacity, Theory of Mind and logical reasoning ability, which are

hypothesized to modulate reference game performance. We also collect participants' strategies using the elicitation method introduced in Chapter 3 in order to better understand the strategies corresponding to the three reasoning types, as well as to exclude participants who misunderstood the task to ensure a more accurate estimate of the relationship between pragmatic reasoning in preference games and the individual difference measures.

This chapter is based on, and in parts identical to, the following publication:

- **Mayn, A.**, & Demberg, V. Sources of individual variability in a pragmatic reference game: Effects of reasoning and Theory of Mind. To appear in *PLOS One*.

Data availability

Preregistration for the experiment presented in this chapter is available at `https://osf.io/392pk/`. Data and analysis scripts are available at `https://osf.io/5ab3f/`.

4.1 Introduction

The cooperative principle, formalized by Paul Grice (1975), states that communicators should behave cooperatively and adhere to Maxims of Conversation, such as Maxim of Quantity, which requires the speaker to be as informative as is required by the situation. Listeners are then expected to derive inferences based on the assumption that the speaker is behaving in accordance with Grice's rules. For example, upon hearing a guest say "It's cold in here", the listener might infer that the guest is politely asking her to turn on the heating or close the window as opposed to merely making a statement about room temperature (Trott & Bergen, 2019).

While Gricean principles decribe the ideal pragmatic communicator, there is a growing body of work showing that, in practice, people differ vastly in their tendency to derive pragmatic inferences and that some people tend to consistently favor literal interpretations. For example, Heyman & Schaeken (2015) identified three groups of responders in a scalar implicature derivation task: consistently pragmatic, consistently literal and inconsistent. Variability has also been observed in sensitivity to context when deriving inferences (Yang et al., 2018), in metaphor (Fairchild & Papafragou, 2021) and indirect request comprehension (Trott & Bergen, 2019), and in ad-hoc inferences about redundant mentions of typical events (Ryzhova et al., 2023).

Probabilistic pragmatic models, such as the Rational Speech Act framework (Frank & Goodman, 2012) and closely related Iterated Best Response models (Franke, 2009), aim to formalize Gricean principles and make quantitative predictions for various pragmatic phenomena. In these models, a listener and a speaker recursively reason about each other's intent, as well as potential alternative meanings and utterances, to arrive at an interpretation. Such models have been shown to provide a close fit to

population-level data for various phenomena, including scalar implicature and speaker knowledge (Goodman & Stuhlmüller, 2013), hyperbole (Kao et al., 2014b), politeness (Yoon et al., 2016), and rational overspecification (Degen et al., 2020). These models have mostly been fit to population-level data, averaging over participants and disregarding differences in tendency to draw pragmatic inferences.

Franke & Degen (2016) (henceforth, F&D) showed that individual differences in participants' performance on a pragmatic reference game can be explained by positing three probabilistic pragmatic models of different reasoning depth. They conducted an experiment where participants were asked to reason about potentially ambiguous messages allegedly sent by a participant of an earlier study, and showed that participants naturally fell into groups whose performance lined up with the predictions of three probabilistic models for listeners of three different reasoning depths, from the simplest literal model to a pragmatic Gricean model.

To better understand the nature of the observed individual variability and to be able to build more plausible cognitive models of pragmatic processing, it is valuable to investigate factors which drive this variability. While F&D showed that a model where each participant's data was captured by one of the three probabilistic pragmatic models of different complexity was able to better account for the data than a homogeneous model which assumed that all participants had the same reasoning depth, their model is agnostic about the factors which underlie these differences in pragmatic performance.

The current work adds to the growing body of evidence suggesting that individual differences in pragmatic reasoning are systematic and related to cognitive traits. We investigate the nature of individual variation on the pragmatic reference game and relate it to individual differences in three cognitive traits: working memory capacity, logical reasoning ability, and Theory of Mind. We replicate the variability on the reference game observed by F&D on a much larger sample size (254 participants for the reference game, 167 for the individual differences analysis, as opposed to 51 participants in F&D) and using a different set of stimuli, which we found in Chapter 3 to be less susceptible to biases than the original stimuli. In addition to the reference game, we collect a battery of individual differences, with two measures per construct of interest – logical reasoning ability, working memory, and Theory of Mind – in order to ensure robustness of the obtained results. Moreover, we ask participants to report their strategy and inspect the relationship of the reported strategies to groups identified based on performance.

Our analyses reveal a positive effect of logical reasoning ability. In addition to the main effect, we find an interaction with condition, suggesting that logical reasoning ability differentially affects performance on implicatures of different complexity. In addition, we find a positive effect of Theory of Mind, whereby participants with higher mentalizing skills are more likely to draw implicatures. We do not find evidence for an effect of working memory.

4.2 Background

4.2.1 Reference game

Our task of interest is a pragmatic reference game, a signaling game which has been repeatedly used to verify predictions of probabilistic pragmatic models (e.g., Frank & Goodman (2012); Akogul & Erisoglu (2017); Sikos et al. (2021b); Zhou et al. (2022)). It involves a speaker sending a message to refer to one of the objects in the shared view and the listener trying to identify the intended referent based on the message.

The task is a replication of the comprehension experiment (Experiment 1) of F&D, with two modifications motivated by the findings of Chapter 3. First, instead of using the stimuli from F&D (pictures of monsters and robots with accessories), we opted for abstract stimuli (shapes and colors) since in the previous chapter we showed them to be less susceptible to biases leading participants to select the target due to superficial similarity as opposed to successful reasoning. Secondly, at the end of the task, we elicited participants' reasoning strategies in order to identify participants who misunderstood the task and to gain insight into whether participants' reported strategies reflect the assumed reasoning or whether some participants used a different strategy. Strategy elicitation also allowed us to gain more insight into participants' performance on the task and monitor possible biases in the stimuli.

On each experimental trial, participants saw three objects on the display. The objects differed along two dimensions – shapes (triangle, square and circle) and colors (red, green and blue). Participants also saw a message that they were told had been sent by the previous participant, which was either a color or a shape, in order to get them to pick out one of the three objects. Participants were also told that only two of the three shapes and two of the three colors were available to the previous participant to send as messages (square and blue were not available; we'll call them *inexpressible features*), and the available messages were always displayed at the bottom of the screen.

In the beginning of the task, participants completed 4 practice trials which took the speaker's perspective: participants had to select a message to refer to one of three objects on the screen, which was highlighted. Then the actual task began, which consisted of 66 trials, of which 24 were critical and 42 were fillers, distributed as follows.

Half of the critical trials were *simple implicature trials*. On those trials, the target contained the sampled message and the inexpressible feature along the other dimension. On complex implicature trials, the target contained the message along with an expressible feature along the other dimension. An example of both types of critical trials is shown in Figure 4.1. As described in detail in the theoretical background (Section 2.7), the complex implicatures are predicted to be more difficult than the simple ones based on the fact that a probabilistic Bayesian model can be defined that can solve the simple but not the complex implicatures: a pragmatic listener model L_1

who reasons about a literal speaker S_0 only succeeds at simple implicatures, whereas a pragmatic listener model L_2 reasoning about a pragmatic speaker S_1 can solve both implicature types.

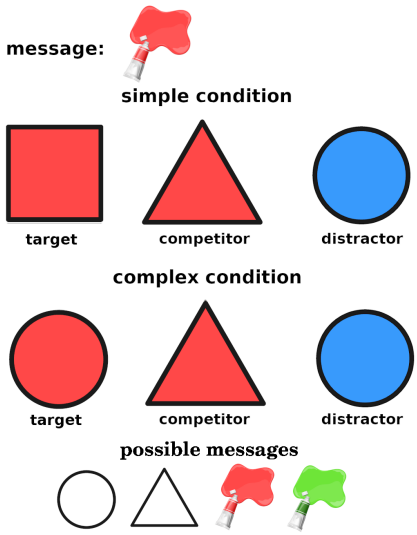


Figure 4.1: An example of a simple and a complex reference game trial.

Of the 42 filler trials in the study, 24 were identical to the critical trials, except the target was the (6) competitor or (6) distractor from the simple condition, or the (12) competitor from the complex condition, unambiguous given the message. For example, a filler trial would have a display identical to the top panel of Figure 4.1, except the target would be the red triangle, and the (unambiguous) message would be "triangle". Of the remaining 18 trials, 9 were completely unambiguous, meaning that they contained each shape and each color only once, and 9 were completely ambiguous, where two of the three objects were completely ambiguous and the message was equally likely to refer to both of them. Ambiguous trials were included as the random baseline, and unambiguous fillers served for determining exclusion. Participants who had accuracy below 80% on the unambiguous trials were excluded from analysis.

4.2.2 Elicitation of reasoning strategies

After participants completed the 66 reference game trials, they were shown one simple and one complex item again, in that order. While the order of other items was randomized for every participant, the two items presented at the end were kept the same for all participants. In those two trials, once participants made their choice, a red box appeared around their chosen object, a textbox popped up, and they were asked to briefly explain their decision.

The purpose of this was twofold: On the one hand, it could help us get additional insight into the way participants performed the task. We expected the different reasoning strategies corresponding to different reasoning depths hypothesized by F&D

to be reflected in the provided explanations. On the other hand, this allowed us to identify participants who might have misunderstood the instructions and therefore needed to be excluded.

Two annotators, one of whom was blind to the purpose of the study, annotated participants' explanations based on the annotation scheme from Chapter 3, with some modifications which will be described below. Inter-annotator agreement was substantial, Cohen's $\kappa = 0.76$, 95% CI [0.72, 0.80]. All disagreements were resolved jointly: both annotators made a case for their label, and if agreement could not be reached after that, the explanation was labeled *unclear*.

The version of the annotation scheme used in this chapter is summarized in Table 4.1. We will discuss the differences from the annotation scheme proposed in Chapter 3 (see Table 3.1 for comparison) below. The complete annotation scheme with more examples for each tag can be found in Appendix A.

Like in Chapter 3, the *correct_reasoning* tag to denote correct counterfactual reasoning about a message the speaker could have chosen and *guess* denotes guessing. However, here we add two subtags to the *guess* tag: Responses which reported deciding between two objects based on their position on the screen (e.g., picking the first one or the middle one) were additionally assigned the subtag *screenloc*. Explanations which reported picking one of the two objects "to mix things up" since the participant had chosen the other shape or color more often in the past were assigned the subtag *variability*.

As in Chapter 3, the *other_reason* tag was used when an explanation revealed a different strategy from correct reasoning and guessing, with a subtag describing the particular strategy. In addition to the subtags *salience*, *preference*, *visual_resemblance* and *other*, the subtag *multiturn* was added, which was assigned when a participant reported selecting a certain object based on how often certain features have appeared earlier in the experiment. *multiturn* is close to *variability*, with the difference being that in the case of *multiturn*, the participant is trying to look for some pattern in speaker behavior, as opposed to guessing.

Whereas in Chapter 3 *odd_one_out* was a type of the *other_reason* category, for the current experiment, we made it its own category for the following reason. In the study described in this chapter, we are interested in relating reference game performance to cognitive traits. *odd_one_out* reasoning reflects a fundamentally different strategy, which results in never selecting the target but does not reflect a lack of reasoning sophistication, unlike guessing. Therefore, we excluded participants whose explanation for either trial revealed that they were applying the *odd_one_out* strategy from all main analyses but we examined them more closely in the analysis of annotations. Likewise, we excluded participants whose explanation for either trial revealed that they misunderstood the experiment and were thus labeled *misunderstood_instr* from all analyses.

Finally, the tag *mistake/changed_mind* was added to label explanations which reported the participant making a mistake. Participants with that label, as well as

Annotation tag	Subtag	Example
correct_reasoning		The participant could have chosen red to uniquely determine the other triangle, but didn't, indicating it must be the blue triangle.
guess		It's a 50/50 guess between the two triangles.
guess	screenloc	There are two triangles, and there's no way for me to know which one is correct, so I chose the first one.
guess	variability	I knew it would be red and the last image was a triangle, so I tried a circle this time.
other_reason	salience	I feel like red is more likely as it's a brighter colour.
other_reason	preference	[I picked the blue triangle and not the red one] because I like blue more than red.
other_reason	visual_resemblance	The image must be red and the paint splodge looks more like a circle.
other_reason	multiturn	I think red images are shown more often so I have a better chance of being right if I choose the red triangle.
other_reason	other	There is no answer option for the blue circle, so I would go for it if I were the previous participant (as a reason to select the distractor).
odd_one_out		Because it was the only shape that was not red.
misunderstood_instr		Because the color blue is not a choice therefore it must be the green or red shape.
unclear		Because it felt right.
mistake/changed_mind		I meant to select the blue one. I clicked on the red one by accident.

Table 4.1: Annotation tags used to label participants' reported reasoning strategies with examples.

those whose responses were labeled *unclear* were kept in the regression analysis.

We would expect the following reported strategies to correspond to the three reasoning types: participants who belong to the L_0 reasoning class should report a *guess* strategy in both the simple and the complex conditions; those in the L_1 class should report *correct_reasoning* in the simple condition and *guess* in the complex condition;

and those in the L_2 class should report *correct_reasoning* in both conditions.

4.2.3 Investigated individual differences

We hypothesize that three individual difference traits may modulate people's pragmatic responding on the reference game: working memory capacity (WMC), Theory of Mind and logical reasoning. A review of related work on the effect of these individual differences for other pragmatic phenomena is included in the theoretical background (Section 2.4). Here, we motivate our choice of these three individual differences.

We hypothesized that working memory capacity (WMC) would have a positive effect on implicature derivation in the reference game given that studies on other pragmatic phenomena, such as scalar implicature (De Neys & Schaeken, 2007; Antoniou et al., 2016), found a role of working memory. If pragmatic reasoning in the reference game is effortful and taxes cognitive resources, we may likewise expect an effect of WMC to emerge. In an early-stage pilot study for this experiment (Mayn & Demberg, 2022), we did not find an effect of WMC, as measured by Operation Span. However, there are several potential reasons why a true effect of WMC did not come out. First, most of the participants of the pilot had a mean score of 0.93 out of 1 on OSpan, meaning that most of their participants were at ceiling, which means that there may not have been sufficient variability for the effect to emerge. Second, the effect may have been too small to be identified given the small sample size (N=68). Third, in the pilot, we only used one measure for WMC, which may have made it harder to detect the effect if the measure is noisy and additionally measured factors unrelated to working memory.

We also investigated the effect of Theory of Mind (ToM), also sometimes called mentalizing (Trott & Bergen, 2019) or mindreading, on pragmatic responding in the reference game. It may be instrumental in pragmatic inferencing, as inferencing may involve taking the speaker's perspective into account and reasoning about why they chose the utterance they did. Studies such as Fairchild & Papafragou (2021) and Trott & Bergen (2019) found an effect of ToM on other pragmatic phenomena, such as scalar implicatures and indirect requests.

Finally, we hypothesized that the ability to reason logically and detect patterns may modulate performance in the reference game. Both in our early-stage pilot in the reference game (Mayn & Demberg, 2022) and in a study of individual differences in atypicality inferences (Ryzhova et al., 2023), we found evidence of the effect of logical reasoning, measured by the Cognitive Reflection Test (Frederick, 2005) and Raven's Progressive Matrices (Raven & Court, 1938). The extent to which CRT and RPM tap the same versus distinct constructs has been debated in the literature. Some studies suggest that the CRT is primarily a cognitive measure strongly associated with intelligence (Otero et al., 2022; Welsh, 2022), whereas others argue that it captures distinct traits, including the tendency towards miserly processing (Toplak et al., 2014)

and the ability to inhibit the intuitive response (Campitelli & Gerrans, 2014). Given the results of the Principal Component Analysis, in this chapter we mostly treat the two measures as reflecting the same ability, which we call *logical reasoning*, but in the discussion we return to this distinction and additionally examine the effects of these two individual difference measures separately.

Other individual differences, such as vocabulary knowledge (Matthews et al., 2018) and print exposure (Scholman et al., 2020), as well as personality and neurodevelopmental factors (Verhoeven & Vermeer, 2002; Yang et al., 2018; Skalicky, 2019), have been shown to play a role in pragmatic processing for other phenomena. We do not investimate them here given the nonlingusitic nature of this task and because of the focus of this thesis on cognitive as opposed to personality or neurodevelopmental traits. These may also play a role, but we leave their investigation to future work.

4.3 Experiment

This study was conducted with the approval of the Ethical Review Board (ERB) of the Department of Computer Science at Saarland University (Approval No. 20-10-2). All participants gave written consent according to the policies set forth by the ERB.

4.3.1 Participants

300 self-reported native English speakers with an approval rate of 95% or higher were recruited via the platform Prolific. Recruitment took place on December 6 and 7, 2022, and the first study session was conducted on the same days due to the efficiency of the platform. In order to ensure variability in the sample, participants were recruited from 3 streams: 100 participants who have a university degree, 100 participants who are in the process of obtaining a technical or community college or a university degree, and 100 participants who completed high school or did not have formal qualifications. 222 participants returned to participate in the second session.

4.3.2 Procedure

The experiment was conducted in two sessions. In the first session, which took place on December 6 and 7, 2022, participants completed the reference game, Backward Digit Span, Cognitive Reflection Test, and Raven's Progressive Matrices, in that order. The first session took participants about 25 minutes to complete, for which participants were compensated £3.16. Only participants who did not meet the exclusion criteria for any of the tests from the first session were reinvited for the second session, which took place one week later, between December 12 and 15, 2022. In the second session, participants completed the OSpan task, the Short Story Task, and the Reading the Mind in the Eyes task, in that order. The second session took

participants about 45 minutes to complete, for which participants were compensated £6.44.

Individual difference measures

Operation Span (OSpan) As one of the measures of working memory capacity, we used the automated online version of operation span (OSpan) based on Unsworth et al. (2005). Participants were asked to verify the correctness of a total of 75 basic math equations, presented in sets of 3 to 7, while holding letters in memory, which appeared one by one at the end of each equation. Participants were instructed to not write anything down and to not use any aids for solving the equations or memorizing the letters. The optimal set length and equation difficulty was determined by conducting a set of pilot studies. This was done to avoid a ceiling effect. First, participants saw the left hand-side of the equation, e.g., $(4 + 3) \times 2 =?$. Then, once they had solved it, they clicked a button to proceed to the next screen, on which a number was presented, and participants had to judge whether the number is the correct solution of the equation or not by clicking on one of two buttons. After that, a letter appeared on the screen for 800 ms, after which either the next equation was shown or, if it was the end of the set, the participant was prompted to type in all the letters. To ensure that participants did not take too long on the equations, a brief training phase in the beginning was used to estimate the time a participant needed to solve the equations. The maximum allowed time was set to the mean + 2.5 standard deviations of the time the participant took during the training phase. If the participant did not verify an equation during that time, a timeout message appeared for 800 ms, after which the letter was shown. After each set, the participants were asked to type in the letters in the order in which they had appeared. The score was computed using the Partial Credit Unit Score (PCU) method from Conway et al.(2005) as the mean proportion of letters within each set which were recalled correctly (i.e., the correct letter was recalled in the correct position). Participants whose accuracy on the equations was below 80% were excluded from analysis.

OSpan has demonstrated strong psychometric properties. Internal consistency, measured by Cronbach's α, typically ranges fron .7 to .8 (Conway et al., 2005; Unsworth et al., 2005; Kane et al., 2004), while test-retest reliability over multiple weeks has found to be within the .7-.83 range (Conway et al., 2005; Unsworth et al., 2005). Construct validity is supported by positive correlations with other working memory measures, such as Reading Span and Counting Span (Kane et al., 2004), as well as with fluid intelligence tasks like Raven's Progressive Matrices (Unsworth et al., 2005). These findings indicate that OSpan is a reliable measure of working memory capacity.

Backward Digit Span (BDS) (Jensen & Figueroa, 1975) was included as as the other measure of working memory capacity. Participants were presented with digit sequences of increasing length, from 2 to 8. There were 2 trials for each sequence

length. Digits were presented one by one for 1000 ms. At the end of each set, participants were asked to type in the digit sequence in reverse. Participants were instructed to not write down the digits and not use any aids to help them remember. Copying, pasting, or adding characters anywhere except the end of the string was disabled in order to ensure that participants were retrieving the digit string from memory and performing the reversal in their heads.

If the participant got both sequences of the same length wrong, the experiment ended. The score is the highest sequence length for which a participant got both trials right. Sequence length ranged from 2 to 8. Participants who did not recall both trials correctly for any length were excluded from analysis.

BDS has been shown to have adequate psychometric properties. Internal consistency, measured by Chronbach's α, ranges from .59 to .93 (Waters & Caplan, 2003; Gignac & Weiss, 2015), and test-retest reliability in the range between .64 and .83 has been reported (Blackburn & Benton, 1957; Waters & Caplan, 2003; Gignac & Weiss, 2015). De Paula et al. (2016) reported moderate split-half reliability of .598. Construct validity of BDS is supported by positive correlations with other memory span measures (Waters & Caplan, 2003), as well as with fluent intelligence (Gignac & Weiss, 2015).

While OSpan is considered a complex span task, since it involves performing a secondary task while keeping information in memory, BDS is typically considered a simple span task. Simple span tasks are thought to primarily engage short-term memory and emphasize maintenance rather than processing, whereas complex span tasks like OSpan involve both storage and processing components (Redick & Lindsey, 2013; Engle et al., 1999). However, it has been argued that BDS can be considered a measure of working memory as opposed to only short term memory, as it requires the manipulation of information in memory and not just storage, unlike the forward digit span (Redick & Lindsey, 2013; Kessels et al., 2008; Gignac & Weiss, 2015; Gignac et al., 2019). Despite these distinctions, simple and complex span tasks have been consistently shown to be closely related. Meta-analyses suggest that both types of tasks reflect the same underlying cognitive mechanisms – rehearsal, storage and updating – but to different degrees, with simple spans placing greater emphasis on storage (Unsworth & Engle, 2007; Colom et al., 2006). The Principal Component Analysis conducted in this study supports this view, as both OSpan and BSD loaded strongly onto the same principal component.

Raven's Progressive Matrices (RPM) (Raven & Court, 1938) were included as a test of logical reasoning ability. On each trial, participants were presented with an incomplete pattern and 6-8 options out of which they needed to select the missing piece to complete the pattern.

We used a shortened 10-question version of the full Raven's Standard Progressive Matrices Test (RSPM), since a shortened 9-question version of the test has been found to correlate almost perfectly with a full-length RSPM test (Bilker et al., 2012). The

10-question version consisted of the 9-question Form A test from Bilker et al.(2012) and the most difficult item from the Advanced Progressive Matrices Test (RAPM) (item 36), which was added as the final question to account for possible ceiling effects, following Scholman et al.(2024).

The questions were presented in ascending order of difficulty. While participants were solving the task, a timer remained on the top of the screen, which showed the total amount of time participants had spent on the task up until that point. Participants were told that while their time isn't limited, they should avoid thinking too long. The score was computed as the number of correctly solved items.[1]

The 9-item version of RSPM, which is identical to the version we use with the exception of the last item, has been shown to have good reliability (Cronbach's α=.80, compared to α=.96 for the full 60-questions RSPM) (Bilker et al., 2012).

Cognitive Reflection Test (CRT) (Frederick, 2005) was used as the other measure of logical reasoning ability. CRT has been argued to measure reflexivity, that is, how likely someone is to override the first intuitive response and engage in deeper reasoning (Toplak et al., 2014; Pennycook et al., 2016), although recent studies suggest that it is primarily a cognitive measure strongly associated with intelligence (Otero et al., 2022; Welsh, 2022). It has been shown to correlate robustly with measures of intelligence and in addition to that be a unique predictor in heuristics and biases tasks (Toplak et al., 2011).

In our early-stage pilot study (Mayn & Demberg, 2022), we found a correlation of r=0.45 between RPM and CRT, and Ryzhova et al. (2023) found that their participants' scores on RPM and CRT loaded onto the same principal components. Therefore we hypothesize that CRT also taps into logical reasoning ability and will load onto a principal component corresponding to logical reasoning in our study as well.

We used a 10-question version of CRT, with 6 critical questions, 3 verbal and 3 computational, and 4 decoy questions, presented in random order. This version was developed by Scholman et al. (2024) and consists of questions from earlier versions of CRT (Baron et al., 2015; Primi et al., 2016; Sirota & Juanchich, 2018; Thomson & Oppenheimer, 2016; Toplak et al., 2014). The critical questions were "trick questions", where the intuitive answer does not match the correct answer. An example of a question is "If you were running a race, and you passed the person in 2nd place, what place would you be in now?". The intuitive answer is 1st place but the correct answer is 2nd place.

[1]In the preregistration, we stated that we would exclude participants who had taken an abstract reasoning test of this type in the last 6 months. However, due to experimenter error we did not ask participants that question. However, we collected their responses about whether they recalled seeing any of the questions before. 20 out of 300 participants responded "Yes" or "Not sure", 8 of whom did not meet exclusion criteria for other tasks. The results section of the chapter reports analyses which include those 8 participants. Excluding these participants does not substantially change the posterior estimates of the effects or their credible intervals. Regression results with those participants excluded are reported in Appendix C.

After each question, participants were asked whether they had seen it before. The score was computed as the proportion of correctly answered previously unseen critical questions. Since CRT has been shown to be affected by familiarity (Stieger & Reips, 2016; Thomson & Oppenheimer, 2016), participants who reported having seen 3 or more critical questions were excluded from analysis.

CRT has adequate psychometric properties. Moderate to good internal consistency has been reported by previous studies (Cronbach's α ranging from .51 to .88 for different versions (Toplak et al., 2014; Thomson & Oppenheimer, 2016; Sirota & Juanchich, 2018; Primi et al., 2016; Baron-Cohen et al., 2001)). Construct validity has been supported by positive correlations with other measures of rational thinking, such as the belief bias task, temporal discounting, denominator neglect and conditional reasoning, as well as with Need for Cognition (Toplak et al., 2014; Sirota & Juanchich, 2018; Thomson & Oppenheimer, 2016; Primi et al., 2016).

Short Story Task (SST) (Dodell-Feder et al., 2013) was used as a measure of Theory of Mind. It has been found to predict comprehension of indirect requests (Trott & Bergen, 2019) and reading time of texts containing irony (Valles-Capetillo et al., 2022).

In the task, participants read the short story "The End of Something" by E. Hemingway. After reading the story, participants were asked if they had read the story before, and if so, how well they remembered it and whether they had read it in school or for pleasure. Then participants were asked to briefly summarize the story, after which they answered 8 open-ended questions about characters' mental states along with 5 comprehension questions. Participants typed their answer to each of the comprehension questions in a textbox. An example of a mental state question is "Why does Marjorie take the boat and leave and what is she feeling at that moment?".

The response to each question was scored 0-2 using the rubric from Dodell-Feder et al. (2013). The total score was the sum of the scores for the individual mental state inference questions. Thus, the possible range of scores was 0-16. Participants who scored less than half (i.e., fewer than 5 out of 10 points) on the comprehension questions were excluded from analysis.

Given the broad range of content covered by individual items in the SST, its internal consistency is moderate, with Cronbach's α ranging from .27 to .51 for comprehension questions and from .54 to .63 for mental state reasoning questions (Dodell-Feder et al., 2013; Giordano et al., 2019; Jarvers et al., 2025). Construct validity has been supported by positive correlations with the Reading the Mind in the Eyes Test (RMET) and the fantasy subscale of the Interpersonal Reactivity Index (IRI) (Davis, 1983; Giordano et al., 2019). Additionally, positive associations have been reported with both verbal and non-verbal intelligence by Dodell-Feder et al. (2013) and Giordano et al. (2019), although these findings were not replicated by Jarvers et al. (2025).

Reading the Mind in the Eyes Test (RMET) (Baron-Cohen et al., 2001) was

used as the second measure of Theory of Mind. It has been suggested that RMET may tap more into emotion recognition than into mentalizing (Oakley et al., 2016). It has also been observed that RMET is strongly related to vocabulary knowledge and moderately related to emotion perception and self-reported cognitive empathy (Olderbak et al., 2015), suggesting that RMET may be reflective of those individual differences, possibly to a larger extent than of mentalizing (Quesque & Rossetti, 2020).

Despite these limitations, we opted for using it as a ToM measure for several reasons. It is the most commonly used test for Theory of Mind, and unlike most ToM tests, it does not show ceiling effects for neurotypical adults (Yeung et al., 2024). Also, it has been shown to correlate with other Theory of Mind measures, including our other measure, the Short Story Task (Dodell-Feder et al. (2013) report a correlation of r=.49, 95% CI = [.28, .68]).

In this task, participants are presented with 36 pictures of people's eyes, presented one-by-one, and are asked to select the emotion which the person is experiencing from the 4 available options. The score is the number of correctly answered questions.

According to a review of 1,461 studies (Higgins et al., 2024), the psychometric properties of the RMET are generally reported to be mixed but generally acceptable. Internal consistency, measured by Cronbach's α across 132 studies, ranged from .17 to .96, with 57.9% of estimates falling below .70. McDonald's ω, reported in 18 studies, ranged from .43 to .85 (mean = .65), with 61.1% of values below .70. Split-half reliability, assessed in 8 studies, ranged from .50 to .75, with 55.6% of estimates below .70. The validity of the RMET is commonly supported by its ability to discriminate between neurotypical adults and those with autism (151 studies), schizophrenia (29 studies), and between males and females (19 studies). Convergent validity has been assessed using other ToM measures, as well as measures of empathy and emotion recognition. Discriminant validity has been demonstrated by the RMET's lack of correlation with fluid intelligence, verbal ability (though see Olderbak et al. (2015)), and executive functioning (Higgins et al., 2024).

4.3.3 Results

We first examine the results of the reference game to verify that the effect of condition (simple and complex implicature) observed in the comprehension experiment of F&D and in Experiment 4 of Chapter 3 holds for our sample and to confirm that there is individual variability.

Next, we conduct an analysis where we add the individual differences we collected as predictors to the regression model to examine whether they explain performance on the reference game.

Participant exclusion

Of the 300 participants who participated in the first session, 8 were excluded for accuracy below 80% on the unambiguous trials on the reference game. Further 27 participants were excluded because they were familiar with 3 or more of the 6 critical questions on the Cognitive Reflection Task. Further 8 participants were excluded because they did not correctly recall both sequences for any length.

The remaining 257 out of 300 participants were reinvited to participate in the second session, and 222 of them did. Of those 222, 194 participants passed exclusion criteria for all of the tasks.

Thus, 292 participants had passed the exclusion criterion for the main task, 194 of whom also passed exclusion criteria for all individual difference tasks.

Annotations of participants' explanations revealed that a total of 25 participants (23 of whom were part of the 292 retained for the main task, and 15 of whom were part of the 194 retained for the individual differences analysis) misunderstood the instructions of the main task, and 15 participants (all of whom were part of the 292 retained for the main task, and 12 of whom were part of the 194 retained for the individual differences analysis) applied the odd-one-out (inverse reasoning) strategy not captured by any of the RSA models. Those people were also excluded from analysis.

As a result, 254 participants (149 female, 105 male; mean age = 35.6 years (SD = 13.6, 3 participants did not report their age)) entered the replication analysis without individual differences, and 167 participants (97 female, 70 male; mean age = 36.3 years (SD = 13.6, 2 participants did not report their age)) entered the main analysis with individual differences.

Replication of effects from F&D (2016)

First, we examine the performance on the reference game task to make sure that we replicate the effect of condition found in F&D, as well in Experiment 4 of Chapter 3, which used the same stimuli as we use in the current experiment: both studies found better performance on simple trials compared to complex trials and better performance on complex trials compared to the chance baseline.

Figure 4.2 shows average proportions of choices (target, competitor and distractor) for each trial type. Since the proportions are averaged across participants, they may not necessarily sum to one for each trial type. As expected, performance on ambiguous trials is at chance (mean proportion of target responses = 0.50, SD=0.17) since there are two identical objects and which one of them is target is determined via a coin flip. Performance on unambiguous trials is at ceiling (mean proportion of target responses = 0.99, SD=0.03), which indicates that participants understood the task. Finally, performance on the critical simple condition (mean proportion of target choices = 0.72 (SD=0.27)) appears to be better than on the complex condition (mean proportion of

target choices = 0.62 (SD=0.21)). We will confirm that the difference between the two conditions is significant in a regression analysis.

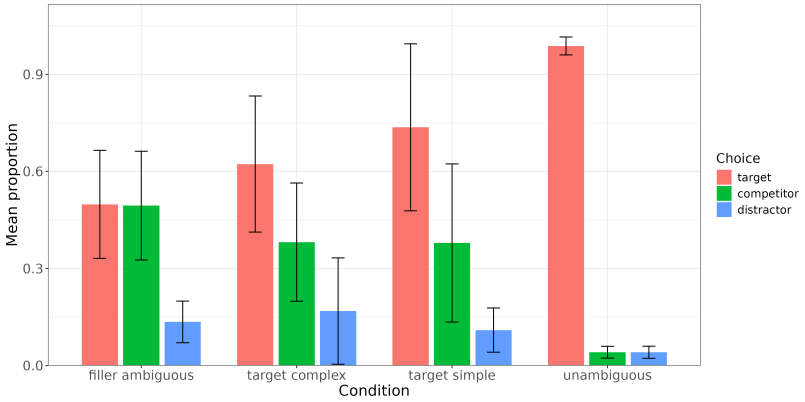


Figure 4.2: Average proportions of responses in the reference game per utterance type. Error bars represent standard deviations.

These results are consistent with the original study by F&D, whose results, however, showed a somewhat numerically stronger difference between the simple and the complex conditions (mean proportion of target choices in the simple condition = 0.79 (SD=0.19, in the complex condition = 0.57 (SD=0.18)[2].

F&D observe somewhat higher performance on the simple condition and lower performance on the complex condition than the current study. That is not very surprising given the findings of the previous chapter that performance on the simple condition is inflated for the original stimuli because of reasons related to the stimuli and unrelated to reasoning.

Interestingly, we observe more individual variability in our sample than F&D and in Experiment 4 of Chapter 3. Figure 4.3 shows proportion of target selection in the two critical conditions by participant. We find that the effect is quite graded and, interestingly, that quite a few participants appear to perform somewhat better on the complex condition than on the simple condition, corresponding to the points above the diagonal. That is something that none of the RSA models predict. We will look at those individuals more closely in our analysis of annotations. However, consistent with the RSA's predictions, there are many more participants on or below the diagonal than there are above it (70% vs. 30%); and, of course, some variability can be attributed to guessing.

Next, in order to verify the difference in performance on the simple and complex conditions, we conduct Bayesian logistic regression using the *brms* package in R (Bürkner, 2017). Following F&D, we exclude unambiguous trials, as well as trials on

[2]These averages were computed by us using F&D's data. In their paper, F&D report pooled proportions of choices (77% target choices in the simple condition and 57% in the complex condition; Fig 2 in F&D). However, we opted for reporting average proportions here instead as they provide a clearer view of the data's spread.

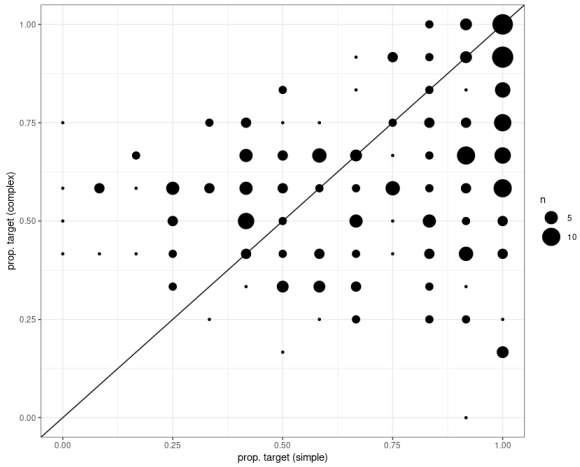


Figure 4.3: Proportion of target choices on simple and complex trials by participant. Dot size denotes number of participants.

which participants chose the distractor (which amounts to 1.2% of trials), from this analysis. Trial correctness, a binary variable, is regressed onto condition (helmert-coded, simple vs. rest and complex vs. ambiguous), trial number (mean-centered), participant age (mean-centered), target position (left, middle or right, dummy-coded with left as the reference level), message type (sum-coded, -1 = shape, 1 = color), and the interaction of trial number with condition. The random effect structure included per-participant and per-item random intercepts and per-participant random slopes for condition, message type, and trial number.

For all effects, very wide weakly informative priors centered at 0 were set. The full model equation and a discussion of the priors is included in Appendix C. We ran four chains of 6000 iterations each, with the first 1000 iterations of each chain discarded as warm-up.

To assess the evidence for the effects, we examined whether the credible intervals for the parameter estimates included zero. If the credible interval did not include zero, we concluded that there was a meaningful relationship between the predictor and the dependent variable. The regression results are reported in Table 4.2.

The model confirms that participants' performance on the simple condition was better than on the complex condition ($\hat{\beta}$ = 1.30, 95% CrI = [1.03, 1.58]), and performance in the complex condition was above chance ($\hat{\beta}$ = 0.62, 95% CrI = [0.44, 0.80]). There was a very small positive effect of trial number ($\hat{\beta}$ = 0.00, 95% CrI = [0.00, 0.01]). Since the effect size is so small, we believe that we can still say that the assumption that participants retain one reasoning complexity type throughout the experiment holds. Participants were less likely to interpret the message correctly if the message was a color as opposed to a shape ($\hat{\beta}$ = -0.08, 95% CrI = [-0.14, -0.01]), and they were more likely to select the target if it was in the middle as opposed to on the left ($\hat{\beta}$ = 0.36, 95% CrI = [0.24, 0.49]), and on the left compared to the right ($\hat{\beta}$ = -0.20, 95% CrI = [-0.33, -0.08]).

	Model without IDs (*N*=254)		**Model with IDs (*N*=167)**	
Effect	**Estimate**	**95% CrI**	**Estimate**	**95% CrI**
Intercept	**0.63**	[0.49, 0.77]	**0.46**	[0.23, 0.68]
condition1 (simple vs. rest)	**1.30**	[1.03, 1.58]	**0.97**	[0.60, 1.34]
condition2 (complex vs. ambig)	**0.62**	[0.44, 0.80]	–	–
trial number	**0.00**	[0.00, 0.01]	**0.01**	[0.01, 0.02]
participant age	-0.00	[-0.01, 0.00]	**-0.02**	[-0.03, -0.01]
msgtype (color vs. shape)	**-0.08**	[-0.14, -0.01]	-0.10	[-0.23, 0.02]
targetpos (middle vs. left)	**0.36**	[0.24, 0.49]	**0.53**	[0.33, 0.74]
targetpos (right vs. left)	**-0.20**	[-0.33, -0.08]	**-0.21**	[-0.40, -0.03]
condition1 : trial	-0.01	[-0.01, 0.00]	-0.01	[-0.02, 0.00]
condition2 : trial	0.01	[-0.00, 0.01]	–	–
logical reasoning	–	–	**0.22**	[0.06, 0.38]
memory	–	–	0.06	[-0.10, 0.22]
ToM	–	–	**0.26**	[0.10, 0.42]
condition1 : logical reasoning	–	–	**0.45**	[0.13, 0.78]
condition1 : memory	–	–	0.24	[-0.07, 0.56]
condition1 : ToM	–	–	0.16	[-0.16, 0.48]

Table 4.2: Regression model output for the model without individual differences (N=254) and for the model with individual differences (N=167). Effects for which the 95% CrI does not include 0 are bolded.

Overall, we see that we very closely replicate the original study by F&D, and that the effect sizes for condition are very close to the those reported in the original study (see Table 4.3 for model results of F&D).

Analysis of annotations

Next, we inspect the strategies reported by participants on the main task and how those strategies related to participants' performance.

First, we examine the relationship between their reported strategy and their tendency to select the target (i.e., pragmatic responding). For this analysis, we also include the strategies *misunderstood_instr* and *odd_one_out*, which we excluded for the regression analysis. Thus, this analysis only excludes the 8 participants whose accuracy on the unambiguous condition was below 80%, with N=292 participants

	F&D (2016) (*N*=51)		**Exp. 4 of Ch. 3 (*N*=60)**	
Effect	**Estimate**	**SE; *p***	**Estimate**	**95% CrI**
Intercept	-0.15	0.11; 0.18	0.20	[-0.20, 0.61]
condition1 (simple vs. rest)	**1.28**	0.12; <0.0001	**1.24**	[0.67, 1.86]
condition2 (complex vs. ambig)	**0.44**	0.13; <0.001	**0.42**	[0.03, 0.85]
trial number	0.00	0.00; <0.3	**0.01**	[0.00, 0.02]
msgtype (color vs. shape)	-0.02	0.12; <0.85	0.22	[0.03, 0.85]
targetpos (middle vs. left)	**1.28**	0.14; <0.0001	**0.38**	[0.11, 0.64]
targetpos (right vs. left)	**0.74**	0.13; <0.0001	-0.04	[-0.31, 0.23]
condition1 : trial	0.00	0.01; <0.9	**-0.02**	[-0.03, -0.00]
condition2 : trial	0.01	0.01; <0.9	**0.02**	[0.00, 0.03]

Table 4.3: Regression model output from Franke & Degen (2016) (F&D; N=51) and from Experiment 4 of Chapter 3 of this thesis (N=60). Effects for which p<0.05 for F&D and effects for which the 95% CrI does not include 0 for our experiment from Chapter 3 are bolded.

remaining.

Figure 4.4 shows the mean proportions of target selections (averaged across participants) and the frequency of each annotation tag. We can see that the two annotation tags that were by far most common in both conditions are *correct_ reasoning* and *guess*. In the simple condition, there is a higher proportion of *correct_ reasoning* responses than in the complex condition (50.7% (148 out of 292) vs 34.2% (100 out of 292)) and a lower proportion of *guess* responses (25% (73 out of 292) vs. 44.2% (129 out of 292). We also see that the mean proportion of target selections corresponding to the *correct_ reasoning* tag is higher in the simple condition than in the complex condition (0.90 (SD=0.13) vs. 0.75 (SD=0.21)), suggesting that even when people reported a correct strategy in the complex condition, at the population level, they applied it less consistently than in the simple condition. We see that the strategy labeled *odd_ one_ out* corresponds to a relatively low proportion of target responses: 0.15 (SD=0.13) in the simple condition and 0.36 (SD=0.29) in the complex condition. That is because *odd_ one_ out* strategy picks out the object which is unlike the other two, which is most often the distractor. *misunderstood_ instr* only appears in the simple condition, since it involves misunderstanding the meaning of unavailable messages, and in the complex condition, messages for all features of the three objects are available. *misunderstood_ instr* has an average proportion of target responses of 0.23 (SD=0.17), meaning that people who misunderstand the instructions might occasionally select the target but most of the time select the competitor or the distractor. *other_ reason* responses – salience, preference and other strategies fall into that category – are very infrequent and correspond to about half of target selections (0.44 (SD=0.19) in the simple condition and 0.55 (SD=0.20) in the complex condition), which is close to guessing.

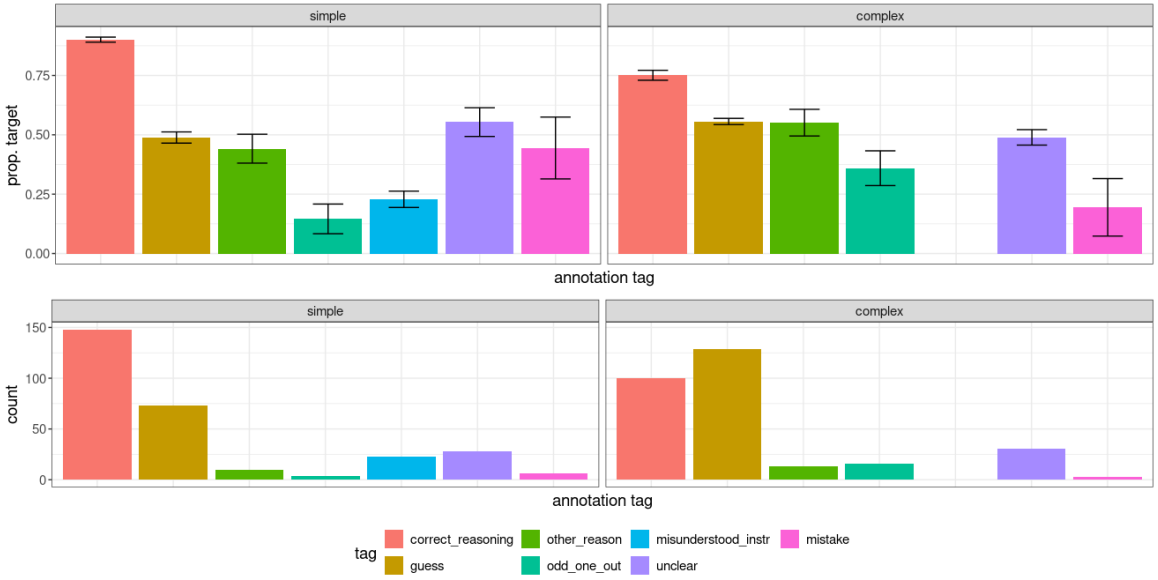


Figure 4.4: Mean proportion of target selections per annotation tag and the frequency of that tag in the two critical conditions. Error bars represent standard deviations.

Overall, at least at the population level, reported strategies appear to match average proportion of target selections, with a high proportion of target selections for *correct_reasoning*, target proportion of near 50% for *guess* and a low proportion of target selections for *misunderstood_instr* and *odd_one_out*. We will examine the relationship between individual participants' performance and their reported strategies in more detail in the next section.

Grouping participants using LPA and reported strategies

As can be seen in Figure 4.3, there is a lot of individual variability in our data and participants' performance appears to be gradient as opposed to building three clear clusters corresponding to the three reasoning types (L_0, L_1 or L_2). However, since the three reasoning types serve as the theoretical explanation of the variation we observe, we start by conducting a clustering analysis where we examine whether such clusters can be identified and whether they have distinct profiles with respect to the individual differences.

We assessed the clustering tendency of the reference game data using the Hopkins statistic(Hopkins & Skellam, 1954), which is an established measure for distinguishing data with meaningful cluster structure from randomly distributed data (Banerjee & Dave, 2004; Panayirci & Dubes, 1983). Values close to 0.5 indicate randomness, whereas values close to 1 indicate a cluster structure. Since the Hopkins statistic is stochastic, we computed it 10 times with different random seeds. The mean value was 0.93 (SD=0.05), indicating a strong clustering tendency. This suggests that conducting a clustering analysis on this data is appropriate.

We perform latent profile analysis using the *tidyLPA* package in R on participants' average performance on the simple and the complex conditions in order to identify clusters of participants in our data for the main task. LPA identifies clusters while remaining agnostic to origins of those clusters. A 3-class model obtained a better fit to the data (AIC = -108.69, BIC = -73.31) than models with 1 (AIC = 1.55, BIC = 15.70) or 2 (AIC = -41.53, BIC = -16.77) classes. A 4-class model obtained a better fit than the 3-class model according to AIC but not according to BIC (AIC = -112.60, BIC = -66.61). Since we have a theoretical motivation to posit 3 groups of participants corresponding to the three RSA models, we examine the model with 3 classes.

Figure 4.5 shows the obtained classes. They correspond approximately to the predictions of the 3 idealized reasoning types. Participants whose performance on the simple is 0.6 or below are clustered together, roughly corresponding to the predictions of the L_0 model. Another cluster includes participants with scores above 0.6 on the simple condition and below about 0.7 on the complex, corresponding to being able to solve simple but not complex trials, as the L_1 model predicts. Finally, the third cluster represents participants with high performance on both conditions, corresponding to the predictions of the pragmatic listener model L_2.

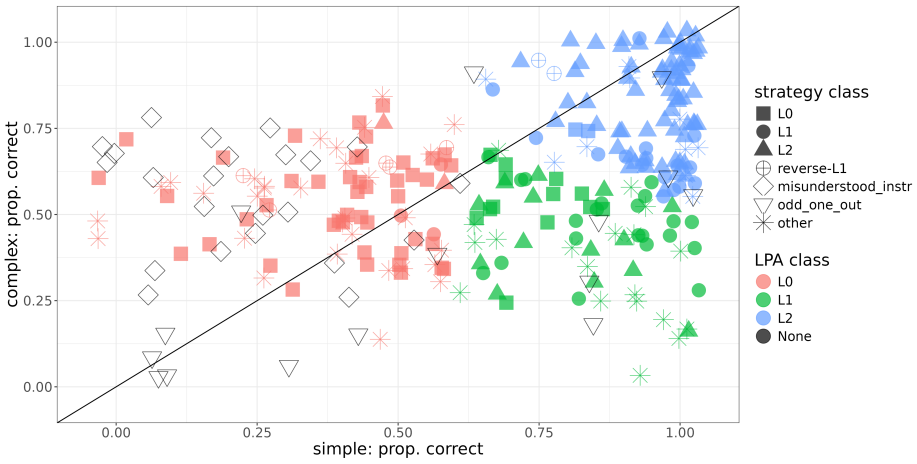


Figure 4.5: LPA classes vs. classes based on annotations.

How do the classes identified by LPA based on performance relate to strategies that participants reported? In addition to the class assigned based on performance, Figure 4.5 also shows classes based on strategy. If the participant reported guessing on both conditions, we assign them to the L_0 class based on their reported strategy. If their strategy was *correct_reasoning* in the simple condition and *guess* on the complex, they were assigned to the L_1 class, whereas if their strategy was labeled *correct_reasoning* in both conditions, they were labeled L_2. Note that participants assigned to the L_1-class based on annotations perform counterfactual reasoning about *the speaker's intent* ("if they had wanted to communicate X, they would have chosen the message Y"), but they only do so in the simple condition. We do not observe

explanations referring to the distribution of messages available to a literal speaker, as the literal interpretation of the L_1 RSA model would suggest. This suggests that the L_1 model should be viewed as a stable outcome capturing the difference in complexity between the two reference game conditions – only the simpler one of which can be solved correctly by participants belonging to this class – and not as an accurate processing-level account of people's actual reasoning process. We will test this hypothesis more directly in one of the experiments in Chapter 5, where participants will be told that they are interacting with a speaker who is explicitly literal, a simple computer program.

Additionally, there were a few people who reported guessing in the simple condition but reasoning correctly on the complex – something that no RSA model predicts given that any model which is able to solve the complex condition should also be able to solve the simple condition – those people are assigned to the *reverse-*L_1 class. Participants who employed a strategy other than *guess* and *correct_reasoning* on one or both conditions were assigned to the *other* class. Those strategies were: *other reason – salience* (5 in the simple condition, 3 in the complex condition), *other reason – preference* (4, 5), *other reason – multiturn* (2, 5), *other reason – visual resemblance* (0, 2), *other reason – other* (1, 2), *unclear* (32, 34), and *mistake/changed_mind* (7, 3).

Additionally, we visualize participants who misunderstood the instructions or applied the *odd_one_out* strategy. Those participants were not used for the regression analysis or for the LPA clustering.

First, we examine the relationship between classes assigned by LPA based on performance and those assigned based on reported strategies. Table 4.4 is a confusion matrix which compares the two sets of classes. Overall, there is a decent degree of consistency between the two sets of classes. 45 out of 55 participants (81.8%) (who could be classified based on annotations, i.e., not counting participants in the *other* annotation class) in the L_0 class based on LPA reported guessing as their strategy on both conditions. Only 21 out of 48 (43.75%) of participants who were assigned to the L_1 class based on LPA reported correct counterfactual reasoning in the simple condition and guessing on the complex condition. 13 participants (27.08%) reported guessing on both conditions, so their relatively high performance on the simple condition is presumably caused by lucky guessing, and 14 participants (29.17%) reported performing correct counterfactual reasoning on both conditions, despite their relatively low performance on the complex condition.

There are two possible explanations for this misalignment. Recall that participants were asked to reported their strategies at the very end, first for a simple trial and then for a complex trial. It could be that those participants figured out how to solve the complex condition during the task, resulting in lower performance on complex than on simple. It could also be that being asked to reflect on how they solved the simple condition made participants more likely to apply this reasoning to the complex condition on the final trial.

Finally, 70 out of 89 participants assigned to the L_2 class based on LPA (78.65%) reported applying correct reasoning in both conditions. 17 (19%) participants assigned to the L_2 class reported guessing on the complex condition, corresponding to the L_1 model, i.e., they were likely lucky in the complex condition.

	Annotation Classes					
LPA Classes	L0	L1	L2	Reverse-L1	Other	Total
L0	**45**	3	2	5	31	86
L1	13	**21**	14	0	19	67
L2	2	17	**70**	2	10	101
Total	60	41	86	7	60	254

Table 4.4: Confusion matrix comparing classes of participants based on LPA (rows) and based on annotation of reported strategies (columns).

Interestingly, the L_1 class is the least populous, both based on LPA (67 out of 254 participants, 26.38%) and based on reported strategies (41 out of 254, 16%) and also has the lowest consistency between classes based on performance and based on annotations. This contrasts with the findings of F&D, for whom L_1 was the largest class among their participants. One possible reason for that might be sampling chance since F&D had a much smaller sample size (51 participants, as opposed to 254 in the current study). Another reason may be related to the difference in stimuli. In Chapter 3, we showed that the original stimuli used by F&D, led to inflated performance in the simple condition. Thus, it may be that some people whose reasoning is actually more accurately described by the L_0 model were classified into the L_1 class. Some support for that hypothesis comes from comparing clusters obtained in Experiment 1 and Experiment 4 in Chapter 3, which use F&D's original stimuli and the geometric stimuli used in the current study respectively. In Experiment 1, there is a smaller L_0 class and a larger L_1 class, whereas in Experiment 4, it is the other way around.

Also, it is notable that there are quite a few people above the diagonal line, corresponding to higher performance on the complex condition than on the simple one, a pattern not predicted by the RSA. In particular, quite a few participants have a low proportion of target selections (below 0.3) on the simple condition. A large portion of those participants misunderstood the instructions, taking the fact that the target in the simple condition had an inexpressible feature to mean that it could not be referred to. There were presumably some unlucky guessers as well. Finally, there might have been participants who initially misunderstood the instructions but in the course of the experiment converged on a guessing strategy, thus reporting guessing at the end.

Individual difference measures

The summary statistics of the individual differences measures are reported in Table 4.5, and the correlation matrix, including the two conditions of the reference game, is included in Figure 4.6.

Measure	Mean	SD	Observed range	Possible range	Skewness	Kurtosis
OSpan	0.79	0.16	0.19-1	0-1	-1.33	4.69
BDS	4.75	1.51	2-8	0-8	0.34	2.39
RPM	5.38	2.14	0-9	0-10	-0.13	2.23
CRT	0.35	0.25	0-1	0-1	0.45	2.42
RMET	27.58	3.82	15-36	0-36	-0.59	3.44
SST	8.88	2.81	2-14	0-16	-0.23	2.32

Table 4.5: Summary statistics for the individual difference measures (N=167).

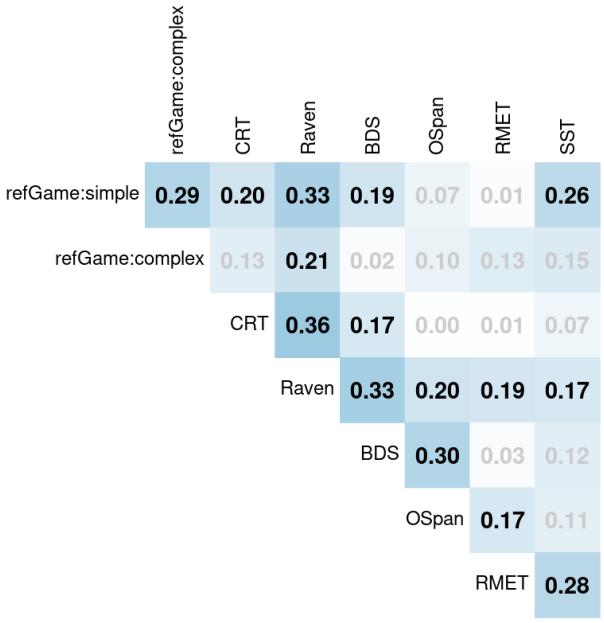


Figure 4.6: Correlations of the individual difference measures. Insignificant correlations are greyed out.

The distributional statistics of the individual differences show that there is variability in all of the measures that none of the measures are at ceiling.

The correlation matrix only reports significant correlations, corrected for multiple comparisons using FDR (False Discovery Rate).

The correlation matrix reveals that the measures which we selected to reflect same underlying construct are moderately correlated with each other (CRT and RPM at r=0.36, p<0.0001, BDS and OSpan at r=0.30, p=0.001, and RMET and SST at r=0.28, p=0.003).

Examining correlations of the individual differences measures with average performance on the two reference game conditions – simple and complex – reveals a distinct pattern for the two conditions. While performance on the complex condition only correlates with RPM (r=0.21, p=0.02), the simple condition correlates positively with a larger number of measures: both measures of logical reasoning, CRT (r=0.20, p=0.03) and RPM (r=0.33, p=0.0001), as well as with BDS (r=0.19, p=0.03) and SST (r=0.26, p=0.002). When we conduct the regression analysis later, we include interaction terms of condition with individual difference measures to further statistically test this pattern.

The only other significant correlations are relatively weak correlations between CRT and BDS (r=0.17, p=0.046) and between OSpan and RMET (r=0.17, p=0.046), and between RPM and all other individual difference measures. The correlation between CRT and BDS is not surprising given that digit span tasks have been shown to have a linear positive relationship with general intelligence (Gignac & Weiss, 2015), which is related to reflective logical thinking which CRT taps into. The significant, though weak, correlation between OSpan and RMET is more surprising as we would not expect RMET to rely on working memory. This correlation may be attributed to the fact that the two tasks were both included in Session 2, and this correlation may simply be reflective of participants' attention and disposition on that day. The fact that all cognitive tasks correlate significantly with RPM is not very surprising since all these tasks involve an element of reasoning and pattern recognition.

Since we collected two measures per construct (RPM and CRT for logical reasoning ability, OSpan and BDS for working memory, and RMET and SST for Theory of Mind), we conduct principal component analysis (PCA) in order to test whether these measures will indeed load onto three components representing these three constructs. We opted for PCA as opposed to CFA because some of the tasks which are supposed to tap into the same construct are not necessarily very similar or strongly correlated. For instance, while the Short Story Task and the Reading the Mind in the Eyes test are both said to measure Theory of Mind, SST involves interpreting intention from text and RMET involves emotion recognition, and they correlate at r=0.28.

We conducted a Principal Component Analysis (PCA) with varimax rotation on the standardized individual differences measures (centered and scaled to a 0–1 range). Three principal components (PCs) were retained based on eigenvalues greater than 1; the remaining components had eigenvalues of 0.74 or lower. The selected PCs accounted for 23%, 23%, and 22% of the total variance, respectively; all together, they accounted for 68% of the variance. The factor loadings are presented in Table 4.6. The components align closely with the intended cognitive constructs: PC1 reflects logical reasoning ability (CRT and RPM), PC2 reflects Theory of Mind (RMET and SST), and PC3 reflects working memory (OSpan and BDS). In each case, the two measures which are thought to tap into that construct had loadings above 0.70, while the remaining measures loaded weakly ($\leq$0.35, most below 0.20), indicating that the components represent their respective constructs with minimal overlap. We will

therefore use the obtained factors as predictors of participants' performance on the reference game.

	PC1 (Logical reasoning)	PC2 (ToM)	PC3 (Memory)
CRT	**0.87**	-0.03	-0.07
RPM	**0.70**	0.23	0.33
RMET	-0.02	**0.82**	0.08
SST	0.14	**0.75**	0.04
OSpan	-0.12	0.17	**0.84**
BDS	0.35	-0.06	**0.73**

Table 4.6: PCA factor loadings.

Before progressing to a continuous analysis, we look at the distribution of the three individual difference constructs for the three reasoning type clusters obtained by LPA from the previous section. Boxplots of the distributions of memory, logical reasoning and Theory of Mind for each of the three clusters identified by LPA are shown in Figure 4.7. While the distribution is fairly wide for all measures, we can see a general increasing trend from L_0 to L_2 for all three measures, where the mean memory, logical reasoning and ToM score for L_0 is lower than for L_2. This pattern appears clearest for logical reasoning and least clear for memory.

To verify these visually evident patterns, we conduct a simple ordinal regression using the *ordinal* package in R, with the assigned reasoning class (L_0, L_1 and L_2) as the dependent variable and the three individual difference constructs as predictors. The model reveals a positive effect of logical reasoning (β =0.58, SE=0.15, p<0.001) and of Theory of Mind (β =0.38, SE=0.15, p=0.01). The estimate for memory is positive but the effect does not reach significance (β =0.17, SE=0.15, p=0.24).

Thus, belonging to a more sophisticated reasoning class is associated with better logical reasoning ability and Theory of Mind skills. When we examine **??**, it appears that there may also be groupings between classes such that L_1 and L_2 are closer together in terms of logical reasoning than L_0, and that L_0 and L_1 are closer to each other in terms of ToM than L_2. However, we do not find statistical evidence for that distinction, as the Brant test indicated that neither predictor violated the proportional odds assumption (omnibus $\chi^2(3)$ =3.00, p=0.39).

The observed differences between the identified classes in terms of Theory of Mind are in line with the findings of Battaglini et al. (2025), who used a data-driven approach to identify groups of responders in a metaphor interpretation task. One of the two groups they identified, which they term *mentalizers*, used more cognitive and affective terms in their metaphor interpretation, suggesting heavier ToM use. Their mentalizers group may correspond to the L_2 class we identify, which is associated with higher ToM scores.

Next, we will conduct a full continuous analysis, which additionally includes the relationship between the individual differences and the two conditions in the reference

game separately, as well as covariates and random effects.

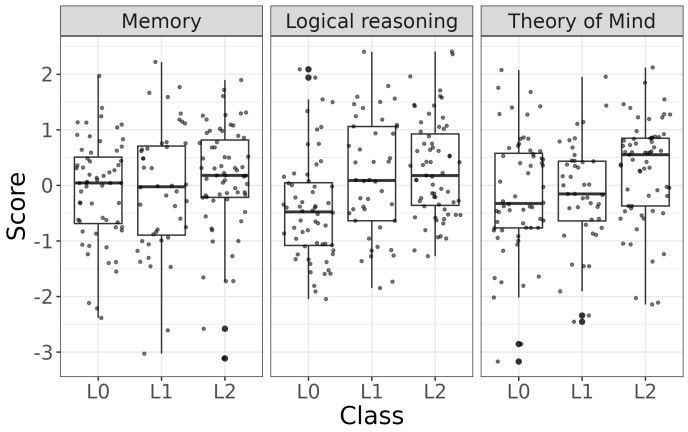


Figure 4.7: A boxplot showing the distribution of the three individual differences for each reasoning class identified by LPA. The y-axis represents the composite score for the corresponding individual difference obtained using PCA. Ordinal regression revealed the effect of logical reasoning and ToM but not of memory on reasoning class.

Regression analysis with individual difference measures

We again conduct Bayesian logistic regression. For this analysis, we keep only critical trials (simple and complex), since ambiguous trials were included in the original model only as a chance baseline to compare the complex condition and there is no theoretical reason to believe that individual differences should affect performance on the ambiguous trials. Trial correctness, a binary variable, is regressed onto condition (dummy-coded with complex as reference), trial number (mean-centered), participant age (mean-centered), target position (left, middle or right, dummy-coded with left as reference), message type (sum-coded, -1 = shape, 1 = color), the interaction of trial number and condition, main effects of individual differences (logical reasoning, memory and ToM) and interactions of individual differences with condition. Like before, the random effects structure included per-participant and per-item random intercepts and per-participant random slopes for condition, message type and trial number.

For all effects, very wide weakly informative priors centered at 0 were set. The full model equation is included in Appendix C. We ran four chains of 6000 iterations each, with the first 1000 iterations of each chain discarded as warmup. The regression results are reported in Table 4.2.

Main effect of condition persists: participants performed better on simple trials than on complex trials ($\hat{\beta} = 0.97$, 95% CrI = [0.60, 1.34]). There was again a small positive effect of trial number ($\hat{\beta} = 0.01$, 95% CrI = [0.01, 0.02]). Like in the first model, participants were more likely to choose the target if it was in the middle

compared to the left ($\hat{\beta}$ = 0.53, 95% CrI = [0.33, 0.74]) and on the left compared to the right ($\hat{\beta}$ = -0.21, 95% CrI = [-0.40, -0.03]). Unlike the model without individual differences, this model does not show an effect of message type ($\hat{\beta}$ = -0.10, 95% CrI = [-0.23, 0.02]) but it does show a small negative effect of age, such that older participants are less likely to respond correctly ($\hat{\beta}$ = -0.02, 95% CrI = [-0.03, -0.01]).

In terms of individual differences, there is a positive main effect of logical reasoning ($\hat{\beta}$ = 0.22, 95% CrI = [0.06, 0.38]) and ToM ($\hat{\beta}$ = 0.26, 95% CrI = [0.10, 0.42]), meaning that participants with higher logical reasoning and ToM selected the target more often. There is no main effect of memory ($\hat{\beta}$ = 0.06, 95% CrI = [-0.10, 0.22]).

There is also a positive interaction of logical reasoning with condition, whereby the effect of logical reasoning is stronger in the simple condition ($\hat{\beta}$ = 0.45, 95% CrI = [0.13, 0.78]), but no interaction of memory or ToM with condition. Figure 4.8 visualizes the effects of logical reasoning and ToM in the two implicature conditions.

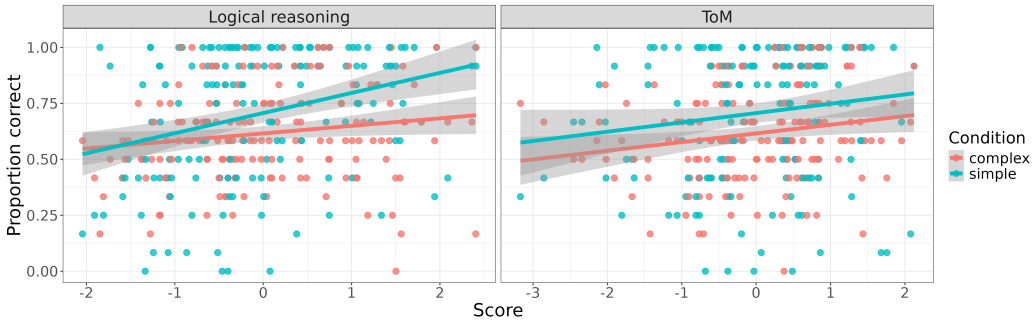


Figure 4.8: Proportion of correct responses of individual participants in the two implicature conditions as a function of logical reasoning and ToM.

4.4 Discussion

In this chapter, we conducted the pragmatic reference game task from F&D on a much larger sample and systematically investigated substantial individual variability in performance, relating it to individual differences in cognitive traits.

We identified groups of participants using Latent Profile Analysis and found, consistent with F&D, that they approximately correspond to the predictions of the three idealized reasoning types formalized by three RSA models of different reasoning depth – literal listener L_0, pragmatic listener reasoning about a literal speaker L_1, and Gricean listener L_2. Unlike F&D, for whom only a small proportion of their participants fell into the L_2 class, in our data, that is the largest identified class. This could be related to the difference in the stimuli – whereas F&D used images of mosters, robots and accessories for their study, we opted for the more abstract shapes and colors since in the experiments described in the previous chapter we found that the original stimuli were more susceptible to biases. It could also be that in F&D's sample there were so few participants who could solve the task due to sampling chance,

as their sample size was significantly smaller than ours (51 vs. 254 participants for the main task). Additionally, at the end of the experiment, we asked participants to report their strategy for one simple and one complex implicature trial as a way to gain additional insight into how participants performed the task and how strategy relates to performance. We found that overall, there was a decent degree of alignment between participants' performance and their reported strategy: most participants assigned to the literal L_0 class reported guessing on both types of implicature trials and most participants assigned to the Gricean L_2 class reported applying correct counterfactual reasoning for both trial types. However, there was some misalignment, presumably related at least in part to lucky guessing (i.e., having high performance on a condition despite applying a guessing strategy) or to figuring out how to solve the critical trials during the task. In the regression model, we find a small effect of trial number, suggesting that some participants may have learned how to solve the implicatures during the task. Finally, there were two groups of participants who could be identified based on annotations of their reported strategy who used different strategy which lead them to have a very low performance: one group of participants misunderstood the meaning of available messages, and the other applied inverse reasoning, e.g., interpreting the message "red" to mean "the only *non-red* object". Those participants were excluded from the LPA and regression analyses since the relationship between strategy and performance is different for them than for the rest of the participants.

Having established that there is sizeable variability in performance which reflects application of different strategies, we investigated the source of that variability by relating it to measures of cognitive traits. In addition to the reference game, we collected six individual differences measures belonging to three constructs: logical reasoning (Cognitive Reflection Test and Raven's Progressive Matrices), working memory (Operation Span and Backward Digit Span) and Theory of Mind (Reading the Mind in the Eyes and the Short Story Task). We then conducted a principal component analysis and found that the tasks indeed loaded into the corresponding components. We observed an effect of logical reasoning: participants with better logical reasoning skills responded more pragmatically. This effect was more pronounced for simple implicature trials compared to complex trials. We also found an effect of Theory of Mind, whereby participants with better mentalizing skills performed better on the reference game. We did not find an effect of working memory.

Notably, the observed effects are additive in nature: i.e., the effect of logical reasoning persists after WMC and Theory of Mind are controlled for, and the same holds for the effect of Theory of Mind. This is important to stress since the three cognitive constructs we investigated are not entirely independent – working memory and fluid intelligence, which is related to our logical reasoning component, are known to correlate and to draw on higher-level executive processes, such as attention and disengagement (Conway et al., 2003; Shipstead et al., 2016). Theory of Mind has also been shown to be related to intelligence and executive function: previous work

has observed a weak positive relationship between RMET and SST and fluid intelligence measures (RPM, as well as standardized tests), suggesting that mentalizing as measured by these tasks draws on general cognitive abilities (Baker et al., 2014; Coyle et al., 2018). Our findings suggest that logical reasoning and mentalizing have distinct contributions to pragmatic processing.

One could ask whether the significant effect of logical reasoning ability on reasoning strategy might be related to the fact that the reference game is a nonlinguistic, somewhat puzzle-like task and whether this effect of logical reasoning would generalize to linguistic pragmatic tasks. In a study I collaborated on (Ryzhova et al., 2023), we found an effect of logical reasoning, also measured using Raven's Progressive Matrices and Cognitive Reflection Test, in a linguistic atypicality inference task, where better reasoners were more likely to interpret a redundant utterance pragmatically. This suggests that logical reasoning may modulate pragmatic performance beyond nonlinguistic reference games. However, Heyman & Schaeken (2015) found that performance on the Cognitive Reflection Test did not predict the rate of scalar implicature derivation for their participants. It is therefore possible that logical reasoning is needed for novel settings, whereas for more conventionalized expressions the reasoning may be amortized (White et al., 2020). If this were the case, then there would more likely be an effect of logical reasoning for novel metaphors than conventionalized metaphors, for instance. The effect of logical reasoning ability on other pragmatic tasks could be investigated in future work.

Interestingly, the observed effect of logical reasoning was stronger in the simple implicature condition than in the complex implicature condition. Given that performance was generally lower on the complex condition, it is possible that the complex condition was too complex for some people to solve even despite good reasoning skills, whereas in the simple condition, better reasoners were more likely to select the right approach to the task.

The fact that an effect of Theory of Mind emerges suggests that the framing of the reference game as a communication game, where participants are told that they are interpreting messages sent by a participant of a previous experiment, was sufficient to induce perspective-taking and reasoning about why the speaker might have said what they did. Notably, the tasks we used for measuring Theory of Mind, Reading the Mind in the Eyes and the Short Story Task, are both quite different from the reference game, and were conducted in a different session. The fact that an effect still emerges suggests that there is an element of mentalizing about the speaker to the pragmatic reasoning involved in the reference game.

We did not observe an effect of working memory capacity (WMC), which is not particularly surprising given mixed evidence of the role of WMC in other pragmatic phenomena. Since we did not observe a direct effect of memory, we hypothesized that an indirect effect of memory might emerge if working memory is employed for supporting logical reasoning or Theory of Mind computations – that is what Fairchild & Papafragou (2021) observed for scalar implicatures, metaphors and indirect re-

quests. Since the principal components which we used as predictors are decorrelated, in an additional analysis, we computed composite scores of each individual difference construct by summing up z-scores for the corresponding tasks and used those as predictors instead. We then compared a full model with all three individual difference predictors and models with one or both of logical reasoning and Theory of Mind taken out to see if an effect of memory emerges, which would provide indirect evidence for a possible role of WMC. In the full model with composite scores, exactly the same effects were observed as in the model with principal components which we reported in the chapter. Moreover, even in the model where both logical reasoning and Theory of Mind were taken out, the credible interval for the effect of WMC contained zero, meaning that we do not observe even indirect evidence for the role of working memory on this task.

We opted for using principal components for our main analysis since individual tasks may be more prone to task impurity (Friedman et al., 2008). However, given the observation we made when examining correlations of the individual differences with the two reference game conditions – that only Raven's Progressive Matrices correlated significantly with the complex condition while a larger number of measures (both CRT and RPM, as well as SST and BDS) correlated with the simple condition, we conducted an additional analysis where we included the logical reasoning and ToM measures individually as predictors instead of the principal components. A further motivation is that the individual difference measures we selected for each construct tap into distinct constructs in addition to the shared construct: it has been suggested that CRT taps into reflective thinking and ability to override the intuitive response (Pennycook et al., 2016; Toplak et al., 2014), whereas RPM is a measure of fluid intelligence and abstract reasoning; RMET is thought to tap emotion recognition (Quesque & Rossetti, 2020), while SST may be more representative of ability to reason about others' mental states. The model estimates are reported in Appendix C. In the model with individual predictors, the main effect of CRT and interactions of condition with RPM and SST come out; there is no effect of RMET. This might suggest that reflective thinking and abstract reasoning both make a distinct contribution to pragmatic processing, and that cognitive Theory of Mind, i.e., ability to understand other people's mental states and thoughts, is more relevant to this task than emotion recognition. That would not be surprising since this a pragmatic reasoning task in which emotions should not play a large role.

There is ample evidence that different pragmatic tasks vary in the degree to which they require logical reasoning, working memory and mentalizing (see e.g. Bambini & Lecce (2025)) and that a more differentiated, task-specific view of the role of individual differences in Pragmatics is necessary – indeed, the fact that studies with different pragmatic tasks yield different findings is evidence of that. In addition, it has been argued that the interpretive strategy which is employed on a given task – and different strategies may tax cognitive resources to different extent – is also dependent on the comprehender's profile (Katsos, Napoleon and Kissine, Mikhail, 2025). Our findings

that different individuals recruit different strategies for solving the reference game and that those strategies are predicted by Theory of Mind and logical reasoning align with that view.

We would like to briefly discuss where the difference in complexity between simple and complex implicature trials might stem from. As we noted earlier, the L_1 model is able to solve the simple trials better than the complex ones, and, consistent with that model, the majority of our participants have better performance on the simple trials (Figure 4.3). Additionally, we observed a stronger effect of logical reasoning ability for simple implicature trials. Taken literally, the L_1 model's ability to solve simple trials is based on reasoning not about the speaker but about the available messages: it assumes that the speaker is literal, and the listener reasons that since there is only one way to refer to the target and two ways to refer to the competitor, and the probability mass is split between them, the message is twice as likely to be referring to the target. In an experiment which will be described in the next chapter, we showed that when participants were presented with an explicitly literal speaker, they failed to reason about alternative messages and behaved as L_0s and not as L_1s themselves, suggesting that reasoning about the distribution of available signals comes less naturally to people than reasoning about the speaker's intent. Indeed, when we examine participants' reported strategies for the simple condition in this experiment, they also overwhelmingly refer to the speaker's intent – that they would have used a different message if they had meant a different referent – as opposed to the distribution of available messages. This suggests that L_1 grounds the complexity of simple implicatures themselves rather than accurately describing the reasoning process used by participants to derive them. A processing-level modeling approach may be needed to construct a plausible model of how participants solve simple vs. complex implicatures. In a co-authored paper which is not part of this dissertation (Duff et al., 2025), we propose a model of the pragmatic reference game in ACT-R, where pragmatic reasoning needed to solve simple and complex implicatures – which in formal probabilistic models is captured by different recursion depth – is associated with distinct interpretation strategies, which differ in their expected utility.

We also observe that, in the simple condition, it is sufficient to pay attention to the target and the competitor to solve the implicature since the distractor is completely irrelevant to the computations. In the complex condition, on the other hand, the distractor needs to be taken into account when deriving the inference, which means that reasoning involving more objects needs to be performed.

Reference games are widely used for testing predictions of Rational Speech Act models, with the claim that this minimal signaling game captures some important properties of pragmatic communication using language "in the wild" (Frank et al., 2016). However, reference games are in many ways unlike communication using language, and it is a warranted question to what extent findings in this setting generalize to language use. The effect of Theory of Mind which we observe suggests that reasoning about the speaker's perspective, which is assumed to be an essential part of

pragmatics of language, is at least in part preserved in the reference game setting. Also, exactly because reference games have been so widely used in probabilistic pragmatics, it is valuable to better understand their properties and their relationship to individual difference in cognition. Insights gained from this individual difference analysis can be useful for developing processing-level models of pragmatic reasoning.

4.5 Conclusion

In this chapter, we investigated whether the variation in performance on the pragmatic reference game can be explained by individual differences in cognitive traits. We found that two individual differences, logical reasoning and Theory of Mind, are related to reference game performance, while there was no evidence of an effect of working memory capacity. These findings suggest that the variation we observe is stable and that it reflects differences in pragmatic reasoning ability.

By identifying cognitive traits which modulate reasoning in the pragmatic reference game, the research discussed in this chapter provides a foundation for processing level models. In a paper I co-authored (Duff et al., 2025), we build on these findings and propose a sketch of a processing-level model that associates distinct strategies with different reasoning types and makes predictions about their time course.

Notably, while the variation in reference game performance is theoretically motivated by the predictions of three distinct RSA models, the variability in participants' performance in our data appears to be quite gradient. This is not necessarily a contradiction since these three reasoning types are idealized, and the RSA framework can accommodate deviations from them through hyperparameters like the temperature parameter α and the cost term. However, this suggests a picture where there are these three broad categories of reasoners but the boundaries between those categories may be flexible. This interpretation is supported by the substantial alignment between participants' reasoning types based on performance and their self-reported strategies. Much of the observed misalignment arises from participants who are classified as intermediate L_1 reasoners based on performance, but who report strategies corresponding to one of the other two classes.

More generally, the L_1 model warrants further investigation. Taken at face value, it describes a listener's reasoning about a literal speaker, and the derivation of simple implicatures is achieved by reasoning not about the speaker's intent but about the relative probabilities of alternatives. However, we this type of reasoning is not reflected in participants' explanations, suggesting that the L_1 model should be thought about more as a way of formalizing problem complexity than as an accurate description of people's actual reasoning process. We investigate this further in one of the experiments in the next chapter, where participants are tasked with reasoning about an explicitly literal speaker, a simple computer program which randomly selects any true message to send, thus corresponding to an S_0 speaker model. We ask whether

participants will derive an inference based on the distribution of available alternatives or whether they will fail to do so and behave like literal listeners themselves. We find that the latter is the case.

The results presented so far suggest that pragmatic reasoning in this paradigm is relative stable within individuals, as evidenced by the predictive relationship between reference game performance and cognitive traits. In the remaining chapters of this thesis, we turn to the role of conversational context in shaping how this reasoning is deployed. In particular, we ask whether participants' tendency to derive implicatures is conditioned on their beliefs about the speaker's pragmatic competence, and whether certain situations are more conducive to deeper pragmatic reasoning than others. These questions are addressed in the next chapter, which focuses on the role of speaker identity in implicature derivation.

Chapter 5

Effects of speakers' perceived pragmatic competence on inferences

The previous chapter demonstrated substantial variability in pragmatic reasoning across individuals and showed that this variability is predicted by cognitive traits, adding to a growing body of work indicating that the assumption of perfect rationality often does not hold in practice. While Chapter 4 focused on differences between comprehenders, it left open an important question: how is such variability in pragmatic competence accommodated during communication?

In this chapter, we shift perspective and investigate whether comprehenders' pragmatic inferences are influenced by the identity and perceived pragmatic competence of the speaker. We ask whether people's beliefs about a speaker's pragmatic sophistication influence whether a pragmatic inference is drawn, and how strong that inference is. Prior work, reviewed in Section 2.5 of the theoretical background, suggests that characteristics of the speaker – such as their language fluency, knowledge and persona – can affect the interpretation of their utterances. Building on this literature, we examine whether comprehenders adapt their pragmatic reasoning depending on the perceived pragmatic competence and cooperativity of the speaker.

There also remains a question about the flexibility of pragmatic reasoning types identified in Chapter 4. While the results of the previous chapter suggest that individuals' reasoning types are relatively stable and that they correspond to distinct cognitive profiles, it remains unclear whether depth of reasoning is sensitive to contextual expectations about the speaker: comprehenders may reason more or less deeply depending on how pragmatically competent or cooperative they perceive the speaker to be.

To address these questions, this chapter investigates pragmatic interpretation for a range of different speaker types who are hypothesized to vary in their presumed reasoning sophistication: children, adults, artificial agents, and explicitly literal speakers. We also explore the implications of our findings for cognitively plausible models of

pragmatic reasoning and propose a sketch of how this reasoning about the speaker's pragmatic competence can be modeled in the RSA framework.

This chapter is based on, and in parts identical to, the following publications:

- **Mayn, A.**, Loy, J. E., & Demberg, V. (2025). Beliefs about the speaker's reasoning ability influence pragmatic interpretation: Children and adults as speakers. *Open Mind*, 9, 89-120.
- **Mayn, A.**, Loy, J., & Demberg, V. (2024). Capable but not cooperative? Perceptions of ChatGPT as a pragmatic speaker. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).
- **Mayn, A.**, Duff, J., Bila, N., & Demberg, V. (2025). People do not engage in ad-hoc reasoning about alternative messages when interacting with a literal speaker. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 47).

Data availability

Preregistrations, data and analysis scripts for Experiments 1 and 2 are available at `https://osf.io/f5nmv/`.

Preregistration for Experiment 3 is available at `https://osf.io/7ycrq`. Data and analysis scripts are available at `https://github.com/sashamayn/perceptions_of_chatgpt_cogsci2024/`.

Preregistration, data and analysis scripts for Experiment 4 are available at `https://osf.io/erbn3/`.

5.1 Introduction

In the previous chapter, we observed substantial variability in participants' performance on the pragmatic reference game, which was predicted by their logical reasoning ability and Theory of Mind. In this chapter, we turn to the question how such differences in pragmatic reasoning sophistication are accommodated by the comprehender: Are comprehenders sensitive to such differences and adjust their interpretations in a speaker-specific manner, or do they have a generic model of the speaker, who they may or may not assume to be perfectly rational? There is experimental evidence that, in some situations, the latter is the case: that people's interpretations of the same utterances change depending on the identity of the speaker. Related work on the effects of speaker identity in Pragmatics and language comprehension more generally was discussed in detail in Section 2.5, so here we only give a brief overview of the most relevant papers.

Previous work showed that speaker persona (chill vs. nerdy) affected the interpretation of the precision of quantity expressions (Beltrama & Schwarz, 2021, 2024).

Language proficiency and accent have been found to influence people's perceptions of speaker trustworthiness (Ip & Papafragou, 2021), as well as their interpretations of metaphor (Puhacheuskaya & Järvikivi, 2025) and irony (Puhacheuskaya & Järvikivi, 2022). Grodner & Sedivy (2011) found that speaker reliability influenced the way scalar adjectives were interpreted: their participants interpreted scalar adjectives contrastively by default (i.e., "a short glass" means that a tall glass must be present in the shared view as well), but this interpretation did not arise when participants were told that the speaker had an impairment "that caused language and social problems".

Previous studies also showed that people may interact differently with a human and a computer partner, which may be related to the perception of the agent's competence, cooperativity, or both. Branigan et al. (2011) found that participants exhibited greater lexical alignment – repeating the interlocutor's word choices – when interacting with a computer than with a human partner, and with an older computer compared to a more modern one, presumably reflecting participants' beliefs about the limited communicative competence of computers. Fischer (2005) found that people gave simpler, more streamlined instructions with more limited vocabulary to a non-verbal robot than to a verbal robot. Interestingly, a recent visual perspective taking study by Loy & Demberg (2023) found the opposite result, where participants were more egocentric with a computer partner. This discrepancy in findings may be explained by a shift in people's perception of competence and cooperativity of artificial agents related to technological developments.

The studies described above show that *explicit information about a speaker's identity or characteristics* can influence comprehension and pragmatic inference. The influence of such processes based on comprehenders' beliefs about the speaker is the focus of this chapter. However, relevant characteristics of the speaker are not always explicitly marked. Speakers may differ in their pragmatic reasoning ability, as shown in the previous chapter, or be under cognitive load, and such differences may only become apparent through interaction. A growing body of work shows that comprehenders can construct speaker-specific models *based solely on behavioral evidence*, without external cues to speaker identity. People derive distinct speaker-specific representations of the speaker's use of uncertainty expressions such as "might" and "probably" (Schuster & Degen, 2020; Schuster et al., 2023) and pick up speaker-specific communicative styles (Loy & Demberg, 2024). Ryskin et al. (2019) and Gardner et al. (2021) followed up on Grodner & Sedivy (2011) and showed that, given sufficient exposure, similar effects arise from interaction with a speaker who uses scalar adjectives non-contrastively.

This chapter presents four experiments examining how comprehenders adapt their pragmatic interpretations across a range of speaker types based on their beliefs about those speakers' reasoning ability and cooperativity. We investigate whether ambiguous messages are more or less likely to be interpreted pragmatically – i.e., as implicatures rather than literally – depending on the speaker's perceived level of pragmatic competence. The first two experiments investigate the differences in the interpreta-

tion of utterances produced by child versus adult speakers. The third experiment extends this approach to artificial agents, focusing on how participants reason about the competence and cooperativity of large language models as speakers. The final experiment investigates participants' interpretations in the case of an explicitly literal speaker.

We discuss the implications of these findings for theories of perspective taking and propose a model of comprehenders' reasoning about the speaker's pragmatic competence in the RSA framework.

5.1.1 Experimental paradigm

For the experiments described in this chapter, we again use the reference game paradigm. The critical trials in this experiment correspond to the simple condition in Franke & Degen (2016) and in Chapters 3 and 4. We decided to use only the simple implicature trials in the experiments presented in this chapter for the following reason. Our hypothesis is that listeners may perform meta-reasoning about the pragmatic competence of the speaker, which may then influence the listener's interpretation. However, if the task is too complex for the listener to solve, they are unlikely to additionally engage in meta-reasoning. Therefore, we only use simple implicature trials for this task which participants have been shown to be more successful at solving.

While experiments in Chapters 3 and 4 used a forced choice task between the three possible referents, in the experiments described in this chapter, we ask participants to distribute 100 points between the three referents using sliders with the aim of collecting more informative graded judgments.

An example of a critical trial is shown in Figure 5.1. The message (the color red), if taken literally, matches two of the objects – the square and the triangle. However, one may resolve the ambiguity by reasoning about the possible alternative utterances the speaker could have used. If they had wanted to refer to the red triangle, they could have used the unambiguous message "triangle". Since the speaker chose not to use this message, one may reason that they meant to refer to the red square (the pragmatic target) because there is no unambiguous way to refer to that object as "square" is not an available message. However, one may also either fail to or decide not to draw that implicature and interpret the message literally, resulting in assigning equal probability to the target and to the competitor. Finally, the betting paradigm allows for capturing implicatures of varying strength: for instance, assigning a rating of 90 to the target plausibly reflects a stronger implicature than a rating of 70. In the experiments described in this chapter, we investigate whether the identity of the speaker influences whether an implicature is drawn and how strong that implicature may be.

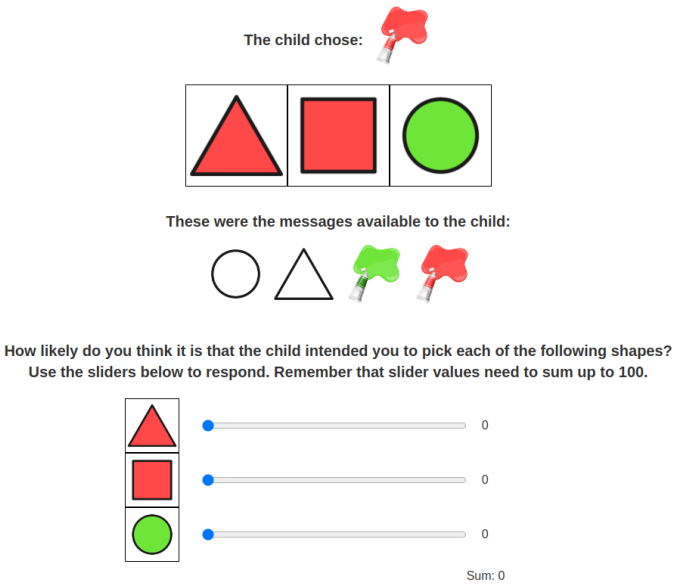


Figure 5.1: An example of a critical trial in the child speaker condition (Experiments 1 and 2).

5.2 RSA model of reasoning about the speaker's pragmatic competence

5.2.1 Model sketch

We hypothesize that, at the population level, participants will be more likely to interpret the same message pragmatically as opposed to literally when it is uttered by a speaker whose pragmatic reasoning ability they believe to be more advanced. When the speaker's pragmatic competence is believed to be insufficient for deriving an optimal message, we expect participants to be more likely not to derive an implicature or to derive a less strong one. Here, we propose a model sketch in the Rational Speech Act framework that formalizes this idea.

A detailed review of the RSA framework is included in Section 2.6, and predictions of models of different recursion depths are included in Section 2.7. Recall that the literal speaker S_0 will use any true message with equal probability, whereas the pragmatic speaker S_1 is a more complex model, which reasons about the literal speaker L_0 and seeks to maximize informativity while minimizing utterance cost.

We propose to model the effect of the speaker's pragmatic competence on inferences as the listener's uncertainty about the speaker's recursion depth. When faced with a speaker, the listener will be uncertain about whether they are dealing with a literal speaker S_0 who sends literal messages, or a pragmatic speaker S_1 who always chooses optimally. The listener's beliefs about the reasoning depth of the speaker they are dealing with are expressed with the weight λ:

$$L_2^{belief-driven}(o|u,\lambda) \propto S_{weighted}(u|o,\lambda) \cdot P(o)$$
$$\text{where} \quad S_{weighted}(u|o,\lambda) \propto \lambda \cdot S_1(u|o;\alpha \to \infty) + (1-\lambda) \cdot S_0(u|o) \quad \text{and } 0 \leq \lambda \leq 1$$

If λ=1, the listener is confident that the speaker is fully rational and always selects the optimal utterance. At λ=0, the listener is confident that the speaker is literal and randomly selects a true message to send to refer to an object. For in-between λ-values, the listener has some uncertainty about the speaker's rationality. Derivation of the model's predictions for different λ values is included in Appendix D.

The mixture mechanism in the proposed implementation is inspired by the rational mixture mechanism which has been used to model spatial perspective-taking (Hawkins et al., 2021), but here it has a different interpretation. In Hawkins et al. (2021)'s model, the mixing weight w is shared between the two speaker perspectives, egocentric and allocentric, and the speaker chooses how heavily to weigh each of the perspectives based on an informativity-cost tradeoff. Here, the mixing weight λ reflects the *listener's* existing beliefs and potential uncertainty about the recursion depth of the speaker. Our proposed use of a mixture weight to represent uncertainty is closest to Schuster & Degen (2020)'s model of uncertainty expressions, whose expected pragmatic speaker model is a weighted average over different speaker models which differ in terms of thresholds and cost functions.

When λ=1, reflecting full listener certainty that they are dealing with a rational speaker, $L_2^{belief-driven}(target|u)$ will be 1 as well, whereas when λ=0, $L_2^{belief-driven}$ will essentially be equivalent to an L_1 model and $L_2^{belief-driven}(target|u) = \frac{2}{3}$, and for intermediate λ values, the probability assigned to the target will be between $\frac{2}{3}$ and 1. Let us explain why $L_2^{belief-driven}(target|u;\lambda = 0)$ is $\frac{2}{3}$ and not $\frac{1}{2}$. It is because the RSA assumes that the pragmatic listener is rational and, in addition to reasoning about the speaker rationality, also reasons about alternative utterances.

Let's again consider the example in Figure 5.1. If S_0 wants to refer to the red square, they will use the message "red" 100% of the time since that is the only literally true message. If they want to refer to the red triangle, they will choose the message "red" half of the time and the message "triangle" half of the time. Therefore, L_1 reasons that, if the message "red" is used, it is twice as likely to be referring to the red square. Hence, the model predicts that, even if the listener is convinced that the speaker is fully literal, the $L_2^{belief-based}$ model will assign a probability of at least $\frac{2}{3}$ to the target. In Experiments 1, 2, and 4 described in this chapter, we find that this is not always the case in practice and that people often fail to take probabilities of alternative messages into account when reasoning about a literal speaker, behaving more like literal listeners L_0 themselves. This is discussed in more detail later in the chapter.

5.2.2 Predictions for our experiments

We expect that individual participants will vary in terms of their beliefs, which can be expressed with different settings of λ, but also that participants will, on average, interpret messages uttered by an adult more pragmatically than those uttered by a child in Experiments 1 and 2, reflecting the assumption of greater pragmatic reasoning sophistication, corresponding to a higher λ. Since it is less clear what people's beliefs about artificial agents are, which is the subject of investigation in Experiment 3, we do not have a clear prediction of where the population-level λ would fall in relation to the adult and child conditions in the first two experiments.

While we expect people at the population level to reason about the speaker and the speaker's reasoning sophistication, we know from the findings of Franke & Degen (2016) and from the results of Chapter 4 that some people interpret messages literally and appear not to take the speaker's perspective into account at all, as described by the literal L_0 model. We expect that there will be some such participants in our experiments, but since participants are assigned to conditions randomly, there should be a roughly equal number of them in all conditions, which means that we should still be able to detect differences between conditions.

5.3 The influence of beliefs about the speaker's reasoning ability on pragmatic inferences: Child vs. adult speaker

In the first two experiments, participants are asked to identify the referent of a message which, they are told, was sent to them by a participant of a previous experiment. We manipulate the identity of the speaker between participants: in one condition, it is a 4-year-old child from a local kindergarten, while in the other condition, it is an adult. We hypothesize that when the speaker is believed to be a child, ambiguous messages will be interpreted less pragmatically than when the speaker is believed to be another adult as we expect people to have more uncertainty about whether 4-year-old children possess sufficient reasoning sophistication to act as pragmatic speakers.

Evidence that adults have beliefs about limitations of children's communication and reasoning abilities comes from studies in which participants were explicitly asked to give an estimate of the age at which children acquire communicative abilities. Becker & Hall (1989) collected beliefs about the age of acquisition of communicative behaviors that rely on the understanding of context and appropriateness, such as appropriate turn-taking and explaining the reason for making a request. They found that for most of these skills, participants believed them to be acquired around the age of 6. Miller et al. (1980) presented their participants with a Piaget task battery, which assessed abilities including simple perspective taking based on asymmetric perspective, as well as understanding of physical properties of objects such as conservation

of weight and height, and asked them to estimate at what age children acquire these abilities. The average estimated ages of acquisition of the two perspective-taking tasks, which involved realizing that the person sitting opposite of the child sees a picture from a different perspective, were 4.30 and 5.11 years. These findings suggest that adults have beliefs about pragmatic and communicative competence of a child at a particular age and that there would be uncertainty about a child's pragmatic competence at the age of 4. However, these studies asked participants about a child's abilities explicitly, whereas in our experiments, we ask whether adults will deploy their beliefs about children when interpreting signals purportedly sent by them.

5.3.1 Experiment 1

In a between-subjects design, participants were told they would interpret pictorial messages sent either by another adult or by a four-year-old child and indicated their beliefs about the intended referent by distributing 100 points between 3 objects – target, competitor and distractor – using sliders.

We hypothesized that there would be an effect of speaker identity: in the adult speaker condition, we expect participants to interpret ambiguous messages more pragmatically than in the child speaker condition, resulting in higher ratings assigned to the target on critical trials. Additionally, it is possible that target ratings in the unambiguous condition, where the message only matches the target, are higher in the adult speaker condition as well if participants have some uncertainty about the child being able to reliably select a message that matches one of the features of the referent even in the unambiguous context.

Participants

80 native speakers of English, 40 per speaker condition, with an approval rating of at least 95% were recruited via the crowdsourcing platform Prolific. 6 participants (3 in the adult condition, 3 in the child condition) were excluded because their average target rating on the unambiguous trials was below 80, suggesting that they may have misunderstood the task or randomly clicked through the experiment. 3 participants (1 in the adult condition, 2 in the child condition) were excluded because their reported strategy suggested that they had misunderstood the setup of the experiment. 1 participant in the adult condition was excluded because they had had a technical issue. New participants were recruited in their place, resulting in 40 participants in the adult speaker condition and 39 in the child speaker condition. There are 39 and not 40 participants in the child speaker condition because we excluded an additional participant upon closer inspection of the data once we'd already finished data collection. Excluding this participant from the analysis or including them does not affect any of the results.

Speaker identity manipulation

The critical manipulation was the identity of the speaker whose messages participants were asked to interpret. Participants were randomly assigned to either the adult speaker or the child speaker condition. In the adult speaker condition, we told participants that the messages had been sent by a participant of a previous experiment. In the child speaker condition, we used the cover story that we had recently conducted a study with 4-year-old children at a local kindergarten. In Experiment 1, the cover story included children being told that they were scientists communicating with an alien and selecting a message to send to the alien so that the alien could identify the correct object. We told participants in that condition that they would be interpreting messages that one of the children chose to send to the alien. In Experiment 2, we made the child speaker condition more comparable to the adult speaker condition by removing the mention of the alien and instead including a photo of a 4-year-old child.

We hypothesized that listeners may not derive an implicature or derive a less strong one if they believe that the speaker's reasoning ability may be insufficient for determining what the optimal message to send is. We hypothesized that participants would assign lower probability to the target in the child speaker condition because they would think that a 4-year-old's reasoning may not be sophisticated enough to reason that, if there are two messages that apply to the target and one of them is ambiguous while the other one is unambiguous, the unambiguous message is preferable.

Procedure

At the start of the experiment, participants briefly, for 3 trials (two completely unambiguous and one completely ambiguous), took the perspective of the speaker. This was done in order to make sure that participants understood the relationship between the messages and the referents and the fact that not all messages were available to the speaker.

The main task was a reference game described in Section 5.1.1. There were a total of 24 experimental trials, of which 8 were critical, corresponding to the simple implicature condition from Chapters 3 and 4, and 16 were unambiguous control trials. Each trial display consisted of the target, competitor and distractor, presented in random order. Trial order was randomized for every participant.

Of the 16 fillers, 4 were completely unambiguous, where the features of every object were unique, and 4 were completely ambiguous, meaning that they contained two identical objects which are equally likely to be the referent. The remaining 8 trials are unambiguous and have the same display as the 8 critical trials but the target was the competitor from the corresponding critical item (4 trials) or the distractor from the corresponding critical item (4 trials).

After completing the 24 experimental trials, each participant saw the first critical trial they had seen in the experiment again and, once they had given a response using sliders, the sliders were disabled and the question "Why did you decide to put the sliders in those positions?" appeared on the screen next to the sliders along with a textbox. Like in the previous chapter, we included this open-answer question to get an additional insight into how participants went about solving the task. Additionally, we are interested in whether the meta-reasoning about the speaker we are expecting is a conscious process. If so, we would expect it to be reflected in participants' explanations.

The whole experiment took participants about 10 minutes to complete.

Annotation of reasoning strategies

We annotated participants' responses to the question "Why did you decide to put the sliders in those positions?" using an annotation scheme based on the one introduced in Chapter 3. The annotation scheme is summarized in Table 5.1; the complete annotation scheme with more examples for each tag can be found in Appendix A. The annotated dataset is included in the repository with data and scripts.

Like in Chapter 3, *correct_reasoning* describes hypothetical reasoning about alternatives and *guess* denotes random guessing between the target and the competitor. In this chapter, we merge *salience* and *preference* into one tag *salience/preference* since the difference between these tags is quite subtle and not the focus of investigation of the current chapter. This tag was assigned to responses which indicated that a participant chose an object due to a preference for one shape or color over another, or because the shape or the color stood out to them. Unclear answers were labeled *unclear*.

Like in Chapter 3, the tag *misunderstood_instr* was assigned when a participant's reported strategy indicated that they had misunderstood the instructions of the experiment, taking the fact that a feature is inexpressible (e.g., that there's no message "square") to mean that an object with that feature cannot be referred to. Participants whose explanation was assigned this tag were excluded from all analyses, and a new participant was recruited in their place.

Answers indicating that a participant changed their mind were labeled *changed_mind* – a tag introduced in the previous chapter – and were not considered in the analyses of annotations since they indicated that the participant's post hoc reasoning did not match their reasoning in the moment.

None of the other tags present in the annotation scheme introduced in Chapter occurred in the experiments in this chapter.

In addition, we introduce one new tag in this chapter: *meta_reasoning*. It was assigned when the answer included reasoning about the alternative messages available to the speaker, as well as explicitly stated uncertainty about whether the speaker would have been capable of selecting the optimal message. An example from the

annotation tag	example
correct_reasoning	If the speaker had wanted to refer to the triangle, they could have used the message triangle, which would have been unambiguous.
meta_reasoning	I am in two minds if a 4-year-old would be smart enough to choose the triangle shape if the target is the red triangle, rather than the color red.
guess	There are 2 red objects, so it's 50-50.
salience/preference	The triangle may be more popular.
unclear	This is what I would have meant if I were the speaker.
changed_mind	I picked the triangle but now I think it's the square.
misunderstood_instructions	The square was not available so the triangle must be the target.

Table 5.1: Annotation scheme which was used to label participants' explanations.

child speaker condition is "I am in two minds if a 4-year-old would be smart enough to choose the triangle shape if the target is the red triangle, rather than the color red."

It should be noted that even if a participant's open-ended answer does not include meta-reasoning about the speaker, that does not necessarily mean that no meta-reasoning occurred. It may be that either this meta-reasoning is not entirely conscious, at least for some people, or they engaged in meta-reasoning but did not report it.

Results

First, we look at the average probability (out of 100 points) assigned to the target for the unambiguous, ambiguous and critical trials. Top panel of Figure 5.2 shows that in both the child and the adult speaker conditions, participants' performance was at ceiling on unambiguous trials and at chance between the two identical options on ambiguous ones. This strongly suggests that participants understood the task and that participants judged 4-year-olds to be capable at least of feature matching. Accuracy on ambiguous trials is at chance because on those trials, there are two identical objects and a coin is flipped to decide which of them is the target. On critical trials, average target probability is 70.8 (SE=1.3) in the adult speaker condition and 57.3 (SE=0.98) in the child speaker condition. This suggests that participants indeed tend to be more likely to draw the inference if they believe that the speaker's reasoning is sufficiently sophisticated to select the optimal message.

To verify this apparent effect, we fit a Bayesian regression model to the data using the *brms* package in R (Bürkner, 2017). For this analysis, we only used critical and unambiguous trials because ambiguous trials were mainly included for counterbal-

ancing purposes and do not pertain to the question of interest. We regressed the probability assigned to the target onto the speaker condition (sum-coded, levels: -1 = child, 1 = adult), trial type (sum-coded, levels: -1 = unambiguous (control), 1 = critical), the interaction between speaker condition and trial type, target position (three-level factor: left, middle, right; dummy-coded with middle as reference), trial number (mean-centered), and message type (shape or color; sum-coded, levels: -1 = shape, 1 = color).

The random effect structure included per-participant and per-item random intercepts as well as per-participant random slopes for trial type, message type and trial number.

For all effects, wide weakly informative priors were set. Specifically, for the effects of interest – the effect of speaker condition and interaction of speaker condition with trial type – we intentionally set very wide priors to avoid biasing the model in any direction. For the speaker condition, we set the prior to be a normal distribution centered around 0 with a standard deviation of 12.5 so that the whole probability range 50-100 (50 corresponding to literal responding and guessing between target and competitor and 100 corresponding to perfectly rational responding) is covered within 2 standard deviations. The same wide prior of a normal distribution centered around 0 with a standard deviation of 12.5 was applied to the interaction of speaker condition and trial type. The full model equation as well as a detailed discussion of all priors is included in Appendix D. We ran four chains of 4000 iterations each, with the first 1000 iterations of each chain discarded as warm-up.

To assess the evidence for the effects, we examined whether the credible intervals for the parameter estimates included zero. If the credible interval did not include zero, we concluded that there was a meaningful relationship between the predictor and the dependent variable. The results are reported in Table 5.2. Unsurprisingly, much lower target ratings were assigned on critical trials compared to unambiguous trials ($\hat{\beta}$ = -17.25, 95% CrI = [-19.18, -15.27]). There is a very small but robust effect of trial number, suggesting that participants on average became slightly better over the course of the experiment ($\hat{\beta}$ = 0.14, 95% CrI = [0.03, 0.24]).

Crucially, participants assigned higher probability to the target in the adult speaker condition than in the child speaker condition ($\hat{\beta}$ = 3.37, 95% CrI = [1.41, 5.34]). There was also an interaction of speaker condition with trial type, whereby the difference in target ratings was larger on critical trials than on unambiguous ones ($\hat{\beta}$ = 3.17, 95% CrI = [1.35, 5.00]). Therefore, we can conclude that speaker's identity and perceived pragmatic reasoning complexity influences the interpretation of ambiguous messages.

Next, we take a look at the annotations of participants' strategies. In the adult speaker condition, there's a much higher proportion of *correct_reasoning* responses (17 out of 40, 43.6%) than in the child speaker condition (6 out of 39, 15.4%). Interestingly, even in the adult condition, there were quite a few guesses (16 out 40, 41.1%, vs. 23 out of 39, 59%, in the child speaker condition). This suggests that at the population level, people are more likely to derive an implicature in the adult condition but

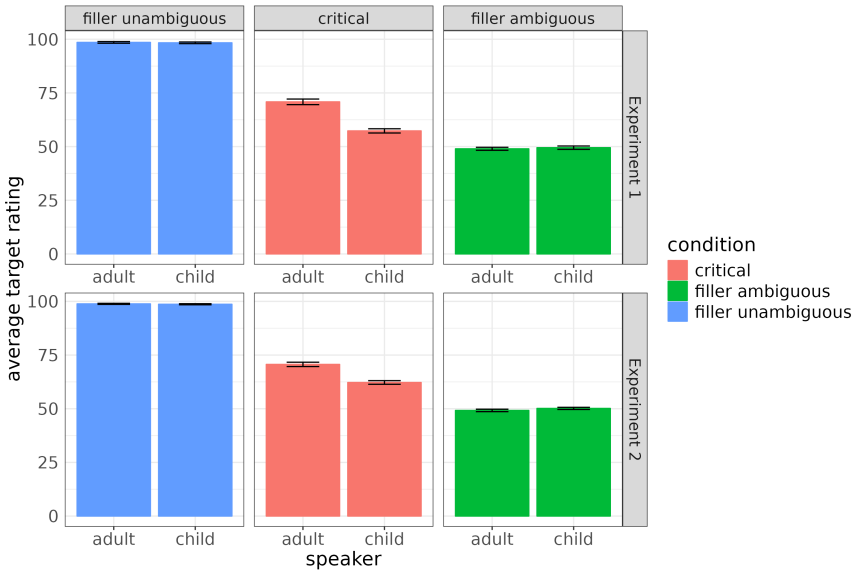


Figure 5.2: Average target ratings in the two speaker conditions for all trial types.

some individuals do not do so. This is not surprising given Franke & Degen (2016)'s original findings that participants could be divided into reasoning complexity types based on their performance and some participants fell into the literal listener (L_0) reasoning type. If someone is not able to reason pragmatically themselves, they will also not be able to meta-reason about whether the speaker is able to do so.

In both conditions, there are a few *meta_reasoning* responses (4, 10.3%, in the adult and 5, 12.8%, in the child speaker condition), suggesting that people are sometimes aware of deriving a less strong implicature based on considerations about the speaker. We take a closer look at reasons that were named for providing a lower target rating in the explanations in the *meta_reasoning* category. In the adult speaker condition, the reasons were the following: "some allowance for the participant thinking illogically or mistakenly", "I think some people may miss this", "there is a small chance they didn't think of that" and "it can't be fully ruled out [that the speaker meant the competitor]". In the child speaker condition, the following reasons were mentioned: "I think there's a chance (25%) that the child operated entirely off color", "There's some doubt [that the child selected the message pragmatically]. So I accounted for that not very likely possibility", "There's a chance they were just picking by colour without making that calculation", "I think there is a 70% chance that the child would be smart enough to do that", and "A child's thought may not follow this process, hence the 20% error". It is noteworthy that in the adult condition, too, some participants acknowledge the possibility of error or not paying attention resulting in non-optimal responding. Another interesting observation is that in some of these explanations, people directly refer to the provided ratings and describe the rating assigned to the competitor as the chance that the speaker was not thinking optimally (e.g., 25% chance). From the perspective of probability theory, this is not correct since even the literal speaker will select the pragmatically optimal message half of

the time by chance. This kind of probabilistic reasoning errors has been widely discussed for other phenomena (e.g., Fox & Levav, 2004a). We will return to potential probabilistic reasoning errors in this study in the discussion.

There is only one response in the *salience/preference* category, in the child speaker condition, which is "the red option may be more popular [than the blue option]". The fact that there was only one explanation falling into this category, together with the fact that there was no effect of message type (color vs. shape) in the regression model, suggests that salience did not have a large effect on participants' responses. However, even if salience did have a small effect, we assume that it would affect participants in the two speaker conditions in the same way.

It is highly unlikely that there were a lot more literal reasoners in the child speaker condition due to chance. Instead, it seems to be more likely that some of the *guess* responses in the child speaker condition involved not deriving the implicature because of considerations about the child's reasoning ability but either the participants were not aware of the meta-reasoning they were performing or they did not report it in their explanations.

Additionally, when examining the average target probability associated with each annotation tag, we see that, both in the case of *correct_ reasoning* and *meta_ reasoning*, the associated target probability is higher in the adult speaker condition than in the child speaker condition (88.0 (SE=3.43) vs. 73.8 (SE=6.82) for *correct_ reasoning* and 87.5 (SE=5.7) vs. 72.0 (SE=3.14) for *meta_ reasoning*), suggesting that even when people believe that the child might be able to solve the task, they are less certain, and in the cases where participants explicitly reason about the speaker's reasoning sophistication, they assign only a small probability to an adult not thinking optimally and a larger probability to a child not thinking optimally.

Since this experiment's sample size is not very large, we cannot draw definitive conclusions based on the differences in annotation tag counts. Instead, we view this analysis as supplementary to the main regression analysis, and since the annotations of reasoning strategies point to the same conclusions as the regression analysis, we view them as additional support for our findings.

Taken together, the results suggest that, provided that the listener themself has sufficient reasoning ability, they will take the reasoning ability of the speaker into account and may not derive an implicature or derive a weaker one if they believe the speaker's reasoning ability to be insufficient for selecting the optimal message.

Discussion of Experiment 1

This experiment provided evidence that, at the population level, people take incorporate their beliefs about the speaker's reasoning ability into the inferences they derive. On critical trials, participants were more likely to derive an inference if they believed the speaker was an adult, as indicated by higher target ratings.

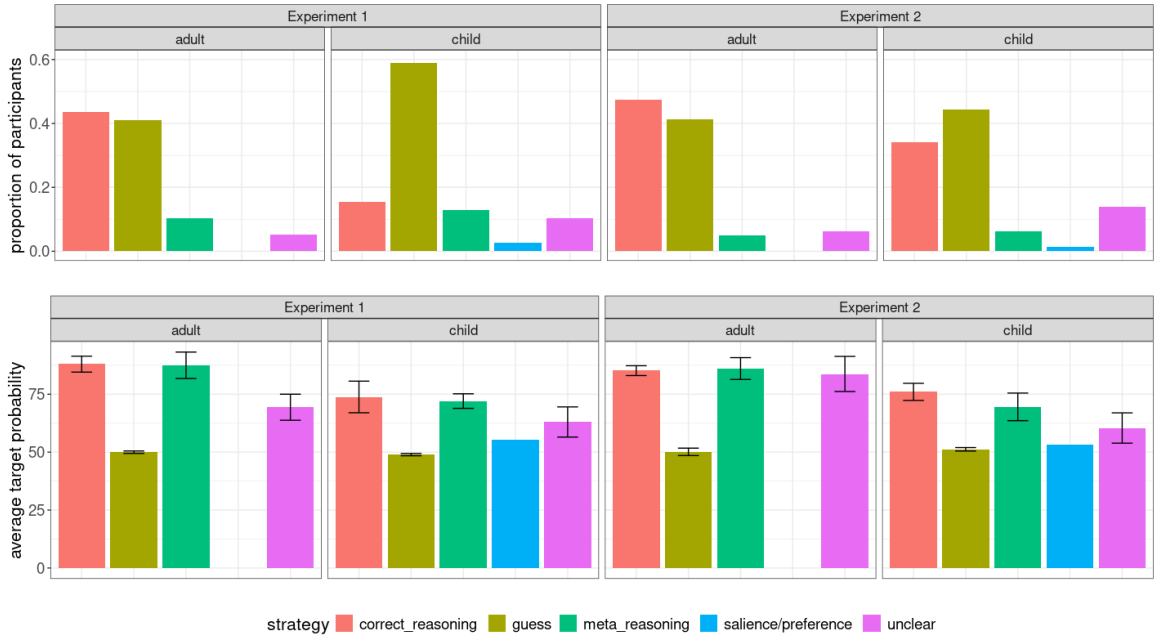


Figure 5.3: Frequency of each strategy tag per speaker condition and the corresponding target probability for that tag in Experiments 1 and 2.

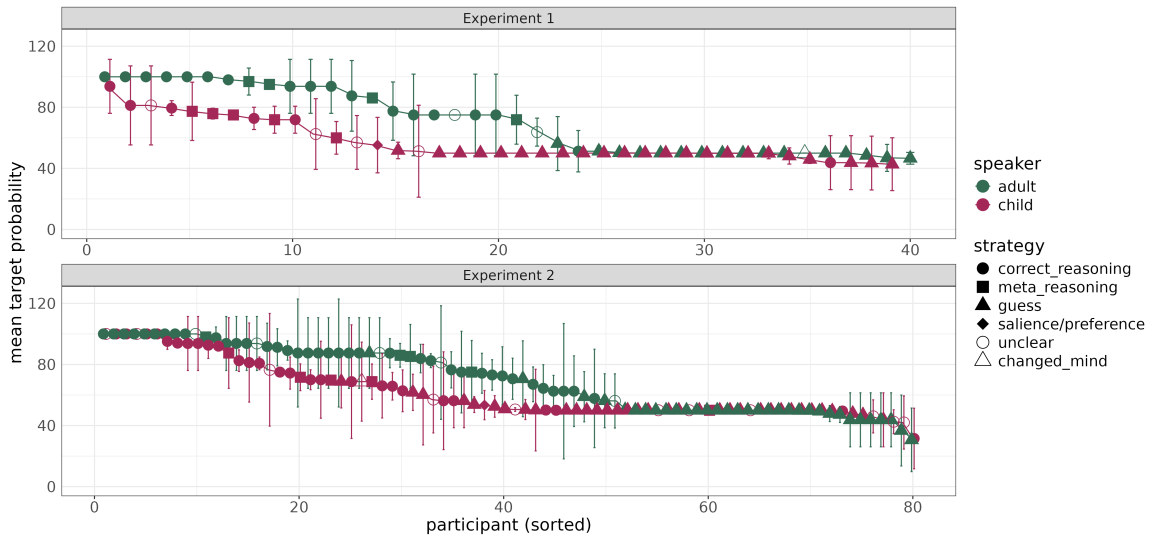


Figure 5.4: Individual participants' average target ratings by speaker condition in Experiments 1 and 2, sorted from high to low, with reported strategies.

	Experiment 1 (N=79)		Experiment 2 (N=160)	
Effect	**Estimate**	**95% CrI**	**Estimate**	**95% CrI**
Intercept	81.81	[79.63, 84.03]	82.67	[81.07, 84.22]
speaker (adult vs. child)	3.37	[1.41, 5.34]	1.90	[0.41, 3.39]
trial type (critical vs. control)	-17.25	[-19.18, -15.27]	-16.19	[-17.62, -14.79]
trial number	0.14	[0.03, 0.24]	0.17	[0.10, 0.24]
message type (color vs. shape)	0.51	[-0.16, 1.19]	-0.03	[-0.46, 0.40]
targetpos (left vs. center)	-0.48	[-1.61, 0.69]	0.34	[-0.60, 1.27]
targetpos (right vs. center)	-1.14	[-2.27, 0.00]	-0.67	[-1.60, 0.25]
speaker : trial type	3.17	[1.35, 5.00]	1.83	[0.42, 3.27]

Table 5.2: Effect estimates and 95% credible intervals (CrI) for the two experiments. Effects for which the 95% CrI does not include 0 are bolded.

The effect that we reported is a population-level effect. Top panel of Figure 5.4 displays individual participants' average target ratings and shows that there is a lot of individual variability. The effect appears to be driven by two factors: participants in the adult condition assigning higher ratings to the target, corresponding to drawing a stronger inference, and more participants in the child speaker condition assigning a 50% rating to the target, corresponding to not drawing an inference.

An anonymous reviewer pointed out a potential confound in the setup of the experiment: in the child speaker condition, the cover story involved the children being told to imagine that they are scientists communicating with an alien. Our intention in constructing the instructions that way was to make the cover story believable by presenting the instructions in the format of a game that would be accessible and engaging for children. However, as the reviewer rightly pointed out, this may have had unintended effects related to the perception of aliens. Participants' lower target ratings in the child speaker condition may have reflected not their beliefs about children but their beliefs about aliens: aliens may not be perceived as pragmatic communicators, and one may communicate differently with an alien than with another human.

In order to address this concern, we conducted a follow-up experiment where we took out communication with the alien from the cover story, making the two speaker conditions more directly comparable.

5.3.2 Experiment 2

This experiment was conducted to address the potential concern that the effect in Experiment 1 may have arisen not from the speaker identity manipulation but from the alien cover story in the child speaker condition.

This experiment also has a larger sample size than Experiment 1 (80 participants

per condition in Experiment 2 vs. 40 in Experiment 1) so that the effect of speaker can be robustly detected since it may be somewhat smaller if the difference between the speaker conditions in Experiment 1 was driven in part by the mention of an alien in the cover story in the child speaker condition.

Participants

160 native speakers of English, 80 per speaker condition, with an approval rating of at least 95% who did not participate in the first experiment were recruited via the crowdsourcing platform Prolific. 19 participants (7 in the adult speaker condition and 12 in the child speaker condition) were excluded because their average target rating on the unambiguous trials was below 80, suggesting that they may have misunderstood the task or randomly clicked through the experiment. 6 participants (1 in the adult condition and 5 in the child condition) were excluded because their reported strategy suggested that they had misunderstood the experiment. 1 participant in the adult speaker condition was excluded because on 6 out of 8 critical trials they assigned a 100%-rating to the distractor, suggesting that they were applying a very different strategy which does not correspond either to literal or to pragmatic responding (odd-one-out, e.g., message "red" means the only *non-red* object). New participants were recruited in their place, resulting in 80 participants per condition.

Design and procedure

Design and procedure of the experiment were identical to Experiment 1, except for the instructions introducing the speaker.

Participants in the child speaker condition were told, as before, that we had conducted a study with 4-year-old children at a local kindergarten. Instructions that were allegedly given to the children mirrored those in the adult condition but used simpler language and shorter sentences. Additionally, on the instructions screen, a picture of a 4-year-old child (daughter of one of the authors) was added, showing a computer screen with three shapes corresponding to an unambiguous speaker trial and four cards with the possible messages. The child is holding up the message with the correct answer.

On the actual listener trials where participants needed to give ratings, we included a picture of the same child looking at the four messages, where the screen with the trial which the child is solving is not visible.

This was done in order to help participants imagine a 4-year-old child. Before designing this experiment, we also ran a version where we did not include pictures of a child but only the simplified instructions. In that version of the experiment, we observed a trend in the same direction as in Experiment 1 but neither the effect of speaker condition nor the interaction of speaker condition with trial type showed strong evidence for an effect, as the 95% credible intervals for both included zero. We

regard that as a failed manipulation and hypothesize that simply mentioning that the speaker is a child may not be enough for the listener to take their perspective and imagine how the child would do the task. We expect that the cover story about interaction with an alien in Experiment 1 achieved the purpose of making it easier to imagine a 4-year-old and making it clear that the speaker was a small child. We hypothesize that including pictures of an actual 4-year-old will achieve that purpose while eliminating the potential confound of an alien interlocutor.

In the adult speaker condition, we also rephrased the instructions introducing the speaker slightly to make them more clear and closer to the child speaker condition. The instructions, along with the images used in the child speaker condition, can be found in Appendix D.

Results

We again first inspect the average probability (out of 100 points) assigned to the target for each speaker type. The bottom panel Figure 5.2 shows the same pattern as in Experiment 1: participants are at ceiling on unambiguous trials and at chance on ambiguous ones, suggesting that they understood the experiment and that they considered a 4-year-old to be capable at least of feature matching. On critical trials, the average target probability is 70.7 (SE=1.01) in the adult speaker condition and 62.2 (SE=0.85) in the child speaker condition. In the adult speaker condition, target probability is nearly identical to that in Experiment 1 (70.8 vs. 70.7), while in the child speaker condition, it is somewhat higher (62.2 vs. 57.3). This seems to suggest that the effect of speaker condition persists but that it is somewhat smaller than in Experiment 1, possibly due to the fact that in Experiment 1 it was partially driven by the mention of the alien in the cover story.

To verify this effect, we fit a Bayesian regression model in the same way as for Experiment 1. The results are reported in Table 5.2. We can see that Experiment 2 closely replicates Experiment 1. Again, much lower ratings were assigned to the target on critical trials compared to unambiguous trials ($\hat{\beta}$ = -16.19, 95% CrI = [-17.62, -14.79]). Likewise, like in Experiment 1, there is a small but robust effect of trial number, suggesting that participants on average became slightly better over the course of the experiment ($\hat{\beta}$ = 0.17, 95% CrI = [0.10, 0.24]). There was no effect of message type or target position.

The main effect of interest, the effect of speaker, persists and the credible interval does not include 0 ($\hat{\beta}$ = 1.90, 95% CrI = [0.41, 3.39]). The estimate for the effect of speaker is smaller than in Experiment 1 (1.90 vs. 3.37) but it does fall inside the 95% CrI for the estimate of the effect in Experiment 1. Similarly, there is an interaction of speaker and condition ($\hat{\beta}$ = 1.83, 95% CrI = [0.42, 3.27]), which is again smaller than in Experiment 1 (1.83 vs. 3.13) but does fall inside the 95% CrI for the estimate in Experiment 1. This provides more evidence for the effect of speaker identity on the derivation of pragmatic inferences.

Next, we look at the annotations of participants' strategies, shown in the right panel of Figure 5.3. As in Experiment 1, there are quite a few *guess* responses in both conditions (33 out of 80, 41.2%, in the adult condition, and 35 out of 80, 44.3%, in the child speaker condition), suggesting that even in the adult condition, there were some participants who were behaving like literal listeners themselves. While in Experiment 1 there was a much higher proportion of *correct_reasoning* responses in the adult speaker condition compared to the child speaker condition, in Experiment 2 this pattern persists but it is much less pronounced (47.5% vs. 34.2%), suggesting that the effect may be less categorical and more graded.

There are again a few *meta_reasoning* responses, 4 in the adult speaker condition and 5 in the child speaker condition, corresponding to making a less strong inference when considering the speaker. In the adult speaker condition, the following explanations are provided: "I can't discount that they didn't act overly quickly without thinking", "They may not have noticed it", "They may have rushed and not noticed" and "There's a small chance the participant giving the answer wasn't paying attention". In the child speaker condition, the reasons given were "I don't know how good kids are at thinking logically like that", "This depends on the mental age of the child", "I have some doubt in the child's ability to make this more complex decision so I've left 25% of room for doubt", "There's always a chance as a 4-year-old the will have just picked the shape", "[If they had wanted to refer to the competitor], they would (probably, if their reasoning skills were advanced enough) have chosen the triangle message". There seems to be a distinction that in the adult speaker condition, when people reason about the speaker, they draw a less strong inference because they account for the person rushing or not paying attention in the moment, whereas in the child speaker condition, the discounting happens because of uncertainty about the child's *ability* to reason optimally.

Like in Experiment 1, there is only one explanation that falls into the *salience/preference* category, also in the child speaker condition. It is "the blue seems more striking so I thought maybe they would have noticed it more than the red". Interestingly, this explanation is exactly the opposite of the salience explanation in Experiment 1, which supposed that children may have a preference for red, suggesting that people may differ in their perception of salience.

When we examine average target ratings, we find, like in Experiment 1, that for the same strategy tag, the associated probability is higher in the adult speaker condition than in the child speaker condition: 85.2 (SE=2.13) vs. 76 (SE=3.69) for *correct_reasoning*, 86.1 (SE=4.70) vs. 69.5 (SE=5.96) for *meta_reasoning* and 83.7 (SE=7.57) vs. 60.4 (SE=6.53) for *unclear*. This suggests that even when people make a pragmatic inference in the child speaker condition, they are less certain about it and assign a higher probability to the child not choosing an optimal message than to an adult not doing so.

Discussion of Experiment 2

Experiment 2 successfully replicated the effect of speaker identity on the strength of pragmatic inferences while removing the potential confound introduced by the mention of an alien in the instructions of the child speaker condition in Experiment 1, with a larger sample size.

The estimates of the effects of interest (main effect of speaker and interaction of speaker with trial type) were smaller than in Experiment 1 but fell within the 95% CrI of the estimates for those effects from Experiment 1. This suggests that a part of the effect in Experiment 1 may have been driven by the fact that people were interpreting children's messages less pragmatically because the child was said to be communicating with an alien. Therefore the estimates of the effect of speaker identity and the interaction of speaker identity with trial type obtained in Experiment 2 are likely to be more accurate.

The effect that we observed is a population-level effect. If we examine the performance of individuals in each condition, we again observe a lot of variability. Figure 5.4 shows mean probability assigned to the target on critical trials by individual participants in the two speaker conditions, sorted from high to low. This individual variability indicates that different people are likely to have different beliefs about rationality of speakers, and possibly also that people utilize the scale differently.

Is the observed effect categorical or gradient in nature? In other words, do fewer people in the child speaker condition derive an inference, or do comparably many people derive a less strong inference? Examining individual participants' performance (Figure 5.4) suggests that it is likely to be both: more participants consistently assign a 50% probability to the target, consistent with literal responding, in the child speaker condition than in the adult speaker condition, but there is also a gradient component, where people in both conditions draw an inference but it is less strong in the child speaker condition, as indicated by the line corresponding to individual participants' performance in the adult speaker condition being above the line for the child speaker condition. Additional evidence for a less strong inference in the child speaker condition comes from lower target ratings corresponding to the reported *correct_reasoning* and *meta_reasoning* strategies in the child speaker condition compared to the adult speaker condition.

We also see that, even in the adult speaker condition, there are some people who assign the probability of near 50% to the target, consistent with a literal interpretation. We do not expect people to believe that another adult is unable to reason about alternatives. Therefore, those people likely had no internal model of the speaker and just interpreted messages literally themselves, which corresponds to the literal listener RSA model L_0. This is consistent with Franke & Degen (2016), who found that a literal listener model fit the data of a subset of their participants best, which we also observed in Chapter 4.

5.3.3 Discussion of Experiments 1 & 2

In Experiments 1 and 2, we found that participants were more likely to interpret an ambiguous message pragmatically if they thought that the speaker was another adult than if it was a four-year-old child.

We observed that there is likely both a categorical and a gradient component to this effect: more people appear to not derive inference in the child speaker condition, and participants in the child speaker condition who do derive an inference derive a less strong inference than those in the adult speaker condition.

Here, we briefly consider how this reasoning about the speaker can best be captured in the Rational Speech Act framework; this question will be discussed in more detail in General discussion. In both the child and the adult speaker conditions, participants appear to belong to a mixture of reasoning types. See Appendix D for an additional analysis which compares experimental data to simulated participants drawn from a normal distribution with the mean and standard deviation from each speaker condition, which represents the assumption that all participants in each speaker condition used the same strategy and that individual differences are due to sampling chance. It illustrates that, in both speaker conditions, the differences between subsets of participants are more extreme than would be expected under a unimodal normal distribution, suggesting that multiple strategies were used in each speaker condition.

In both speaker conditions, there are participants who assigned the probability of about 50% to the target, consistent with the predictions of the L_0 RSA model. Since we do not expect adults to assume that other adults are completely unable to reason rationally, we assume that all participants who assigned a rating near 50% to the target in the adult speaker condition are literal L_0 listeners who do not have an internal model of the speaker and were not able to solve the task pragmatically themselves.

In the child speaker condition, a higher proportion of participants assigned an equal probability to the target and to the competitor than in the adult speaker condition. There are two possible explanations for this, which are not mutually exclusive. First, it is possible that some such literal-looking responding in the child speaker condition comes from reasoning about the child speaker and deciding that the speaker is literal, therefore deciding not to derive the inference. It is possible that some of the 50% responding in the child speaker condition may result from deciding that the child is a literal speaker but making an error in probabilistic computations: recall that a pragmatic listener model reasoning about a literal speaker will assign a rating of $\frac{2}{3}$ and not $\frac{1}{2}$ to the target by considering probabilities of alternatives. We think that this is possible but not very likely since all instances of participants reporting a *meta_reasoning* strategy corresponded to higher target ratings, i.e., participants appear to be relatively certain that children are fairly capable of behaving as pragmatic speakers, whereas an average target probability of 50% corresponds to the

reported strategy *guess*. We return to the question of whether people are able to perform pragmatic reasoning about alternative utterances when the speaker is literal in Experiment 4.

The second and, in our view, more likely explanation for there being more literal-appearing responders in the child speaker condition is that the situational context of reasoning about a child is more conducive to an L_0-level reasoning setting. Participants may choose to expend less effort on reasoning if they suspect low pragmatic competence of the speaker. We discuss this point in more detail in General discussion.

The gradient part of the observed effect, i.e., that participants in both conditions derive a pragmatic inference but it is stronger in the adult speaker condition, can be modeled using the belief-driven pragmatic listener model $L_2^{belief-driven}$ proposed in Section 5.2, whose beliefs about whether the speaker will select the optimal message are governed by the weight parameter λ. In the adult condition, the λ would be higher, reflecting higher certainty that the speaker is pragmatic and capable of selecting the optimal message.

Interestingly, the effect of speaker on the derived pragmatic inferences only emerged when the instructions included a visual which made the child speaker setting more salient – a drawing of an alien in Experiment 1 and a photo of a 4-year-old in Experiment 2. Before designing Experiment 2, we ran a version of the experiment where we took out the alien communication game cover story from the child speaker condition and simply told participants that the messages came from a 4-year-old. The effect was much smaller and less evident in that version, suggesting that this kind of perspective taking is difficult in the abstract – when not immersed in the situation of actually communicating with the speaker directly – and requires sufficient salience of the speaker's characteristics to be able to take their perspective. This is consistent with findings about other perspective-taking phenomena, such as visual perspective taking in a director-matcher task (Keysar et al., 2000) and interpretation of sarcastic messages (Epley et al., 2004).

The first two experiments showed that participants' inferences were influenced by the presumed pragmatic sophistication of the speaker. In the next experiment, we extend the investigation of speaker effects to artificial agents and explore another perceived speaker characteristic which may affect interpretation, namely, cooperativity. As discussed in the theoretical background, human communication is thought to be governed by the cooperative principle (Grice, 1975), which states that the interlocutor behaves cooperatively and observes conversational maxims. However, it is much less clear whether the cooperative principle holds for artificial agents. Therefore, people may not draw the same inferences with an artificial agent as they might with another human, even if they believe it to be similarly competent. Experiment 3 thus investigates the inferences drawn by participants when they believed the messages to be selected by the popular chatbot ChatGPT.

5.4 Reasoning about artificial agents as speakers: The case of ChatGPT

Pragmatic implicature derivation presupposes that Grice's cooperative principle (1975) is observed and crucially depends on interlocutors expecting each other to behave cooperatively. It is much less clear, however, whether people extend this assumption to communication with artificial agents. People might therefore not draw the same pragmatic inferences when interacting with an artificial agent as they would with other conversationally competent humans, even if the agent is in principle believed to be similarly competent.

With rapid developments in technology and with human-computer interaction becoming more and more commonplace, it is of interest to which extent collaborative effects widely attested for humans also extend to artificial agents. The "Computers as Social Agents" (CASA) framework (Nass et al., 1994) developed in the 1990s states that human-computer interaction is social and that humans readily exhibit social behavior when interacting with artificial agents, such as politeness and personality attribution. The question whether, when interacting with a computer, people tend to take its perspective into account to a lesser or a greater degree than when interacting with another human, has been investigated in the domain of spatial perspective taking, where instructions may be ambiguous when the perspectives of the speaker and the listener do not match (e.g., "Give me the book on the right" could mean the speaker's or the listener's right). Interestingly, while some studies found that participants were more egocentric in their interpretation when the speaker was another human and took the robot's perspective more willingly (Fischer, 2006; Branigan et al., 2011), other studies found the opposite, that is, participants expecting robots to adopt their perspective (Carlson et al., 2014; Loy & Demberg, 2023). The perceived competence of the artificial agent also appears to matter. Fischer (2005) found that participants always took the perspective of a nonverbal robot and gave it streamlined instructions, whereas with a verbal robot they used a more varied vocabulary and syntax and included some egocentric descriptions. Similarly, Branigan et al. (2011) investigated lexical alignment in dialogue with humans and artificial agents and found that their participants tended to align – i.e., repeat the interlocutor's choice of words – more with a computer than with a human, and with an older computer more than with a more modern computer, presumably reflecting participants' beliefs about computers' limited capability.

More recently, Loy & Demberg (2023) found that participants provided more egocentric descriptions in a spatial perspective taking task with a computer than with a human partner and more egocentric descriptions with a modern compared to an older computer. They explain the discrepancy between their findings and earlier studies, which had mostly found that people were more willing to take the perspective of a computer than that of another human, in terms of a shift in perception of artificial

agents' competence: whereas a couple of decades ago, artificial agents were perceived as having very limited capabilities, they are now perceived as much more complex agents with high intelligence and more collaborative capacity. Similarly, in a communication game where participants' task was to get their partner to pick out objects from a grid by referring to them, Peña et al. (2023) found that participants were more likely to consider their partner's visual perspective when their partner was a human than when it was a computer. However, this difference disappeared when it was emphasized to participants that the computer was a collaborative agent with a shared communicative goal and separate from the system which the experiment was run on. This suggests that not only the artificial agent's capability but also their willingness to collaborate may influence how they are perceived and treated by humans.

In Experiment 3, we investigate people's perception of the influential LLM-based chatbot ChatGPT as a cooperative agent. In a short time, ChatGPT has come to occupy an important position in a variety of domains and is claimed to have the potential to have a transformative impact in areas such as healthcare (Younis et al., 2024) and education (Grassini, 2023).

While surveys have been conducted examining people's attitudes towards ChatGPT (Singh et al., 2023; Shoufan, 2023; Ngo, 2023), to our knowledge, this is the first study to examine people's perception of ChatGPT as a collaborative agent and to investigate the perceived pragmatic abilities of a modern artificial agent.

Like in the previous two experiments, we use the reference game paradigm with sliders described in Section 5.1.1. In this experiment, we tell participants that the messages were sent by ChatGPT. We compare our results to the two speaker conditions from Experiment 2 and find that, at the population level, participants interpret ChatGPT's messages less pragmatically than those of an adult, similarly to those of a 4-year-old child.

Note that the publication that Experiment 3 is based on (Mayn et al., 2024) compares the ChatGPT conditions to those of Experiment 1 and finds, in addition to a difference between the adult and the ChatGPT conditions that we report here, a difference between the child and the ChatGPT conditions. The reason the publication used Experiment 1 and not Experiment 2 for comparison was that Experiment 2 had not been conducted at the time. In this chapter, we compare the results of Experiment 3 to those of Experiment 2 since we expect it to have produced more accurate effect size estimates since Experiment 1 included the alien manipulation in the child speaker condition which may have affected the results. Additionally, while in Mayn et al. (2024) we fit a linear mixed-effects model, here we report a Bayesian regression model with the same priors as Experiments 1 and 2 since it allows for better comparability to those experience and since it converged with a full random effect structure.

Also, we observe a notable discrepancy between participants' performance and their explicit estimates of ChatGPT's pragmatic ability. When asked explicitly in a post-test questionnaire, participants overwhelmingly expressed high confidence in

ChatGPT's ability to solve the task, including those participants who interpreted ChatGPT's messages literally in the pragmatic task itself. We discuss possible reasons for this discrepancy, as well as the implications of our findings for our understanding of human perception of modern artificial agents as communicative partners.

5.4.1 Experiment 3

Participants

40 native speakers of English with an approval rating of at least 95% were recruited via the crowdsourcing platform Prolific. 4 participants needed to be excluded because their performance strongly suggested that they were responding randomly or because their post-hoc explanation of their reasoning strategy suggested that they had misunderstood the setup of the experiment. New participants were recruited in their place, resulting in a total of 40 participants.

Materials

We used the reference game paradigm described in Section 5.1.1 and told participants that the messages were generated by ChatGPT. Since not all participants may be familiar with ChatGPT, they were first given a short description of what ChatGPT is and an example of a prompt and ChatGPT's response unrelated to the topic of the experiment ("What desserts should I try in Paris?").

To make the cover story as convincing as possible, participants were told that since ChatGPT 3.5 only accepts textual input (which was the case in November 2023 when the study was conducted), it completed the speaker task in textual form which is completely equivalent to the visual form. They were shown the speaker task in the visual form side by side with the task in textual form. The textual form was as follows:

> Your task is to send a message so that someone, let's call them the interpreter, is able to identify a particular object from the set of three objects.
>
> You are only allowed to send one of four messages, and the interpreter knows this: the color red, the color green, circle and triangle. You can only send one message.
>
> The objects the interpreter will see: red square, green circle, blue triangle. You need the interpreter to pick out: blue triangle. Which of the four available messages (the color red, the color green, circle or triangle) will you send?

Since the visual scene in this task is very simple, the textual and visual instructions are equivalent and therefore this cover story is unlikely to have raised participants' suspicion or significantly influenced their behavior.

Effect	Estimate	95% CrI
Intercept	79.67	[76.60, 82.76]
speaker1 (adult vs. ChatGPT)	4.79	[1.07, 8.47]
speaker2 (child vs. ChatGPT)	0.95	[-2.73, 3.57]
trial type (critical vs. control)	-18.62	[-21.48, -15.79]
trial number	0.14	[0.08, 0.20]
msg type (shape vs. color)	-0.03	[-0.78, 0.73]
targetpos (left vs. center)	0.43	[-0.41, 1.24]
targetpos (right vs. center)	-0.59	[-1.41, 0.25]
speaker1 : trial type	4.24	[0.79, 7.72]
speaker2: trial type	0.50	[-2.94, 3.89]

Table 5.3: Regression results for Experiment 3.

Procedure

Like in Experiments 1 and 2, participants first completed 3 trials from the speaker's – i.e., ChatGPT's – perspective, after which they completed the main task described in Section 5.1.1, consisting of 8 critical trials and 16 unambiguous control trials.

After completing the main task, participants saw the first critical trial they had seen again and were prompted to explain their response. Participants' responses to this question were annotated using the same annotation scheme as for Experiments 1 and 2 (Table 5.1), and the distribution of responses was compared to that in the adult and child speaker conditions from Experiment 2.

Finally, since we were interested in how participants' responding related to their experience with ChatGPT and their beliefs about its competence, participants completed a short post-test questionnaire intended to tap into those questions. Participants were asked how much experience using ChatGPT (on a scale from 1 to 5) and how much knowledge of Machine Learning and Natural Language Processing (also 1-5) they had. They were then shown an item from the speaker's perspective (Figure 5.7) and asked to explain in their own words, by referring to that example, how they think ChatGPT determines which message to send. Finally, once a participant provided an answer, we explained to them that in order to solve this task, the speaker needed to reason about alternatives and that the unambiguous message "color green" is more optimal than "triangle" since upon hearing "triangle", the listener may select the blue triangle. We then asked the participants how likely they thought it was, on a 5-point scale, that ChatGPT is capable of performing that kind of reasoning.

Results

Throughout this section, we plot the results of this experiment alongside those from Experiment 2 for ease of comparison.

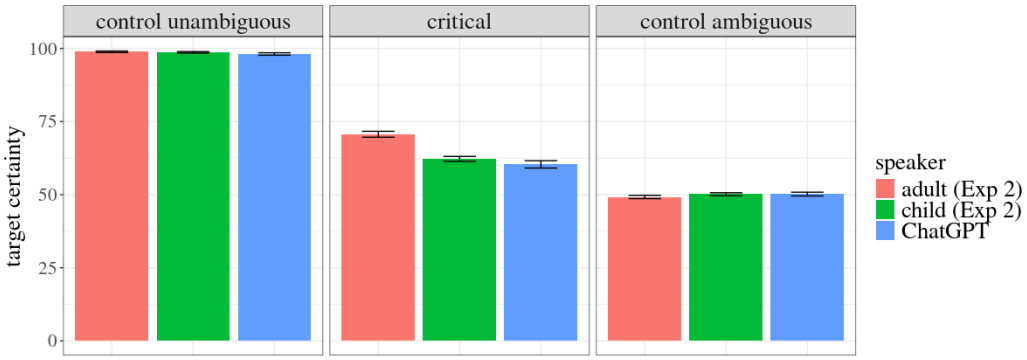


Figure 5.5: Average target rating ($\pm SE$) per trial type for each speaker condition (child and adult speaker from Experiment 2 and ChatGPT speaker from Experiment 3).

We first look at the average target ratings per trial type (unambiguous, critical and ambiguous), shown in Figure 5.5. We see that, similarly to the child and adult speaker conditions from Experiment 2, in our experiment, participants performed at ceiling in the unambiguous control condition. This shows that participants understood the task and consider ChatGPT to be at least capable of feature matching. In the ambiguous control condition, the target ratings are at chance, as predicted.

In the critical condition, the average probability assigned to the target in the ChatGPT speaker condition is 60.36 (SE=1.26), which is considerably lower than in the adult speaker condition from Experiment 2 (70.7, SE=1.01) and slightly lower than in the child speaker condition from Experiment 2 (62.2, SE=0.85).

In order to verify the significance of these apparent differences between the speaker conditions, we pool our data with the data from Experiment 2 and fit a Bayesian regression model in the same way as for the first two experiments. The same wide weakly informative priors as for the first two experiments were used.

The ambiguous control condition is removed from analysis. We regress the target probability on each trial onto the speaker identity (ChatGPT, child or adult; dummy-coded with ChatGPT as reference level since we want to compare ChatGPT to the two speaker conditions from Experiment 2), item type (sum-coded, levels: -1 = unambiguous, 1 = critical), interaction between speaker and item type, target position (left, middle or right; dummy-coded with middle as reference), mean-centered trial number, and message type (color or shape; dummy-coded with shape as reference). The random effect structure included per-participant and per-item random intercepts, and per-participant random slopes for trial type, message type and trial number.

The results are reported in Table 5.3. Much lower target ratings were assigned in the critical condition compared to the unambiguous control ($\hat{\beta}$ = -18.62, 95% CrI = [-21.48, -15.79]). There was a main effect of speaker (adult vs. ChatGPT), whereby the target ratings were higher in the adult speaker condition from Experiment 2 compared to the ChatGPT condition regardless of trial type ($\hat{\beta}$ = 4.79, 95% CrI =

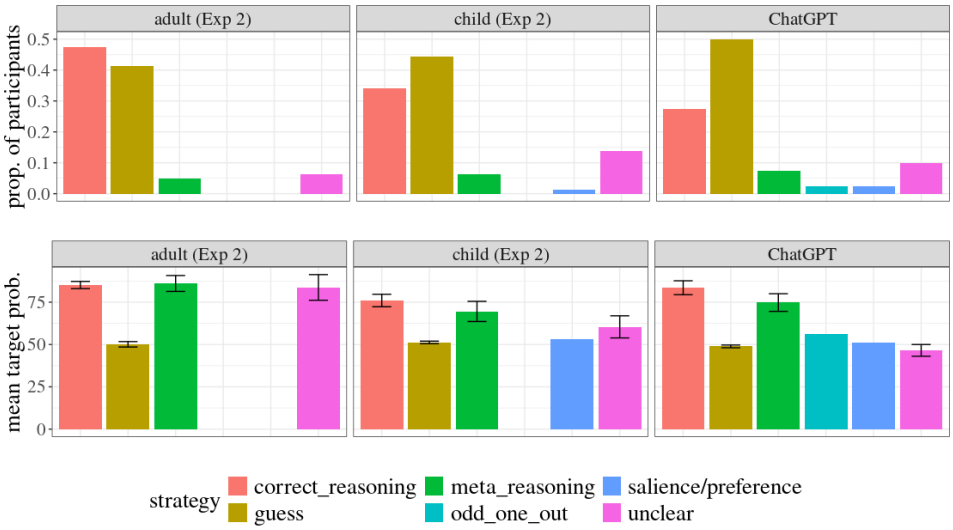


Figure 5.6: Frequency of each annotation tag (top panel) and corresponding average target ratings (bottom panel) for the three speaker conditions: child and adult from Experiment 2 and ChatGPT from Experiment 3.

[1.07, 8.47]). There was also a significant interaction of speaker (adults vs. ChatGPT) and trial type, indicating that the difference between target ratings between the adult and the ChatGPT speaker was stronger in the critical condition than in the control condition ($\hat{\beta}$ = 4.24, 95% CrI = [1.07, 8.47]). There was no evidence for a main effect of the difference between the child speaker and the ChatGPT speaker or an interaction with trial type. This suggests that, at the population level, ChatGPT's perceived reasoning ability or cooperativity is inferior to that of an adult speaker. Finally, the effect of trial number ($\hat{\beta}$ = 0.14, 95% CrI = [0.08, 0.20]) persists.

Next, we examine the annotations of participants' explanations. The top panel of Figure 5.6 shows the distribution of annotation labels for the three speaker conditions. The bottom panel shows the corresponding average target ratings. Similarly to the adult and child conditions from Experiment 2, the majority of explanations in the ChatGPT condition fall either into the *correct_reasoning* or into the *guess* category. The proportion of *correct_reasoning* responses in the ChatGPT condition is lower than in the adult and the child speaker conditions from Experiment 2 (ChatGPT: 27.6%, adult: 47.5%, child: 34.18%). On the other hand, the proportion of *guess* responses in the ChatGPT speaker condition is somewhat higher than in the other two speaker conditions (ChatGPT: 50.0%, adult: 41.25%, child: 44.3%). There are also 3 *meta_reasoning* responses where the explanations explicitly express doubts about how likely ChatGPT is to be capable of reasoning necessary to select the optimal message. Interestingly, the average target ratings corresponding to the *correct_reasoning* and *meta_reasoning* are lower than in the adult condition but higher than in the child condition from Experiment 2. Thus, there are fewer correct reasoning responses than in the child condition but the associated target rating is higher. Notably, the average ratings are consistent with the reported strategies: the average ratings in

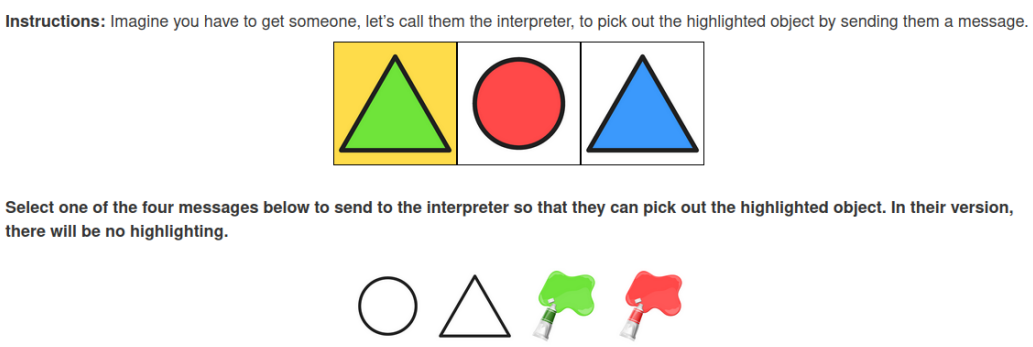


Figure 5.7: Item from the speaker's (i.e., ChatGPT's) perspective shown to participants in the post-test questionnaire of Experiment 3.

the *guess* category are at chance, as we would expect, whereas for *correct_reasoning* and *meta_reasoning* they are much higher but not at ceiling, especially in the child and ChatGPT conditions, reflecting population-level uncertainty that those speakers would behave optimally or collaboratively. The results of both this supplementary analysis of annotations and the main regression analysis suggest that, at the population level, utterances produced by ChatGPT are interpreted less pragmatically than those of an adult and more so than those of a child. However, as we will see, there is a lot of individual variability in the responses.

We now turn to participants' responses to the post-test questionnaire. Our participants had limited experience with ChatGPT (*mean*=2.08, *SD*=1.10) and almost no knowledge of machine learning and NLP (*mean*=1.33, *SD*=0.53).

The speaker item that was shown to the participants in the post-test questionnaire (Figure 5.7) doesn't exactly correspond to the critical items from the listener's perspective. If it had, the highlighted object would have been the blue triangle. However, in that case, from the speaker's perspective, the task is near-trivial since the only way to refer to the blue triangle is the "triangle" message. Therefore, in this question, the highlighted object is the green triangle (which corresponds to the competitor on critical listener trials). There are two ways to refer to the highlighted object, the color green and the triangle, and one needs to reason that the unambiguous message (the color green) is preferable. Interestingly, more than half of the participants (25 out of 40) in their open-ended answers to the question how ChatGPT likely decides which message to send described a process of elimination by which the unambiguous message is preferred to the ambiguous message. This is likely a lower bound since some explanations indicated high perceived competence without specifying the strategy, e.g., "ChatGPT is a master at this".

Likewise, when explicitly asked the likelihood that ChatGPT is capable of reasoning about alternatives needed to select an optimal message, participants gave high ratings (*mean*=4.05 out of 5, *SD*=0.85). Notably, there appears to be no relationship between people's performance on the task itself and their estimates of ChatGPT's competence on this question. Figure 5.8 shows average target ratings in the critical

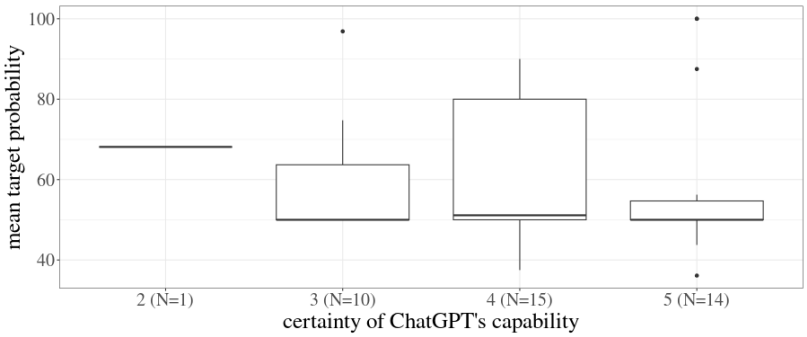


Figure 5.8: Average target ratings for each response to the question "How capable do you think ChatGPT is of performing this task?"

condition corresponding to each of the Likert scale responses. It can be seen that the majority of participants who rated ChatGPT's capability of solving the speaker task at a 5 in the post-test questionnaire gave very low target ratings in the actual task. This is an interesting discrepancy which we discuss below.

5.4.2 Discussion of Experiment 3

In Experiment 3, we investigated the perceived pragmatic ability of ChatGPT and compared participants' performance to the child and adult speaker conditions from Experiment 2.

We found that the ratings that participants assigned to the target, which we use a proxy for the perceived pragmatic competence of the speaker, are consistently and significantly lower than the ratings that were assigned in the adult speaker condition in Experiment 2. Indeed, the responses in our study pattern more closely with the child speaker condition from Experiment 2. At first glance, this seems to suggest that participants perceive ChatGPT to be less pragmatically capable than an adult human. This finding seemingly at odds with the results obtained by Loy & Demberg (2023), who found that participants expected a computer to take their perspective more often than another human, and a modern computer to take their perspective more often than an older computer. Our findings are, in fact, more in line with earlier work which had found that people more willingly took the computer's perspective than another human's (Fischer, 2006; Branigan et al., 2011).

The reason for the apparent discrepancy between our findings and those of Loy & Demberg (2023) may lie in the difference between the two tasks. In spatial perspective taking, which Loy & Demberg (2023)'s study investigated, the two perspectives are arguably quite salient, and we would expect any adult participant to be aware that there are two perspectives, even if adopting one of them is effortful. The pragmatic perspective taking task in the current study, on the other hand, is arguably much more difficult. Even in the adult condition in Experiment 2, some participants perform at chance on the critical items (Figure 5.4), suggesting that they interpreted the messages

literally and did not consider the speaker's perspective at all. Therefore, it may be less effortful to consider how an artificial agent would behave in a spatial perspective taking task than in the pragmatic reference game. Moreover, participants may assume that perspective taking is built into the computer program that is performing the spatial perspective taking task and that it had been specifically trained to do it, conceivably in order to be accommodating to humans. In the current study, on the other hand, ChatGPT would have needed to perform the task ad hoc, making it more likely that participants would not assume that ChatGPT would behave pragmatically by default.

This brings us to the other interesting finding of Experiment 3: the mismatch in participants' behavior on the task and their explicit ratings of ChatGPT's competence in the post-test questionnaire. When asked explicitly, participants overwhelmingly gave very high confidence ratings of ChatGPT's ability to do the reasoning needed to send the optimal message, including those participants who had just given chance-level responses on the reference game. One factor that likely contributed to this effect is that some participants initially could not solve the task pragmatically themselves, but then, when it was explained to them in the post-test questionnaire, they were forced to think about ChatGPT's ability to perform the task and rated it as high. Another explanation for this mismatch could lie in the distinction between participants believing that ChatGPT is *capable* of performing optimally and them believing that it would actually do so and select a message that is maximally helpful to the listener. Earlier, we hypothesized that the difference between the child and adult speaker conditions that we observed could have to do with the communicative context being more or less conducive to deeper reasoning: in the adult speaker case, it could be that participants are more likely to assume intentionality of the speaker's actions and therefore be more motivated to figure out the intended meaning. The same may be applicable to the distinction between ChatGPT and an adult human speaker. In the case of a human speaker, there is a much more direct parallel to self and one's own reasoning than when ChatGPT is the speaker. Since we know that the task is nontrivial and potentially effortful, sufficient motivation may be needed to engage in reasoning about the speaker's intent, which may be more readily available in the case of a fellow human interlocutor. This explanation is in line with findings of Peña et al. (2023), who showed that participants initially behaved more egocentrically with computers than with humans in a referring task but this difference disappeared when the computer was explicitly framed as a collaborative agent with a shared goal. It could be that in this task, since participants were merely told that ChatGPT generated the messages, the interaction was not direct enough to assume cooperative intention. Future work could investigate whether utterances by ChatGPT are interpreted more pragmatically when the interaction is framed as more direct and collaborative, for instance, when ChatGPT is said to generate messages in real time or when ChatGPT additionally "addresses" the participant directly.

Finally, while we asked participants about ChatGPT's capability on this specific

task and about their experience with AI and NLP, we did not elicit their opinions of AI more generally. Future work could include additional questions targeting people's opinions of AI and investigate how those opinions influence their interpretations.

The first three experiments investigated how pragmatic inferences are affected by the identity of a speaker who participants may not believe to be completely rational or cooperative. As discussed in Section 5.2.1, even with a perfectly literal speaker, any pragmatic listener RSA model always favors a pragmatic interpretation by considering the relative probabilities of alternative messages available to the speaker. In the Experiment 4, we investigate whether these predictions are consistent with what people actually do – whether people are able to reason about probabilities of alternatives when the speaker is fully literal, a simple computer program which randomly selects any true message to send. This design allows us to separate participants' reasoning about alternative utterances from any assumptions about the speaker's intent or rationality.

5.5 Reasoning about an explicitly literal speaker

In the previous three experiments described in this chapter, we found that participants' pragmatic interpretations were influenced by their beliefs about the speaker: participants were more likely to interpret an ambiguous message literally when they thought they were interpreting messages sent by a child or a large language model, but drew an implicature when they thought the messages had been sent by another adult. We proposed a model in the Rational Speech Act framework capturing this reasoning about the speaker's level of pragmatic competence in Section 5.2.1. According to this model, the listener has some uncertainty about whether they are dealing with a literal or a pragmatic speaker, and weighs the two speaker models according to their beliefs. However, even when the speaker is assumed to be perfectly literal, the pragmatic listener model always favors a pragmatic interpretation by considering the relative probabilities of alternative messages available to the speaker. As shown by the results of child and ChatGPT speaker conditions in the previous three experiments, this seems to be at odds with what many people do: they derive literal intepretations, consistent with the predictions of the literal listener model L_0.

While in Experiments 1 and 2 most participants seemed to perceive the child speaker as lacking advanced pragmatic reasoning skills, substantial individual variation suggests that people hold different beliefs about children's communicative abilities. In the current experiment, we investigate people's reasoning about a speaker who is explicitly presented as literal, a simple computer program which randomly chooses any true message to refer to an object. This design allows us to separate participants' reasoning about alternative utterances from any assumptions they may have about the speaker's intent or rationality.

We ask whether participants will successfully incorporate the information about

alternative messages available to the speaker into their inference, leading them to favor a pragmatic interpretation, or whether they will derive a literal intepretation instead. A pragmatic interpretation, in this context, requires adjusting the inferred likelihood of each referent based on how many literally true messages apply to it, corresponding to $P(object|utterance) \propto P(utterance|object)$, whereas a literal interpretation corresponds to assigning equal probability to all literally true options.

Some evidence that people might struggle to correctly incorporate probabilities of alternative messages into their interpretation comes from literature on Bayesian reasoning errors in other domains. It has been shown that people often fail to apply Bayesian principles when solving probability problems and instead rely on heuristics (Tversky & Kahneman, 1993) or combine different pieces of information, such as base rates and false alarm rates, linearly, instead of applying Bayes' rule (Stengård et al., 2022). Fox & Levav (2004b) showed that their participants overwhelmingly did not solve Bayesian reasoning problems correctly due to incorrectly partitioning the probability space, but nudging them to consider all possible events significantly increased Bayesian reasoning.

We find that, when presented with an explicitly literal speaker, people overwhelmingly do not perform ad-hoc reasoning about alternative messages and instead provide responses consistent with a literal interpretation. Furthermore, nudging people to consider all possible messages does not result in more pragmatic responding. Our findings suggest that while reasoning about speaker's intent is natural, reasoning about probabilities of alternative messages, at least in the absence of speaker's intent, may be effortful and may not be done by default.

We discuss the fact that the RSA framework currently does not allow for decoupling reasoning about alternative utterances and their probabilities from reasoning about the speaker's intent or rationality and that it may be useful to model the situation where the speaker's intent is successfully considered but the probabilities of alternatives are not.

5.5.1 Model predictions and possible comprehender strategies

Like in the previous experiments, we used the reference game paradigm described in Section 5.1.1, where participants distributed 100 points between the three possible referents using sliders. In this experiment, the speaker was explicitly presented as literal. Participants were told that they would play a communication game with a simple computer program called *basic_ message_ picker*, which randomly selects any true message to refer to an object. Thus, participants should not have had any doubts about the speaker's pragmatic reasoning type.

Here, we will briefly describe possible comprehender behaviors and provide the corresponding theoretical explanation.

As explained in Section 2.5.1, if participants behave in the way a pragmatic RSA model of a listener reasoning about a literal speaker would predict, they will assign the probability of $\frac{2}{3}$ to the target, corresponding to the slider value of 66 or 67. This is predicted by all pragmatic listener models more complex than L_0 since these models consider the probabilities of alternative utterances and there is only one way to refer to the target but two to refer to the competitor.

It is also possible that a listener does not consider the distribution of alternative messages and interprets the message literally instead, resulting in assigning an equal probability (50) to the target and to the competitor. This is what the literal listener model L_0 would predict.

Finally, it is possible that a listener tries to perform pragmatic reasoning but makes an error in probabilistic computations about alternative messages. There is a lot of evidence from other domains that people struggle with performing Bayesian computations, therefore, it could also be the case in this setting. This would also result in assigning equal ratings to the target and to the competitor (50%) but it is a distinct process from L_0 in that pragmatic reasoning is attempted but it involves a probabilistic computation failure, whereas L_0 models a case where no pragmatic reasoning is attempted.

There are two sources of information one could consider when deriving an inference: the speaker's intent or rationality and the available alternative messages and their probabilities. A literal listener model L_0 considers neither, and a pragmatic listener model L_2 considers both. Importantly, any recursive listener RSA model which has an internal speaker model – i.e., a model more complex than L_0 – will automatically correctly incorporate reasoning about alternative messages into the inference. Thus, modeling a listener who successfully reasons about the speaker's intent but does not about alternative messages presents a challenge for the RSA.

Importantly, we do not claim that interacting with a literal (or otherwise not very rational) speaker is a case RSA should necessarily be equipped to model. It is, after all, a *Rational* Speech Act model, which formalizes the cooperative principle, at the core of which lies the idea that both interlocutors are being cooperative, and it models an idealized scenario of near-perfect rationality of both interlocutors. Still, we consider the question of how rational, or near-rational, listeners accommodate not-so-rational speakers to be relevant for a realistic model of pragmatic reasoning, since there are many situations where speakers may not be in a position to behave rationally given limited language proficiency, high cognitive load or not sufficiently developed reasoning ability. Because this is such a commonplace setting, it may be one which we may want to model, whether in the RSA model or in another framework. Currently, in RSA there is no way to derive the prediction that people may reason about the speaker but not correctly incorporate alternative messages into their reasoning. We return to this point in the discussion.

5.5.2 Experiment 4

The experiment investigated participants' interpretations of messages uttered by a literal speaker.

Participants

94 native English speakers with an approval rate of 95% and above, who were recruited on the crowdsourcing platform Prolific, completed the experiment.

Procedure

Participants were told that they would play a communication game with a simple computer program called *basic_ message_ picker*, which randomly selects any true message to refer to an object. We used the cover story that we are interested in comparing how the advanced LLM-based chatbot ChatGPT communicates compared to a basic computer program and that participants were randomly assigned to play the game with a simple program. This was done to emphasize the simplicity and literalness of the speaker and to avoid people ascribing rationality and pragmatic competence to it.

In the beginning, participants briefly practiced selecting messages as if they were the computer program and received feedback to ensure they understood its expected behavior.

The remainder of the experiment consisted of three phases: Block 1, training, and Block 2. Participants were randomly assigned to one of two conditions, No training or Training. The order of trials was the same for all participants.

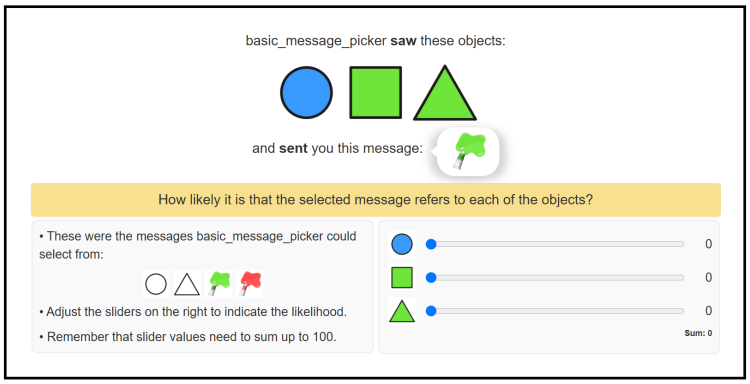


Figure 5.9: Example critical trial in Experiment 4 (Block 1).

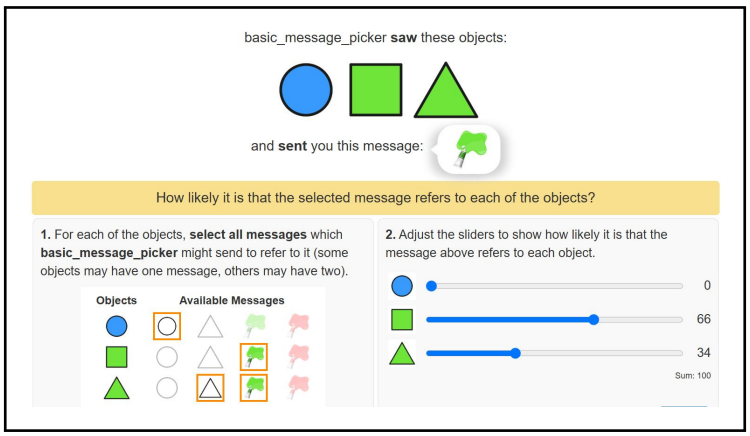


Figure 5.10: Example critical trial in the Training condition of Experiment 4 (training and Block 2).

Block 1. All participants completed 8 critical trials, 12 unambiguous (where the message clearly identified a single referent) and 4 ambiguous (where the target and competitor were identical, making it impossible to distinguish between them based on the message) filler trials, interleaved. Because Block 1 was practically identical across groups[1], performance on the 8 critical trials in this block serves as a baseline to evaluate the effect of the subsequent manipulations. At the end of Block 1, participants saw one critical trial again and were prompted to briefly explain, in a textbox, why they decided to put the sliders in those positions.

Training block. After Block 1, participants in the Training condition completed 4 additional reference game trials containing a secondary task, intended to increase awareness of alternative messages which could be used to refer to each object. Before assigning probabilities using sliders, participants had to complete an additional step. For each object, they were presented with a list of all the messages that were available to *basic_message_picker*. Their task was to select all messages that were true for each object (Figure 5.10). Participants received immediate feedback indicating whether their selection was correct, and they were unable to proceed to the probability sliders unless they had correctly identified all true messages. They only received feedback on the message selection and not on the sliders. If some participants provide low target ratings on critical trials simply because of inattention to the asymmetrical distribution of available messages, this extra step should support

[1]There was a slight layout difference between the No training and Training conditions. In No training, the available messages were shown below the three possible referents, whereas in Training, they were positioned on the left-hand side of the screen. This change was implemented after an initial test of the Training condition, which used the same layout as No Training. We observed a drop in participants' performance during Block 2, suggesting the training might have been confusing. To address this, we adjusted the layout to reduce clutter on the screen during training and Block 2. The layout changes were small, and the fact that performance in Block 1 is very similar across both conditions suggests that the two conditions remained comparable. Images of the two layouts can be found at `https://osf.io/erbn3`.

improved performance by drawing attention to that distribution. In the No training condition, participants completed an unrelated task (a 10-question version of Raven's Progressive Matrices (Raven & Court, 1938)) in place of training.

Block 2. All participants then completed a further sequence of 8 critical trials and 16 filler trials, interleaved. In this block, apart from 10 ambiguous and unambiguous filler trials similar to those in Block 1, participants saw 6 filler trials which were more complex. In these trials, the probabilities of the three objects when considering the distribution of available messages should be $^2/_5 : {}^2/_5 : {}^1/_5$ (2 trials), $^1/_3 : {}^1/_3 : {}^1/_3$ (2 trials), and $^1/_2 : {}^1/_4 : {}^1/_4$ (2 trials). These trials were added so that, if there is a positive effect of training, we could investigate whether the understanding generalizes to other situations.

In the No training condition, the procedure in this block was identical to that in Block 1. In the Training condition, participants again had to complete the additional step of choosing all possible messages for every object, as introduced in the training block. However, in this block, they did not receive any feedback on their selections.

Annotation of strategies

Participants' open-ended explanations of their response after Block 1 were annotated using the same adaptation scheme as for Experiments 1-3 (Table 5.1), with an addition of one annotation tag, *acribe_rationality*. It was assigned when a participant's explanation revealed that they expected *basic_message_picker* to behave rationally, despite the instructions explicitly stating that it is literal. An example from this category is "It doesn't make much logical sense to choose green for [referring to] the green triangle when there are two green shapes and you could have chosen the triangle symbol instead."

Results

12 participants were excluded for accuracy below 80% in unambiguous filler trials in either block, as this may indicate a general lack of attention. One further participant was excluded because their reported strategy suggested that they had misunderstood the instructions. Those two exclusion criteria were preregistered. Two participants were excluded because they appeared to have had technical issues: their answers were saved on the server multiple times and were slightly different every time. The remaining 79 participants (41 in the No training condition, 38 in Training) entered the analysis.

First, we classified participants' performance in each block based on the responses on critical trials. We computed the likelihood of participants' responses coming from normal distributions centered around 50, corresponding to a literal interpretation, 66 (or $^2/_3$), corresponding to a pragmatic interpretation, and 100, corresponding to ascribing rationality to the speaker, with SD=2. Participants whose likelihood for

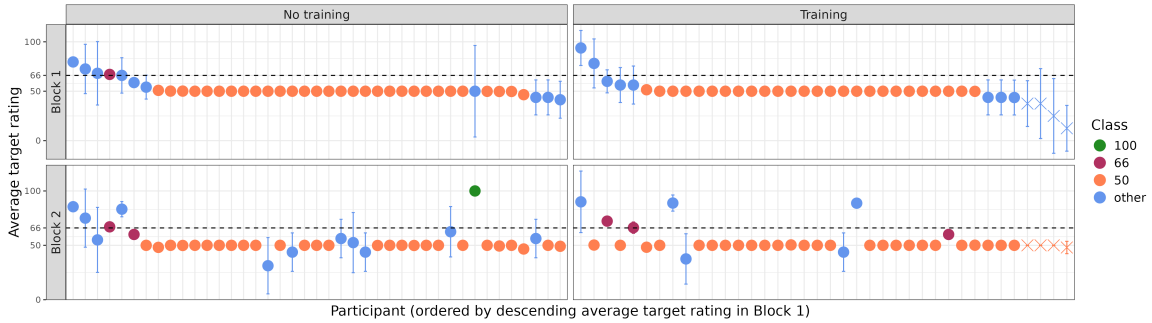


Figure 5.11: Individual participants' average target ratings on critical trials across blocks in the two conditions, ordered by their average rating in Block 1 (in descending order), with their assigned class.

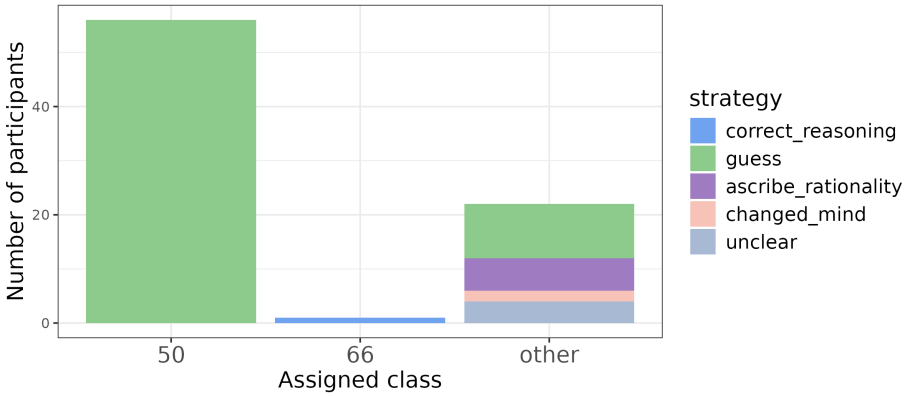


Figure 5.12: Alignment between the class assigned based on performance and participants' reported strategies for Block 1.

all three classes was at or below the threshold set based on a pilot study (10^{-30}) were classified as "other". This threshold controls how conservative our classification approach is, i.e., how much deviation from the above-mentioned means we can still confidently consider to belong to that class. We chose a fairly conservative threshold by visually inspecting the data.

Figure 5.11 shows participants' mean target ratings in each block with their assigned class and annotation of their reported strategy for Block 1. Due to implementation error, we did not obtain participants' reported strategies after Block 2. Analysis of annotations is only supplementary to the main analysis and is meant to provide more support for ratings aligning with people's actual strategies.

In both blocks in both conditions, most participants (65% or more) were assigned to the "50" class, indicating literal responding. We see that very few people were assigned to the "66" class, corresponding to correct reasoning about probabilities of alternative messages: in Block 1, only one person in the No training condition and none in the Training condition, and in Block 2, two people in the No training condition and three people in the Training condition.

It appears that people overwhelmingly interpreted ambiguous messages sent by

	Estimate	SE	t	p
Intercept	16.92	0.81	20.87	$<$ **0.001**
Trial Number	0.06	0.10	0.59	0.55
Block (2 vs. 1)	0.66	0.53	1.26	0.21
Condition (Tr. vs. No tr.)	0.24	1.13	0.22	0.83
Target Position (Left vs. Ctr.)	–0.29	0.49	–0.58	0.56
Target Position (Right vs. Ctr.)	0.15	0.49	0.31	0.76
Message Type (Color vs. Shape)	–0.11	0.22	–0.49	0.63
Block : Condition	–1.28	0.79	–1.63	0.10

Table 5.4: Regression model output for Experiment 4.

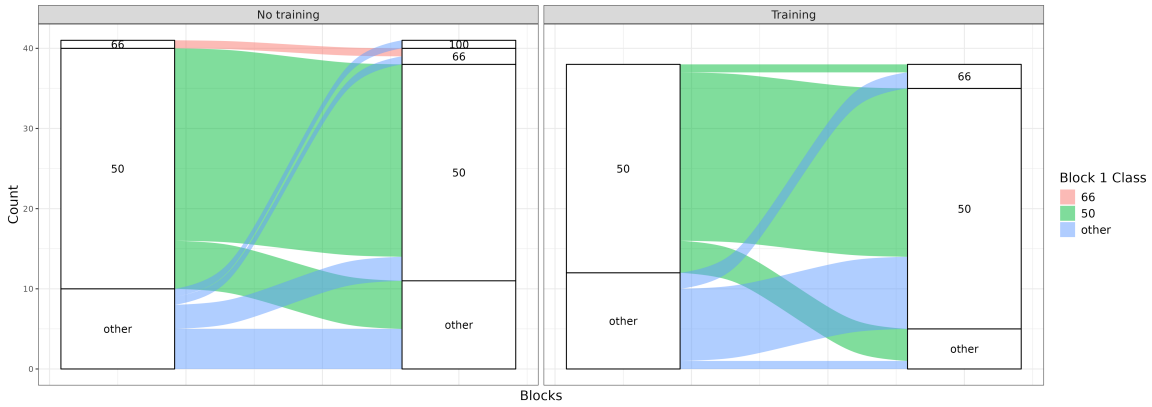


Figure 5.13: Changes in participants' class assignments based on performance across the two blocks in both conditions.

basic_message_picker literally, consistent with predictions of an L_0 model, and that nudging participants to consider alternative messages in training did not improve performance. Participants in the Training condition did not struggle with the added message selection step, as evidenced by their high accuracy on message selection in Block 2 (mean proportion correct = 0.95, $SD = 0.07$). This suggests that the lack of improvement is not caused by confusion about the added message selection step, or a failure to accurately enumerate alternative messages, but crucially by not exploiting those alternative messages in the main task.

To verify this observation and gain more insight into the effect of training, we fit a linear mixed-effects regression model to the data. We only used critical trials. The dependent variable was the absolute distance of the rating given to the target from the "correct" answer of $\frac{2}{3}$, computed as the minimum absolute distance from 66 and 67. The dependent variable was regressed onto block (1 or 2, dummy-coded with 1 as reference), condition (no training or training, dummy-coded with no training as reference), trial number (mean-centered), position of the target (left, middle, or right, dummy-coded with middle as reference), message type (color or shape, sum-coded with shape as reference), and the interaction between block and condition. The

random effect structure consisted of random intercepts per participant and per item.

For this analysis, we removed the four people in the Training condition marked with crosses in Figure 5.11 because their mean accuracy was below 40 and their by-trial performance showed inconsistent responding likely due to not paying attention or misunderstanding the task.

The results are reported in Table 5.4. There is no effect of condition ($\beta = 0.24$ (0.13), $p = 0.83$) or of block ($\beta = 0.66$ (0.53), $p = 0.21$), suggesting that people were not more likely to derive a correct interpretation with greater exposure. Importantly, the interaction between block and condition is not significant ($\beta = -1.28$ (0.79), $p = 0.10$): participants in the Training condition did not improve more than in the No training condition, indicating that the training did not make participants more likely to correctly weigh alternative messages.

As we mentioned above, for the regression analysis we excluded the four people with suspiciously low target ratings (below 40). If we keep those people in[2], the effect of condition and the interaction of block and condition become significant, suggesting that people in the Training condition were initially further away from the correct rating of $^2/_3$, but got closer after training. However, closer inspection of those four people's performance (Figure 5.11) reveals that this interpretation is misleading: these participants gave low ratings in Block 1, then consistently rated the target at 50 in Block 2, indicating a shift to a literal interpretation (50), not a pragmatic one (66). Therefore, making the conclusion that the training was helpful is not warranted. We interpret this as evidence that training improved task understanding but not the tendency to derive the pragmatic interpretation.

Next, we take a look at the changes in individual participants' class assignment based on performance between blocks in the two conditions, shown graphically in Figure 5.13. We see that, in both conditions, there is a large degree of consistency between the assigned classes, suggesting that most participants do not figure out the correct interpretation during the task. In both conditions, the majority of participants fall into the "50" class, and in the Training condition, the main change is that most participants previously classified as "other" join the "50" class, becoming more consistently literal responders. This supports the interpretation that while training helped clarify possible confusion about the task, it mostly led to strengthening of literal responding as opposed to an increase of pragmatic responding.

Finally, we examine the alignment of participants' reported strategies for Block 1 with the class assigned based on their ratings, to assess how closely participants' explanations match their behavior. Figure 5.12 shows the distribution of reported strategies by class. Since Block 1 is identical across conditions, we pool the data. We see that participants' explanations are generally consistent with class assignment based on ratings, but that a small part of guessers, as well as all participants who expected *basic_message_picker* to behave rationally (*ascribe_rationality*) were

[2]This exclusion criterion was not preregistered; we did not anticipate participants giving ratings below 50.

classified as "other". The reason that participants whose responses were labeled *ascribe_rationality* were assigned to the "other" and not to the "100" class seems to be that people provided lower ratings than 100, reflecting uncertainty about whether *basic_message_picker* would select the optimal message. This aligns with the observation in Experiments 1-3 that even people whose strategy was *correct_reasoning* had some uncertainty about their interlocutors' ability to reason rationally. Overall, this suggests that the classification based on ratings is a fairly reliable proxy for strategy.

5.5.3 Discussion of Experiment 4

In this experiment, we investigated whether people would correctly perform ad-hoc reasoning about alternative messages available to the speaker when the speaker was explicitly literal, a simple computer program which is equally likely to select any true message to refer to an object. We found that participants overwhelmingly did not incorporate probabilities of alternatives into their inferences and instead provided a fully literal interpretation, consistent with predictions of the literal listener RSA model. These findings are in line with those of the first two experiments, where we found that when interpreting messages purportedly sent by a child speaker, participants often behaved as literal comprehenders. This is noteworthy, as prior work in the same paradigm, including Chapter 4 of this thesis, has repeatedly shown that with a rational speaker, most participants behaved as pragmatic comprehenders (Degen et al., 2012; Franke & Degen, 2016).

Interestingly, nudging people to attend to alternative messages through training did not increase the rate of pragmatic interpretations. This suggests that weighing probabilities of messages ad-hoc is a difficult task which is not caused by inattention. This interpretation aligns with literature on errors in Bayesian reasoning in other domains, which shows that people often make errors in probabilistic computations (Fox & Levav, 2004b; Starns et al., 2019).

Of course, questions remain about the extent to which these findings transfer to real-life language use. It could be that people do not engage in this kind of reasoning about alternative messages ad-hoc, especially in an unfamiliar scenario, but that they are more successful with familiar linguistic alternatives. It is also possible that when dealing with a not-so-rational speaker, people choose to expend less effort and reason less deeply and fall back on a literal interpretation.

We also observed that some participants expected the speaker to behave rationally and ascribed intention to it despite having been explicitly told that the program is fully literal, again supporting the observation that reasoning about a speaker's intent comes very naturally and sometimes may be the default that needs to be suppressed.

The assumption behind assigning a $\frac{2}{3}$ probability to the target under a pragmatic interpretation is that each of the three objects is equally likely to be referred to. In our design, participants were told that the *basic_message_picker* was given a specific object to refer to on each trial, but they were not told how these targets were selected.

As a reviewer pointed out, participants might instead assume that the likelihood of an object being referred to is proportional to the number of expressions available to describe it: the competitor, which can be referred to in two ways, might be assumed to be twice as likely to be referred to as the target, which can only be described with one message. Under this alternative assumption, a target rating of 50 would align with a pragmatic rather than a literal interpretation. While this is a plausible interpretation, we consider it unlikely that most participants adopted such a complex prior. Future work should test whether explicitly stating that all three objects are equally likely to be referred to reduces the proportion of target ratings of 50.

Finally, since it appears that participants generally do not take probabilities of alternatives into account when performing ad-hoc reasoning about a literal speaker, it may be important to develop a cognitive model of reasoning that captures this tendency. While the Rational Speech Act model is primarily a model of pragmatic competence (though there have been some efforts of extending it to account for performance limitations, see e.g., Hawkins et al., 2021) and describes what humans are in theory able to do, it is also important to be able to accurately model performance, especially in cases where it seems to diverge significantly from the predicted competence.

5.6 General discussion

In this chapter, we presented four experiments where we investigated the influence of comprehenders' beliefs about the speaker on their pragmatic interpretation across several different speaker types: adults, 4-year-old children, artificial agents and an explicitly literal speaker. We hypothesized that people would not derive an implicature or derive a less strong one if the speaker to be less pragmatically competent. At the population level, we consistently found the expected effect: participants assigned higher ratings to the pragmatic interpretation when they thought that the ambiguous message came from an adult as opposed to the other speaker types. In the case of a child speaker, the effect seems to be related to beliefs about the child's pragmatic reasoning sophistication. In the case of the LLM-based chatbot ChatGPT, it is less clear that the beliefs about pragmatic competence give rise to the effect: when asked directly, participants rated ChatGPT's competence very highly. We hypothesized that in the case of ChatGPT the effect was more related to its perceived cooperativity: even if an artificial agent is capable of solving the task, they may not necessarily be motivated to do so in the most helpful way (Peña et al., 2023).

In the discussion of the individual experiments, we already touched on the nature of this speaker identity effect. Here, we discuss it in more detail. First of all, we asked whether this effect is gradient or categorical in nature: do fewer people derive an inference when interacting with a less pragmatically competent or less cooperative speaker, or do people derive a less strong inference? Indeed, both seem to be the case:

there were more literal-like responses in the child and ChatGPT speaker conditions than in the adult speaker condition (categorical), and those who derived an inference tended to derive a less strong one, evidenced by lower target ratings (gradient). The gradient part of the effect – namely, that participants derive a less strong inference – can be described by the RSA model proposed in Section 5.2.1, where the listener has beliefs about how likely the speaker they are interacting with is to be literal vs. pragmatic.

For the categorical part of this effect – namely, that there are more literal responders in the child and ChatGPT speaker conditions than in the adult speaker condition – there are several explanations which are not mutually exclusive.

First, we hypothesized that comprehenders may be less motivated to reason deeply when they do not believe the speaker to be very pragmatically competent or cooperative, behaving as literal listeners L_0 instead. It is known that we can process information more or less deeply depending on the context. In dual-process theories of cognition, these distinct processing mechanisms are framed as two systems: a quick, heuristic System 1, and a slower System 2 that engages in deeper processing (Kahneman, 2011; Stanovich, 2012; Evans & Stanovich, 2013). In language comprehension research, the influential *good enough processing* account states that in everyday language processing comprehenders often rely on heuristics and compute more coarse-grained, "good enough" representations unless the task requires them to expend more effort and create a more fine-grained interpretation (Ferreira et al., 2002; Ferreira & Patson, 2007). In line with these accounts, it could also be hypothesized that people, as comprehenders and reasoners, have different reasoning settings, a shallower reasoning setting that approximately corresponds to the L_0 listener model and a deeper reasoning setting that corresponds to the L_2 model. It could be, then, that the conversational context may motivate the comprehender to shift into one setting or the other: in the case when the interlocutor is another adult, people may be more likely to expect intentionality and cooperativity of their interlocutor, which motivates them to reason more deeply about possible intentions behind the utterance. This would suggest that reasoning types, which we expect to be relatively stable within individuals based on the findings of the previous chapter, may be flexible depending on the setting, and the same person may be an L_0 or an L_2 in different situations. Since our experiments were conducted between participants, we do not have a way of directly testing this. Thus, we leave it to future work to investigate the flexibility of pragmatic reasoning types depending on interlocutor type.

The second explanation of the fact that there are more literal-like responses in the child and ChatGPT speaker conditions, which may be complementary to the first, is that participants may be reasoning about the speaker's pragmatic competence or cooperativity but not correctly incorporating the probabilities of alternative utterances into their reasoning. As shown in Section 5.2.1, any pragmatic RSA listener model automatically correctly incorporates probabilities of alternatives into the reasoning. Thus, the L_1 listener model which reasons about a literal speaker predicts a target

probability of $\frac{2}{3}$ since in this setup, the target is twice as likely as the competitor purely based on the probabilities of alternatives. However, in Experiment 4 we showed that this prediction is at odds with what people do: when interpreting messages uttered by literal speaker, they overwhelmingly provided interpretations consistent with the predictions of the literal RSA model L_0, not L_1. Thus, people may have categorized the child or ChatGPT as a literal speaker but did not derive the probabilistically correct inference, which would be consistent with literature on reasoning errors in other domains. This also provides evidence for our assumption in the previous chapter that L_1 may not be an accurate description of people's reasoning process in the case of a literal speaker. We pointed out that the RSA in its current form does not allow to decouple reasoning about the probabilities of alternatives, which people appear to struggle with, from reasoning about the speaker, which people appear to do naturally. Developing a model which captures reasoning about rationality but does not correctly weigh the probabilities of available alternatives is a potential direction for future work.

For the reasons discussed above, the RSA mixture model of reasoning about the speaker's pragmatic competence described in Section 2.5.1 only covers the target probability range from $\frac{2}{3}$ to 1, preferring the target even when the speaker is believed to be fully literal. One relatively straightforward extension which would allow for the whole target probability range from $\frac{1}{2}$ to 1 to be covered would be to model the listener as a mixture model as well, namely a mixture of L_0 and L_2. This would be a way of formalizing the hypothesis that the listener's pragmatic type may vary depending on the perceived characteristics of the speaker, resulting in shallower or deeper reasoning.

In terms of generalizability of these results, reasoning needed to solve the critical trials in this study relies on reasoning about messages which are available and unavailable to the speaker: if both features of the target were expressible as messages, the message on critical trials would be fully ambiguous between the target and the competitor. One could argue that having not all features be expressible as words or utterances is an artificial assumption which does not hold for the way we communicate using language. That is likely the case. However, this derivation process needed to resolve the ambiguity – considering the alternative meanings and that there might be better utterances to express an alternative meaning – is very similar to the one that is assumed for implicatures in language (e.g., Noveck, 2001). Also, there may be situations in linguistic communication where there is another way to express a feature of a referent but it is less accessible because it is, for example, much less frequent. In this chapter, as discussed in Section 5.1.1, we decided to use stimuli corresponding to the simple condition from Franke & Degen (2016) and from Chapters 3 and 4, which relies on some features not being expressible as messages. We did so because we showed in the previous chapter that people perform worse in the complex condition, where all object features are expressible, than in the simple condition. If the task is too complex for participants to solve themselves, they would be unlikely

to additionally reason about the speaker. Future work could extend our findings by conducting similar studies in a setting where all object features are expressible. For instance, one could have an adult and a child record underinformative sentences and compare whether people draw less strong pragmatic inferences when hearing an utterance spoken by a child.

5.7 Conclusion

In this chapter, we investigated how variability in speakers' pragmatic competence is accommodated by comprehenders. The findings of this chapter contribute to the existing theories of pragmatic processing by providing evidence that the derived meaning is sensitive to the comprehenders' beliefs about the speaker's perceived attributes. Across four experiments with four different speaker types – adults, 4-year-old children, artificial agents and explicitly literal speakers – pragmatic inferences were shown to be less likely to be derived if the speaker was believed to be insufficiently competent or cooperative.

In addition, our findings point towards flexibility of listeners' reasoning depth in response to contextual factors: reasoning about a less pragmatically competent speaker resulted in a significantly increased number of participants who behaved as literal comprehenders, suggesting more shallow reasoning. Together, these results support a context-dependent view of meaning, where pragmatic meaning is not accessed automatically but is contingent on the listener's beliefs about the communicative situation. In terms of existing theoretical frameworks, our findings align with the core claims of relevance theory (Sperber & Wilson, 1986), which states that pragmatic meaning is only derived when sufficient contextual support justifies the additional processing effort.

We also proposed a model in the Rational Speech Act framework that formalizes how beliefs about the speaker's pragmatic competence can be incorporated into pragmatic inference, offering a computational account of the observed variability in comprehenders' interpretations.

This chapter established that speaker characteristics influence derived inferences through beliefs about the speaker's pragmatic competence. In this chapter, speaker characteristics were explicitly communicated through experimental framing. In the next chapter, we extend the investigation of the effect of speakers' perceived pragmatic competence on inferences to the setting where the speakers' characteristics are not explicitly communicated and ask whether participants will adjust their interpretations dynamically based on feedback during interaction.

Chapter 6

Dynamic adaptation to speakers' pragmatic competence

In the previous chapter, we presented evidence from four experiments that people's interpretations are sensitive to their beliefs about the speaker and the speaker's presumed pragmatic competence: if the speaker is not expected to be a sophisticated reasoner, a pragmatic inference may not be derived or a less strong one may be derived. In the previous chapter, speaker characteristics were explicitly communicated through experimental framing: participants were told what kind of speaker they were interacting with, such as a four-year-old child or an artificial agent, which activated the beliefs associated with that type of speaker. However, in many communicative situations, the speaker's characteristics such as pragmatic competence and cooperativity are not directly observable. In that case, the comprehender needs to infer these properties dynamically during interaction.

Such dynamic adaptation to the speaker during interaction is the focus of the current chapter. We investigate whether people dynamically construct a model of their interlocutor's pragmatic competence and adjust their interpretations if they receive feedback during interaction suggesting that the speaker is not behaving in a pragmatically optimal way. We provide experimental evidence that in a very short time period, with very few examples of non-pragmatic behavior, people adjust their interpretations. Since individual differences are a key focus this thesis, we also examine differences in adaptation behavior among participants and provide evidence for different adaptation patterns.

We extend the RSA model of comprehenders' reasoning about the speaker's pragmatic competence proposed in the previous chapter to incorporate dynamic belief updating.

This chapter is based on, and in parts identical to, the following publication:

- Naszádi, K.*, **Mayn, A.***, Duff, J., Monz, C., & Demberg, V. Listeners Adapt

to Speakers' Pragmatic Competence. (in prep.) *Equal contribution

Data availability

The experimental data, analysis scripts and RSA model code for this chapter are hosted on OSF at `https://osf.io/5d2f6/`.

6.1 Introduction

Previous work described in Section 2.4 suggests that comprehenders' derivation of pragmatic inferences is sensitive to their beliefs about the speaker. In Chapter 5, we provided evidence that interpretations are influenced by the speaker's perceived pragmatic competence. For instance, we showed in the previous chapter that participants are less likely to derive an implicature when they think that they are interpreting messages sent by a 4-year-old child as opposed to another adult.

However, in everyday interaction, relevant speaker characteristics may not be observable and may need to be inferred during interaction. For instance, we may be interacting with a person we just met and therefore know very little about. Franke & Degen (2016) showed that participants vary in terms of their reasoning depth; while they observed more variation in comprehension than in production, there was some variation in production as well, with some people being better described by the literal as opposed to the pragmatic Rational Speech Act speaker model. Moreover, a speaker's pragmatic competence may be temporarily reduced by factors such as fatigue or cognitive load. In such cases, assuming a high degree of speaker rationality may lead to misinterpretation – for instance, deriving an implicature when the speaker intended a literal meaning. Thus, listeners need to calibrate their interpretation to particular speakers over the course of the interaction.

There is some existing work suggesting that people construct a model of the interlocutor during interaction and derive partner-specific interpretations in the absence of salient characteristics given sufficient exposure. Ryskin et al. (2019) and Gardner et al. (2021) showed that people stop interpreting adjectives contrastively ("tall glass" meaning that there is a short glass in the common ground) given enough exposure to non-contrastive adjective use by a speaker. Roettger & Rimland (2020) found that comprehenders make use of intonation clues in discourse when the speaker uses intonation reliably but discard them with an unreliable speaker.

It has also been shown that participants construct speaker-specific models of the use of modals which express uncertainty, such as "might" and "probably" (Schuster & Degen, 2020) and learn to associate speakers with distinct communicative styles (Loy & Demberg, 2024).

In the experiment described in this chapter, we investigate whether participants will dynamically adjust their interpretation of ambiguous utterances, which can be

interpreted either literally or as an implicature, based on feedback during interaction, which would suggest rapid adaptation to the speaker's pragmatic reasoning depth. Our participants interact with two simulated partners, one of whom behaves like an optimally rational pragmatic speaker, while the other produces literal utterances. We find that the majority of our participants adjust their confidence in pragmatic interpretations based on accumulated evidence about the behavior of each of the two speakers.

We then model the dynamics of the adaptation process as Bayesian belief updating by extending the RSA model of reasoning about the speaker's pragmatic competence proposed in the previous chapter to capture learning over time. Our modeling results suggest that comprehenders may adapt not by directly observing speaker behavior but rather by reasoning about their own interpretations, as evidenced by the fact that adaptation occurs on trials which are ambiguous from the listener's perspective, even if they are unambiguous from the speaker's perspective. Additionally, we observe evidence for different adaptation strategies, which may be explained by different prior beliefs about the speaker and tendencies to generalize from one conversation partner to the next.

6.2 Background

6.2.1 Experimental paradigm

The experiment described in this chapter, like the previous chapters, uses a variation the reference game paradigm. Like in the previous chapter, the critical trials in this experiment correspond to the simple condition of F&D and in Chapters 3 and 4. The rationale for only using the simple condition is the same as for the previous chapter: we know from F&D and Chapter 4 that people's performance on the complex condition is lower than on the simple condition, and if the task is too complex for participants themselves to solve, they are unlikely to additionally engage in reasoning about the speaker's pragmatic competence.

Participants interacted with two simulated speakers, literal, whose behavior corresponded to the predictions of the literal RSA speaker model S_0, and pragmatic, whose behavior corresponded to the predictions of the idealized pragmatic RSA speaker model S_1 with $\alpha \to \infty$.

Like Chapters 3 and 4, this chapter used a forced-choice version of the reference game paradigm: on each trial, participants made their selection by clicking on one of the three referents. After that participants were asked to indicate their confidence that their selection is the referent that the speaker intended to communicate on a four-point scale. Finally, participants received feedback and were told whether they selection was correct and what the intended referent was.

In each speaker block, there were 24 critical trials, on which the target was ambigu-

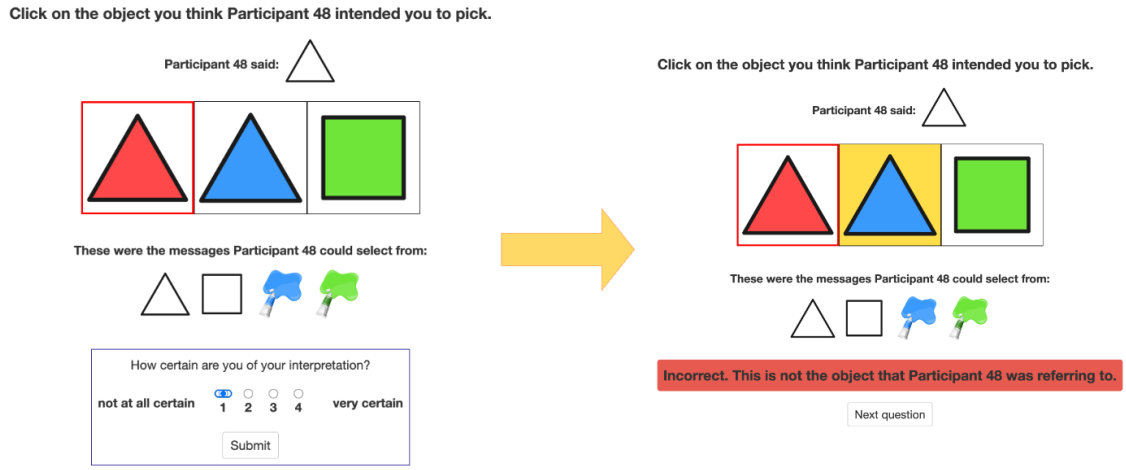


Figure 6.1: Example critical trial in the literal speaker condition. If the pragmatic target was selected, participants sometimes (in one third of cases) received feedback that it was incorrect.

ous. An example of a critical trial is presented in Figure 6.1. The message "triangle" is literally true of the red triangle and the blue triangle. The red triangle is the *pragmatic target* because its other feature, the color red, is not expressible as a message, whereas the other object, blue triangle, could be more optimally and unambiguously referred to using the message "triangle". In the pragmatic speaker condition, the pragmatic target is always the intended referent, corresponding to the predictions of the idealized S_1 RSA model. In the literal speaker condition, on one third of critical trials (8 out 24), the intended referent is not the pragmatic target but the other literally possible object, in this case, the blue triangle. This corresponds to the predictions of the literal speaker RSA model S_0. As discussed in detail in Section 2.6.2 of the theoretical background and in the previous chapter, the RSA predicts that when the speaker is literal, the ambiguous message will refer to the pragmatic target two thirds of the time. That is because there is only one way to refer to the target and two ways to refer to the competitor, so the literal speaker, who uses every literally true message with equal probability, will use the message to refer to the target twice as often. Therefore, with the literal speaker, if the participant selected the pragmatic target, on one third of critical trials they received feedback that their selection was not correct. Figure 6.1 shows an example trial from the literal speaker block where the participant selected the pragmatic target and received negative feedback.

We investigate whether participants' confidence in their interpretations will change over time based on the feedback they receive. We hypothesized that people's confidence in pragmatic interpretations with the literal speaker will decrease over time, and possibly that it will increase over time with the pragmatic speaker.

Unlike in the previous chapters, where the set of available messages remained the same – i.e., the same two shapes and two colors – throughout the experiment, in the current experiment, the set of available messages changed on every trial, but it always

included two shapes and two colors. This was done in order to increase the number of possible distinct critical trials since we wanted to have a high enough number of trials for participants to be able to learn. We also hypothesized that the fact that the available messages changed on every trial would make their relevance more salient to participants, which in this case is what we wanted: in previous chapters, we saw that fairly many participants behaved like literal L_0 listeners, indicating that they are unlikely to be able to additionally adapt to the speaker. We hypothesized that the current design – different available messages in combination with feedback on every trial – would lead to the majority of participants being able to figure out how to derive the implicature and thus be able to additionally reason about whether the speaker is likely to have intended it.

6.2.2 Modeling preliminaries

Model of uncertainty about the speaker's pragmatic competence

In Section 5.2.1 of the previous chapter, we proposed a model in the RSA framework which formalizes the comprehender's uncertainty about the speaker's pragmatic competence. We will use this model as the starting point and extend it to incorporate belief updating over the course of the interaction.

In that model, the listener is unsure whether the speaker is literal or pragmatic. Their beliefs are represented as a weighted combination of the literal speaker S_0 and the pragmatic speaker S_1:

$$L_2^{belief-driven}(o|u,\lambda) \propto S_{weighted}(u|o,\lambda) \cdot P(o)$$
$$\text{where} \quad S_{weighted}(u|o,\lambda) \propto \lambda \cdot S_1(u|o;\alpha \to \infty) + (1-\lambda) \cdot S_0(u|o) \quad \text{and } 0 \leq \lambda \leq 1$$

When $\lambda = 0$, the listener is certain that the speaker is fully literal; when $\lambda = 1$, the listener is certain that the speaker is fully pragmatic. Intermediate values reflect graded uncertainty about the speaker's reasoning sophistication.

While in Chapter 5 we treat speaker characteristics (e.g., child vs. adult) as determinants of a fixed λ value, the goal of the current chapter is to model how such beliefs are learned and updated dynamically through interaction.

Adaptation as Bayesian belief updating

Learning about the partner type from observations is often modeled as Bayesian belief updating (Schuster & Degen, 2020; Hawkins et al., 2021, 2017; Roettger & Franke, 2019; Kleinschmidt & Jaeger, 2015). After making a set of observations $D = \{(u_1, o_1), (u_2, o_2), .., (u_i, o_i)\}$, where each observation is an utterance-referent pair, the listener updates their beliefs about some parameter or set of parameters Θ according to Bayes' rule:

$$P(\Theta|D) \propto P(D|\Theta) \cdot P(\Theta)$$

In our case, the parameter of interest is the mixture weight λ from the model proposed in the previous chapter, which expresses the listener's beliefs that they are dealing with a speaker of a certain type. The prior $P(\lambda)$ encodes the listener's expectations about speaker rationality before interaction. We aim to capture the intuition that the listener will update their prior beliefs about the speaker's pragmatic competence during interaction, which will in turn impact their interpretations.

6.3 Experiment

In this experiment, we investigate listeners' adaptation to pragmatic competence of two simulated speakers in the reference game paradigm. The two speakers' behavior corresponds to the predictions of RSA models S_0, literal speaker, and S_1, pragmatic speaker who reasons about the literal listener L_0 and seeks to maximize communicative success, respectively.

6.3.1 Participants

101 native English speakers based in the UK with an approval rate of at least 95% were recruited via the crowdsourcing platform Prolific.

6.3.2 Procedure

Participants completed the reference game described in Section 6.2.1. They interacted with two simulated speakers, literal (corresponding to the predictions of the S_0 RSA model) and pragmatic (corresponding to the idealized S_1 speaker model with $\alpha \to \infty$). The order of speaker blocks was counterbalanced.

Each speaker block consisted of 32 trials, of which 24 were critical and the remaining 8 were unambiguous control trials. As described in Section 6.2.1, in the pragmatic speaker condition, on critical trials the pragmatic target was always the intended referent. In the literal speaker condition, on 8 out of 24 critical trials, the intended referent was the competitor.

The 8 unambiguous trials were included as a baseline. There were 3 types of unambiguous trials: 3 trials were completely unambiguous where no feature was repeated between the three possible referents; the 5 remaining trials had the same displays as critical trials except the target was the competitor (3 trials; filler type A) or distractor (2 trials; filler type B) from the critical condition.

The order of target, competitor and distractor on the screen was randomized for each trial and trial order was randomized for every participant, with the first trial always being an unambiguous control trial.

6.3.3 Results

4 participants whose accuracy on the unambiguous trials in either block was below 80% were excluded from analysis, resulting in 97 participants.

Figure 6.2 shows the change in distribution of participants' confidences by block over time for the two speaker orders. We see that, even in the literal speaker condition, participants predominantly chose the pragmatic target. Participants' reported confidence appears to follow a different pattern in the two blocks, increasing over time in the pragmatic speaker block and decreasing in the literal speaker block, indicating adaptation.

To verify this apparent effect, we fit a Bayesian ordinal regression model to the data using the *brms* package in R (Bürkner, 2017). Only critical trials were used for this analysis. Trials on which participants chose the distractor were excluded; those only constituted a very small proportion of total responses (10 out of 4656; 0.2%). The dependent variable was an 8-level ordinal variable (choosing pragmatic target vs. competitor × 4-point confidence scale). This way, 1 corresponds to choosing the competitor with a confidence rating of 4, and 8 corresponds to choosing the target with a confidence rating of 4. This variable was regressed onto speaker type (sum-coded, levels: -1 = literal, 1 = pragmatic), block order (sum-coded, levels: -1 = S_0 first, 1 = S_1 first), quarter of block (mean-centered), as well as all two- and three-way interactions. Target position (left, middle, or right; dummy-coded with middle as reference) and message type (sum-coded, levels: -1 = shape, 1 = color) were included as covariates. The random effect structure included per-participant and per-item random intercepts as well as per-participant random slopes for speaker type, block number, quarter of block, and interactions between speaker type and quarter of block and block number and quarter of block.

For all effects, very wide weakly informative priors were set to avoid biasing the model. For the intercepts, we set the prior to be a normal distribution with a mean of 0 and a standard deviation of 5. For all main effects, we set the prior to be a normal distribution with a mean of 0 and a standard deviation of 1.

We ran 6 chains of 12000 iterations each, with the first 2000 iterations of each chain discarded as warmup.

The regression results are reported in Table 6.1. Effects are taken as noteworthy if the 95% credible interval excludes 0. There is a main effect of speaker type ($\hat{\beta}$ =0.74, 95% CrI = [0.56, 0.92]), indicating that confidence is lower in the S_0 condition compared to the S_1 condition. There is also a positive effect of trial number ($\hat{\beta}$ =0.15, 95% CrI = [0.06, 0.23], indicating a general increase in confidence over time. This may appear puzzling given that we expected the confidence to decrease in literal speaker condition. Figure 6.2 suggests that, in the literal speaker block, there is a decrease in competitor responses (the blue part of the bars becomes smaller) and an increase in low-confidence target responses (the light red part of the bars becomes larger). That indicates that in the literal speaker block, participants shifted over

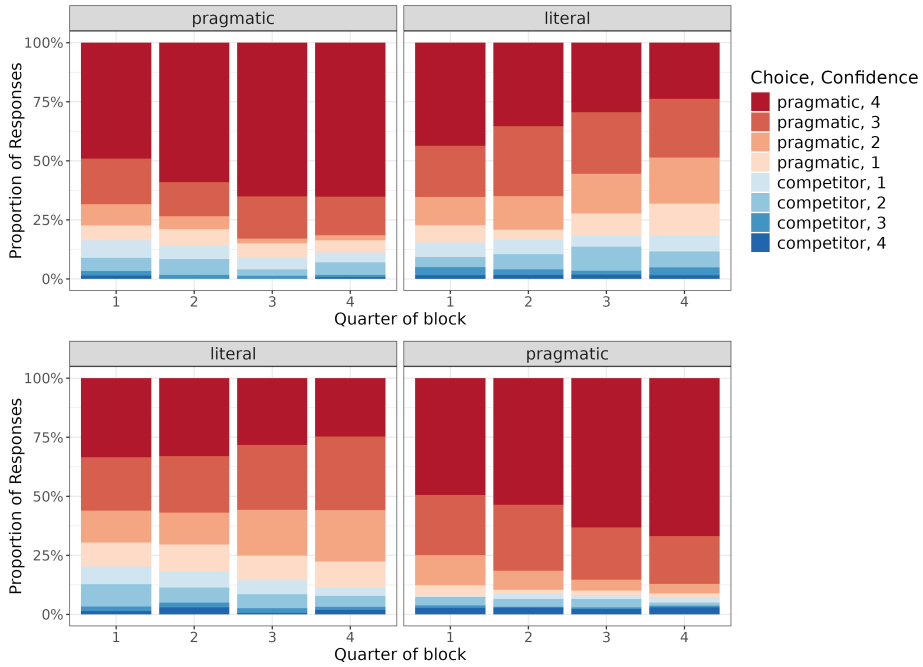


Figure 6.2: Confidence change by block for the two speaker orders.

time from selecting the competitor to selecting the target more often but with low confidence. This could be interpreted as participants figuring out what the "correct" answer should be – the pragmatic target – but lowering their confidence that the speaker will behave optimally. There is an interaction of speaker type with quarter of block ($\hat{\beta}$ =0.24, 95% CrI = [0.15, 0.32]), providing evidence for our observation that confidence develops differently for the two speakers. In the pragmatic speaker condition, confidence increases over time, whereas in the literal speaker condition, pragmatic interpretations with low confidence become more common. Finally, there is a block number by quarter of block interaction ($\hat{\beta}$ =-0.08, 95% CrI = [-0.16, -0.00]), which provides evidence that the order in which a speaker is seen (first or second) matters, although we do not find evidence for a three-way interaction of speaker type, block number, and quarter of block ($\hat{\beta}$ =0.00, 95% CrI = [-0.07, 0.08]).

6.3.4 Discussion

The presented experiment provided evidence that, when interacting with two interlocutors who differ in their pragmatic competence, people dynamically adjust their interpretations to the specific speakers. Importantly, they do so in the absence of any explicit mention of speaker characteristics, purely based on feedback on their interpretations. We find evidence of confidence developing differently in the two speaker conditions, with a gradual increase of confidence in the pragmatic speaker block. In the literal speaker block, participants seemed to converge on a pragmatic interpretation over time but give it a low confidence rating.

Parameter	$\hat{\beta}$	95% CrI
Speaker Type (S_1 vs. S_0)	0.74	[0.56, 0.92]
Block Number (Block 2 vs. 1)	-0.01	[-0.18, 0.17]
Quarter of Block	0.15	[0.06, 0.23]
Message type (color vs. shape)	-0.01	[-0.05, 0.03]
Target position (left vs. center)	-0.05	[-0.14, 0.04]
Target position (right vs. center)	-0.06	[-0.16, 0.02]
Speaker Type : Block Number	-0.04	[-0.33, 0.25]
Speaker Type : Quarter of Block	0.24	[0.15, 0.32]
Block Number : Quarter of Block	-0.08	[-0.16, -0.00]
Speaker Type : Block Number : Quarter of Block	0.00	[-0.07, 0.08]

Table 6.1: Bayesian ordinal regression model results.

It is interesting to note that, even in the literal speaker condition, participants by and large consistently chose the pragmatic interpretation but with low confidence, as opposed to guessing between the pragmatic target and the competitor. Consistently applying the pragmatic strategy is in this case indeed a more optimal comprehension strategy than guessing, as it will result in $\frac{2}{3}$ of correct interpretations because in $\frac{2}{3}$ of trials in the literal speaker block the pragmatic target is the intended object, as predicted by the S_0 RSA model. However, it is unclear whether the participants would be able to pick up on those statistics online; indeed, it seems more plausible that they do not, but that this preference for the pragmatic strategy is the result of them having figured out the way of solving the task that can be viewed as "correct" in that it results in ambiguity resolution.

Everything we've discussed so far has concerned adaptation at the population level. Next, will propose a Bayesian belief updating model of this adaptation process and examine the underlying mechanisms more closely, including strategies employed by individual participants.

6.4 Modeling perspective-taking underlying adaptation

In the experiment, we observed that participants track the speakers' pragmatic competence, as evidenced by lowered confidence in the pragmatic interpretation when interacting with a literal speaker. To zero in on the factors underlying this adaptation process, we now turn to modeling it in the RSA framework.

6.4.1 Deriving the likelihood function

In Section 6.2.2, we characterized learning about partner competence as Bayesian belief updating over a mixture of partner types. We now turn to defining potential likelihood functions $P(D|\lambda)$ that can serve as the basis for learning. Following Schuster & Degen (2020) and Hawkins et al. (2021), a natural starting point is to define the likelihood in terms of the speaker function $S_{weighted}$: the speaker behaves differently depending on the λ parameter; thus, the speaker's behavior provides information about their λ, i.e., how pragmatically competent they might be:

$$P(\lambda|D) \propto P(D|\lambda) \cdot P(\lambda) = \prod_i S_{weighted}(u_i|o_i, \lambda) \cdot P(\lambda)$$

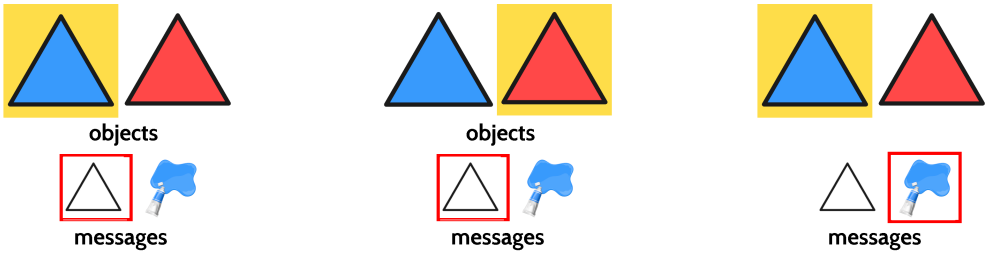
An alternative possible likelihood function is the listener function $L_2^{belief-driven}$ – that would correspond to the listener learning from *their own interpretations* as opposed to by observing speaker behavior directly:

$$P(\lambda|D) \propto P(D|\lambda) \cdot P(\lambda) = \prod_i L_2^{belief-driven}(o_i|u_i, \lambda) \cdot P(\lambda)$$

This is a subtle difference, which, however, results in distinct predictions with respect to the types of trials on which learning should occur. Thus, we have two competing updating models which differ in their likelihood functions, corresponding to learning by observing the weighted speaker $S_{weighted}$ (we'll call it the *speaker updating model*) or the pragmatic listener $L_2^{belief-driven}$ (the *listener updating model*), which can be teased apart based on their predictions. Bayesian updating models predict that updating occurs every time something can be learned from the observation: in our case, if the trial is unambiguous and every λ value would lead to the same outcome, no inference can be made about λ, and thus no update occurs.

There are three trial types on which one or both of the above adaptation models predict that updating would occur (Figure 6.3). Each of these trials shows two referents and two messages: we simplified the trials we used in our experiment by removing the distractor and the messages which are not true of either the target or the competitor since those have zero probability and do not affect the predictions of the model.

Figure 6.3a corresponds to the nonpragmatic critical trials in our experiment in the S_0 speaker block where the intended target was not the pragmatic target. Under both updating models, a negative λ update is predicted. Under the speaker updating model, the speaker selects the triangle rather than the blue paint to refer to the blue triangle. This is something that only S_0 might do - thus, after an update lower λ values are more likely. Under the listener updating model, the listener chooses the blue triangle upon receiving the message "triangle", which is also a lot more likely under lower λ values. Thus, on non-pragmatic critical trials the two updating models



(a) Trial type 1 (critical nonpragmatic). **(b)** Trial type 2 (critical pragmatic). **(c)** Trial type 3 (filler type A).

Figure 6.3: Three trial types for which the speaker updating model and the listener updating model make distinct predictions.

make the identical prediction that the listener should become more certain that they are dealing with an S_0.

For the other two trial types the two updating models diverge. Figure 6.3b shows a critical trial which corresponds to the critical trials in our experiment where the intended target is the pragmatic target (that is, all critical trials in the S_1 block and 16 out of 24 critical trials in the S_0 block). From the speaker's perspective, this trial is unambiguous: the only message which can be used to refer to the red triangle is "triangle". Thus, since both S_0 and S_1 would always behave the same way on this trial type, no updating is predicted to occur under the speaker updating model. In contrast, from the listener's perspective this trial is ambiguous: the triangle message is compatible with both referents, but the red triangle is more likely under higher λs; thus, a small positive update is predicted.

Finally, Figure 6.3c shows a trial where the opposite pattern is predicted: the speaker updating model predicts a positive λ update and the listener updating model predicts no update. This trial corresponds to one of our unambiguous trial types, Filler type A, of which there were only 3 in each block. From the speaker's perspective, selecting the blue paint to refer to the blue triangle is more likely for S_1 than for S_0. Thus, a small positive λ update is expected under the speaker updating model. From the listener perspective, this trial is completely unambiguous: the blue paint can only be referring to the blue triangle object. Thus, since with either underlying speaker model, S_0 or S_1, the listener will always select the blue triangle when receiving the message "blue", no update is predicted from the listener's perspective.

For the other two filler types (type B and unambiguous), neither model predicts an update since they are unambiguous from both the speaker's and the listener's perspective. Table 6.2 summarizes updates which are predicted for different trial types.

In order to determine which of these two models is a more accurate reflection of the adaptation process in our experiment, we examine the data and focus on whether people's confidence on critical trials improved following critical pragmatic trials, consistent with the listener updating model, or following filler type A trials,

Trial type	Updating using S_{mixed}	Updating using $L_2^{belief-driven}$
Critical pragmatic	no update	S_1 **slightly more likely**
Critical nonpragmatic	S_0 **much more likely**	S_0 **much more likely**
Filler type A	S_1 **slightly more likely**	no update
Filler type B	no update	no update
Unambiguous	no update	no update

Table 6.2: Predictions of the two updating models for the different trial types in the reference game.

	β	SD	p
Intercept	0.16	0.04	<0.0001
Filler type A	-0.15	0.13	0.26
Block (S_1 vs. S_0)	-0.17	0.05	<0.001
Filler type A : Block	0.23	0.17	0.17

Table 6.3: Estimates of a regression model verifying which trial type leads to confidence increase.

consistent with the speaker updating model. To verify which of the two updating models' predictions describe our data better, we fit a simple regression model. The dependent variable is the difference in confidence between the current critical trial and the previous critical trial (for this analysis, only the subset of the data where the previous critical trial was critical pragmatic is used). This variable was regressed onto a binary predictor corresponding to whether there was a filler type A trial in between the current pragmatic trial and the previous pragmatic trial or not (dummy-coded). Since learning from positive feedback may look different with the two speaker types, we also included speaker type (S_1 vs. S_0) and the interaction between block and filler type A as predictors. A significant positive intercept would indicate that participants' confidence improves following critical pragmatic trials, consistent with the listener updating model. A significant positive effect of filler type A would suggest an improvement following filler type A trials, consistent with the speaker updating model.

The model reveals a significant positive intercept (β =0.16 (0.04), p<0.0001) and no effect of filler type A (β =-0.15 (0.13), p=0.26), suggesting that participants' confidence increased following critical_pragmatic trials, not filler type A trials, providing evidence for the updating model which uses the listener model $L_2^{belief-driven}$ as the likelihood function. Notably, the model also reveals an effect of block (β =-0.17 (0.05), p<0.001, indicating that the learning effect from critical pragmatic trials was stronger in the S_0 speaker block than in the S_1 speaker block. Examination of data suggests that this is because in the S_1 block participants often start out with maximum confidence, and it does not change over the course of the experiment.

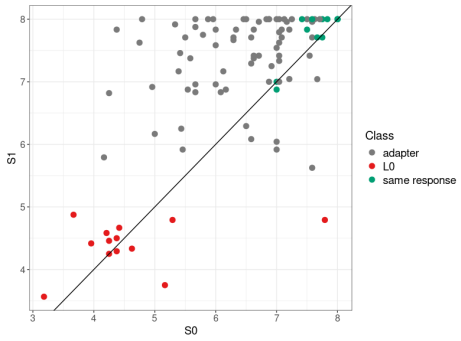


Figure 6.4: Individual participants' average confidence on the two speaker blocks.

In summary, it appears that, in contrast to how belief updating has been implemented in related work on pragmatic adaptation (Schuster & Degen, 2020; Hawkins et al., 2017, 2021), an adaptation model with the pragmatic *listener* model and not the speaker model as the likelihood function is supported by our data. Thus, people appear to be learning about the speaker not by making inferences about the speaker's behavior directly but about their own interpretations: they learn from trials which are ambiguous from their own perspective but unambiguous from the speaker perspective. We discuss the implications of this finding in Section 6.5.

Now all preliminaries are in place to fit the updating model to the experimental data.

6.4.2 Fitting the proposed adaptation model to the data

Now, we fit the Rational Speech Act adaptation model to the experiment data. Because not every subject showed an adaptation effect – some participants behaved like literal listeners, suggesting that they were not able to solve the task themselves, while others always gave an identical response – we first identify a subset of 72 "adapters" who our model is fit to describe. We fit the model with different hyperparameter combinations for each participant and discuss the model fit and the interpretation of those hyperparameters.

Identifying adapters

While so far we have focused on the population-level adaptation effect, it is important to note that there are individual differences in participant performance and that not every participant showed an adaptation effect. Figure 6.4 shows individual participants' average confidence on the 8-point confidence scale (the dependent variable from our regression models, where 1 corresponds to choosing the competitor with high confidence and 8 corresponds to choosing the target with high confidence) on the two speaker blocks.

First, we see that there is a cluster of participants whose average confidence on

both blocks is low (the red cluster in Figure 6.4). These participants are grouped into a separate class by the best-fitting Latent Profile Analysis model based on BIC and an analytic hierarchical fitting procedure based on AIC, AWE, BIC, CLC, and KIC (Akogul & Erisoglu, 2017). The resulting LPA has 3 classes; the other 2 classes are not so clearly interpretable. When we inspect these participants' responses per trial, it becomes apparent that they are highly inconsistent, often oscillating between the target and competitor. This suggests that these participants were not able to solve the task pragmatically and thus did not adapt. These participants can be thought of as being literal listeners L_0, consistent with Franke & Degen (2016), who observed that a subset of their participants fell into that class, as well as with the findings of Chapter 4. Since our model assumes that participants are pragmatic listeners reasoning about speakers, it is not meant to model the behavior of literal listeners; therefore, we exclude those participants from the modeling analysis.

Second, we exclude participants who gave the same response on every or almost every trial (up to 3 trials different from the rest across the two blocks) across the two blocks (marked by green dots in Figure 6.4).

We thus fit the model to the resulting subset of 72 of 97 participants. We return to the discussion of excluded participants and the strategies they use in Section 6.5.

Hyperparameters: shape of the prior and degree of generalization between speakers

There are three hyperparameters in our proposed model: α, *favor* and β. We will later compare models with different hyperparameter settings.

Hyperparameter α and *favor* control the shape of the prior over mixing weights, $P_{Prior}(\lambda)$. $favor$ is binary, 0 or 1, and it determines whether the prior favors S_0 (corresponding to lower values of λ being more likely) or S_1 (corresponding to higher λ values being more likely). α, in turn, expresses the strength of prior beliefs, with higher α values corresponding to a stronger preference. The prior is described by a Beta distribution:

$$P_{Prior}(\lambda|\alpha, favor) = \begin{cases} Beta(\alpha+1, 1) & \text{if favor} = 1 \\ Beta(1, \alpha+1) & \text{if favor} = 0 \end{cases}$$

Thus, for example, the combination of $\alpha = 5$ and $favor = 0$ describes moderate prior beliefs that the speaker is an S_0, and the combination of $\alpha = 50$ and $favor = 1$ corresponds to strong prior beliefs that the speaker is an S_1.

β is a hyperparameter which controls *generalization between speakers* – that is, how strongly interacting with the first speaker influences the expectations about the pragmatic competence of the second one. It can also be thought of as the degree of generalization from a specific partner to the population. β ranges between 0 and 1. At $\beta = 0$, the two speaker blocks are entirely independent, and at the beginning of Block

2, the prior probability distribution $P_{Prior}(\lambda)$ is reset back to the shape defined by α and *favor*. At $\beta = 1$, there is complete generalization: the posterior over λ at the end of Block 1 becomes the prior for Block 2. Intermediate values of β result in a mixture of these two sources: $P_{Prior}(\lambda_{Block2}) = \beta \cdot P_{Posterior}(\lambda|D_{Block1}) + (1 - \beta) \cdot P_{Prior}(\lambda)$.

Fitting the model

In order to make sure that we sufficiently explored the hyperparameter space, we fit models with the following hyperparameters: α from 0-50 with the interval of 5 (11 values), $favor$ – 0 or 1, β from 0-1 with the interval of 0.2 (6 values). Since $\alpha = 0$ corresponds to a flat prior, $P_{Prior}(\alpha = 0, favor = 0) = P_{Prior}(\alpha = 0, favor = 1)$. Thus, this results in a total of $11 \cdot 2 \cdot 6 - 6 = 126$ hyperparameter combinations.

Next, we computed the predictions of the adaptation model for each subject and hyperparameter combination. Since the model outputs the $P(target)$ for each trial, whereas participants in the task gave responses on a 4-point confidence scale, we needed to map model probabilities to the confidence scale. We did so by fitting a simple Bayesian ordinal probit model for each subject and hyperparameter combination, with the 8-point response (the measure from the regression model for the experiment, 4 confidences × target or competitor)[1] as the dependent variable and the RSA probabilities as the only predictor. The ordered probit regression model split the continuous latent variable ($\eta_i = a + b \cdot logit(x_i)$, where we defined $a \sim Normal(0, 2)$ and $b \sim HalfNormal(2)$) into 8 categories with 7 cutpoints using an ordinal probit likelihood, with an ordered transform to ensure that the cutpoints are strictly increasing, i.e., that a higher RSA probability corresponds to a higher confidence. The spacing between the cutpoints was allowed to vary. We fit 4 chains for 2000 iterations each, with the first 1000 iterations discarded as warmup.

Since the ordinal model was fit for each subject, the same RSA probability may correspond to a different confidence value for different subjects. This does not seem implausible to us, since it is known that people may differ in the way they utilize a scale (Van Vaerenbergh & Thomas, 2013; Ulitzsch et al., 2024) and a rating of e.g. 3 may correspond to different underlying probabilities for different individuals.

Next, we assigned one hyperparameter combination to each subject that resulted in the highest possible prediction accuracy. Since models sometimes tied for accuracy for a subject, we broke the ties in a way that resulted in the smallest number of hyperparameter combinations across the population. In total, there were 16 different hyperparameter combinations, with 11 of those 16 combinations assigned to only 1 or 2 subjects. The retained hyperparameter combinations assigned to at least 1 subject are shown in Figure 6.5. We can see that all but two participants were assigned a nonnegative α, indicating that at the population level, there was generally a prior preference for S_1 over S_0, or a flat prior. Most α values are ≤ 15, indicating that this

[1]For ease of model fitting, distractor selections were treated as competitor selections. This only affected 0.2% of trials.

preference was often moderate and overridable with experience. β values across the whole spectrum are represented, suggesting that participants varied in their tendency to generalize from the first speaker to the second. Individual subject assignments are reported in Figure 6.6.

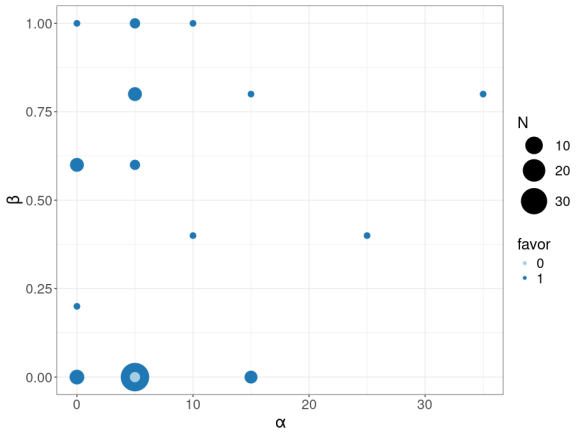


Figure 6.5: Retained hyperparameters assigned to at least one subject.

We will discuss the 5 most common hyperparameter combinations (assigned to more than 2 subjects); for the rest, the reader is invited to consult Figure 6.6. The five most common hyperparameter combinations were: (1) moderate prior preference for S_1, no generalization across speakers ($\alpha = 5$, *favor*$= 1$, $\beta = 0$; 37 subjects), (2) flat prior, no generalization across speakers ($\alpha = 0$, $\beta = 0$; 6 subjects), (3) moderate prior preference for S_1, strong generalization ($\alpha = 5$,*favor*$= 1$, $\beta = 0.8$; 5 subjects), (4) flat prior, fairly strong generalization ($\alpha = 0$, $\beta = 0.6$; 5 subjects), and (5) strong preference for S_1, no generalization across speakers ($\alpha = 15$, *favor*$= 1$, $\beta = 0$; 4 subjects).

Given that about half of our sample (37 out of 72 participants) is best described by the same hyperparameter combination ($\alpha = 5$, *favor*$= 1$, $\beta = 0$), how much evidence is there for individual-level hyperparameter settings? While the gain in accuracy at the population level is not very large (71.8% if we assign the most common hyperparameter combination to everyone vs. 76.0% with the individual hyperparameter combinations), it is significant ($t(71) = 5.02$, $p<0.0001$). Thus, positing five individual-level hyperparameter combinations as opposed to just one is justified.

An example of a participant for whom the most common hyperparameter combination ($\alpha = 5$, *favor*$= 1$, $\beta = 0$) yields a much worse fit that their assigned combination ($\alpha = 10$, *favor*$= 1$, $\beta = 1$), is shown in Figure 6.7. Prediction accuracies for these combinations are 58.3% and 87.5% respectively. The best hyperparameter combination at the population level assumes that there is no generalization from the first speaker to the second ($\beta = 0$). It fails to account for the fact that the participant's confidence decreased from 8 to 7 during the first speaker block (S_0) and remained at that level for the majority of the second speaker block (S_1). In other words, the

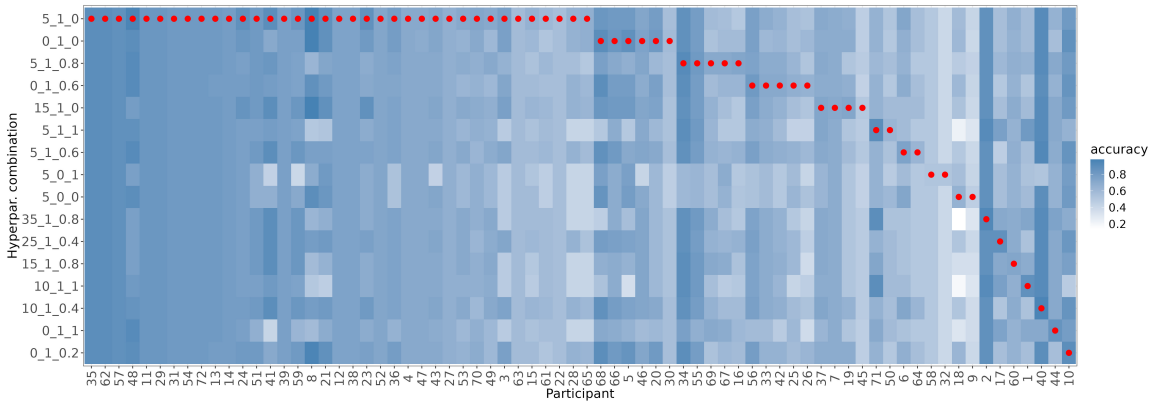


Figure 6.6: A heatmap of RSA model accuracies for every participant and hyperparameter combination $(\alpha, favor, \beta)$. The hyperparameter combination assigned to each subject is marked with a red dot.

participant needed to see a lot of evidence of the second speaker's rationality before their confidence went back up, which is consistent with the high β value.

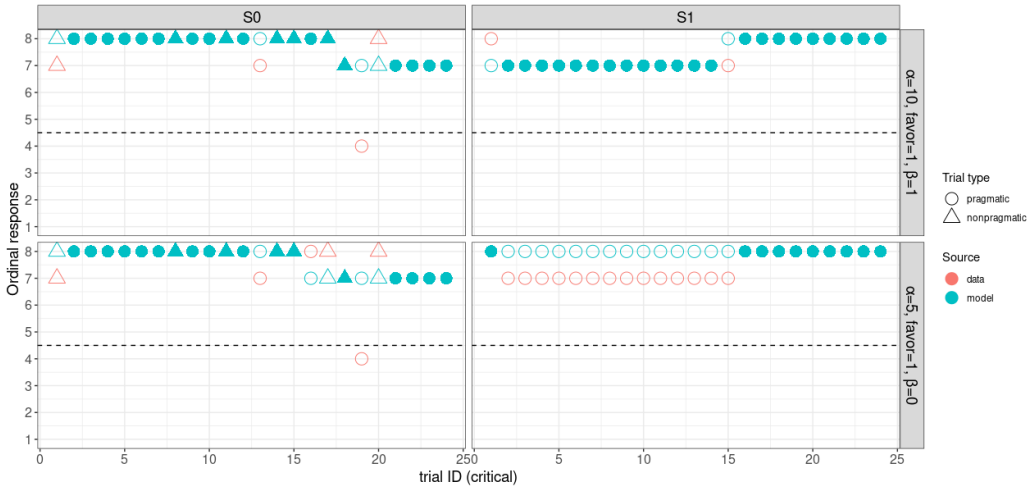
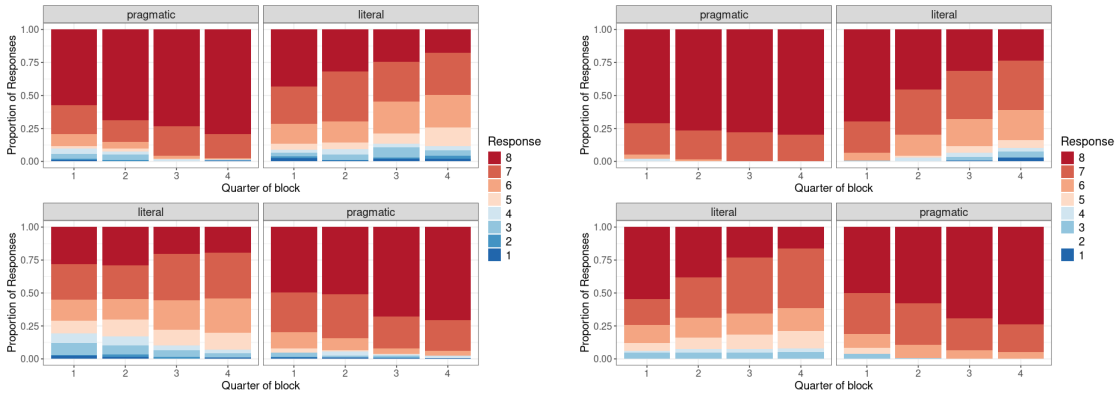


Figure 6.7: Model predictions for Participant 1 of models with the best hyperparameter combination for that participant ($\alpha = 10$, $favor$=1, $\beta = 1$; top panel) vs. population-best hyperparameter combination ($\alpha = 5$, $favor$=1, $\beta = 0$; bottom panel). Filled shapes denote correct predictions. Points above the dashed line indicate target selections, and points below the line indicate competitor selections.

6.4.3 Discussion of the model

As mentioned in the previous section, our proposed adaptation model with individual hyperparameter combinations achieves an average prediction accuracy of 76.0%. This suggests that the model is able to successfully predict important trends in par-



(a) Confidence change by block in the data.

(b) Confidence change by block predicted by the model.

Figure 6.8: Confidence change in the data (top panel) vs. model predictions (bottom panel) for the 72 adapters.

ticipants' adaptation behavior. The fact that an individual-level model yielded a significantly better accuracy than a population-level model provides more evidence for the role of the prior (modeled by the hyperparameters α and *favor*) and of generalization from the first partner to the next (modeled by β). Here, we would like to discuss the patterns of participant behavior which the model successfully captures, as well as patterns we observe which our model does not predict.

Figure 6.8 shows a barplot comparing the change in confidence during the experiment between the actual data and the model predictions for the two speaker block orders. We see that the model is able to capture the trend of increasing confidence in the pragmatic speaker block and decreasing confidence in the literal speaker block quite well. The distribution of confidences is also quite similar.

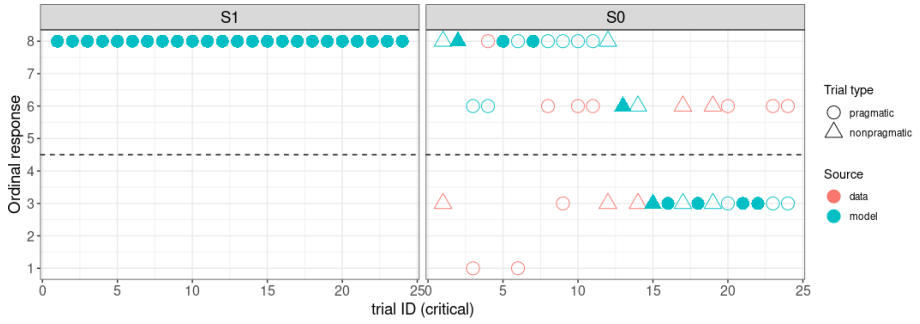


Figure 6.9: Example of a participant who appears to employ a guessing strategy on the S_0 block. Filled shapes denote correct predictions. Points above the dashed line indicate target selections, and points below the line indicate competitor selections.

However, we see that the model overpredicts the confidence on the pragmatic speaker block when it comes first: it barely predicts any responses except 7 and 8,

and no competitor selections. We assume that the reason for competitor selections in the participant data was due to the fact that it took them a while to figure out the task and optimal strategy, which our model does not capture.

Relatedly, we notice that several subjects' choices oscillate between the speaker and the competitor on the S_0 block, suggesting that they employed a guessing strategy. Figure 6.9 shows a participant for whom that is the case. They seem to have realized that the speaker is not behaving optimally, but that led them to match the speaker's guessing strategy. We observed this L_0 behavior when interacting with an S_0 in Experiment 4 of the previous chapter, where we showed that when interacting with a speaker who is explicitly literal – a simple computer program which randomly selected any fitting message – participants tended to fall back on the L_0 guessing strategy as opposed to favoring the pragmatic interpretation, even though it was statistically more likely. Our proposed model currently does not account for this behavior since it assumes that the listener is a pragmatic $L_2^{belief-driven}$ listener with an internal speaker mixture model $S_{weighted}$.

Future work could extend the proposed model by assuming that the listener is also a mixture model of L_0 and L_1 with a certain mixing weight and that the listener weight gets updated in addition to the internal speaker weight. Hawkins et al. (2021)'s model of perspective taking implements the idea that both communication partners have a certain mixing weight, although their interpretation of those weights is a bit different. This extension would additionally require relaxing the strictly increasing assumption which we used for fitting the RSA model prediction to confidences since the same $P(target)$ would result in guessing, i.e., multiple confidence values.

It can be seen in Figure 6.9 that the model predicts a large number of competitor responses on the S_0 block. One may ask why that would ever happen given that, as we explained in Section 5.2.1 of the previous chapter, even with a completely literal speaker, the model predicts that the listener will always favor the target with $P_{L_2}(target) \geq \frac{2}{3}$. That is the result of how we performed ordinal model fitting. In the ordinal model, the 8-point ordinal response variable, which encodes both choice and confidence (target vs. competitor $\times$ 4-point confidence response), is mapped onto the probability of target predicted by the RSA model, with the only restriction being that the mapping should be strictly increasing, i.e., that a higher response value should correspond to higher target probability. Since we do not specify in this mapping procedure which ordinal response values correspond to the target vs. the competitor, there may be mappings where some portion of the target probability range are mapped to the competitor selection, as is the case in Figure 6.9. If we interpret the fact that the predicted $P(target)$ is always above $\frac{2}{3}$ to mean that the competitor should never be selected, one could restrict the fitting procedure to map the predicted probability only to the upper half of the response range; however, that would result in competitor responses never being predicted, which is not supported by the data. As mentioned above, future work could explore alternative techniques for mapping the probability of target predicted by the RSA onto participants' responses.

6.5 General discussion

In this chapter, we provided evidence that people rapidly adapt their interpretations to speakers' pragmatic competence based on feedback and decrease their confidence about a pragmatic interpretation when their experience interacting with a communication partner suggests that they are not behaving optimally. We demonstrated that this happens in the absence of any explicit mention of the speaker's characteristics, purely based on feedback about whether the participants' interpretation was correct.

We also proposed a model of the mechanisms underlying this adaptation process in the RSA framework, where we formalized it as a Bayesian belief updating process. Our results suggest that, in contrast to previous work which has modeled listener adaptation as learning from speaker behavior (Schuster & Degen, 2020; Hawkins et al., 2017, 2021), in some settings people may be learning by reasoning about their own interpretations: we find evidence of learning on trials which are ambiguous from the listener perspective while unambiguous from the speaker perspective. Since perspective-taking is known to be effortful (Hawkins et al., 2021), and presumably all the more so in a novel setting like that of reference games, it may be effort-preserving to reason about one's own interpretations.

We find that our model, where a prior over speaker types, a literal speaker S_0 and a pragmatic speaker S_1, is adjusted over the course of the interaction, provides a good fit to the data and successfully captures the patterns of confidence change. Fitting the model also provided insights into adaptation strategies which differ among individuals. About half of the participants were best described by a model which has a moderate prior preference for a pragmatic speaker over a literal one and treats the two interlocutors as independent, indicating that those participants' the experience with the first speaker did not affect their expectations for the second one. This is consistent with work on convention formation which has observed that while people quickly converge on shorter referring expressions with a partner, when interacting with a new partner, they return to longer, more descriptive referring expressions, treating the new partner as independent from the previous one (Wilkes-Gibbs & Clark, 1992).

However, we observe evidence that this does not hold for all participants: in the other half of our sample, different degrees of generalization are represented, such that some participants' confidence judgements when interpreting the utterances of the second speaker appear to be influenced by their experiences with the first one. This suggests a generalization process where those participants generalize what they learn from a specific partner onto the whole population, a process which has also been observed in convention-formation (Hawkins et al., 2017). This can be explained via a dual-pathway interpretation model, similar to one that has been proposed for spoken word processing (Wu & Cai, 2024): it states that when interpreting speech, the listener is guided by both a speaker-specific model and a population model. When interacting with a new speaker, a generic population-level model is used, and as the listener gains more experience with a specific speaker, the listener develops a speaker-

specific model. Under this account, the two models interact with each other. Our results seem to suggest that a similar account may hold for pragmatic adaptation, and also that for different people, the way in which these models interact is different: some update their population-level model much more strongly or rapidly than others.

While our work sheds light on the dynamic adaptation of speakers' reasoning sophistication and different strategies people may be applying, some open questions remain. We found, for instance, that while at the population level there was clear evidence of adaptation, some participants did not adapt. What distinguishes adapters from non-adapters? In a closer look at individual participants' data, we saw that some people responded in a way that indicated guessing, suggesting that they did not figure out what the pragmatic interpretation should be. In Chapter 4, we observed that participants with lower Theory of Mind and logical reasoning abilities, measured by Raven's Progressive Matrices and Cognitive Reflection Test, were more likely to adopt literal as opposed to pragmatic interpretations. The same factors are likely to be at play in the current experiment.

In addition, there were some participants who consistently favored the pragmatic interpretation and did not cancel the inference when interacting with a literal speaker despite negative feedback. It is less clear what distinguishes those participants from adapters. It may be that they are less attuned to tracking speaker-specific behavior, which would be in line with previous work on individual differences in humor comprehension (Loy & Demberg, 2024) and uncertainty expressions (Schuster et al., 2023), which found that participants with higher working memory updating abilities were more likely to adapt differentially to specific speakers. Another explanation – and the two are not mutually exclusive – may be that some participants had very strong beliefs about the pragmatic interpretation being "correct" since it resulted in ambiguity resolution and were unwilling to step away from it despite evidence to the contrary. Future work could investigate whether willingess to adapt to a less pragmatically sophisticated interlocutor is related to cognitive or personality individual differences, such as, for instance, openness. Finally, we observed that people differed in their prior beliefs about speaker rationality and the degree to which they made generalizations from the first speaker to the second one.

6.6 Conclusion

In this chapter, we demonstrated evidence of rapid adaptation to two speakers' pragmatic competence during interaction based on feedback, in the absence of any explicit cues to those speakers' pragmatic abilities. When interacting with a literal speaker, participants lowered their confidence in pragmatic interpretations. These findings contribute to the growing body of work suggesting that comprehenders construct a mental model of a specific interlocutor during interaction which guides their interpretations.

A further contribution of this chapter is the formalization of this adaptation process as a Bayesian belief updating model. Our modeling results suggested that when perspective-taking is effortful, people may learn by reflecting on their own interpretations as opposed to through direct observation of the speaker.

Finally, since one of the focus areas of this thesis is individual differences, we examined differences in adaptation behavior and found evidence that not all participants adapted in the same way. About a quarter of our sample did not adapt, either behaving as literal comprehenders, consistent with the findings of Chapter 4, or consistently interpreting the message pragmatically despite repeated feedback that such interpretations were sometimes incorrect. As for the rest, modeling results suggest that most participants had the prior belief that their interlocutor is pragmatic, but the strength of this belief differed between individuals, and that they varied in how strongly they generalized from one interlocutor to the next. This suggests an account, similar to those that have been proposed for spoken word processing (Wu & Cai, 2024) and convention formation (Hawkins et al., 2017), where people have a speaker-specific and a population-level mental model of the interlocutor and those two models may interact. Our results suggest that the degree to which they interact varies among individuals.

Up until this point, all chapters of this dissertation have used a variant of the reference game paradigm. This was done due to the paradigm's simplicity and flexibility and because it lends itself well to modeling since the sets of possible utterances and meanings are enumerable and shared between the speaker and the comprehender. However, as discussed in Section 2.2.3 of the theoretical background, reference games are a highly simplified communicative setting. Notably, they use symbols instead of words, which may raise concerns about the generalizability of the results to more complex communicative situations. Therefore, in the next chapter of the thesis, we extend the adaptation experiment described in this chapter to a more complex paradigm with linguistic stimuli, a block building game where participants interpret underspecified building instructions.

Chapter 7

Beyond reference games: Pragmatic adaptation in the block building paradigm

One of the central research questions of this thesis is how listeners accommodate differences in speakers' pragmatic competence. In Chapter 5, we showed that comprehenders' beliefs about the pragmatic abilities of different types of speakers – including children, adults and artificial agents – influence their interpretation of ambiguous utterances. Chapter 6 further suggests that comprehenders construct a mental model of the speaker's pragmatic competence and dynamically update it over the course of an interaction. In that study, participants interacted with two speakers: one who always selected the pragmatically optimal message (pragmatic speaker), and another who produced true but sometimes suboptimal utterances that resulted in ambiguity when a better, unambiguous message was available (literal speaker). We observed that participants learned from feedback, which was reflected in systematic changes in participants' confidence: confidence in the pragmatic interpretation decreased over time with the literal speaker, and increased with the pragmatic speaker.

All previous chapters of this thesis, including Chapters 5 and 6, have used variants of the reference game paradigm. As discussed in Section 2.2.3 of the theoretical background, reference games are an easy-to-implement, flexible experimental paradigm that lends itself naturally to modeling within the Rational Speech Act framework. However, the same features that make reference games so easy to use and to model – the limited set of possible utterances and meanings that is shared between the interlocutors – also constitute an important limitation. Natural communication using language is substantially more complex, and the set of possible meanings and alternatives is very large and potentially unbounded. This raises the question of whether the findings about adaptation in the reference game generalize to more naturalistic communicative contexts.

The present chapter builds on the findings of Chapter 6 by extending the investigation of dynamic adaptation to speakers' pragmatic competence to a more naturalistic paradigm with more complex linguistic stimuli. The paradigm used in this chapter is a block building game, where participants build 3D block structures following potentially underspecified instructions provided by simulated interlocutors. Observing similar adaptation effects to those found in the reference game setting in the previous chapter would provide further evidence for the generalizability of our findings.

This chapter is based on, and in parts identical to, the following publication:

- Naszádi, K.*, **Mayn, A.***, Duff, J., Monz, C., & Demberg, V. Listeners Adapt to Speakers' Pragmatic Competence. (in prep.) *Equal contribution

Data availability

The experimental data and analysis scripts for the experiment presented in this chapter are hosted on OSF at `https://osf.io/g6ktz/`.

7.1 Introduction

When faced with an ambiguous utterance, the comprehender has to decide how to interpret it, using contextual clues to infer the intended meaning. According to the cooperative principle (Grice, 1975), interlocutors behave cooperatively and divergence from maxims of communication allows comprehenders to draw inferences. However, the assumptions of cooperativity and rationality may often not hold in practice, for instance, if the speaker is distracted or fatigued, or if they are not very pragmatically competent. Thus, if an ambiguous utterance *can* be interpreted nonliterally – for instance, as an implicature or an ironic statement – that doesn't necessarily mean that it *should be*, as the speaker may have meant the literal meaning (Elder & Haugh, 2023). Thus, comprehenders need to calibrate their interpretations to the particular speaker and situational context.

Previous work, reviewed in Section 2.4, suggests that comprehenders use speaker characteristics – such as persona, language fluency, and domain knowledge – to guide their interpretation. In many real-world situations, however, characteristics that influence a speaker's communicative behavior are not directly observable. In such cases, comprehenders must adapt to the speaker over the course of the interaction. As discussed in the previous chapter, previous work shows that listeners are sensitive to speaker-specific language use, including idiosyncratic word meanings (Schuster & Degen, 2020; Grodner & Sedivy, 2011) and intonation patterns (Roettger & Rimland, 2020).

In the previous chapter, we contributed to this literature by demonstrating that comprehenders dynamically adapt their interpretations to individual speakers' pragmatic competence in a reference game. When participants interacted with a literal

speaker, who was equally likely to produce any true utterance, their confidence in the pragmatic interpretation decreased over time. By contrast, with a pragmatic speaker, confidence in the pragmatic interpretations increased. The study presented in the previous chapter employed the reference game paradigm, which uses a highly constrained setup with a small number of possible references and non-linguistic alternative utterances that are known to both interlocutors. In this chapter, we investigate whether the same adaptation effect will emerge in a more naturalistic paradigm with linguistic stimuli.

The experimental paradigm used in this chapter is inspired by collaborative building games, such as the Minecraft Collaborative Building Task (Narayan-Chen et al., 2019), which have been used for creating datasets for grounded language understanding for artificial agents (Mohanty et al., 2022; Kiseleva et al., 2022, 2023). There are two players, the instructor and the builder. The instructor sees the target structure and provides written instructions to the builder, whose task is to reconstruct the target structure based on the provided instructions. The game is successful if the built structure matches the intended one. In our experiment, the speaker's utterances sometimes omit information and require the comprehender to make an inference about the color or height of a part of the structure.

As an example, consider the instructions "Place a green block on the existing green block. Then stack 3 blocks immediately to the left of these". The last sentence contains underspecification: the color of the blocks is not explicitly mentioned. However, the listener may infer that the color of the second set of blocks is the same as the one mentioned in the first part of the utterance, green, since that is the only color that has been mentioned. This follows from Grice's maxim of informativity (Grice, 1975): the comprehender should be able to infer the relevant property of the structure – color or height – from the context, otherwise the utterance is underinformative, and the relevant property is only in the common ground if it is the same as the one that has been previously mentioned.

Like in the experiment presented in the previous chapter, participants interact with two simulated speakers. One of the speakers behaves pragmatically, i.e., they only leave out properties when they can be inferred from context – we'll call this speaker the *pragmatic speaker*. The other speaker leaves out relevant properties even when they cannot be reliably inferred: in the example above, the unmentioned color of the stack may be something other than green. This speaker we'll call *literal*, for consistency with the terminology used in the previous chapter, but they could also be thought of as a less pragmatically competent speaker or an inattentive speaker. Like in Chapter 6, we investigate whether participants' confidence in the pragmatic interpretation will decrease over time with the literal speaker and increase with the pragmatic speaker. In addition to reporting confidence in their interpretation, in this experiment participants, could optionally ask a clarification question.

Consistent with the findings of the previous chapter, the results of the current experiment indicate an adaptation effect, as well as some generalization from the

first speaker to the next, providing further evidence for the generalizability of these findings. Additionally, we find that people are more likely to ask for clarification with a less pragmatic speaker. This suggests that, in interactive settings, if comprehenders have strong doubts about the pragmatic competence of the speaker, they may respond not only by adjusting their interpretation, but also by requesting clarification. These findings are consistent with previous work on clarification requests, which has argued that clarifications are made when the instruction follower lacks sufficient certainty to carry out the action (Schlangen, 2004; Madureira & Schlangen, 2023).

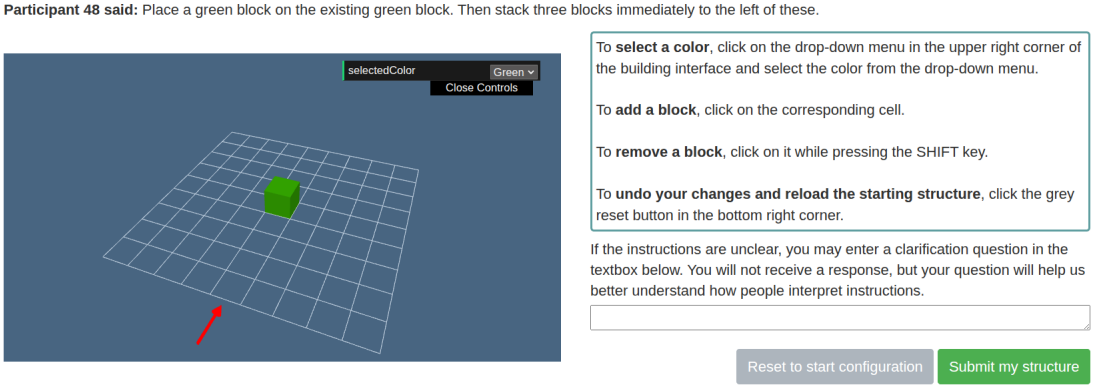
7.2 Experimental paradigm

The paradigm used for the current experiment is a block building game, where the builder follows the instructions of the instruction giver to construct a structure. Since we are interested in pragmatic adaptation of comprehenders, participants of our experiment played the role of builders. They were told that they would be following instructions written by two participants of an earlier experiment. Participants were told that previous participants had been shown series of images corresponding to building steps and had had to write building instructions based on the images. To make this cover story more convincing and to familiarize participants with the experimental setup, participants were shown one example of previous participants' supposed task and then had to take the perspective of the instruction giver and write instructions for an additional example.

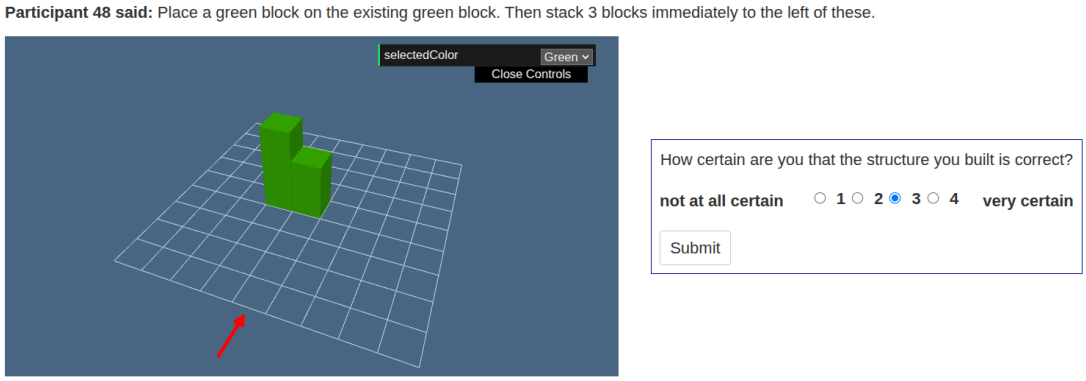
On each trial, participants built a structure on a 9x9 grid. On critical trials, important information regarding either the color of a part of the structure ("color-underspecified") or the height of a stack ("number-underspecified") was omitted from the instruction. The instructions on all trials consisted of either 2 or 3 building steps.

Like in the adaptation experiment in the previous chapter, participants interacted with two simulated speakers who differed in their pragmatic competence. In this experiment, pragmatic competence was operationalized as underspecification: one of the speakers, the *pragmatic speaker*, only omitted relevant information – the height or the color of a stack – when it could be reliably inferred from the context: when the color or the height of a stack was not mentioned, it was always the same as the one mentioned in the previous building step. The other speaker – we call them the *literal speaker* for consistency with previous chapter, but they can also be thought of as a less pragmatic speaker or an inattentive speaker – sometimes omitted information which could not be inferred from context.

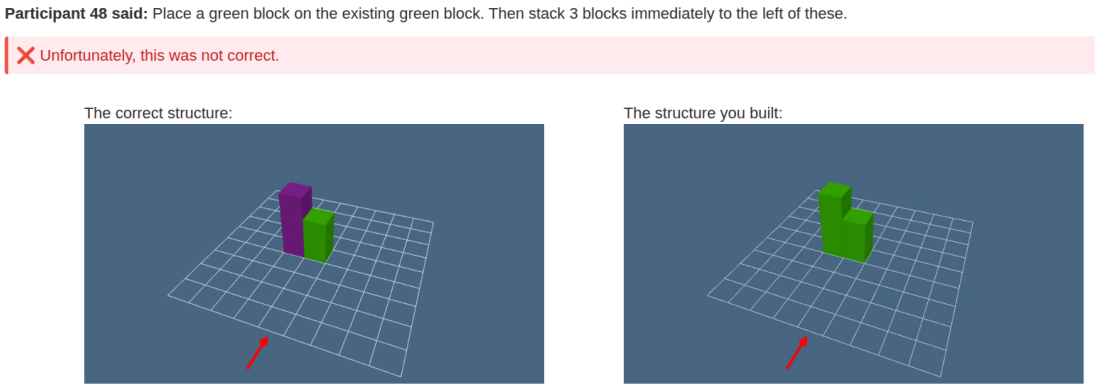
Each speaker block consisted of a total of 20 trials, 12 of which were critical (6 number-underspecified and 6 color-underspecified) and 8 were unambiguous control trials. In the literal speaker condition, on 8 out of 12 critical trials (4 number-underspecified and 4 color-underspecified), the pragmatic solution was incorrect. On non-pragmatic trials in the number-underspecified condition, the stack of underspec-



(a) Step 1: Building a structure based on instructions.



(b) Step 2: Reporting confidence.



(c) Step 3: Feedback.

Figure 7.1: Example critical trial in the literal speaker condition. If the built structure reflected the pragmatic interpretation, participants sometimes (in two thirds of cases) received feedback that it was incorrect.

ified height was always either one block taller or one block shorter than the stack described in the previous building step. In the color underspecified condition, the part of the structure whose color was not mentioned was one of the other four colors.

We adopted a 2:1 ratio of non-pragmatic to pragmatic trials in the literal speaker condition based on pilot testing. In an earlier version of the experiment, the literal

speaker behaved non-pragmatically on all critical trials, and participants' post-test questionnaire responses indicated that some perceived the speaker as not merely inattentive but uncooperative. In contrast, such interpretations were rarely observed in a pilot version using a 2:1 split, as in the present experiment. While the adaptation experiment reported in Chapter 6 employed a 1:2 ratio of non-pragmatic to pragmatic trials for the literal speaker, this was not feasible here because each trial involved building a structure and was therefore much more time-consuming.

We created two stimuli lists and counterbalanced whether the critical trials and fillers from list 1 or list 2 were used to ensure that there were no systematic differences in stimuli difficulty for the two speakers. Additionally, in order to make it more believable that participants were following instructions of two different speakers, we manipulated the style of the stimuli slightly: one of the speakers always used digits to denote stack height ("tower of 2 red blocks") while the other wrote the numbers out ("tower of *two* red blocks"), also counterbalanced. This difference in style was minimal since more major linguistic differences could have had the unintended effect of inviting inferences about the speaker's persona, which, in turn, might have affected the expectations of the speaker's pragmatic competence (Beltrama, 2020). For instance, if one of the speakers had not used any capitalization, that could have resulted in them being perceived as putting in less effort. Responses to the post-test questionnaire suggested that none of the participants suspected that the speakers were simulated, indicating that the cover story was believable.

On every trial, in addition to building a structure, participants had the option of entering a question in a textbox if they felt that the instructions were unclear. Participants were told that they were not required to do so but that it would help us learn more about how people understand written instructions. After submitting the structure, participants reported their confidence on a 4-point scale that the structure they built was correct.

On every trial, after submitting the structure and reporting their confidence, participants received feedback on whether the structure they built was correct and what the intended structure was. On critical trials in the literal speaker condition, the pragmatic inference was incorrect 8 out 12 times. An example of a nonpragmatic trial, including building, confidence reporting and feedback, is shown in Figure 7.1.

Thus, the structure of the trials, which consisted of (1) interpreting the instructions and building the structure, (2) reporting confidence on a 4-point scale, and (3) receiving feedback, matches that of the reference experiment in the previous chapter, making the results comparable.

7.3 Experiment

7.3.1 Participants

80 native English speakers based in the UK with an approval rate of at least 95% were recruited via the crowdsourcing platform Prolific. Participants who experienced technical issues (12 participants) and those whose accuracy on the unambiguous control trials was below 75% (12 participants) were excluded from analysis and new participants were recruited in their place. The exclusion criterion for this task was slightly less strict compared to the reference game adaptation experiment in the previous chapter (75% vs. 80% accuracy). We set this threshold based on a pilot study and based on the fact that this is a more challenging task with more potential for accidentally making an error compared to the reference game.

7.3.2 Materials and procedure

Participants completed the block building task described in the previous section, where they interacted with two speakers, literal and pragmatic.

Trial order was pseudo-randomized: the order of *trial types* – i.e., of control, critical-pragmatic and critical-nonpragmatic trials – was fixed across all participants in order to keep the learning trajectory and the course of negative and positive feedback similar across participants. The order of individual items, i.e., which particular control item was followed by which critical-nonpragmatic item, was randomized across participants.

After completing the two blocks, participants filled out a short post-test questionnaire where they were asked what they thought about the way the two speakers provided instructions, as well as whether they had experienced any technical difficulties, and had the opportunity to leave any comments.

7.3.3 Results

Defining nonpragmatic interpretation: Analysis of mistakes and guesses

Unlike in the reference game used in the previous chapters, where there are only two types of non-pragmatic behavior – selecting the competitor or the distractor (the latter was extremely rarely done in practice) – in this paradigm, the space of non-pragmatic interpretations is very broad. Thus, in order to analyze the change in participants' interpretations, we first need to define what constitutes pragmatic vs. non-pragmatic behavior for our purposes. We therefore first examined all critical trials on which participants built a structure other than the pragmatic structure and annotated them. Such trials constituted only a small proportion of critical trials (264 out of 1920; 13.75%), indicating that, overall, participants understood the instructions correctly

and built structures reflecting the pragmatic interpretation (i.e., inferring the underspecified color or height to be the one previously mentioned in the instruction). We distinguish between *guesses* - where participants attempted to guess a height or a color other than the one that was pragmatically inferrable - and *mistakes*, where participants misinterpreted the instructions or placed blocks in the wrong position by accident. The guess trials (N=113; 5.89%) are retained in the analysis, whereas the mistake trials (N=151; 7.86%) are excluded. The annotation scheme, which includes the different mistake types attested in the data, is presented in Table 7.1. Interestingly, while guesses constitute a low proportion of trials, the majority of participants (57 out of 80; 71.25%) made at least one guess.

Tag	**Subtag**	**Description**
guess	diff_color	Participant guessed the omitted color and used a different color than the pragmatically inferrable one.
guess	diff_number	Participant guessed the omitted number of blocks and used a different color than the pragmatically inferrable one.
mistake	blocks_missing	The built structure is missing some blocks.
mistake	gap	Participant left a gap between rows or stacks instead of building them next to each other.
mistake	row_stack_confused	Participant built a row instead of a stack or the other way around.
mistake	too_many_blocks	The built structure includes additional blocks.
mistake	wrong_position	The structure is in the wrong position on the grid.
mistake	wrong_side	Part of the structure was built on the wrong side (e.g., on the left instead of on the right).
mistake	wrong_structure	Wrong structure not covered by the above.

Table 7.1: Annotation scheme for non-pragmatic structures.

The distribution of annotation tags in the two speaker conditions is shown in Figure 7.2. More than half of nonpragmatic structures include mistakes, whereas the rest corresponds to guesses.

To statistically test whether participants' likelihood of making a nonpragmatic guess was affected by the speaker type, trial type (color- or number-underspecified) or block number, we fit a Bayesian logistic regression model. Only critical trials were used for this analysis, and mistake trials were excluded. A binary variable denoting whether a guess was made was regressed onto speaker type (sum-coded: -1 = literal, 1 = pragmatic), trial type (sum-coded: -1 = diff_color, 1 = diff_number), block number (sum-coded: -1 = 1, 1 = 2), and all their interactions.

The random effect structure included random intercepts for participants and items and per-participant random slopes for speaker type, trial type and block number.

	Estimate	95% CrI
Intercept	-3.98	[-4.77, -3.34]
Speaker type (pragmatic vs. literal)	-0.73	[-1.19, -0.31]
Trial type (number vs. color)	0.62	[0.06, 1.25]
Block number (2 vs. 1)	-0.07	[-0.42, 0.28]
Speaker type : Trial type	0.05	[-0.37, 0.47]
Speaker type : Block number	0.30	[-0.04, 0.66]
Trial type: Block number	-0.06	[-0.38, 0.27]
Speaker type : Trial Type : Block Number	-0.11	[-0.47, 0.23]

Table 7.2: Bayesian logistic regression model of number of (non-pragmatic) inferences: guessing different color or number.

For all effects, wide weakly informative priors were set to avoid biasing the model. For the intercepts and the main effects, we set the prior to be a normal distribution with a mean of 0 and a standard deviation of 3.

We ran 6 chains of 9000 interactions each, with the first 2000 interactions of each chain discarded as warmup.

Results are reported in Table 7.2. Effects are taken as noteworthy if the 95% credible interval excludes 0. The analysis revealed that fewer attempts to guess the underspecified color or number of blocks are made in the pragmatic speaker condition than in the literal speaker condition ($\hat{\beta}$ = -0.73, 95% CrI=[-1.19, 0.31]). Additionally, the number of blocks is guessed more often than color ($\hat{\beta}$ = 0.62, [0.06, 1.25]). This is likely due to the fact that there were fewer height options than color options: on nonpragmatic trials, in the literal speaker block the underspecified height was always one above or one below the pragmatically inferrable height.

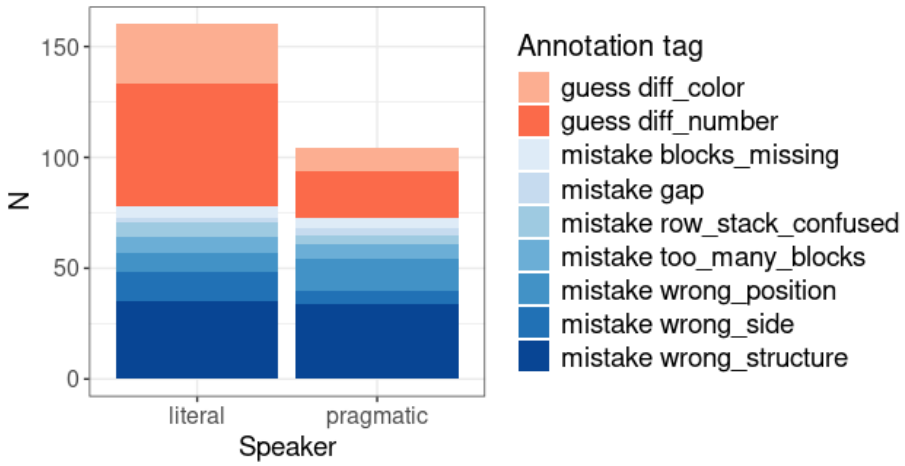


Figure 7.2: Distribution of non-pragmatic responses (mistakes and guesses) for the two speakers.

Analysis of confidence change

Having defined what counts as a pragmatic interpretation, we proceed to the main analysis of change in participants' confidence over time. Figure 7.3 shows the change of participants' confidence over the course of the two blocks for the two speaker orders. "other", i.e., non-pragmatic, interpretations include guesses but exclude mistakes. The first observation that can be made is that, with both speakers, participants overwhelmingly chose the pragmatic interpretation and only very occasionally resorted to guesses. Second, confidence appears to evolve differently depending on speaker type: it increases over time in the pragmatic speaker block and decreases in the literal speaker block, indicating adaptation. Finally, there appears to be the effect of speaker order, where participants' confidence is higher on the literal speaker block when preceded by the pragmatic speaker block, and confidence in the pragmatic speaker block appears to be lower when preceded by the literal speaker block, indicating generalization of expectations from the first speaker to the second one.

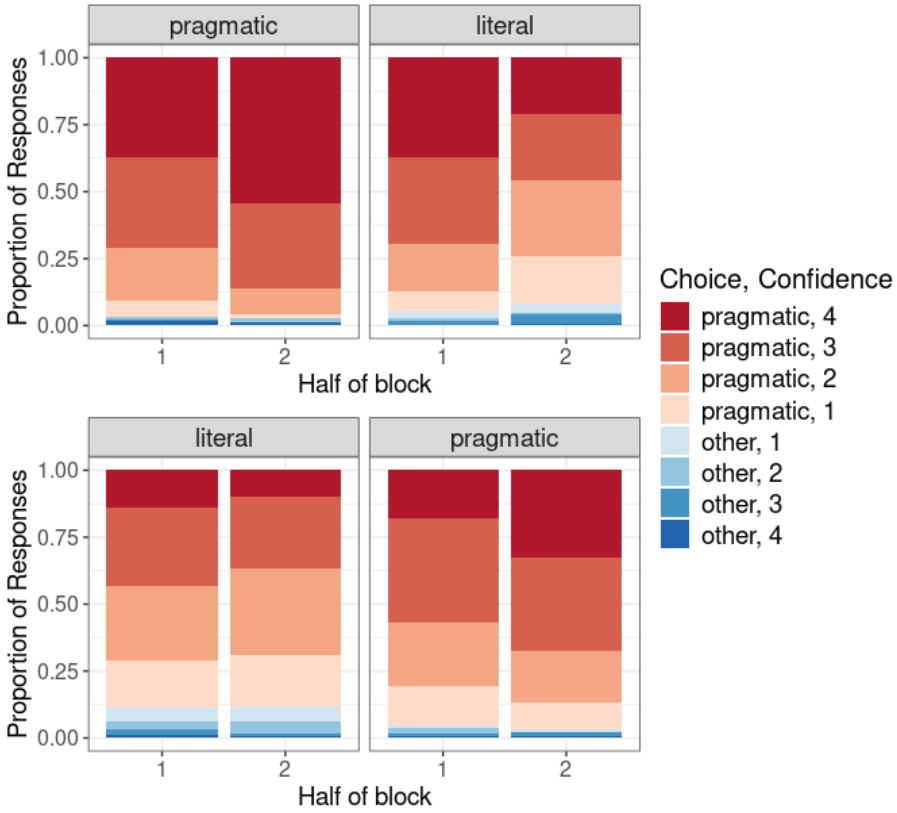


Figure 7.3: Confidence change by block for the two speaker orders.

To verify these apparent effects, we fit a Bayesian ordinal regression model using the *brms* package in R (Bürkner, 2017). Only critical trials were used for this analysis; the trials which were annotated as participants having made a mistake in building were excluded. As in Chapter 6, the dependent variable was an 8-level ordinal variable (whether the built structure was the structure corresponding to the pragmatic

interpretation or not × 4-point confidence scale). Thus, 1 corresponds to building a non-pragmatic structure and providing a confidence rating of 4, and 8 corresponds to building a pragmatic structure and providing a confidence rating of 4. This variable was regressed onto speaker type (sum-coded: -1 = literal, 1 = pragmatic), block order (sum-coded: -1 = 1, 1 = 2) and trial number within block (mean-centered), all interactions of those three variables, and the writing style of the speaker (whether the speaker the numbers out vs. used digits in their instructions; sum-coded: -1 = words, 1 = digits).

We included the maximal effect structure that allowed the model to converge without divergent transitions. It included per-participant and per-item random intercepts, as well as per-participant slopes for speaker type, block order and trial number within block.

For all effects, wide weakly informative priors were set to avoid biasing the model. For the intercepts, we set the prior to be a normal distribution with a mean of 0 and a standard deviation of 5. For all main effects, we set the prior to be a normal distribution with a mean of 0 and a standard deviation of 1.

We ran 6 chains of 12000 iterations each, with the first 2000 iterations of each chain discarded as warm-up.

Effect	Estimate	95% CrI
Speaker type (pragmatic vs. literal)	0.39	[0.27, 0.50]
Block number (2 vs. 1)	0.01	[-0.09, 0.11]
Trial number within block	0.01	[-0.01, 0.02]
Speaker style (digits vs. words)	0.05	[-0.03, 0.15]
Speaker type : Block number	-0.34	[-0.54 -0.14]
Speaker type : Trial number	0.06	[0.05, 0.07]
Block number : Trial number	-0.02	[-0.03, -0.01]
Speaker type : Block number : Trial number	0.01	[-0.01, 0.02]

Table 7.3: Bayesian ordinal regression model of confidence change.

The results are presented in Table 7.3. Effects are taken to be noteworthy if the 95% credible interval excludes 0. There is a main effect of speaker type ($\hat{\beta}$=0.39, 95% CrI = [0.27, 0.50]), indicating that confidence in the pragmatic interpretation was generally higher with the pragmatic speaker. There was also an interaction of speaker type with block number ($\hat{\beta}$=-0.34, 95% CrI = [-0.54, 0.14]), which suggests that confidence was not the same for the two speaker blocks depending on whether that speaker appeared first or second, indicating an expectation transfer effect. There was also an interaction of speaker type with trial number ($\hat{\beta}$=0.06, 95% CrI = [0.05, 0.07]), indicating that confidence developed differently over time for the two speaker types: it tended to decrease for the literal speaker and to increase for the pragmatic speaker. Finally, there was an interaction of block number and trial number ($\hat{\beta}$=-0.02, 95% CrI = [-0.03, -0.01]), indicating that participants' confidence developed

differently in the first and the second block. Taken together, these results provide evidence for adaptation to the two speakers and for a certain degree of generalization from the first speaker to the second one.

To confirm that adaptation reflected participants' beliefs about the speakers' pragmatic competence and their tendency to omit information, rather than doubts about the literal speaker's overall trustworthiness, we conducted the same analysis on the unambiguous control trials. Confidence on the control trials was unaffected by speaker type or any other predictor. This indicates that participants understood that the speakers differed specifically in their tendency to omit relevant information, and that the observed effect reflects beliefs about pragmatic competence rather than about general trustworthiness or cooperativity.

Analysis of clarification questions

Recall that on each trial participants had the option to ask a question in addition to building the structure if they felt that the instructions were unclear (Figure 7.1a). We now conduct an analysis of participants' questions. Questions were asked on 26.9% of critical trials (517 out of 1920), and exactly 50% of participants (40 out of 80) asked at least one question. Unsurprisingly, the questions were about the omitted height on number-underspecified trials and the omitted color on color-underspecified trials.

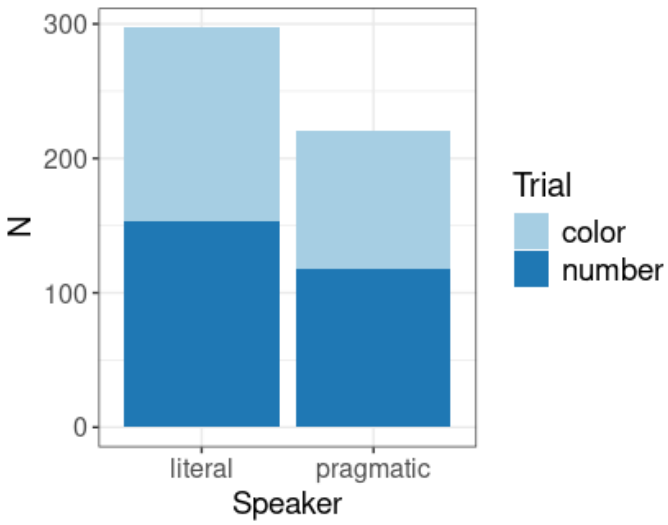


Figure 7.4: Counts of questions asked on critical trials for the two speakers.

Figure 7.4 shows counts of questions asked by speaker type, block and type of trial (color vs. number).

To statistically test which factors affected participants' likelihood of asking a question, we conducted an analysis which mirrors the analysis of guesses in Section 7.3.3, where fit a Bayesian logistic regression model to critical trials, excluding mistake trials with a binary dependent variable denoting whether a question was asked. The

	Estimate	95% CrI
Intercept	-4.07	[-5.67, -2.70]
Speaker type (pragmatic vs. literal)	-0.83	[-1.35, -0.42]
Trial type (number vs. color)	0.24	[-0.21, 0.64]
Block number (2 vs. 1)	0.26	[-0.16, 0.66]
Speaker type : Trial type	0.03	[-0.22, 0.27]
Speaker type : Block number	0.88	[-0.44, 2.27]
Trial type: Block number	-0.04	[-0.25, 0.16]
Speaker type : Trial Type : Block Number	-0.05	[-0.30, 0.20]

Table 7.4: Bayesian logistic regression model of the number of questions asked by participants.

fixed and random effects, priors and the number of iterations were the same as for the regression model for guesses described in Section 7.3.3.

Results are reported in Table 7.4. The analysis revealed an effect of speaker type, where participants asked fewer questions with the pragmatic speaker than with the literal one ($\hat{\beta}$ = -0.83, 95% CrI = [-1.35, -0.42]).

7.3.4 Discussion

In this chapter, we presented a study in the block building paradigm, where participants played the role of builders and followed sometimes ambiguous instructions by two simulated speakers. One of the speakers, the *pragmatic speaker*, only omitted information about the color or height of a part of a structure which could be reliably inferred from the context, whereas the other, the *literal speaker*, sometimes omitted information in a non-predictable way. Participants reported their confidence in the correctness of their built structures and had the option of asking a clarification question, which they knew would not be answered.

We observed evidence of adaptation to the two speakers: participants' confidence in the pragmatic interpretation increased over time with the pragmatic speaker and decreased with the literal speaker. There was also evidence of generalization across speakers, as evidenced by the fact that participants' confidence in their interpretations was affected by the order in which they saw the speakers. Importantly, confidence was only affected on the underspecified trials but not the unambiguous control trials. This indicates that participants recognized that the difference between the speakers was related specifically to the way they omitted information and did not doubt the speakers' overall trustworthiness.

Overall, the results obtained in this study closely match those of the reference game experiment from the previous chapter. The fact that we demonstrated an adaptation effect across two quite different paradigms provides evidence that the results are

likely to generalize to other paradigms and more naturalistic interactions. Below, we compare findings of the current chapter with those of the previous chapter more closely and discuss the differences.

In both experiments, we observed an effect of speaker type, where overall participants had lower confidence in their interpretations of utterances by the literal speaker than by the pragmatic speaker. Likewise, in both experiments, we observed a difference in the learning dynamics with the two speakers, where confidence decreased with the literal speaker and increased with the pragmatic speaker. Furthermore, in both experiments, the change in participants' confidence was different for the two halves of the experiment, independent of speaker type.

There are also several notable differences in the results of the two experiments. First, in the previous experiment, there was a positive effect of trial number, suggesting that people's confidence increased over time independent of speaker type. By contrast, in the current chapter, confidence only increased in the pragmatic speaker condition. We interpret this difference as being related to the specific paradigms we used. The reference game is presumably an unfamiliar setup for most participants, as a result of which it was likely not obvious to most participants from the start what the pragmatic interpretation was, and they needed time to adjust to the task. In addition, given the unfamiliar setting, the implicatures derived in the reference game are ad-hoc implicatures, which are not derived by default, as evidenced by substantial variability of people's success on these implicatures, shown by Franke & Degen (2016) and in Chapter 4 of this dissertation, as well as by the fact that people's success on these implicatures is predicted by their logical reasoning and perspective-taking abilities, as also shown in Chapter 4. Therefore, some participants applied the guessing strategy in the beginning and reported low confidence in their selections until they figured out what the pragmatic interpretation should be. In contrast, in the current experiment, the pragmatic interpretation is likely the relatively effortless default one, and no such adaptation to the task is required. Taken together, these results suggest that the accessibility of the pragmatic interpretation affects the dynamics of comprehenders' adaptation.

Another difference in the results of the two experiments is that, in the current experiment, we observed an interaction of speaker type with speaker order, indicating that people's confidence in the interpretation of utterances of a certain speaker type was different depending on whether this speaker appeared first or second. This indicates generalization from the first speaker to the second one. In contrast, in the reference game experiment, we did not observe evidence for such a generalization tendency at the population level, but a subset of the participants showed such a tendency to various degrees. This could be explained by the difference in accessibility of the pragmatic interpretation which we mentioned above. In the current experiment, if the pragmatic interpretation is the default one, people may not even be considering the nonpragmatic interpretation until they encounter it. Once they recognize the possibility of the nonpragmatic interpretation, which can be explained by inattention,

they may be more likely to question the second speaker's instructions in the way they wouldn't have done otherwise.

Additionally, we conducted an analysis of non-pragmatic interpretations, i.e., when participants guessed a different color or height than the one that would have been pragmatically inferrable. Unsurprisingly, people were more likely to resort to guessing the underspecified color or height with a literal speaker. There was only a small proportion of guesses, which can be explained by the fact that the probability of guessing correctly is low given that there are multiple possible heights and colors. In contrast, the guessing strategy was applied much more often in the reference game experiment in the previous chapter. This can be explained by the fact that the reference game is an unfamiliar setting where the pragmatic interpretation is not as much of a default, as discussed above. Additionally, the space of possible alternatives is lower in the reference game, so that a guess is more likely to be correct. Taken together, these findings suggest that the likelihood of adopting a nonpragmatic interpretation depends on both the degree to which the pragmatic interpretation is the default one and on the number of alternative interpretations.

We analyzed participants' tendency to ask clarification questions and found that they were more likely to ask for clarification when interacting with a literal speaker. This effect emerged despite the fact that the task was not interactive and there was no obvious benefit to asking a question: participants were told that they were following instructions provided by participants of a previous study, so it was clear to them that they would not receive a response. Still, low confidence appeared to justify the effort of asking a clarification question. These findings are in line with existing work on clarification requests in dialogue and instruction following settings. It has been proposed that requests for clarification are motivated by low confidence in the interpretation (Schlangen, 2004) and initiated when the lack of clarity of the instructions prevents someone from performing an action with sufficiently high certainty (Madureira & Schlangen, 2023). Our findings suggest that when the instruction is underspecified, this certainty depends on the comprehender's beliefs about the speaker's pragmatic competence.

The observed adaptation effect was very similar across the two paradigms – the reference game in the previous chapter and the block building game in the current chapter – indicating similar underlying processes. Therefore, the Bayesian belief updating model proposed in the previous chapter should be able to capture the patterns observed in this experiment as well. However, it remains an open question whether similar individual differences in adaptation patterns are present in the block building paradigm as in the reference game, where the modeling results suggested that people differ in their prior beliefs and the degree of generalization between the two speakers. As discussed above, in the current study, there appears to be a stronger population-wide tendency to generalize across speakers compared to the reference game experiment, so individuals may not differ as strongly on that parameter. Future work could extend the adaptation model proposed in the previous chapter to the

block building paradigm and investigate whether the result from the previous chapter concerning individual variability – that assuming different prior beliefs and degrees of generalization between speakers results in a better fit than assuming that every participant has the same parameters – holds for this paradigm as well.

7.4 Conclusion

In this chapter, we presented an experiment in the block building paradigm, where participants interpreted underspecified instructions by two speakers, one of whom only omitted information which could be reliably inferred from context, while the other one sometimes omitted information which could not be inferred. We found that people adapted to the speakers' pragmatic competence by adjusting their confidence in the pragmatic interpretation. The results closely match those of the reference game adaptation experiment in the previous chapter. The fact that results were very similar across two quite different paradigms provides evidence for the generalizability of the findings.

The differences in the findings of the experiments presented in this chapter and in the previous chapter are informative as well. In the current experiment, participants were much less likely to resort to guessing than in the reference game. This can be explained by the fact that in the current experiment, the pragmatic interpretation was much more accessible and was presumably the default, in contrast to the reference game, where implicature derivation is presumably much more effortful, as evidenced by vast differences in people's performance on the reference game and the positive relationship between performance on the reference game and logical reasoning skills shown in Chapter 4. Additionally, in the current experiment, there was a larger space of alternative meanings than in the reference game, reducing the chance of guessing correctly. This difference suggests that the tendency to adopt a non-pragmatic interpretation depends on the accessibility of the pragmatic interpretation and the space of alternatives.

Another difference between the findings of the current chapter and those of the previous chapter is related to generalization across speakers. In the current chapter, we observed a population-wide tendency to generalize beliefs about the first speaker's pragmatic competence to the next speaker, whereas in the reference game this tendency was evident only in a subset of participants. Future work could examine whether individual differences in adaptation patterns which we observed in the reference game also emerge in the block-building paradigm, as well as in other more naturalistic communicative settings.

Taken together, the findings of this chapter and the previous chapter suggest that adaptation to a speaker's pragmatic competence follows general principles that hold across a range of phenomena, but that this adaptation process is shaped by factors specific to different phenomena and contexts, such as accessibility of the pragmatic

interpretation and the space of alternatives.

The task participants had to complete in this chapter, as well as in all other experiments presented in this dissertation, was a non-interactive comprehension task. Therefore, the ways participants could react and adapt were limited to changing their interpretation. However, in a dialogue setting, where one is actively interacting with an interlocutor, comprehenders have the option of asking for clarification. In this chapter, we found that even in this non-interactive setting more questions were asked with a less pragmatic speaker. This finding suggests that, when hearing an ambiguous utterance by a speaker whose pragmatic competence the listener is uncertain about, they are more likely to ask a clarification question. Future work could extend the investigation of sensitivity to the speaker's pragmatic competence to the interactive setting to confirm that this question-asking tendency is preserved in dialogue.

Chapter 8

Discussion and conclusion

This dissertation aims to contribute to theories of pragmatic processing by providing insights about the cognitive factors underlying variability in pragmatic reasoning and about ways in which comprehenders accommodate such differences in speakers by adjusting their interpretations based on their beliefs about the speaker's pragmatic competence. The current chapter discusses main findings of the thesis in the context of related work and outlines possible directions for future work.

8.1 Discussion of main findings

This thesis focused on four research goals, which were formulated in Introduction. Below, we review the progress made towards each of these goals and discuss the contributions of the thesis in the context of existing research.

8.1.1 Proposing a strategy elicitation method for pragmatic tasks

The first research goal of this dissertation was to develop a method for eliciting participants' strategies on the pragmatic reference game in order to ensure that participants' performance on the task reflects pragmatic reasoning as accurately as possible, minimizing biases related to experimental design.

The proposed strategy elicitation method consisted of asking participants to describe their strategy after they had completed the last experimental item of the task, annotating those strategies according to the annotation scheme we proposed, and comparing performance against annotated strategies. This method is quite simple, but it proved effective in several important ways throughout the dissertation.

First, we showed in Chapter 3 that the application of the proposed strategy elicitation method allowed us to identify biases in the stimuli: in the original reference

game stimuli from Franke & Degen (2016), certain visual characteristics of the stimuli resulted in some participants giving the correct answer "for the wrong reason" – not because of successful pragmatic reasoning, but due to accidental visual similarity of some messages to the pragmatic target. We then applied the same method to show that abstract stimuli, shapes and colors, substantially reduce such biases, resulting in significantly better alignment between performance and strategy. For this reason, we used abstract stimuli for subsequent experiments.

Second, we used the proposed strategy elicitation method in the individual differences study in Chapter 4 to filter out participants whose explanations revealed that they had misunderstood the task or applied an entirely different strategy that is not predicted by the formal probabilistic models of different depths, which serve as the theoretical motivation for this variability. This increases confidence that the identified relationship between the individual difference measures – logical reasoning and Theory of Mind – and pragmatic reasoning is not spurious, and that its estimate more accurately reflects the true effect size.

Third, the proposed method proved useful in the experiments on speaker effects described in Chapter 5 for gaining deeper insight into participants' reasoning about the speakers, revealing application of strategies which would not have been discernible from examining performance. In the experiments where participants interpreted messages allegedly sent by children, adults and artificial agents, their explanations sometimes revealed uncertainty about the speaker's pragmatic competence, e.g., whether a 4-year-old child's reasoning abilities are sophisticated enough for selecting the optimal message. This provided additional evidence for the conclusion that participants' lower confidence in the pragmatic interpretation is related to their uncertainty about the speaker's pragmatic competence. Similarly, in Experiment 4, where participants interacted with an explicitly literal speaker, a simple computer program which selected any true message with equal probability, participants' explanations revealed that they sometimes ascribed rationality to the program, expecting it to behave pragmatically despite having been told that it is fully literal.

Together, these applications provide evidence for the usefulness of the proposed method for obtaining more accurate results and gaining more insight into the applied reasoning strategies. This method can be applied to other experimental paradigms. In a paper co-authored during the dissertation period which is not included in this dissertation (Ryzhova et al., 2023), we applied a similar strategy elicitation method to a different pragmatic phenomenon – atypicality inferences following a redundant mention of a highly predictable event (Kravtchenko & Demberg, 2022) – and found that it helped identify individuals who acknowledged the possibility of making the inference but rejected it due to implausibility.

The proposed method has limitations. First, it will only be useful if it is expected that the reasoning performed in the task is relatively conscious. Previous work has shown that certain biases never appear in people' explanations because people not being aware of them (Evans, 2019), and more generally, post-hoc explanations have

been criticized for not always accurately reflecting reasoning in the moment and for involving post-hoc rationalization (Cushman, 2020; Nisbett & Wilson, 1977). Despite these limitations, the applications summarized above show that this method can be useful for the study of pragmatic reasoning. Future work could consider the use of online measures which do not rely on post-hoc explanations, such as eye-tracking, as a way of probing an experimental design for possible biases and gaining additional insight into strategies which people use in the task.

8.1.2 Cognitive factors underlying individual differences in pragmatic reasoning

The second research goal of this thesis was to investigate the cognitive factors underlying variability in ad-hoc pragmatic inferencing. There is a growing body of work exploring cognitive individual differences in pragmatic processing, as well as a tradition of research modeling pragmatic effects using formal probabilistic approaches. However, these two literatures have remained largely separate, with a few recent exceptions (Bohn et al., 2022; van Tiel et al., 2022). One of the contributions of this thesis is bridging these two strains of research by investigating the cognitive factors underlying variability in a pragmatic reference game. Franke & Degen (2016) provided evidence for such variability, demonstrating that people's performance aligns with predictions of three models of different reasoning depths. In Chapter 4, we investigated which cognitive differences predict this variability, laying the groundwork for the study of underlying mechanisms and cognitive modeling of this variability.

We found that people's performance on the pragmatic reference game was predicted by their logical reasoning abilities and Theory of Mind, whereas working memory did not have an effect. While the effects of Theory of Mind in pragmatics are fairly widely studied, few studies have explored the effects of logical reasoning in pragmatic tasks. We found that logical reasoning modulated performance in the pragmatic reference game (Chapter 4) and in atypicality inferences, as reported in a co-authored paper not included in this dissertation (Ryzhova et al., 2023). In contrast, Heyman & Schaeken (2015) did not find such an effect for scalar implicatures. We speculated that these findings may be explained by logical reasoning being required for reasoning in novel settings but not in conventionalized settings where pragmatic reasoning may be amortized. This could be further explored in future work. For instance, it could be that logical reasoning predicts the interpretation of novel but not conventionalized metaphors.

The fact that working memory did not show an effect is somewhat puzzling. Prior work has yielded mixed findings about the role of memory in implicature derivation. While we found no evidence for an effect of working memory, we do not conclude ad-hoc pragmatic reasoning is not effortful. On the contrary, experiments on speaker effects and pragmatic adaptation discussed in Chapters 5 and 6 of this thesis provide evidence that it is likely to be effortful. Findings of Chapter 5 indicate that com-

prehenders may not engage in deeper pragmatic reasoning unless they have sufficient motivation to do so, which suggests that such reasoning is effortful, and the results of the adaptation study in Chapter 6 suggest that people are not easily able to take the speaker's perspective in this task, again pointing to its effortfulness. Why, then, did we not observe an effect of working memory? One relevant observation is that in some related studies, there was no direct effect of working memory (Heyman & Schaeken, 2015; Taguchi, 2008; but see Antoniou et al., 2016; Yang et al., 2018), but people were found to be more literal under cognitive load (De Neys & Schaeken, 2007; van Tiel et al., 2019). Future work could investigate whether performance in this task is sensitive to cognitive load.

The variation that we studied in Chapter 4 is theoretically motivated by the predictions of three RSA models of different reasoning depths. However, we find that there is a lot of gradient variability in our data, as opposed to three clear clusters. Participants' reported strategies largely aligned with their performance, with the largest source of misalignment coming from participants who are classified into the intermediate L_1 class based on their performance but whose reported strategy corresponds to one of the other two classes. This misalignment may be the result of lucky guessing or to figuring out how to solve the implicature during the task.

Taken together, these findings suggest a picture where there are three broad classes of reasoners, but with a lot of variation within these classes and flexible boundaries between them. This view is further supported by our experiments on adaptation to specific speakers, which indicate that the depth of comprehenders' pragmatic reasoning is sensitive to the situational context. In particular, individuals appear to engage in deeper reasoning only when they have sufficient confidence in the speaker's pragmatic competence and cooperativity. We elaborate on this point in the next section.

8.1.3 Adaptation to specific speakers

The third goal of this thesis was to investigate how characteristics of the speaker influence comprehenders' inferences. Previous work has shown that speakers' characteristics, such as the speaker's persona (chill vs. nerdy; Beltrama, 2018, 2020; Beltrama & Schwarz, 2021), language proficiency and accent (Ip & Papafragou, 2021; Fairchild & Papafragou, 2018; Puhacheuskaya & Järvikivi, 2025), and speaker knowledge (Bergen & Grodner, 2012), influenced interpretation. In this thesis, we contributed to this line of work by focusing on the listener's accommodation of the speaker's pragmatic competence.

Influence of comprehenders' beliefs about different speaker types on pragmatic intepretation

In Chapter 4, we found evidence of substantial variability in comprehenders' pragmatic reasoning competence. There is evidence that such variability also exists in speakers. Franke & Degen (2016) showed in their production experiment that while the performance of the majority of the participants aligned with the predictions of the pragmatic Rational Speech Act speaker model S_1, there were some people who behaved more in line with the literal speaker S_0. There is also evidence from other paradigms that speakers often do not behave optimally from the viewpoint of theoretical pragmatics: for instance, they often overspecify attributes in referring expressions (Koolen et al., 2011; Degen et al., 2020) and may fail to take the interlocutor's visual perspective into account (Martin et al., 2019). Besides, situational factors may play a role, resulting in the production of less optimal utterances, such as fatigue or cognitive load.

There is some evidence from other domains that people accommodate limited rationality in the speaker. In the field of visual perspective-taking, Hawkins et al. (2021) showed that when the speaker is less informative and gives ambiguous descriptions, comprehenders' adapt their behavior over time, making fewer errors. In a visual world paradigm, Grodner & Sedivy (2011) showed that their participants interpreted adjectives contrastively be default ("the tall glass" means that there is a short glass in the common ground) but that this contrastive interpretation did not arise when participants were told that the speaker had an impairment that caused "language and social problems". The findings of this thesis contribute to this line of work, showing that people's interpretations of implicatures are sensitive to their beliefs about the speaker's pragmatic competence and by providing insight into the mechanisms underlying this effect.

In Chapter 5, we investigated how the beliefs about the pragmatic competence of specific speaker types affect the derived inferences. The investigated speaker types included adults, 4-year-old children, artificial agents, and explicitly literal speakers. Experiments 1 and 2 compared the interpretation of ambiguous messages which participants believed were sent by 4-year-old children as opposed to adults. We found that with a child speaker, people were more likely to interpret the message literally and not derive an implicature or derive a less strong one.

The observed effect had both a gradient and a categorical component: people derived less strong inferences in the child speaker condition, reflecting comprehenders' lower certainty in the pragmatic interpretation, and more people provided literal interpretations, not deriving an inference. The gradient component of the observed effect appears to quite clearly reflect participants' uncertainty about the pragmatic sophistication of children.

The categorical component warrants more explanation. The fact that there are more literal responders in the child speaker condition seems to suggest that with a

speaker who people expect to be less pragmatically competent, listeners may be less likely to engage in deep reasoning. Participants' reported strategies support that interpretation: explanations labeled *meta_reasoning*, which correspond to people explicitly expressing uncertainty about the speaker's pragmatic competence, were associated with relatively high ratings, indicating that when people consciously reflected on children's reasoning competence, they often concluded that children are most likely capable of behaving like pragmatic speakers; the literal interpretations were associated with reporting guessing. Further evidence for this interpretation comes from Experiment 3, where people interpreted ambiguous messages, which they believed to have been selected by the LLM-based chatbot ChatGPT: when asked explicitly, people rated ChatGPT's competence highly, but in the experiment itself, they often interpreted its messages literally, indicating less deep reasoning. Taken together, these findings suggest that the depth of comprehenders' reasoning is influenced by their beliefs about the speaker: if they do not believe the speaker to be sufficiently pragmatically competent or cooperative, they may not be motivated to expend energy on complex reasoning.

These findings aligns with the core claims of relevance theory (Sperber & Wilson, 1986), which posits that enrichment of an utterance's meaning from literal to pragmatic only happens given sufficient relevance and if the cognitive effort needed for this derivation is justified. They are also in line with accounts of bounded rationality (Simon, 1990; Gigerenzer & Selten, 2008) and good-enough processing (Ferreira et al., 2002), which hold that people make choices that are "good enough" given their current goals, economizing on cognitive resources, and that effortful reasoning required for the optimal solution is often avoided unless required by the situation. Our findings suggest that adequate motivation – such as being sufficiently certain of the speaker's pragmatic competence and cooperativity – may be needed to justify engaging in deeper reasoning.

The results of Experiment 3 contribute to our understanding of the perception of artificial agents as interlocutors. Previous work yielded conflicting findings, with earlier work suggesting that participants perceive artificial agents as less capable than humans and thus adapt more to them (Fischer, 2006; Branigan et al., 2011) and later work showing the opposite pattern (Loy & Demberg, 2023). This inconsistency may be connected to the shift in people's perception of artificial agents' communicative competence due to technological developments. In our study, we observed an interesting discrepancy: participants rated ChatGPT's pragmatic competence highly when asked explicitly, but in the pragmatic reference game itself, they interpreted messages supposedly sent by ChatGPT less pragmatically than those sent by another human adult. This apparent contradiction may be reconciled by assuming that not only competence but also cooperativity of an agent influences comprehenders' interpretations: even if an agent is perceived to be capable, it may not be expected to observe the cooperative principle.

As discussed above, sufficient motivation may be needed to engage in effortful

pragmatic reasoning about the partner, which may be more readily available with a human interlocutor than with an artificial agent. This interpretation is in line with the findings of Peña et al. (2023), who showed that participants initially behaved more egocentrically with a computer partner than with a human partner but this difference disappeared when the computer partner was framed as a collaborative agent with a shared goal. Participants in our experiment did not interact with ChatGPT directly; instead, they interpreted messages in the pragmatic reference game presented as having been sent by ChatGPT. While this approach allowed for direct comparison with the other experiments in the chapter, it may have affected the results It could be investigated in future work whether ChatGPT's utterances are interpreted more pragmatically in a direct interaction.

Also, our experiment was conducted in the fall of 2023. Since then, ChatGPT's capabilities have increased, as has people's familiarity with ChatGPT. Our participants reported having limited experience with ChatGPT (mean of 2.08 out of 5). Three years later, we would expect higher familiarity, which is likely to have an effect on people's beliefs about ChatGPT's communicative competence. We discuss possible extensions to our study in Section 8.2.

In sum, the findings Chapter 5 provided empirical evidence that comprehenders' beliefs about the speaker's pragmatic competence influence the derivation of inferences. Notably, this does not seem to occur by default but instead requires sufficient salience of the speaker's identity: in Experiments 1 and 2, the effect only emerged when there was a cover story with images emphasizing the fact that participants were dealing with a child speaker. When we conducted a version of the experiment where we simply told participants that they were interpreting messages sent by a child, the effect did not emerge as clearly. This suggests that for comprehenders' beliefs about the speaker to influence their interpretations, these beliefs need to be activated by sufficient salience of the speaker's relevant characteristics.

Dynamic adaptation to the speaker's pragmatic competence

The findings of Chapter 5 suggest that comprehenders' interpretation is guided by their beliefs about the speaker's pragmatic competence or cooperativity. In Chapters 6 and 7, we investigated whether such beliefs can be formed and changed through interaction, with no explicit mention of the speaker's characteristics, motivated by the fact that often relevant characteristics of the speaker which influence the optimality of their pragmatic reasoning are not directly observable.

Existing work on other pragmatic phenomena suggests that people may adapt their interpretations to specific speakers, including learning speaker-specific patterns of using modal adverbs (Schuster & Degen, 2020; Schuster et al., 2023), learning to associate humorous utterances to specific speakers (Loy & Demberg, 2023), and adapting interpretations of object descriptions (Hawkins et al., 2021) and gradable adjectives (Gardner et al., 2021; Ryskin et al., 2019). Notably, previous studies have

observed that this adaptation tendency exists at the population level but individuals vary in the extent to which they adapt (Schuster et al., 2023; Loy & Demberg, 2023).

Chapter 6 investigated adaptation to two speakers in the reference game, one of whom behaved consistently with the predictions of the pragmatic speaker RSA model S_1, and the other – with the predictions of the literal speaker model S_0. Chapter 7 extended the investigation of adaptation to speakers' pragmatic competence to a different domain, a collaborative block building game, where participants built 3D block structures following instructions by two speakers, which were sometimes underspecified and where ambiguity needed to be resolved by inferring the color or height of a part of the structure. In both paradigms, on each trial, participants reported their confidence in their answer and received feedback. We investigated whether participants' confidence would develop differently with the two speakers. Across the two paradigms, we observed rapid adaptation to the two speaker types: participants' confidence in the pragmatic interpretation increased when they interacted with the pragmatic speaker and decreased when they interacted with the literal speaker. These findings suggest that people calibrate their beliefs about speakers' pragmatic competence during interaction and adjust their interpretations accordingly.

Additionally, the findings of Chapter 6 revealed differences in adaptation behavior between participants. While at the population level, we observed strong evidence of adaptation to specific speakers, about a quarter of our sample did not adapt. Some of those participants behaved like literal comprehenders with both speakers; they presumably never figured out what the pragmatic interpretation should be. This is in line with the findings of Chapter 4 on individual differences in comprehenders' pragmatic reasoning: some people are less pragmatically competent and tend to derive literal interpretations. Others consistently provided pragmatic interpretations, refusing to lower their confidence even with a literal speaker despite repeated evidence that their interpretation was incorrect. In Section 8.2, we provide recommendations for future work investigating cognitive individual differences underlying this variability. We observe differences among adapters as well. The proposed computational cognitive model, which we will discuss in the next section, suggests that these differences can be explained by assuming that individuals differ in their prior beliefs about speakers' pragmatic competence and in the degree of generalization across speakers. We comment on this further in the next section, where we discuss the proposed computational cognitive models.

The experiments presented in this dissertation indicate that comprehenders accommodate differences in speakers' pragmatic competence by adjusting their interpretation. Findings of Chapter 7 suggest an additional way in which comprehenders may respond to ambiguous utterances by a speaker whose pragmatic competence they are unsure about, namely, request clarification. In that study, participants had the option of asking a follow-up clarification question if the instructions provided by the speaker were unclear, and they did so more often with the less pragmatically competent speaker. Notably, they did so even though they knew that they would not

receive a response: low certainty in the interpretation appeared to justify the effort of asking a question. Previous work on clarification requests suggests that people are likely to ask for clarification when they have low confidence in their interpretation, which prevents them from carrying out an action with sufficient certainty (Schlangen, 2004; Madureira & Schlangen, 2023). Our findings suggest that this certainty is dependent not just on the ambiguity of the utterance itself but also on beliefs about the speaker's pragmatic competence.

Taken together, the findings of Chapters 6 and 7 provide insights into the dynamics of adaptation to speakers' pragmatic competence. They contribute to theories of pragmatic processing by providing additional evidence that pragmatic meaning is both dynamic and interlocutor-specific and that it evolves over time as comprehenders update their beliefs. Moreover, our results suggest that effectively updating these beliefs depends on the comprehender's own pragmatic competence.

8.1.4 Computational cognitive modeling of comprehenders' reasoning about the speaker

The fourth research goal of this dissertation was to develop computational cognitive models formalizing how comprehenders' reasoning about the speaker influences the derivation of pragmatic inferences. The Rational Speech Act (RSA) framework (Frank & Goodman, 2012) was used in this thesis because it has a long tradition of being used to model various pragmatic phenomena (Degen, 2023), offers considerable flexibility, and is particularly well-suited for modeling reference games.

In Chapter 5, we proposed a model of comprehenders' reasoning about the speaker's pragmatic competence. In the proposed model, the listener holds beliefs about the type of speaker they are interacting with – literal or pragmatic. The speaker is represented as a mixture model (Heller et al., 2016) of these two types, with the mixture weight reflecting the listener's beliefs about the speaker's pragmatic competence. Beliefs about speaker types examined in Chapter 5 – adults, children, and artificial agents – can thus be captured through different mixture weights, with the literal type being weighed more heavily when the speaker is perceived as less pragmatically competent. Since participants were shown to differ in their beliefs about these speakers' pragmatic sophistication, this mixture weight can be set individually.

In Chapter 6, this model was extended to incorporate updating of the comprehender' beliefs about the speaker's pragmatic competence during interaction. This adaptation process is modeled as Bayesian belief updating. The modeling work provided important insights about the mechanisms underlying the pragmatic adaptation process.

First, it suggested that people may be updating their beliefs by reasoning about their own interpretations and not by observing speaker behavior directly: in our experiment, participants' reported confidence in the pragmatic interpretation increased on trials which were ambiguous from the listener's perspective, even when those trials

were unambiguous from the speaker's perspective, as predicted by the Bayesian belief updating model which uses the listener and not the speaker as the likelihood function. This is an interesting finding since it contrasts with how updating of listeners' beliefs has been modeled previously (Hawkins et al., 2017, 2021; Schuster & Degen, 2020; Roettger & Franke, 2019). This may be due to the fact that the task we use is an ad-hoc pragmatic reasoning task in a novel setting, where the speaker's perspective is quite different from the listener's, and participants have almost no experience with the task from the speaker's perspective. Thus, perspective-taking in this setting may be especially effortful. Sikos et al. (2021b) showed that having previous experience in the speaker role in a similar paradigm results in increased pragmatic reasoning as a listener. Therefore, it could be that given enough experience from the speaker perspective, people also more readily consider that perspective during adaptation. We return to this point and outline possible directions for investigation in future work in Section 8.2.

Second, the proposed model provided evidence for individual differences in adaptation behavior, which can be explained by differences in prior beliefs about the speakers' pragmatic competence and tendencies to generalize from the first speaker to the next one. A model where the hyperparameters corresponding to prior beliefs and generalization were set individually for each participant provided a better fit to the data than assuming the same hyperparameters for the whole population, providing evidence for distinct adaptation tendencies. The resulting hyperparameters indicate that most people expected the speaker to behave pragmatically a-priori, but the strength of this expectation varied, and that people differed in the strength of the generalization from one speaker to the next, with some treating the speakers as entirely independent and reverting to their prior beliefs with the second speaker, while others updated their prior beliefs about the second speaker after interacting with the first one. These findings contribute to our understanding of factors at play in pragmatic adaptation and open avenues for future work to explore this individual variability further. Theoretical accounts have been proposed in related domains of language processing, spoken language comprehension (Wu & Cai, 2024) and coordination of reference (Hawkins et al., 2017), which state that people have both a mental model of the population and of the specific interlocutor, and that those models may inform each other. Our findings suggest that the same principle may hold for pragmatic adaptation to the speaker, and that the degree to which these two mental models interact varies across individuals.

In Chapter 5, we discussed the fact that the proposed model predicts that the listener always favors the pragmatic target, even when the speaker is predicted to be fully literal. This is because in the RSA framework, alternative utterances available to the speaker and their corresponding probabilities are always correctly weighed, allowing the listener to make a pragmatic inference not about the speaker's intent but about the probabilities of alternatives. We showed in Experiment 4 of Chapter 5 that this assumption does not hold in practice, at least for the ad-hoc inferences

in the novel paradigm used here: with a literal speaker, participants did not weigh alternative utterances, instead resorting to literal interpretations. In the adaptation experiment in Chapter 6, we also observed the pattern where some people who derived pragmatic interpretations with high confidence with a pragmatic speaker behaved like literal comprehenders with a literal speaker, although the pragmatic interpretation is more likely even with a literal speaker based on the available alternatives. Future work could explore whether people are more successful about reasoning about alternatives with a literal speaker in a more familiar linguistic paradigm. Given how widespread this failure to correctly reason about alternative messages appears to be in our experiment, this is something that may be important to model in the future, whether in RSA or in another framework. One way that this could be expressed in RSA is by expressing not only the speaker as a mixture of a literal and pragmatic model, but the listener as well. This would correspond to the assumption that the belief that the speaker is less confident or cooperative leads listeners to reason more deeply, which is a likely interpretation of the experimental findings of Chapter 5, as discussed in the previous section.

8.2 Directions for future work

The research presented in this dissertation presents a number of important questions which can be investigated in future work.

8.2.1 Extension to other paradigms

The experiments in this dissertation, with the exception of Chapter 7, employed the reference game paradigm, a simple signaling game with pictorial messages. This choice was motivated by the fact that this is a flexible paradigm which is easy to adapt to various settings and lends itself to computational modeling in the RSA framework. However, these advantages of the reference game paradigm also constitute an important limitation: in contrast to reference games, in natural communication using language, the set of possible meanings and alternatives is very large and not necessarily shared between interlocutors. Thus, to ensure the generalizability of the findings, it is important to extend the investigation to more naturalistic paradigms which rely more heavily on language. We did so in the case of the adaptation experiment in Chapter 7, where we extended the reference game adaptation study from Chapter 6 to the more complex block building paradigm and showed that a very similar effect of dynamic adaptation to the speaker's pragmatic competence emerged across the two experimental paradigms.

Future work could extend experiments from Chapters 4 and 5 to other experimental paradigms. For instance, experiments described in Chapter 5, where we investigated the influence of beliefs about the speaker's pragmatic competence on interpretation,

could be conducted in the block building paradigm introduced in Chapter 7: participants could be told that the instructions were provided by a child and transcribed or that they were generated by ChatGPT.

Relatedly, the experiments presented in this dissertation used accuracy and post-hoc explanations as measures. Using online processing measures, such as reading times and eye-tracking, could yield valuable insights into the timecourse and processing mechanisms that underlie the observed effects. A relevant recent application of eye-tracking to the study of pragmatic reasoning in reference games is Mykhalievskyi (2025), who showed that people's accuracy in the reference game is predicted by the amount of time they spend examining the bank of available messages, which may be an indicator of deeper reasoning.

8.2.2 Cognitive individual differences in pragmatics

Chapter 4 investigated the cognitive factors underlying variability in ad-hoc pragmatic reasoning. This variability is theoretically motivated by three formal probabilistic models of different depths – L_0, L_1 and L_2. We identified a link between deeper pragmatic reasoning and logical reasoning ability and Theory of Mind. We also found that the effect of logical reasoning is stronger for complex implicatures than for simple ones. Further investigation of this finding is warranted: Is it related simply to the overall lower success rate on complex implicatures, or is there some fundamental difference in strategies that underlies this interaction effect? Relatedly, open questions remain about how these pragmatic reasoning strategies come about in the first place – do people learn to be better pragmatic reasoners as a result of having access to a certain kinds of data or experience?

Relatedly, investigation of strategies revealed a small group of "odd-one-out" reasoners – participants who interpreted the speaker's messages inversely, i.e., taking the message "red" to mean the only *non-red* object. We excluded such participants from our analyses since their reasoning was fundamentally different than the kind of reasoning predicted by the three formal probabilistic models – L_0, L_1 and L_2 – which theoretically motivated the observed variability. Future work could investigate such reasoners in more detail: How does such a complex pragmatic reasoning strategy come about? Are these cognitive traits which separate such reasoners from others?

In the experiment presented in Chapter 4, we identified the three classes of reasoners using Latent Profile Analysis (LPA), while the related work by Franke & Degen (2016) uses hierarchical Bayesian modeling. Our motivation for using LPA was that it offers additional flexibility in terms of the number of identified classes due to being data-driven, which was an advantage since we showed in Chapter 3 that participants apply a number of different strategies which may result in more than 3 reasoning classes. Future work could compare clusters obtained by hierarchical Bayesian modeling with those obtained using LPA and classification based on reported strategies. We expect that quite similar results would be obtained.

8.2.3 Speaker effects in pragmatic processing

One limitation of the experiments presented in Chapter 5 is that the derivation of implicatures relied on a restricted set of available messages, which is arguably not very natural. Future work could extend the investigation of the effect of speaker identity on pragmatic interpretation to other paradigms, where it could be signaled in other ways. For instance, the same underinformative utterances could be recorded by a child and an adult and their interpretations could be compared.

In Experiment 3, we found that participants interpreted messages allegedly sent by ChatGPT less pragmatically than those sent by another human adult. However, participants' experience with ChatGPT in this experiment – where they were told that ChatGPT selected the messages they were interpreting – was substantially different from the interactive, chat-based use they were presumably used to. The experiment could be repeated by simulating a direct interaction with ChatGPT through a chat window or by including a "screenshot" of sent messages in the familiar chat window.

Also, at the time when the Experiment 3 was conducted in the fall of 2023, familiarity with ChatGPT was considerably lower. Additionally, at the time ChatGPT did not process images directly, which is why we told participants that the task was given to it in written form. It could be that the perception of ChatGPT as a pragmatic speaker has changed in the past 3 years, reflecting increased familiarity and growing complexity of the model. Therefore, it would be worthwhile to replicate the study and compare the findings to those of our study.

Taken together, the findings of the experiments presented in Chapter 5 suggest that people may be engage in more or less deep pragmatic reasoning depending on their beliefs about the speaker, indicating that the same person may behave more or less pragmatically in different settings, as indicated by a larger proportion of literal interpretations with child and ChatGPT speakers than with an adult human speaker. Future work could investigate different settings which may be more or less conducive to deeper or shallower pragmatic reasoning, potentially in a within-participant experiment for more direct evidence. Since our results indicate that sufficient motivation may be required to engage in effortful pragmatic reasoning, it seems likely that people would behave more pragmatically when the situation is framed as especially important and meaningful. This would be consistent with findings from the psychology of reasoning, which indicate that people's reasoning changes in high-stakes situations, with increased use of more complex strategies (Kunda, 1990; Tversky & Kahneman, 1981).

8.2.4 Dynamic adaptation to the speaker

In Chapters 6 and 7 we showed empirical evidence of comprehenders' dynamic adaptation to the speaker's pragmatic competence and proposed a computational model of this adaptation process. There are possible modeling extensions, as well as open

questions about the observed individual variation in adaptation patterns.

We showed that, when interacting with a literal speaker, some participants resorted to behaving like literal listeners, as evidenced by their responses oscillating between the target and the competitor, indicating guessing. Our proposed model does not capture this pattern since it assumes that comprehenders are pragmatic listeners with uncertainty about the speaker's pragmatic competence. Future work could extend the model to capture the intuition that the listener type is a mixture as well and that when dealing with a literal speaker, some listeners reason less deeply and fall back on literal interpretations.

As discussed in Section 8.1.4, our modeling work in Chapter 6 provided evidence that in a setting where perspective-taking is effortful, people may update their beliefs by reasoning about their own interpretations as opposed to by directly observing speaker behavior, which was reflected in the fact that learning occurred on trials which were ambiguous from the listener's perspective but unambiguous from the speaker's perspective. However, our study featured only a small number of trials which were ambiguous from the speaker's perspective and unambiguous from the listener's perspective – those are the trials on which belief updating is expected to occur if listeners learn by observing the speaker. Therefore, the number of such trials may have been too small for participants to recognize their relevance. Future work could conduct an experiment where the number of trials ambiguous from either perspective is balanced and investigate whether adaptation patterns change.

In addition, it has been shown that having experience from the speaker's perspective improves one's pragmatic reasoning as a listener (Sikos et al., 2021b). Future work could investigate this by having people play the role of speakers before switching to being listeners and test whether this results in higher rates of adaptation and more perspective-taking.

The adaptation model proposed in Chapter 6 could be extended to other paradigms, such as the block building game used in Chapter 7. Given that the experimental results of the two experiments are quite similar, we expect that the predictions of the proposed model will provide a good fit to the results of Chapter 7 as well. Extending the model to this setting would require defining the possible alternative utterances and alternative structures. For simplicity, one could start by assuming that the two alternative utterances are the underspecified utterance and the corresponding specified utterance that mentions the stack height or color explicitly. The alternative structures could have all possible colors and stack heights between 1 and 5.

The proposed model formalizes comprehenders' pragmatic adaptation as Bayesian belief updating. Because such updating is multiplicative – to derive the posterior, the prior is multiplied by the likelihood of parameters given the observation – there is no parameter controlling the *speed* of updating. However, it is quite possible that some participants adapt more quickly than others; in our model, this is captured by the strength of prior beliefs – if a person holds stronger beliefs, they will take more exposure to be changed. Since it is plausible that some participants adapt faster than

others, future work could additionally explore approaches which involve an adaptation learning rate.

Finally, the findings of Chapter 6 raise questions about individual differences in pragmatic adaptation. While there was strong evidence of adaptation at the population level, about a quarter of our sample did not adapt: some participants behaved like literal listeners, while others did not lower their confidence in the pragmatic interpretation despite repeated feedback that it was incorrect. Future work could investigate whether these two types of non-adapters can be distinguished from adapters based on cognitive or personality traits. For instance, the findings of Chapter 4 suggest that the literal comprehenders who did not adapt may have lower logical reasoning ability of lower Theory of Mind. The unwillingness to lower confidence in the pragmatically correct response may be related to not being able to track the speaker's responses due to low working memory updating skills (Loy & Demberg, 2024; Schuster et al., 2023); alternatively, it could be related to a personality trait such as openness. Finally, there were differences between adapters as well: we found evidence of differences in prior beliefs and tendency to generalize between speakers. Future work could investigate whether factors underlying these differences may be identified.

8.2.5 Extension to interactive settings

The experiments presented in this dissertation were comprehension experiments in which participants did not have the option of responding to the speaker or asking a clarification question. Thus, comprehenders were limited to adjusting their own interpretations to adapt to a particular speaker. However, much of real-life communication is interactive and involves a dialogue, where the option of requesting clarification exists. Future work could extend the presented experiments to the interactive setting and investigate in which situations comprehenders request clarification or otherwise change the way they communicate.

Interestingly, in Chapter 7, we observed that even though asking a question was optional and did not have an obvious benefit to participants as they did not receive a reply, half of the participants asked questions, and more clarification questions were asked with the literal speaker. Building on these findings, future work could test whether these results generalize to an interactive setting where comprehenders can receive an answer to their question.

8.2.6 Investigation of individual differences and adaptation in production

Finally, this thesis focused on individual differences in comprehenders' pragmatic reasoning and the way they accommodate pragmatic abilities of the speaker. However, there remains the question of the extent to which similar findings would hold for production. There is a large body of work demonstrating that speakers tailor their

utterances to the addressee, including speaking differently to different addressee types (Fischer et al., 2011) and providing different amounts of information depending on what they expect the addressee to know (Krauss, 1987). At the same time, existing research indicates that there are limits to this perspective-taking, e.g., speakers have been shown to overestimate the informativeness of their utterances (Keysar & Henly, 2002) and act egocentrically under time constraints (Horton & Keysar, 1996). Still, open questions remain about pragmatic abilities of speakers, and individual differences in those abilities and in speakers' tendency to adapt to the comprehender.

Franke & Degen (2016), whose comprehension experiment Chapters 3 and 4 of this dissertation build on, observed fewer individual differences in production than in comprehension. However, while most of the participants of their production experiment were best described by the pragmatic speaker model S_1, some participants behaved like literal speakers. Do the same individual differences which were shown in Chapter 4 to be associated with better pragmatic reasoning in comprehension – logical reasoning skills and Theory of Mind – also predict pragmatic reasoning abilities in production?

In Chapters 5, 6 and 7 we showed that listeners accommodate differences in speakers' pragmatic competence by adjusting their interpretation and requesting clarification. Chapter 6 additionally provided evidence for distinct adaptation patterns among individuals. Are speakers equally sensitive to individual differences in listeners' pragmatic competence? Are there individual differences in speakers' tendency to adapt to the listener? These questions could be explored in future work.

8.3 Conclusion

The research presented in this dissertation aimed to advance theories of pragmatic processing by investigating individual variability and adaptation to the speaker through a combination of experimental work and computational cognitive modeling.

We showed that individual differences in pragmatic reasoning are predicted by cognitive traits – specifically logical reasoning and Theory of Mind – thereby laying the groundwork for processing-level models of pragmatic reasoning. In addition, our research revealed that comprehenders' inferences are guided by their beliefs about the speaker's pragmatic competence, and that those beliefs are updated dynamically during interaction. This dissertation also makes a methodological contribution by proposing a method for eliciting strategies that participants use in a pragmatic task, which proved to be effective for revealing and mitigating biases in experimental stimuli and for gaining deeper insight into the strategies participants used in the task.

Taken together, the findings of this dissertation support a nuanced view of pragmatic processing in which individuals' pragmatic reasoning is influenced by both cognitive and situational factors, and in which interlocutors anticipate and adapt to such differences in one another.

Finally, this work opens several avenues for future research, including identifying specific situational factors that are conducive to deeper versus shallower pragmatic reasoning, investigating cognitive and personality factors which may explain individual differences in adaptation to speakers' pragmatic competence, and achieving a more fine-grained understanding of the underlying cognitive mechanisms through online measures such as reaction times and eye-tracking.

Appendices

Appendix A

Annotation scheme for the proposed strategy elicitation method

The following annotation scheme was used for annotating the strategies reported by participants in the reference game in Chapters 3-5. Below, definitions of each tag, as well as examples from the data (slightly edited for clarity) are included. Tags are listed in alphabetical order.

Each of the chapters uses a slightly different version of the annotation scheme that is adapted to its specific goals. The annotation scheme described below includes all tags used across the three chapters. If a tag appears only in one chapter, that is noted in its description. For the subset of tags used in a given chapter, please refer to that chapter.

Annotation tag: ***ascribe_rationality***.

- Definition: Despite being explicitly told that the speaker is literal, a participant's explanation reflects that they expect the speaker to behave pragmatically.
- Examples
  - See examples of *correct_reasoning* below. In the case when the speaker is explicitly framed as literal, such examples should be assigned the tag *ascribe_rationality*.
- Notes: Used only in Experiment 4 of Chapter 5.

Annotation tag: ***correct_reasoning***.

- Definition: The reported strategy describes counterfactual reasoning (if the speaker had meant X, they would have said Y) captured by the pragmatic RSA listener model.

- Examples
  - "The speaker could have picked the triangle to direct me to the triangle but picked red instead so I believe wanted to direct me away from the triangle to the square."
  - "There was a triangle they could have used, but not a square, so it's likely it was the green square they wanted me to pick."
  - "They are definitely indicating a circle, there is no colour swatch for the blue circle so by process of elimination the white circle indicates the blue one."

Annotation tag: *changed_ mind*

- Definition: The participant changed their mind upon reflection, leading to it not matching the ratings they provided using sliders.
- Examples
  - "I've made an error to be honest, after carefully thinking just now it should be blue."
  - "I tried to click the red one but misclicked."
  - "I chose the red tirangle but process of elimination, but now I think I made a mistake and they actually meant the red circle!"
- Notes
  - This tag is called *mistake/changed_ mind* in Chapter 4.
  - Participants who provided this explanation were excluded from analysis of annotations but included in the regression analysis.

Annotation tag: *guess*.

- Definition: The participant says that there is no way to choose between two objects (target and competitor) since both objects have the feature expressed by the message.
- Examples
  - "The clue is a triangle, there are two of them so it's a 50/50 split."
  - "The clue is green so it is equally likely to be the triangle or the square."
  - "Previous participant chose red. I split the slider on the two red shapes given 50/50."

Annotation tag: *guess*. Subtag: *screenloc*.

- Definition: Participant reported decided which referent to select based on its position on the screen (e.g., the object on the left or in the middle).
- Examples
  - "It was the first red object."
  - "I couldn't decide on the shape, so I went with the first of the two red options."
  - "I selected the middle object of the same color as the message."

Annotation tag: *guess*. Subtag: *variability*.

- Definition: The explanation states that the participant chose a certain object "to mix things up" since they had chose a different shape or color more often in the past.
- Examples
  - "It has to be a red shape, and I chose the triangle last time, so this time I chose a square."
  - "There were no blue objects on the previous trial, so I chose a blue object this time."

Annotation tag: *meta_ reasoning*. Used in Experiments 1-3 of Chapter 5.

- Definition: The reported strategy explicitly mentions what the correct pragmatic reasoning would be but also expresses uncertainty about whether the speaker behaved optimally.
- Examples
  - "Because it has to be a circle, and there isn't an option to choose blue. So the empty circle would be the correct choice from the child if the blue circle had the yellow background, but I don't know how good kids are at thinking logically like that."
  - "I have some doubt in the child's ability to make this more complex decision so I've left 25% in room for doubt, and 75% likelihood for the blue triangle, as the child didn't have a blue paint option to differentiate the two triangle options."
  - "The messages available don't have a blue colour to distinguish the different circles if the blue one was highlighted, therefore it is most llikely the blue circle is the right answer. There is a small chance the participant giving the answer wasn't paying attention, therefore I put a 1% chance on the other circle."

Annotation tag: *misunderstood_ instr*

- Definition: The reported strategy suggests misinterpreting the instructions that the speaker could not utter a certain message, e.g. "blue", as meaning that they could not refer to an object if it had that feature (e.g. a blue triangle cannot be the intended referent since "blue" is not an available message).
- Examples
  - "It is more likely they choose green as choose a circle and it could not be a blue circle so it had to be a green circle."
  - "The speaker was not given a square message, only a circular one, which is green." (As a reason to rule out a square as a possible referent)
  - "The speaker chose a circle, and only red was available to match." (As a reason to rule out a blue circle as a possible referent)
- Notes: Participants whose explanations were annotated with this tag were excluded from all analyses because it reflects performing a fundamentally different task.

Annotation tag: *odd_ one_ out.*

- Definition: Selecting an object because it stands out from the rest based on the presence or absence of a feature. Often corresponds to an inverse interpretation, i.e., picking the only object that is *unlike* the message (e.g., the only *non-red* object when receiving the message *red*).
- Examples
  - "Because two of the three objects are red, the message chosen by the participant [the color red] must mean the blue object which is different from the rest."
  - "There is one red triangle but 2 circles (one blue, one red) I assume that sending the circle message is the easiest way to identify the triangle in this situation because it's the odd one out in terms of shape."
  - "There are two red shapes, so I went for the triangle, which is the distinct shape from the rest."
- Notes:
  - In Chapter 3, *odd_ one_ out* was used not as a separate tag but as a subtag of *other_ reason*.

  - Not to be confused with *other reason - salience*. *odd_one_out* is assigned when the object is selected because it stands out based on the presence or absence of a feature (exact match or mismatch), whereas *salience* is assigned when the object stands out based on being bigger or brighter than the rest (relative difference).

Annotation tag: *other_reason*. Subtag: *multiturn*.

- Definition: In the explanation, the participant reports selecting a certain object based on how often certain features (e.g., color or shape) have appeared earlier in the experiment. The explanation suggests that the participant is looking for a pattern based on visual features.
- Examples
  - "I think red images are shown more often so I have a better chance of being right if I choose the red triangle."
  - "Triangle turns up quite often in the choices."
  - "I think it is the triangle because I have kind of noticed a pattern, which is that red color is used for indicating a triangle."

Annotation tag: *other_reason*. Subtag: *preference*.

- Definition: The explanation expresses a personal preference for a certain visual feature of the selected object.
- Examples
  - "[I chose blue] because I like blue more than red."
  - "I like circles better than triangles."
  - "It was more visually appealing."
- Notes: In Chapter 5, *preference* is combined with *salience* and used not as a subtag but as a separate tag: *preference/salience*.

Annotation tag: *other_reason*. Subtag: *salience*.

- Definition: Participant reports selecting an object because it stood out to them visually, e.g. due to being bigger or brighter than the rest.
- Examples
  - "I think the blue triangle stands out more than the red one."
  - "I feel that red is more likely since it's a brighter color."
  - "[I chose] blue because it is the brighter color."

- Notes: In Chapter 5, *salience* is combined with *preference* and used not as a subtag but as a separate tag: *preference/salience*.
- See *odd_one_out* for an explanation of the difference between the two tags.

Annotation tag: *other_reason*. Subtag: *visual_resemblance*.

- Definition: The reported strategy reveals that the participant selected an object because it was visually the most similar one to the message.
- Examples
  - The paint splodge image chosen by the the participant looks like a circle [as a reason for choosing the circle object].
  - The cap [the message] is facing in the same direction [as the cap that the selected referent is wearing] (Example from Experiment 1 of Chapter 3, where the stimuli were creatures and accessories)
  - The previous participant's image is the same except that my chosen image is wearing a scarf but its head is exposed (Also an example from Experiment 1 of Chapter 3)
- Notes: A large number of occurrences of this tag may indicate that the stimuli are biased, as discussed in Chapter 3.

Appendix B

Chapter 3: Annotations of the complex condition and LPA model fit

A Annotations for the complex condition

Figure B.1 shows by-tag results in the complex condition of the 4 experiments in Chapter 3. It is notable that there is a large proportion of *odd_ one_ out* responses. The fact that we observe a smaller proportion of *other_ reason* responses in Experiment 4 is a good sign because the employed strategies are predominantly *correct_ reasoning* and *guess*, as is assumed by formal probabilistic models.

B LPA model fit

Table B.1 and Table B.2 show fits for models with 1 through 6 latent classes for Experiments 1 and 4 respectively, as measured by AIC and BIC. The 4-class model provides the best fit for Experiment 1, and the classes approximately correspond to the predictions of the three probabilistic listener models: L_0, L_1, and L_2. For Experiment 4, the 4-class model has the best fit according to BIC and the 5-class model has the best fit according to AIC. The difference between the 4- and 5-class models is that in the 5-class model, one participant who utilized the *odd-one-out* strategy on all trials, resulting in 0 accuracy on both simple and complex trials, is assigned to their own class. Since that is the only difference, for comparability in the chapter we discuss 4-class models for both experiments.

C LPA classes (complex condition)

The 4 latent classes for Experiments 1 and 4 with annotation tags from the complex condition are included in Figure B.2.

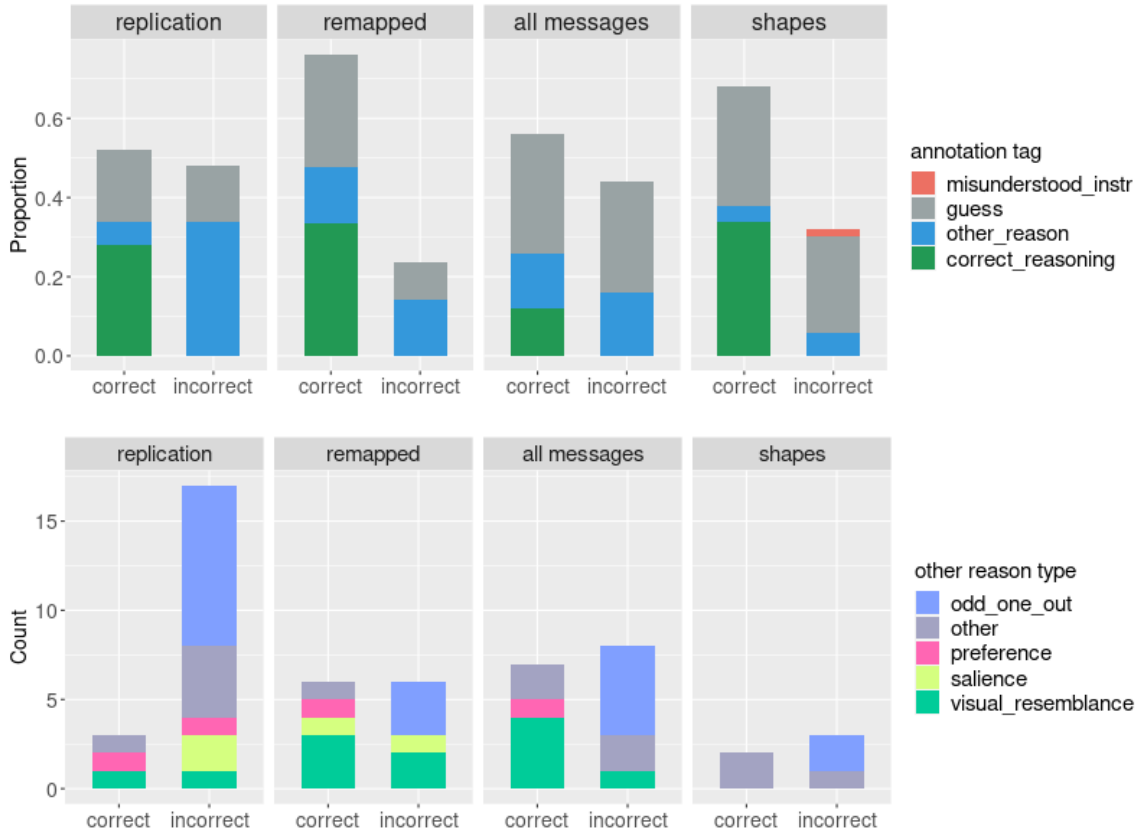


Figure B.1: Participants' explanations by tag in the complex condition (top – all explanations, bottom – *other_reason* explanations).

In Experiment 1 (top panel), we observe a large number of people who use the *odd_one_out* strategy, thus performing below chance. They are assigned to the class corresponding to the L_1 model, where the reasoner can solve simple implicatures but guesses randomly on complex ones. That is not ideal since the *odd_one_out* strategy where the message is interpreted inversely is a different, more complex strategy than guessing. In Experiment 4 (bottom panel), we get cleaner results: there are only a few people who use *odd-one-out* reasoning; for the most part, L_0 and L_1 classes consist of people whose strategy in the complex condition was guessing, and the L_2 class consists of people who applied the *correct_reasoning* strategy.

Num. classes	AIC	BIC
1	13.89	22.06
2	-4.88	9.42
3	-6.37	14.06
4	**-18.83**	**7.73**
5	-16.55	16.14
6	-11.62	27.2

Table B.1: Fit of LPA models for Experiment 1 of Chapter 3.

Num. classes	AIC	BIC
1	30.27	38.65
2	17.96	32.62
3	8.55	29.5
4	-6.16	**21.07**
5	**-9.66**	23.85
6	-4.87	34.92

Table B.2: Fit of LPA models for Experiment 4 of Chapter 3.

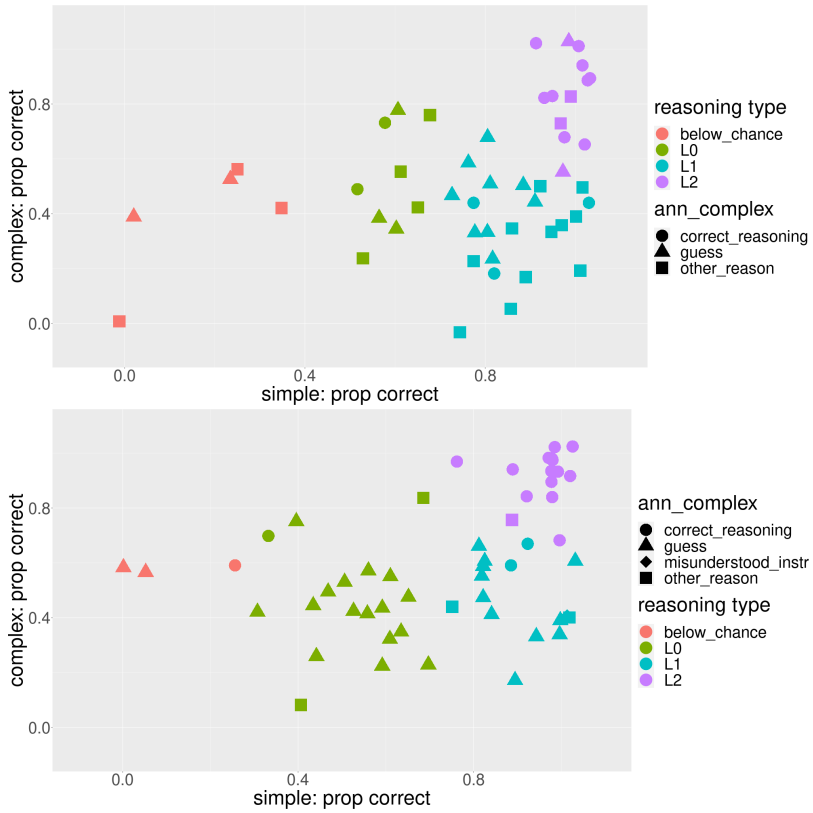


Figure B.2: Classes of participants identified by LPA (top - Experiment 1 (replication), bottom - Experiment 4 (debiased stimuli)). Shapes correspond to the tags assigned to the reasoning strategies in the complex condition.

Appendix C

Chapter 4: Model priors and additional models

A Regression model output with additional exclusions based on Raven's matrices

Table C.1 shows the output of a regression model which additionally excludes 8 people who responded "Yes" or "Not sure" to the question whether they recall seeing any of the questions on the Raven's matrices test before, resulting in a sample size of 159 participants.

The effect size estimates and credible intervals are very similar to those reported in the main analysis, where these participants were included. The interpretation of the results remains unchanged.

B Regression model priors

B.1 Model without individual differences

We fit the following Bayesian logistic regression model to the data using the *brms* package in R:

$$\begin{aligned} binaryCorrectness \sim condition + trialNumber + condition * trialNumber + \\ targetPos + msgType + (1 + condition + msgType \\ + trialNumber | participantID) + (1 | itemID) \end{aligned}$$

For all effects, we set wide weakly informative priors in order to avoid biasing the model. Below we list the priors with a brief rationale.

	Model with Raven's exclusions (*N*=159)	
Effect	**Estimate**	**95% CrI**
Intercept	**0.45**	[0.22, 0.68]
condition1 (simple vs. complex)	**1.01**	[0.62, 1.42]
trial number	**0.01**	[0.00, 0.02]
msgtype (color vs. shape)	-0.09	[-0.22, 0.04]
targetpos (middle vs. left)	**0.54**	[0.34, 0.75]
targetpos (right vs. left)	**-0.23**	[-0.42, -0.03]
condition1 : trial	**-0.01**	[-0.02, -0.00]
logical reasoning	**0.21**	[0.04, 0.39]
memory	0.05	[-0.12, 0.21]
ToM	**0.20**	[0.03, 0.37]
condition1 : reasoning	**0.41**	[0.06, 0.77]
condition1 : memory	0.23	[-0.11, 0.58]
condition1 : ToM	0.18	[-0.17, 0.52]

Table C.1: Output of the regression model with individual differences, where participants who reported being familiar with any of the questions on Raven's matrices were excluded.

- Fixed effects:
  - Prior: Normal(0, 2).
  - Rationale: For all main effects and interactions we set a wide prior centered at 0. The wide standard deviation of 2 was chosen after examining the effect sizes in priors work in similar paradigms (see Table 4.3 in Chapter 4). This wide prior is chosen to be conservative and allow for flexibility in estimating effect sizes.
- Standard deviation of random effects for *participantId* and *itemId*
  - Prior: Student_t(3, 0, 1)
  - Rationale: Student's-t distribution with 3 degrees of freedom, centered at 0 with a scale of 1 was chosen to be robust to outliers and to capture for potential variability between participants and items beyond that accounted for by fixed effects.
- Correlation structure
  - Prior: LKJ(2)
  - Rationale: We set a weakly informative prior, allowing for moderate correlation of random effects.

B.2 Model with individual differences

This model includes the same effects as the model without the individual differences. It additionally includes main effects of the three individual differences, as well as their interactions with condition:

$$binaryCorrectness \sim condition + trialNumber + condition * trialNumber + targetPos + msgType + condition * reasoning + condition * WMC + condition * ToM + (1 + condition + msgType + trialNumber|participantID) + (1|itemID)$$

The same wide regularizing priors as for the model without individual differences (discussed in the previous subsection) were set for all effects.

C Regression model output without principal components

shows the output of a regression model where the individual logical reasoning and ToM measures are used as predictors instead of the principal components (N=167).

	Model with ID measures instead of PCs (N=167)	
Effect	**Estimate**	**95% CrI**
Intercept	-0.88	[-1.55, 1.20]
condition1 (simple vs. complex)	-0.32	[-1.53, 0.88]
trial number	**0.01**	[0.01, 0.02]
participant age	**-0.02**	[-0.04, 0.01]
msgtype (color vs. shape)	-0.10	[-0.22, 0.02]
targetpos (middle vs. left)	**0.53**	[0.33, 0.74]
targetpos (right vs. left)	**-0.21**	[-0.40, -0.03]
condition1 : trial	-0.01	[-0.02, 0.00]
CRT	**0.69**	[0.00, 1.38]
RPM	0.45	[-0.29, 1.18]
RMET	0.82	[-0.08, 1.73]
SST	0.58	[-0.11, 1.27]
condition1 : CRT	0.42	[-0.85, 1.70]
condition1 : RPM	**1.53**	[0.19, 2.89]
condition1 : RMET	-1.02	[-2.64, 0.61]
condition1 : SST	**1.46**	[0.15, 2.80]

Table C.2: Output of the regression model with individual differences in logical reasoning and ToM without principal components.

Interestingly, when individual difference measures are included individually, only CRT has a main effect ($\hat{\beta}$=0.69, 95% CrI = [0.00, 1.38]). There are also noteworthy

interactions between condition and RPM ($\hat{\beta}$=1.53, 95% CrI = [0.19, 2.89]) and SST ($\hat{\beta}$=1.46, 95% CrI = [0.15, 2.80]). This might suggest that the relationship between pragmatic processing in the reference game is primarily driven by fluid intelligence or pattern recognition as opposed to reflective thinking, and by cognitive as opposed to affective Theory of Mind.

Figure C.1 visualizes the effects of RPM and SST in the two implicature conditions.

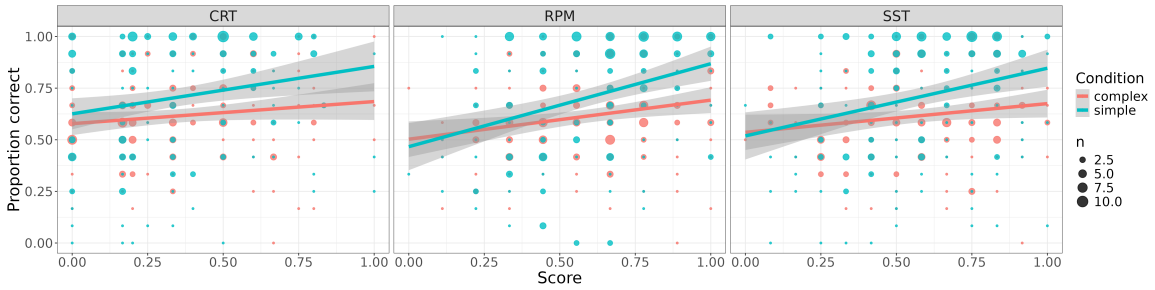


Figure C.1: Proportion of correct responses of individual participants in the two implicature conditions as a function of CRT, RPM and SST.

Appendix D

Chapter 5: Experiment instructions and model priors

A Experiment instructions

Below are the instruction screens that introduced the speaker and the speaker's task in both conditions in Experiments 1 and 2.

A.1 Experiment 1

Child speaker condition:

Instruction screen that introduced the speaker in the child speaker condition in Experiment 1 is shown in Figure D.1.

Adult speaker condition:

Instruction screen which introduced the speaker in the adult speaker condition in Experiment 1 is shown in Figure D.2.

A.2 Experiment 2

Child speaker condition:

The child's face in the photographs is blurred for privacy reasons. It was not blurred in the experiment.

Instruction screen which introduced the speaker in the adult speaker condition in Experiment 2 is shown in Figure D.3.

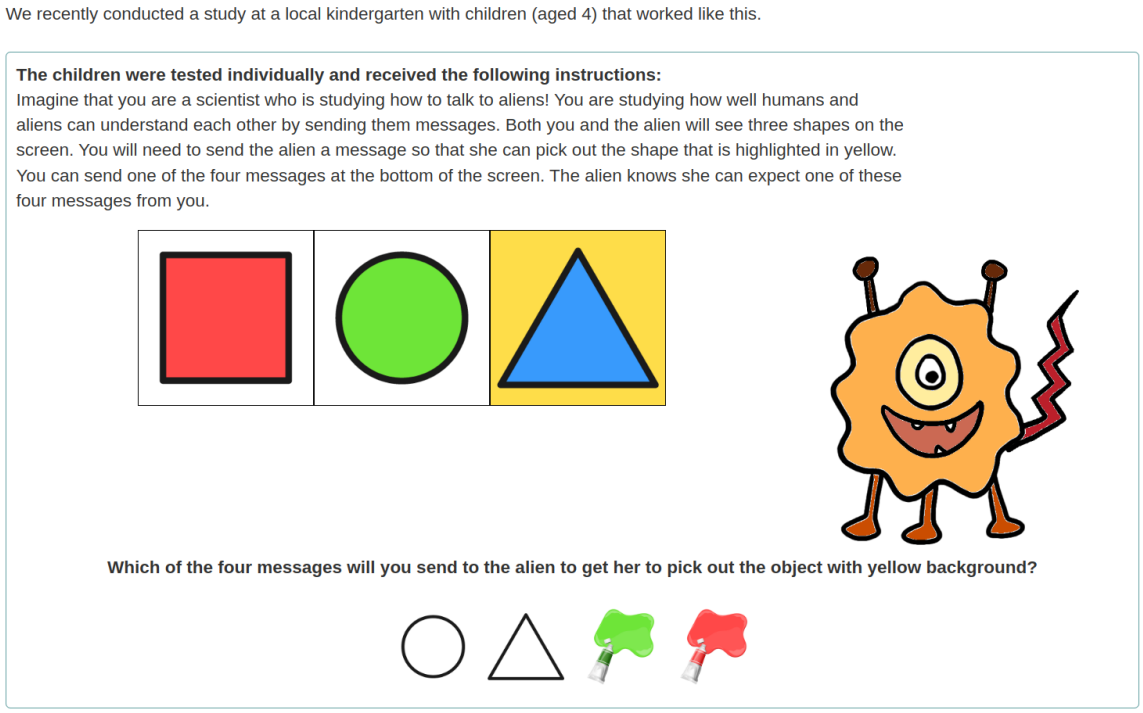


Figure D.1: Instruction screen which introduced the speaker in the child speaker condition in Experiment 1.

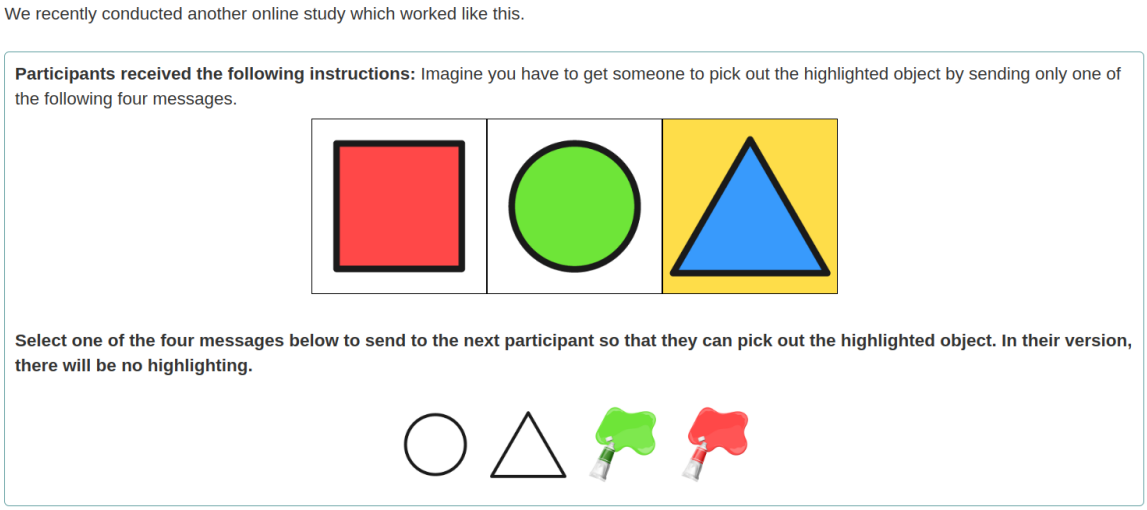


Figure D.2: Instruction screen which introduced the speaker in the adult speaker condition in Experiment 1.

Additionally, a picture where the child is looking at the possible messages was added to the screen with sliders for every trial. The same photograph stayed on the screen for all trials. One of the trials is shown in Figure D.4.

The children were tested individually. The following instructions were explained to them:

We're going to play a fun game! Look at the three shapes on the screen. One of them has a yellow background.

Later, someone else will see the same shapes but they won't see the yellow background. They will need to guess which one of the shapes has the yellow background.

You can choose one hint to help them pick the right shape. Which hint will you choose?

One of the children who participated in the experiment.

Figure D.3: Instruction screen which introduced the speaker in the child speaker condition in Experiment 2.

Adult speaker condition:

In Experiment 2, instructions introducing the speaker in the adult speaker condition were rephrased slightly to be more clear and closer to those in the child speaker condition. The instructions are shown in Figure D.5.

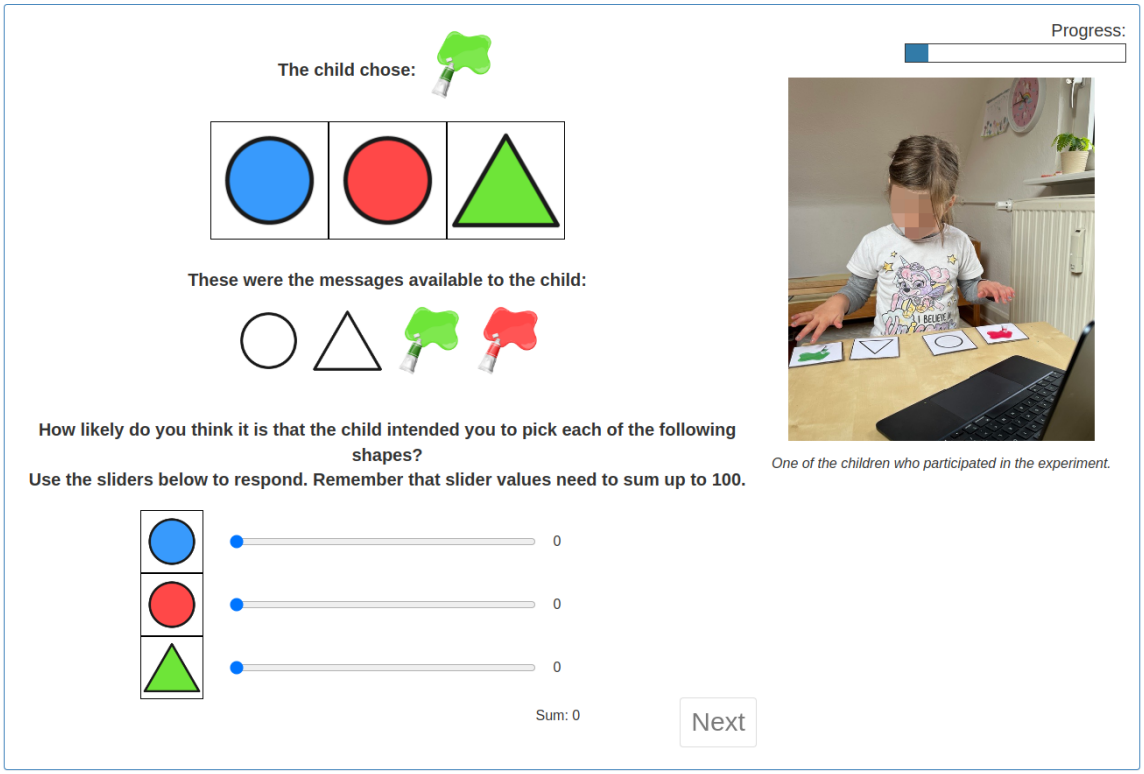


Figure D.4: One of the trials in the child speaker condition in Experiment 2. The photo stayed the same for all trials.

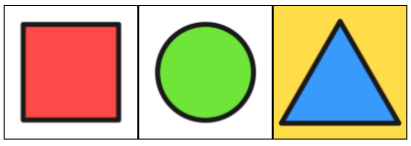
We recently conducted another online study which worked like this.

Participants received the following instructions:

In this study, you will play a communication game.

There three shapes on the screen, one of which is highlighted in yellow. Later, another participant will see the same shapes but without highlighting. Your task is to help them choose the highlighted shape by sending them a message that serves as a hint.

At the bottom, there are four messages to choose from. The other participant will know that they can expect one of these four messages from you.



Which one of the four messages will you send to help the next participant choose the right shape?

Figure D.5: Instruction screen which introduced the speaker in the adult speaker condition in Experiment 2.

A.3 Regression model priors

The following Bayesian regression model was fit to the data using the *brms* package in R:

$$targetRating \sim speaker * trialType + trialID + msgType + targetPos + (1 + trialType + msgType + trialID|participantID) + (1|itemID)$$

For all effects, we set wide weakly informative priors in order to avoid biasing the model. Below we list the priors for all effects with a brief rationale.

Fixed effects:

- Intercept
  - Prior: Normal(87.5, 12.5)
  - Rationale: We expect the ratings on unambiguous trials to be at ceiling (i.e., near 100) for both speaker conditions, and on critical trials, the plausible range of the ratings is 50-100 (50 corresponding to literal responding and 100 corresponding to perfectly pragmatic responding), so we take the middle point, 75, as the hypothesized mean for critical trials. Therefore, we set the mean of the intercept to the midpoint between 75 and 100, with a wide standard deviation.
- Speaker
  - Prior: Normal(0, 12.5)
  - Rationale: We do not want to assume the direction of the speaker effect so we set the mean to be 0. The standard deviation is set to be 12.5 so that the whole plausible range of target ratings, 50-100 (50 corresponding to literal responding and 100 corresponding to perfectly pragmatic responding) is covered by 2 standard deviations from the mean.
- Trial type
  - Prior: Normal(-25, 12.5)
  - Rationale: We expect target ratings on unambiguous trials to be at ceiling, i.e., near 100. Critical trials can plausibly take up the space between 50 and 100, so we take the midpoint as the mean with the standard deviation of 12.5, so that the decrease of ratings of 0 to 50 points compared to the unambiguous trials is covered by 2 standard deviations from the mean.
- Speaker : Trial type interaction
  - Prior: Normal(0, 12.5)

  - Rationale: Here, again, we do not want to assume the direction of the interaction and allow for the range 50-100 to be covered within two standard deviations.

- TrialID
  - Prior: Normal(0,5)
  - Rationale: We assume this to be a very small effect based on the results of previous studies in similar paradigms (e.g., Franke & Degen (2016) and Mayn & Demberg (2023)).

- Message type
  - Prior: Normal(0, 5)
  - Rationale: This is a covariate which we expect to have a very small effect based on the findings of previous studies.

- Target position
  - Prior: Normal(0, 5)
  - Rationale: This is also a covariate which we also expect to have a very small effect based on the findings of previous studies.

Random effects:

- Standard deviation of random effects for participant
  - Prior: student_t(3, 0, 12.5)
  - Rationale: We set a wide prior with a large scale parameter of 12.5 to account for potential variability in participant effects.

- Standard deviation of random effects for item
  - Prior: student_t(3, 0, 2.5)
  - Rationale: We do not expect there to be much variability between items based on previous work in the reference game paradigm with similar items (e.g., Mayn & Demberg (2023)), resulting in a smaller scale parameter of 2.5.

- Residual standard deviation
  - Prior: student_t(3, 0, 2.5)
  - Rationale: The scale parameter of 2.5 reflects the fact that we do not expect wide variation in residuals.

- Correlation structure

- Prior: LKJ(2)
- Rationale: We set a weakly informative prior allowing for moderate correlation of random effects.

A.4 Derivation of RSA model predictions

Let's show what predictions the proposed listener model $L_2^{belief-driven}$ makes for our experiment in different cases. We will use the example from Figure 5.1 but we will not consider the distractor since it is not relevant to the derivation of the inference. Therefore, we will consider two referents (red square and red triangle) and two messages ("red" and "triangle").

The model is defined as follows:

$$L_2^{belief-driven}(o|u) \propto S_{weighted}(u|o) \cdot P(o)$$
$$\text{where} \quad S_{weighted}(u|o) \propto \lambda \cdot S_1(u|o; \alpha \to \infty) + (1-\lambda) \cdot S_0(u|o) \quad \text{and } 0 \leq \lambda \leq 1$$

Let's derive the model's behavior at endpoints, $\lambda = 0$ and $\lambda = 1$, as well as at an intermediate point, $\lambda = 0.5$.

Listener who is certain that the speaker is literal: $\lambda = 0$

When $\lambda = 0$, the $S_{weighted}(u|o)$ equation simplifies to $S_0(u|o)$.

	"red"	"triangle"
red sq.	1	0
red tr.	0.5	0.5

$L_2^{belief-driven}$ with $\lambda = 0$ is essentially L_1. Assuming uniform prior probability over objects, $L_1(o|u) \propto S_0(u|o)$. To obtain the listener probability distribution, we transpose the S_0 matrix and normalize it:

	red sq.	red tr.
"red"	1	0.5
"triangle"	0	0.5

$\rightarrow$

	red sq.	red tr.
"red"	**0.66**	0.33
"triangle"	0	1

We can see that, despite the fact that the speaker is fully literal, L_1 correctly prefers the target (red square) to the competitor (red triangle) because of reasoning about the prior probabilities of alternatives.

Listener who is certain that the speaker is fully pragmatic: $\lambda = 1$

When $\lambda = 1$, the listener model $S_{weighted}(u|o)$ simplifies to $S_1(u|o; \alpha \to \infty)$.

S_1 reasons about literal listener L_0. $L_0(o|u)$ interprets the messages literally and assigns equal probability to all literally true referents. Assuming uniform prior probability of objects, $L_0(o|u) \propto [[u]]$, where $[[u]]$ is a boolean function evaluating whether the utterance is literally true of the object. Thus, for our example, $L_0(o|u)$ is:

	red sq.	red tr.
"red"	0.5	0.5
"triangle"	0	1

The speaker is then defined as $S_1(u|o; \alpha \to \infty) \propto exp(\alpha \cdot (\log L_0(o|u))$, assuming zero utterance costs for simplicity. The speaker temperature parameter α controls the degree to which the speaker maximizes their utility. Since we want the model to represent a speaker who always sends the pragmatically optimal message, we set $\alpha \to \infty$.

To obtain the S_1 distribution, we transform the L_0 matrix and calculate $exp(\alpha \cdot \log L_0)$ for $\alpha \to \infty$:

	"red"	"triangle"
red sq.	0.5	0
red tr.	0.5	1

$\to$

	"red"	"triangle"
red sq.	1	0
red tr.	0	1

As a result, the fully pragmatic speaker $S_1(u|o)$ always uses the message "red" to refer to the red square and the message "triangle" to refer to the red triangle.

Finally, to obtain the listener distribution $L_2(o|u)$, we transpose the $S_1(u|o)$ matrix (which is already normalized):

	red sq.	red tr.
"red"	**1**	0
"triangle"	0	1

Thus, at $\lambda = 1$, the listener will be perfectly certain that "red" refers to the red square.

Uncertain listener: $\lambda = 0.5$

A listener whose lambda is in between the endpoints will have uncertainty about whether the speaker is literal or pragmatic and will weigh the two speaker models equally.

We'll show the computation for $\lambda = 0.5$ but the principle is the same for any intermediate λ value.

$L_2^{belief-driven}(o|u)$ with $\lambda = 0.5$ will weigh the two speaker models, $S_0(u|o)$ and $S_1(u|o; \alpha \to \infty)$, equally.

To obtain $S_{weighted}(u|o)$, we take the $S_0(u|o)$ and $S_1(u|o; \alpha \to \infty)$ matrices, multiply each matrix by $\lambda = 1 - \lambda = 0.5$, add them together and normalize:

$0.5 \cdot$

	"red"	"triangle"
red sq.	1	0
red tr.	0.5	0.5

$+ \quad 0.5 \cdot$

	"red"	"triangle"
red sq.	1	0
red tr.	0	1

$\to$

	"red"	"triangle"
red sq.	1	0
red tr.	0.25	0.75

Finally, to obtain the $L_2(o|u)$ distribution, we transform the $S_{weighted}$ matrix and normalize:

	red sq.	red tr.
"red"	1	0.25
"triangle"	0	0.75

$\to$

	red sq.	red tr.
"red"	**0.8**	0.2
"triangle"	0	1

Thus, $L_2^{belief-driven}$(red square|"red") when $\lambda = 0.5$ is 0.8.

With different λ values, the probability range of $\frac{2}{3}$ to 1 is covered.

A.5 Sorted participant plot with simulated data

We hypothesized that our participants use a mix of different reasoning strategies – literal listeners as well as pragmatic listeners with different beliefs about the speaker's rationality – as opposed to just one strategy. We ask how closely the data we obtained in the experiments aligns with normal distributions with population-level mean and variance in each speaker condition, which would represent the assumption that all participants use the same strategy and the differences between them are due to sampling error.

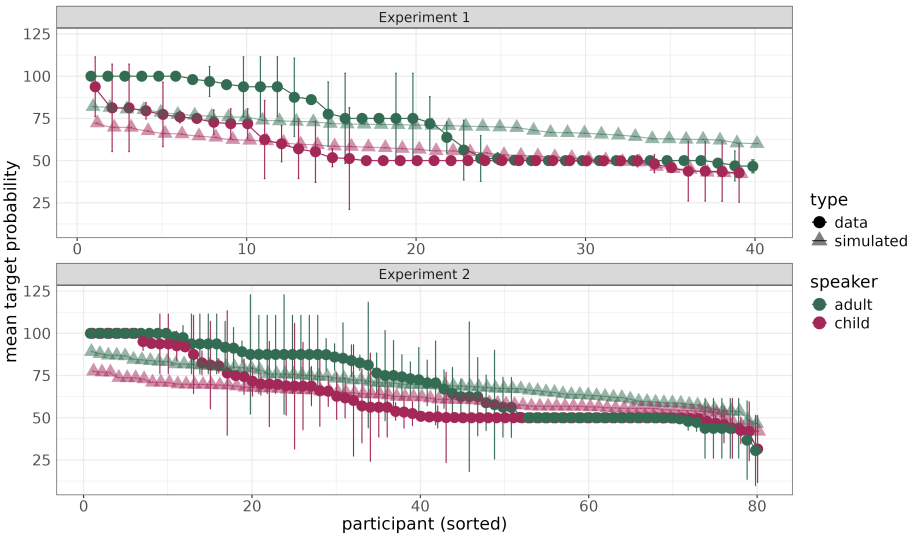


Figure D.6: Individual participants' average target ratings by speaker condition, sorted from high to low. The darker points are the data obtained from the experiment and the lighter points are simulated participants drawn from a normal distribution with the corresponding population-level mean and variance.

To answer that question, we plot actual participants' target ratings, sorted from high to low, as well as simulated participants drawn from normal distributions with population-level mean and variance in each speaker condition (Figure D.6). When we compare empirical data to the simulated data, we see that for both experiments, the alignment is not very good: there are more extreme high ratings and chance-level responses than would be predicted by a unimodal distribution that assumes that all participants used the same reasoning strategy. This suggests that the variation that we observe reflects differences in strategies, which should correspond to distinct models, as opposed to mere sampling chance.

List of Figures

List of Tables

Bibliography

Achimova, A., Scontras, G., Eisemann, E., & Butz, M. V. (2023). Active iterative social inference in multi-trial signaling games. *Open Mind*, *7*, 111–129.

Acton, E. K., & Potts, C. (2014). That straight talk: Sarah Palin and the sociolinguistics of demonstratives. *Journal of Sociolinguistics*, *18*(1), 3–31.

Akogul, S., & Erisoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. *Entropy*, *19*(9), 452.

Albert, D. A., & Smilek, D. (2023). Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports*, *13*(1), 20574.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and brain sciences*, *14*(3), 471–485.

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, *99*, 78–95.

Austin, J. L. (1975). *How to do things with words*. Harvard University Press.

Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes". *Intelligence*, *44*, 78–92.

Bambini, V., Battaglini, C., Frau, F., Pompei, C., Del Sette, P., & Lecce, S. (2025). On the inherent yet dynamic link between metaphor and Theory of Mind in middle childhood: meta-analytic evidence from a research programme bridging experimental pragmatics and developmental psychology. *Philosophical Transactions B*, *380*(1932), 20230489.

Bambini, V., & Lecce, S. (2025). At the heart of human communication: New views on the complex relationship between pragmatics and Theory of Mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *380*(1932), 20230486. doi:10.1098/rstb.2023.0486.

Bambini, V., Van Looy, L., Demiddele, K., & Schaeken, W. (2021). What is the contribution of executive functions to communicative-pragmatic skills? Insights from aging and different types of pragmatic inference. *Cognitive processing*, *22*(3), 435–452.

Banerjee, A., & Dave, R. N. (2004). Validating clusters using the Hopkins statistic. In *2004 IEEE International conference on fuzzy systems* (pp. 149–153). IEEE volume 1.

Barnes, J., & Baron-Cohen, S. (2011). Language in autism: Pragmatics and Theory of Mind. In *The Handbook of Psycholinguistic and Cognitive Processes* (pp. 731–745). Psychology Press.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

Battaglini, C., Frau, F., Mangiaterra, V., Bischetti, L., Canal, P., & Bambini, V. (2025). Imagers and mentalizers: Capturing individual variation in metaphor interpretation via intersubject representational dissimilarity. *Lingue e linguaggio*, *24*(1), 61–81.

Becker, J. A., & Hall, M. S. (1989). Adult beliefs about pragmatic development. *Journal of Applied Developmental Psychology*, *10*(1), 1–17.

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(2), 145–204.

Beltrama, A. (2018). Precision and speaker qualities. The social meaning of pragmatic detail. *Linguistics Vanguard*, *4*(1), 20180003.

Beltrama, A. (2020). Social meaning in semantics and pragmatics. *Language and Linguistics Compass*, *14*(9), e12398.

Beltrama, A., & Schwarz, F. (2021). Imprecision, personae, and pragmatic reasoning. In *Semantics and Linguistic Theory* (pp. 122–144). volume 31.

Beltrama, A., & Schwarz, F. (2024). Social identity affects imprecision resolution across different tasks. *Semantics and Pragmatics*, *17*, 10–EA.

Benz, A., Jäger, G., & Van Rooij, R. (2005). *Game theory and pragmatics*. Springer.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1450.

van Berkum, J. J., van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of cognitive neuroscience*, *20*(4), 580–591.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354–369.

Bischetti, L., Ceccato, I., Lecce, S., Cavallini, E., & Bambini, V. (2023). Pragmatics and Theory of Mind in older adults' humor comprehension. *Current Psychology*, *42*(19), 16191–16207.

Blackburn, H. L., & Benton, A. L. (1957). Revised administration and scoring of the Digit Span Test. *Journal of consulting psychology*, *21*(2), 139.

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*, *151*(11), 2927.

Bonnefon, J.-F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, *17*(9), 747–751.

Bosco, F. M., & Gabbatore, I. (2017). Sincere, deceitful, and ironic communicative acts and the role of the Theory of Mind in childhood. *Frontiers in Psychology*, *8*, 21.

Bosco, F. M., Tirassa, M., & Gabbatore, I. (2018). Why pragmatics and Theory of Mind do not (completely) overlap. *Frontiers in Psychology*, *9*, 1453.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, *7*(4-5), 585–614.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Burle, B., Spieser, L., Roger, C., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. *International Journal of Psychophysiology*, *97*(3), 210–220.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116.

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73–101.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*(3), 434–447.

Carlson, L., Skubic, M., Miller, J., Huo, Z., & Alexenko, T. (2014). Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task. *Topics in Cognitive Science*, *6*(3), 513–533.

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133.

Chapman, S. (2010). Paul Grice and the philosophy of ordinary language. In *Meaning and Analysis: New essays on Grice* (pp. 31–46). Springer.

Chater, N., & Oaksford, M. (2012). *Normative systems: Logic, probability, and rational choice*. The Oxford handbook of thinking and reasoning.

Chierchia, G. et al. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and Beyond*, *3*, 39–103.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive Psychology*, *11*(4), 430–477.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, *34*(1), 158–171.

Constable, R. T. (2023). Challenges in fMRI and its limitations. *Functional Neuroradiology: Principles and Clinical Applications*, (pp. 497–510).

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552.

Conway, A. R., & Kovacs, K. (2013). Individual differences in intelligence and working memory: A review of latent variable models. *Psychology of Learning and Motivation*, *58*, 233–270.

Coyle, T. R., Elpers, K. E., Gonzalez, M. C., Freeman, J., & Baggio, J. A. (2018). General intelligence (g), ACT scores, and theory of mind:(ACT) g predicts limited variance among Theory of Mind tests. *Intelligence*, *71*, 85–91.

Cummings, L. (2013). Clinical pragmatics and Theory of Mind. In *Perspectives on Linguistic Pragmatics* (pp. 23–56). Springer.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, *51*(1), 1–40.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, *44*(1), 113.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128–133.

Degen, J. (2023). The Rational Speech Act Framework. *Annual Review of Linguistics*, *9*(1), 519–540.

Degen, J., Franke, M., & Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 35.

Degen, J., Franke, M. et al. (2012). Optimal reasoning about referential expressions. *Proceedings of SemDIAL*, (pp. 2–11).

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*(1), 172–201.

Dekker, P., & Van Rooy, R. (2000). Bi-directional optimality theory: An application of game theory. *Journal of Semantics*, *17*(3), 217–242.

Del Sette, P., Bambini, V., Bischetti, L., & Lecce, S. (2020). Longitudinal associations between Theory of Mind and metaphor understanding during middle childhood. *Cognitive Development*, *56*, 100958.

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*, 103569.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367.

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: A new task for assessing Theory of Mind in adults. *PloS one*, *8*(11), e81279.

D'onofrio, A. (2018). Personae and phonetic detail in sociolinguistic signs. *Language in Society*, *47*(4), 513–539.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one*, *18*(3), e0279720.

Duff, J., Mayn, A., & Demberg, V. (2025). An ACT-R model of resource-rational performance in a pragmatic reference game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 47.

Ecker, U. K., Lewandowsky, S., Oberauer, K., & Chee, A. E. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 170.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476.

Eckert, P. (2019). The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, *95*(4), 751–776.

Elder, C.-H., & Haugh, M. (2023). Exposing and avoiding unwanted inferences in conversational interaction. *Journal of Pragmatics*, *218*, 115–132.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327.

Evans, J. S. B. (2016). Reasoning, biases and dual processes: The lasting impact of Wason (1960). *Quarterly Journal of Experimental Psychology*, *69*(10), 2076–2092.

Evans, J. S. B. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Fairchild, S., & Papafragou, A. (2018). Sins of omission are more likely to be forgiven in non-native speakers. *Cognition*, *181*, 80–92.

Fairchild, S., & Papafragou, A. (2021). The role of executive function and Theory of Mind in pragmatic computations. *Cognitive Science*, *45*(2), e12938.

Feng, W., Wu, Y., Jan, C., Yu, H., Jiang, X., & Zhou, X. (2017). Effects of contextual relevance on pragmatic inference during conversation: An fMRI study. *Brain and Language*, *171*, 52–61.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Ferreira, F., Engelhardt, P. E., & Jones, M. W. (2009). Good enough language processing: A satisficing approach. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 413–418).

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Fischer, K. (2005). Discourse conditions for spatial perspective taking. In *Proceedings of WoSLaD Workshop on Spatial Language and Dialogue*.

Fischer, K. (2006). The role of users' concepts of the robot in human-robot spatial instruction. In *International Conference on Spatial Cognition* (pp. 76–89). Springer.

Fischer, K., Foth, K., Rohlfing, K., & Wrede, B. (2011). Is talking to a simulated robot like talking to a child? In *2011 IEEE International Conference on Development and Learning (ICDL)* (pp. 1–6). IEEE volume 2.

Fox, C. R., & Levav, J. (2004a). Partition-edit-count: Naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, *133*(4), 626.

Fox, C. R., & Levav, J. (2004b). Partition-Edit-Count: Naive extentional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, *133*(4), 626–642. doi:10.1037/0096-3445.133.4.626.

Frank, M. C., Gómez Emilsson, A., Peloquin, B., Goodman, N. D., & Potts, C. (2016). Rational speech act models of pragmatic reasoning in reference games. PsyArXiv preprint. doi:10.31234/osf.io/f9y6b_v1.

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

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. University of Amsterdam.

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual- vs. population-level probabilistic modeling. *PloS one*, *11*(5), e0154854.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*(2), 201.

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, *25*(3), 203–219.

Gardner, B., Dix, S., Lawrence, R., Morgan, C., Sullivan, A., & Kurumada, C. (2021). Online pragmatic interpretations of scalar adjectives are affected by perceived speaker reliability. *Plos one*, *16*(2), e0245130.

Gibbs Jr, R. W., & Colston, H. L. (2012). *Interpreting figurative meaning*. Cambridge University Press.

Gibbs Jr, R. W., Kushner, J. M., & Mills III, W. R. (1991). Authorial intentions and metaphor comprehension. *Journal of Psycholinguistic Research*, *20*(1), 11–30.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT Press.

Gigerenzer, G., & Selten, R. (2008). Bounded and rational. *Philosophie: Grundlagen und Anwendungen*, (pp. 203–228).

Gigerenzer, G., Todd, P. M., ABC Research Group, t. et al. (2000). *Simple heuristics that make us smart*. Oxford University Press.

Gignac, G. E., Reynolds, M. R., & Kovacs, K. (2019). Digit span subscale scores may be insufficiently reliable for clinical interpretation: Distinguishing between stratified coefficient alpha and omega hierarchical. *Assessment*, *26*(8), 1554–1563.

Gignac, G. E., & Weiss, L. G. (2015). Digit Span is (mostly) related linearly to general intelligence: Every extra bit of span counts. *Psychological Assessment*, *27*(4), 1312.

Giordano, M., Licea-Haquet, G., Navarrete, E., Valles-Capetillo, E., Lizcano-Cortés, F., Carrillo-Peña, A., & Zamora-Ursulo, A. (2019). Comparison between the Short Story Task and the Reading the Mind in the Eyes Test for evaluating Theory of Mind: A replication report. *Cogent Psychology*, *6*(1), 1634326.

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. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Gould, S. J., Cox, A. L., & Brumby, D. P. (2016). Diminished control in crowdsourcing: An investigation of crowdworker multitasking behavior. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *23*(3), 1–29.

Graf, C., Degen, J., Hawkins, R. X., & Goodman, N. D. (2016). Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 38.

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, *13*(7), 692.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Grodner, D., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. In E. Gibson, & N. J. Pearlmutter (Eds.), *The Processing and Acquisition of Reference*. MIT Press.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Happé, F. G. (1994). An advanced test of Theory of Mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154.

Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive Science*, *45*(3), e12926.

Hawkins, R. X., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 39.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104–120.

Heyman, T., & Schaeken, W. (2015). Some differences in Some: Examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, *55*(1), 1.

Higgins, W. C., Kaplan, D. M., Deschrijver, E., & Ross, R. M. (2024). Construct validity evidence reporting practices for the Reading the Mind in the Eyes Test: A systematic scoping review. *Clinical Psychology Review*, *108*, 102378.

Hopkins, B., & Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany*, *18*(2), 213–227.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis University of California, Los Angeles.

Horn, L. R. (1984). Towards a new taxonomy for pragmatic inference: Q-and R-based implicature. In D. Schiffrin (Ed.), *Meaning, Form, and Use in Context: Linguistic Applications* (pp. 11–42). Georgetown University Press.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117.

Hu, M. (2022). When native speakers meet non-native speakers: A case study of foreigner talk. *Journal of Language Teaching & Research*, *13*(4).

Ip, M. H. K., & Papafragou, A. (2021). Listeners evaluate native and non-native speakers differently (but not in the way you think). In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 43.

Ip, M. H. K., & Papafragou, A. (2023). The pragmatics of foreign accents: The social costs and benefits of being a non-native speaker. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(9), 1505.

Jäger, G. (2012). Game theory in semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, *3*, 2487–2516.

Jarrold, C., & Towse, J. N. (2006). Individual differences in working memory. *Neuroscience*, *139*(1), 39–50.

Jarvers, I., Sommer, M., Ullmann, M., Simmel, V., Blaas, L., Gorski, S., Krüger-Lassen, S., Vogel, M., & Langguth, B. (2025). Reading between the lines: Exploring the discriminative ability of the Short-Story Task in identifying autistic individuals within autism outpatient services. *Frontiers in Psychiatry*, *16*, 1500396.

Jensen, A. R., & Figueroa, R. A. (1975). Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology*, *67*(6), 882.

Jeong, S. (2021). Deriving politeness from an extended Lewisian model: The case of rising declaratives. *Journal of Pragmatics*, *177*, 183–207.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.

Kao, J., Bergen, L., & Goodman, N. (2014a). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 36.

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 37.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014b). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Katsos, Napoleon and Kissine, Mikhail (2025). No one-to-one mapping between typologies of pragmatic relations and models of pragmatic processing: A case study with mentalizing. *Philosophical Transactions B*, *380*(1932), 20230501.

Kessels, R. P., van Den Berg, E., Ruis, C., & Brands, A. M. (2008). The backward span of the Corsi Block-Tapping Task and its association with the WAIS-III Digit Span. *Assessment*, *15*(4), 426–434.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., & Henly, A. S. (2002). Speakers' overestimation of their effectiveness. *Psychological Science*, *13*(3), 207–212.

Khorsheed, A., Price, J., & van Tiel, B. (2022). Sources of cognitive cost in scalar implicature processing: A review. *Frontiers in Communication*, *7*, 990044.

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

Kiseleva, J., Li, Z., Aliannejadi, M., Mohanty, S., ter Hoeve, M., Burtsev, M., Skrynnik, A., Zholus, A., Panov, A., Srinet, K. et al. (2022). Interactive grounded language understanding in a collaborative environment: IGLU 2021. In *NeurIPS 2021 Competitions and Demonstrations Track* (pp. 146–161). PMLR.

Kiseleva, J., Skrynnik, A., Zholus, A., Mohanty, S., Arabzadeh, N., Côté, M.-A., Aliannejadi, M., Teruel, M., Li, Z., Burtsev, M. et al. (2023). Interactive grounded language understanding in a collaborative environment: Retrospective on IGLU 2022 competition. In *NeurIPS 2022 Competition Track* (pp. 204–216). PMLR.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Krause, M., & Kizilcec, R. (2015). To play or not to play: Interactions between response quality and task complexity in games and paid crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (pp. 102–109). volume 3.

Krauss, R. M. (1987). The role of the listener: Addressee influences on message formulation. *Journal of Language and Social Psychology*, *6*(2), 81–98.

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, *9*(1), 2–24.

Kravtchenko, E. (2022). *Integrating pragmatic reasoning in an efficiency-based theory of utterance choice*. Ph.D. thesis Saarland University.

Kravtchenko, E., & Demberg, V. (2022). Informationally redundant utterances elicit pragmatic inferences. *Cognition*, *225*, 105159.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Kyriacou, M., & Köder, F. (2024). The cognitive underpinnings of irony comprehension: Fluid intelligence but not working memory modulates processing. *Applied Psycholinguistics*, *45*(6), 1219–1250.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836.

Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing – evidence from event-related brain potentials. *Neuroscience Letters*, *339*(3), 191–194.

Leech, G., & Thomas, J. (2002). Language, meaning and context: Pragmatics. In *An encyclopedia of language* (pp. 94–113). Routledge.

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

Lewis, D. (2008). *Convention: A philosophical study*. John Wiley & Sons.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using Theory of Mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Long, E. L., Catmur, C., & Bird, G. (2025). The Theory of Mind Hypothesis of Autism: A Critical Evaluation of the Status Quo. *Psychological Review*, . doi:10.1037/rev0000532. Published online January 9.

Lopes, L. L., & Oden, G. C. (1991). The rationality of intelligence. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *21*, 225–249.

Love, P. E., Ika, L. A., & Pinto, J. K. (2023). Fast-and-frugal heuristics for decision-making in uncertain and complex settings in construction. *Developments in the Built Environment*, *14*, 100129.

Loy, J., & Demberg, V. (2024). Adaptation to Speakers is modulated by working memory updating and Theory of Mind – a study investigating humor comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 46.

Loy, J. E., Bloomfield, S. J., & Smith, K. (2020). Effects of priming and audience design on the explicitness of referring expressions: Evidence from a confederate priming paradigm. *Discourse Processes*, *57*(9), 808–821.

Loy, J. E., & Demberg, V. (2023). Perspective taking reflects beliefs about partner sophistication: Modern computer partners versus basic computer and human partners. *Cognitive Science*, *47*(12), e13385.

Lüthi, D. (2006). How implicatures make Grice an unordinary ordinary language philosopher. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, *16*(2-3), 247–274.

Madureira, B., & Schlangen, D. (2023). Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the CoDraw dataset. arXiv preprint. doi:https://doi.org/10.48550/arXiv.2302.14406.

Marocchini, E., & Domaneschi, F. (2022). "Can you read my mind?" Conventionalized indirect requests and Theory of Mind abilities. *Journal of Pragmatics*, *193*, 201–221.

Martin, A. K., Perceval, G., Davies, I., Su, P., Huang, J., & Meinzer, M. (2019). Visual perspective taking in young and older adults. *Journal of Experimental Psychology: General*, *148*(11), 2006.

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163.

Marty, P. P., & Chemla, E. (2013). Scalar implicatures: Working memory and a comparison with only. *Frontiers in Psychology*, *4*, 403.

Matsui, T. (2000). *Bridging and Relevance*. Amsterdam: John Benjamins.

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.

Mayn, A., & Demberg, V. (2022). Individual differences in a pragmatic reference game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 44.

Mayn, A., & Demberg, V. (2023). High performance on a pragmatic task may not be the result of successful reasoning: On the importance of eliciting participants' reasoning strategies. *Open Mind*, *7*, 156–178.

Mayn, A., & Demberg, V. (2026). Sources of individual variability in a pragmatic reference game: Effects of reasoning and Theory of Mind. *PloS one*, . To appear.

Mayn, A., Loy, J., & Demberg, V. (2024). Capable but not cooperative? Perceptions of ChatGPT as a pragmatic speaker. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 46.

Mele, A. R., & Rawling, P. (2004). *The Oxford Handbook of Rationality*. Oxford University Press.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Miller, S. A., White, N., & Delgado, M. (1980). Adults' conceptions of children's cognitive abilities. *Merrill-Palmer Quarterly of Behavior and Development*, *26*(2), 135–151.

Milton, D. E. (2012). On the ontological status of autism: The 'double empathy problem'. *Disability & Society*, *27*(6), 883–887.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Mohanty, S., Arabzadeh, N., Teruel, M., Sun, Y., Zholus, A., Skrynnik, A., Burtsev, M., Srinet, K., Panov, A., Szlam, A. et al. (2022). Collecting interactive multi-modal datasets for grounded language understanding. arXiv preprint. doi:https://doi.org/10.48550/arXiv.2211.06552.

Mykhalievskyi, T. (2025). *Understanding pragmatic reasoning through eye movement patterns in reference games*. Bachelor thesis Saarland University.

Narayan-Chen, A., Jayannavar, P., & Hockenmaier, J. (2019). Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5405–5415).

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78).

Neale, S. (1992). Paul Grice and the philosophy of language. *Linguistics and Philosophy*, (pp. 509–559).

Ngo, T. T. A. (2023). The perception by university students of the use of ChatGPT in education. *International Journal of Emerging Technologies in Learning (Online)*, *18*(17), 4.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology*, *7*, 280.

Nilsen, E. S., Glenwright, M., & Huyder, V. (2011). Children and adults understand that verbal irony interpretation depends on listener knowledge. *Journal of Cognition and Development*, *12*(3), 374–409.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Noveck, I. (2018). *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press.

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

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

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

Oakley, B. F., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology*, *125*(6), 818.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*(1), 69–84.

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the Reading the Mind in the Eyes Test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*, 1503.

Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, *90*, 101614.

Panayirci, E., & Dubes, R. C. (1983). A test for multidimensional clustering tendency. *Pattern Recognition*, *16*(4), 433–444.

Panesi, S., Bandettini, A., Traverso, L., & Morra, S. (2022). On the relation between the development of working memory updating and working memory capacity in preschoolers. *Journal of Intelligence*, *10*(1), 5.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

de Paula, J. J., Malloy-Diniz, L. F., & Romano-Silva, M. A. (2016). Reliability of working memory assessment in neurocognitive disorders: A study of the Digit Span and Corsi Block-Tapping tasks. *Revista Brasileira de Psiquiatria*, *38*(3), 262–263.

Peña, P. R., Doyle, P., Edwards, J., Garaialde, D., Rough, D., Bleakley, A., Clark, L., Henriquez, A. T., Branigan, H., Gessinger, I. et al. (2023). Audience design and egocentrism in reference production during human-computer dialogue. *International Journal of Human-Computer Studies*, *176*, 103058.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, *48*(1), 341–348.

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, *26*.

Prasertsom, P., Smith, K., & Culbertson, J. (2025). Testing counterintuitive predictions about cost-based inferences in learning from the Rational Speech Act model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 47.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469.

Puhacheuskaya, V., & Järvikivi, J. (2022). I was being sarcastic!: The effect of foreign accent and political ideology on irony (mis)understanding. *Acta Psychologica*, *222*, 103479.

Puhacheuskaya, V., & Järvikivi, J. (2025). Sorry, you make less sense to me: The effect of non-native speaker status on metaphor processing. *Acta Psychologica*, *254*, 104853.

Qing, C., & Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In *Bayesian Natural Language Semantics and Pragmatics* (pp. 201–220). Springer.

Qing, C., Goodman, N. D., & Lassiter, D. (2016). A rational speech-act model of projective content. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 38.

Quesque, F., Apperly, I., Baillargeon, R., Baron-Cohen, S., Becchio, C., Bekkering, H., Bernstein, D., Bertoux, M., Bird, G., Bukowski, H. et al. (2024). Defining key concepts for mental state attribution. *Communications Psychology*, *2*(1), 29.

Quesque, F., & Rossetti, Y. (2020). What do Theory-of-Mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), 384–396.

Raven, J. C., & Court, J. (1938). *Raven's progressive matrices*. Western Psychological Services Los Angeles, CA.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *20*(6), 1102–1113.

Regel, S., Coulson, S., & Gunter, T. C. (2010). The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain Research*, *1311*, 121–135.

Roettger, T. B., & Franke, M. (2019). Evidential strength of intonational cues and rational adaptation to (un-)reliable intonation. *Cognitive Science*, *43*(7), e12745.

Roettger, T. B., & Rimland, K. (2020). Listeners' adaptation to unreliable intonation is speaker-sensitive. *Cognition*, *204*, 104372.

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information integration in modulation of pragmatic inferences during online language comprehension. *Cognitive Science*, *43*(8), e12769.

Ryzhova, M., & Demberg, V. (2020). Processing particularized pragmatic inferences under load. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 42.

Ryzhova, M., Mayn, A., & Demberg, V. (2023). What inferences do people actually make upon encountering informationally redundant utterances? An individual differences study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 45.

Samuelson, L. (2016). Game theory in economics and beyond. *Journal of Economic Perspectives*, *30*(4), 107–130.

Schlangen, D. (2004). Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004* (pp. 136–143).

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1089.

Scholman, M., Marchal, M., & Demberg, V. (2024). Connective comprehension in adults: The influence of lexical transparency, frequency, and individual differences. *Discourse Processes*, *61*(8), 381–403.

Scholman, M. C., Demberg, V., & Sanders, T. J. (2020). Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse. *Discourse Processes*, *57*(10), 844–861.

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285.

Schuster, S., Mayn, A., & Demberg, V. (2023). Working memory updating modulates adaptation to speaker-specific use of uncertainty expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 45.

Scontras, G., Tessler, M. H., & Franke, M. (2021). A practical introduction to the rational speech act modeling framework. arXiv preprint. doi:https://doi.org/10.48550/arXiv.2105.09867.

Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Seyednozadi, Z., Pishghadam, R., & Pishghadam, M. (2021). Functional role of the N400 and P600 in language-related ERP studies with respect to semantic anomalies: an overview. *Archives of Neuropsychiatry*, *58*(3), 249.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science*, *11*(6), 771–799.

Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey. *IEEE Access*, *11*.

Sikos, L., Venhuizen, N. J., Drenhaus, H., & Crocker, M. W. (2021a). Reevaluating pragmatic reasoning in language games. *PloS one*, *16*(3), e0248388.

Sikos, L., Venhuizen, N. J., Drenhaus, H., & Crocker, M. W. (2021b). Speak before you listen: Pragmatic reasoning in multi-trial language games. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 43.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, (pp. 99–118).

Simon, H. A. (1990). Bounded rationality. In *Utility and Probability* (pp. 15–18). Springer.

Singh, H., Tayarani-Najaran, M.-H., & Yaqoob, M. (2023). Exploring computer science students' perception of ChatGPT in higher education: A descriptive and correlation study. *Education Sciences*, *13*(9), 924.

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two-and four-option multiple choice question version of the Cognitive Reflection Test. *Behavior Research Methods*, *50*(6), 2511–2522.

Skalicky, S. (2019). Investigating satirical discourse processing and comprehension: The role of cognitive, demographic, and pragmatic features. *Language and Cognition*, *11*(3), 499–525.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* volume 142. Harvard University Press.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Psychology Press.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. *The Oxford Handbook of Thinking and Reasoning*, (pp. 343–365).

Stanovich, K. E. (2020). Why humans are cognitive misers and what it means for the Great Rationality Debate. In *Routledge Handbook of Bounded Rationality* (pp. 196–206). Routledge.

Stanovich, K. E., & Toplak, M. E. (2023). Actively Open-Minded Thinking and its measurement. *Journal of Intelligence*, *11*(2), 27.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of educational psychology*, *89*(2), 342.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161.

Starns, J. J., Cohen, A. L., Bosco, C., & Hirst, J. (2019). A visualization technique for Bayesian reasoning. *Applied Cognitive Psychology*, *33*, 234–251. doi:10.1002/acp.3470.

Stein, E. (1996). *Without good reason: The rationality debate in philosophy and cognitive science*. Clarendon Press.

Stengård, E., Juslin, P., Hahn, U., & Van den Berg, R. (2022). On the generality and cognitive basis of base-rate neglect. *Cognition*, *226*, 105160.

Stieger, S., & Reips, U.-D. (2016). A limitation of the Cognitive Reflection Test: familiarity. *PeerJ*, *4*, e2395.

Taguchi, N. (2008). The effect of working memory, semantic access, and listening abilities on the comprehension of conversational implicatures in L2 English. *Pragmatics & Cognition*, *16*(3), 517–539.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113.

van Tiel, B., Franke, M., & Sauerland, U. (2022). Modelling language use in autism: A case study on words of estimative probability. In F. Frau, L. Bischetti, C. Pompei, B. Scalingi, F. Domaneschi, & V. Bambini (Eds.), *Book of Abstracts - XPRAG 2022*. OSF.io. doi:10.17605/OSF.IO/C4KP2.

van Tiel, B., Marty, P., Pankratz, E., & Sun, C. (2019). Scalar inferences and cognitive load. In *Proceedings of Sinn und Bedeutung* (pp. 427–442). volume 23.

van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175.

Tonini, E., Bischetti, L., Del Sette, P., Tosi, E., Lecce, S., & Bambini, V. (2023). The relationship between metaphor skills and Theory of Mind in middle childhood: Task and developmental effects. *Cognition*, *238*, 105504.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168.

Trott, S., & Bergen, B. (2019). Individual differences in mentalizing capacity predict indirect request comprehension. *Discourse Processes*, *56*(8), 675–707.

Turcan, A., & Filik, R. (2016). An eye-tracking investigation of written sarcasm comprehension: The roles of familiarity and context. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(12), 1867.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1993). Probabilistic reasoning. *Readings in Philosophy and Cognitive Science*, (pp. 43–68).

Ulitzsch, E., Henninger, M., & Meiser, T. (2024). Differences in response-scale usage are ubiquitous in cross-country comparisons and a potential driver of elusive relationships. *Scientific Reports*, *14*(1), 10890.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*(6), 1038.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505.

Valles-Capetillo, E., Ibarra, C., Martinez, D., & Giordano, M. (2022). A novel task to evaluate irony comprehension and its essential elements in Spanish speakers. *Frontiers in Psychology*, *13*, 963666.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Verhoeven, L., & Vermeer, A. (2002). Communicative competence and personality dimensions in first and second language learners. *Applied Psycholinguistics*, *23*(3), 361–374.

Von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of Games and Economic Behavior*. Princeton University Press.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 550–564.

Weber, R., & Dawes, R. (2005). Behavioral economics. *The Handbook of Economic Sociology*, *2*, 90–108.

Welsh, M. B. (2022). What is the CRT? Intelligence, personality, decision style or attention? In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 44.

White, J., Mu, J., & Goodman, N. D. (2020). Learning to refer informatively by amortizing pragmatic reasoning. arXiv preprint. doi:https://doi.org/10.48550/arXiv.2006.00418.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194.

Williams, C. C., Kappen, M., Hassall, C. D., Wright, B., & Krigolson, O. E. (2019). Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *NeuroImage*, *189*, 574–580.

Wittgenstein, L. (2009). *Philosophical investigations*. John Wiley & Sons.

Wu, H., & Cai, Z. G. (2024). Speaker effects in spoken language comprehension. arXiv preprint. doi:https://doi.org/10.48550/arXiv.2412.07238.

Wu, H., & Cai, Z. G. (2025). When a man says he is pregnant: Event-related potential evidence for a rational account of speaker-contextualized language comprehension. *Journal of Cognitive Neuroscience*, (pp. 1–16).

Wu, H., Duan, X., & Cai, Z. G. (2024). Speaker demographics modulate listeners' neural correlates of spoken word processing. *Journal of Cognitive Neuroscience*, *36*(10), 2208–2226.

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology*, *9*, 1720.

Yeung, E. K. L., Apperly, I. A., & Devine, R. T. (2024). Measures of individual differences in adult Theory of Mind: A systematic review. *Neuroscience & Biobehavioral Reviews*, *157*, 105481.

Yntema, D. B. (1963). Keeping track of several things at once. *Human Factors*, *5*(1), 7–17.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 38.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87.

Younis, H. A., Eisa, T. A. E., Nasser, M., Sahib, T. M., Noor, A. A., Alyasiri, O. M., Salisu, S., Hayder, I. M., & Younis, H. A. (2024). A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: Applications, considerations, limitations, motivation and challenges. *Diagnostics*, *14*(1), 109.

Yuan, A., Monroe, W., Bai, Y., & Kushman, N. (2018). Understanding the rational speech act model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 40.

Zhou, I., Hu, J., Levy, R., & Zaslavsky, N. (2022). Teasing apart models of pragmatics using optimal reference game design. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 44.