Factors that affect derivation of atypicality inferences in humans

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

vorgelegt von
Margarita Ryzhova
aus Tver, Russland

Saarbrücken, 2026

Dekanin der Fakultät P: Prof. Dr. Nine Miedema
Erstberichterstatterin: Prof. Dr. Vera Demberg
Zweitberichterstatterin: Prof. Dr. Stefania Degaetano-Ortlieb
Tag der letzten Prüfungsleistung: 26 May 2026

Abstract

This dissertation investigates how comprehenders derive **atypicality inferences**, a type of pragmatic inference that arises when speakers communicate information that appears informationally redundant. For example, in a narrative about visiting a restaurant, stating that someone ate there is redundant, as eating is typically part of the restaurant script. Rather than treating this literally, listeners may interpret the explicit mention as pragmatically meaningful and infer that the event is somehow atypical, for instance that the person does not usually eat when going to restaurants.

The main goal of this work is to understand **which properties of the comprehender modulate the derivation of atypicality inferences**. Building on previous work by Kravtchenko (2022), I propose a formal derivation scheme in which listeners (i) detect informational redundancy, (ii) interpret it as pragmatically marked, (iii) revise assumptions about event typicality, and (iv) construct a context in which redundant information becomes meaningful.

Empirically, I first examine how comprehenders engage in the final step of contextual accommodation. The results show that participants indeed generate a wide range of explanations for atypical behavior, but also reveal substantial variability: while some consistently derive atypicality inferences and provide coherent explanations, others either struggle to construct plausible explanations or are not sensitive to redundancy at all. This variability suggests that inference derivation is not uniform but depends on comprehenders.

To account for this variability, I adopt two complementary approaches. The **first approach** tests whether atypicality inferences depend on cognitive resources by imposing attentional and working memory load in a series of dual-task experiments. Contrary to predictions from major accounts of pragmatic processing (Levinson, 2000; Sperber & Wilson, 1996; Degen & Tanenhaus, 2019), the results show no reliable effect of cognitive load: participants derive atypicality inferences to a similar extent under both high and low load conditions, despite successful load manipulations.

The **second approach** examines whether inference derivation is related to individual differences in cognitive and personality-related traits. Across a range of measures, no reliable effects are found for executive functions, socio-pragmatic traits, or linguistic experience. In contrast, reasoning-related abilities show some evidence of an effect, with higher reasoning ability associated with an increased likelihood of deriving atypicality inferences.

Taken together, the findings suggest that variability in atypicality inference derivation is linked to higher-level reasoning abilities rather than to general cognitive resources or affective theory of mind abilities. Atypicality inferences appear to rely on the ability to construct and evaluate contextually appropriate explanations, rather than on mentalizing abilities linked to emotion recognition.

Zusammenfassung

Wenn wir im Alltag kommunizieren, sagen wir oft mehr als unbedingt nötig ist. So wiederholen wir beispielsweise Anweisungen gegenüber Kindern ("*Setz deine Mütze auf, draußen ist es kalt. Achte darauf, dass du deine Mütze trägst!*"), geben Wegbeschreibungen mehrfach, wenn uns jemand nicht gut hören kann, oder erklären Dinge, die offensichtlich erscheinen mögen, etwa beim Unterrichten. In solchen Fällen ist es kein Problem, mehr zu sagen als nötig – tatsächlich erleichtert es oft die Kommunikation und macht sie klarer oder effektiver.

Es gibt jedoch auch Situationen, in denen Sprecher Informationen hinzufügen, die zumindest auf den ersten Blick keinem offensichtlichen Zweck zu dienen scheinen. Manchmal erwähnen Menschen etwas, das bereits erwartet wird oder sich leicht aus der Situation erschließen lässt. Solche Erwartungen beruhen auf unserem allgemeinen Wissen darüber, wie sich vertraute Situationen typischerweise entwickeln. So sind wir alle damit vertraut, was es bedeutet, in ein Restaurant zu gehen: Normalerweise isst man dort. Wenn jemand sagt "*Ich bin in ein Restaurant gegangen und habe dort gegessen!*", kann dies unnötig redundant erscheinen, da das Essen in diesem Kontext bereits erwartet wird. Aus Gesprächstheoretischer Perspektive ist dies überraschend, da von Sprechern im Allgemeinen erwartet wird, informativ zu sein, jedoch nicht informativer als erforderlich (Grice, 1975).

Wenn Zuhörer auf eine solche Redundanz stoßen, fragen sie sich möglicherweise, warum der Sprecher sie überhaupt erwähnt hat. Anstatt sie zu ignorieren, versuchen sie vielleicht, ihr einen Sinn zu geben. Dieser Prozess wird als pragmatische Inferenz bezeichnet, d.h. man geht über das explizit Gesagte hinaus und nutzt Kontext sowie Hintergrundwissen, um zu einer angereicherten Interpretation zu gelangen. Eine Möglichkeit besteht darin, anzunehmen, dass etwas an der Situation nicht ganz typisch ist. So kann der Zuhörer beispielsweise schlussfolgern, dass, obwohl das Essen in einem Restaurant normalerweise erwartet wird, es in diesem Fall irgendwie erwähnenswert war. Wenn eine Äußerung etwas explizit macht, das üblicherweise implizit bleibt, kann dies eine Abweichung vom erwarteten Ablauf signalisieren. Infolgedessen können Zuhörer ihre Vorstellungen darüber revidieren, was in der beschriebenen Situation typischerweise geschieht – sie könnten etwa zu dem Schluss kommen, dass die betreffende Person normalerweise nicht in Restaurants isst, entgegen der üblichen Erwartung. Diese Art pragmatischer Inferenz wird als **Atypikalitätsinferenz (atypicality inferences)** bezeichnet.

Die vorliegende Dissertation untersucht, wie Atypikalitätsinferenzen abgeleitet werden und welche Faktoren beeinflussen, ob Zuhörer sie ziehen. Genauer gesagt besteht das Hauptziel darin zu verstehen, **welche Eigenschaften der Zuhörer die Ableitung von Atypikalitätsinferenzen modulieren**. Obwohl Atypikalitätsinferenzen aus einem scheinbar einfachen Phänomen entstehen – mehr zu sagen als üblicherweise notwendig –, beruht ihre Ableitung auf einem Zusammenspiel von Hintergrundwissen, Erwartungen an Kommunikation und kognitiven

Fähigkeiten der Zuhörer.

Zunächst beschreibe ich in Kapitel 2 frühere Arbeiten von Kravtchenko (2022) zu Atypikalitätsinferenzen und skizziere ihre zentralen Eigenschaften. Anschließend schlage ich in Kapitel 3 ein formales Ableitungsschema für Atypikalitätsinferenzen vor, das als konzeptionelle Grundlage für die empirischen Untersuchungen dient. Dieser Ableitungsprozesslässt sich als Abfolge von Schritten verstehen: Zuhörer erkennen, dass eine Äußerung informationsbezogen redundant ist, interpretieren diese Redundanz als pragmatisch markiert, revidieren ihre Annahmen über die Typikalität des Ereignisses und versuchen schließlich, eine Erklärung zu konstruieren, die die abgeleitete Atypikalität im Kontext eine Bedeutung verleiht. Der letzte Schritt beinhaltet die Suche nach einem Kontext, in dem das explizit erwähnte Ereignis nicht als selbstverständlich angesehen wird, sondern stattdessen etwas über die Situation oder die beteiligte Person aussagt.

In Kapitel 7 teste ich empirisch den letzten Schritt der Kontextakkommodation, indem ich die Teilnehmenden bitte, ihre Typikalitätsurteile zu erklären (d.h. die Frage zu beantworten, wie oft die in der Geschichte beschriebene Person ihrer Meinung nach üblicherweise in Restaurants isst). Ich zeige, dass die Teilnehmenden eine Vielzahl von Erklärungen für atypisches Verhalten entwickeln (z.B. dass die Person normalerweise nicht in Restaurants isst, weil sie dort nur Getränke bestellt, nicht genug Geld hat oder aus anderen Gründen dorthin geht). Entscheidend ist, dass die Teilnehmenden informationsbezogen redundante Äußerungen nicht alle auf die gleiche Weise verarbeiten. Während einige konsistent Atypikalitätsinferenzen ziehen und kohärente Erklärungen liefern, bemerken andere zwar die Redundanz, haben jedoch Schwierigkeiten, eine plausible Erklärung für atypisches Verhalten zu konstruieren, und wieder andere reagieren überhaupt nicht auf Informationsredundanz. Diese Variabilität legt nahe, dass die Ableitung von Atypikalitätsinferenzen nicht einheitlich über Zuhörer hinweg erfolgt, sondern durch deren kognitive und persönlichkeitsbezogene Eigenschaften moduliert sein kann.

Während frühere Arbeiten zu anderen pragmatischen Inferenzen diese Variabilität mit kognitiven und sozio-pragmatischen Faktoren in Verbindung gebracht haben, ist die Rolle dieser Faktoren bei Atypikalitätsinferenzen bislang weitgehend unerforscht. Um dies zu untersuchen, betrachte ich Atypikalitätsinferenzen aus zwei komplementären Perspektiven: **(1) durch die Untersuchung der Effekte extern auferlegter kognitiver Einschränkungen** und **(2) durch die Verknüpfung der Inferenzableitung mit natürlicher Variabilität in kognitiven und persönlichkeitsbezogenen Eigenschaften von Zuhörer**.

Der **erste Ansatz** untersucht, ob die Ableitung von Atypikalitätsinferenzen von der Verfügbarkeit kognitiver Ressourcen abhängt, indem er der Aufmerksamkeit und dem Arbeitsgedächtnis externe Einschränkungen auferlegt. In Kapitel 4 bespreche ich zentrale theoretische Ansätze der pragmatischen Verarbeitung (Grice, 1975; Levinson, 2000; Sperber & Wilson, 1996; Degen & Tanenhaus, 2019) und leite daraus deren Vorhersagen für Atypikalitätsinferenzen ab. Diese Ansätze stimmen darin überein,

dass Atypikalitätsinferenzen kognitiv anspruchsvoll sein sollten, da sie stark kontextabhängig sind und erfordern, dass Zuhörer kontextabhängige Alternativen konstruieren.

Um diese Vorhersagen zu testen, führe ich in Kapitel 9 eine Serie von drei Dual-Task-Experimenten durch, in denen kognitive Ressourcen manipuliert werden, indem entweder aufmerksamkeitsbezogene oder verbale Arbeitsgedächtnisbelastung erzeugt wird (Fairchild & Papafragou, 2021; Cho, 2020). In einem Experiment verarbeiten die Teilnehmenden auditives Material, während sie gleichzeitig eine visuo-motorische Tracking-Aufgabe ausführen, die die Aufmerksamkeitsressourcen beansprucht. In den anderen Experimenten lesen die Teilnehmenden die Texte und bearbeiten eine Sekundäraufgabe, bei der sie verbale Informationen im Gedächtnis behaltenund abrufen müssen, wodurch das Arbeitsgedächtnis belastet wird. In allen Experimenten wird die Stärke der Atypikalitätsinferenzen anhand von Typikalitätsurteilen erfasst. Entgegen den Vorhersagen zeigen die Ergebnisse keinen verlässlichen Effekt kognitiver Belastung auf die Ableitung von Atypikalitätsinferenzen. Die Teilnehmenden leiten diese Inferenzen in ähnlichem Ausmaß unter hoher wie unter niedriger (oder keiner) Belastung ab. Die einzige Ausnahme ist ein kleiner Trade-off-Effekt, der im ersten Experiment beobachtet wurde, bei dem eine erhöhte Belastung zu größeren Tracking-Abweichungen in der Bedingung führt, in der Atypikalitätsinferenzen erwartet werden. Dieser Effekt ist jedoch vermutlich eher auf exklamatorische Prosodie als auf pragmatische Verarbeitung selbst zurückzuführen, da er sich in den Experimenten, in denen die Materialien textuell statt auditiv präsentiert werden, nicht wiederholt. Wichtig ist, dass die Belastungsmanipulationen erfolgreich sind, da die Verständlichkeitsgenauigkeit unter höherer Belastung in allen Experimenten abnimmt. Insgesamt zeigen die Ergebnisse, dass Atypikalitätsinferenzen nicht in der Weise empfindlich auf Beschränkungen des Arbeitsgedächtnisses reagieren, wie es bestehende Theorien vorhersagen.

Der **zweite Ansatz** verfolgt eine Perspektive individueller Unterschiede und untersucht, ob sich die Ableitung von Atypikalitätsinferenzen durch natürliche Variabilität in kognitiven und persönlichkeitsbezogenen Eigenschaften von Zuhörer erklären lässt. In Kapitel 5 gebe ich einen Überblick über bisherige Arbeitenzu individuellen Unterschieden in der pragmatischen Verarbeitung und identifiziere eine Reihe kognitiver und sozio-pragmatischer Faktoren, die verschiedene Phasen des Ableitungsprozesses beeinflussen können (Nieuwland et al., 2010; Antoniou et al., 2016; Yang et al., 2018). Darauf aufbauend untersuche ich in den Kapiteln 8 und 10 die Rolle exekutiver Funktionen (Arbeitsgedächtnis, Inhibition und Gedächtnisaktualisierung), schlussfolgerungsbezogener Fähigkeiten (fluide Intelligenz und kognitive Reflexion), sozio-pragmatischer Fähigkeiten (Theory of Mind[1] und autistische Merkmale) sowie Leseerfahrung als Proxy für sprachliche Erfahrung.

Die Ergebnisse zeigen, dass nicht alle diese Faktoren gleichermaßen dazu beitragen. Insbesondere finden sich keine verlässlichen Effekte für exekutive Funktionen,

[1]Englisch für 'Theorie des Mentalen'

Theory of Mind oder Leseerfahrung, was darauf hindeutet, dass die Ableitung von Atypikalitätsinferenzen nicht primär durch allgemeine Ressourcenbeschränkungen, Mentalisierungsfähigkeiten oder Unterschiede in sprachlicher Erfahrung bestimmt wird. Im Gegensatz dazu finden sich Hinweise auf einen Effekt schlussfolgerungsbezogener Fähigkeiten: Teilnehmende mit höherer Schlussfolgerungsfähigkeit leiten eher Atypikalitätsinferenzen ab. Darüber hinaus sind autistische Merkmale in einem der Experimente mit einer erhöhten Wahrscheinlichkeit verbunden, Atypikalitätsinferenzen abzuleiten. Dieses Ergebnis ist bemerkenswert, da es der Erwartung widerspricht, dass ein höheres Ausmaß anautistischen Merkmalen mit Schwierigkeiten in der pragmatischen Verarbeitung einhergehen sollte. Eine mögliche Erklärung besteht darin, dass das verwendete Maß für autistische Merkmale ein Spektrum von Eigenschaften erfasst, die unterschiedliche Auswirkungen auf die Inferenzableitung haben können, von denen einige die Sensitivität für Mechanismen der Atypikalitätsinferenzen fördern könnten.

Zusammengenommen deuten diese Ergebnisse darauf hin, dass die Variabilität in der Ableitung von Atypikalitätsinferenzen eher mit übergeordneten Schlussfolgerungsfähigkeiten als mit exekutiven Funktionen oder Mentalisierungsfähigkeiten zusammenhängt. Eine Möglichkeit, dieses Muster zu erklären, besteht darin, es mit dem vorgeschlagenen schrittweisen Ableitungsprozess in Verbindung zu bringen. Insbesondere könnten Schlussfolgerungsfähigkeiten eine Rolle in den Phasen spielen, die über die anfängliche Erkennung von Redundanz hinausgehen, nämlich bei der Identifikation ihrer pragmatischen Markiertheit und beim anschließenden Reparaturprozess. Um eine Atypikalitätsinferenz zu ziehen, müssen Zuhörer über die wörtliche Interpretation hinausgehen und einen Kontext berücksichtigen, in dem die redundante Äußerung informativ wird. Dies beinhaltet das Generieren und Bewerten möglicher Erklärungen, das Aufrechterhalten mehrerer Interpretationsmöglichkeiten und die Auswahl einer Interpretation, die die Äußerung mit dem weiteren Diskurs in Einklang bringt. Personen mit höherer Schlussfolgerungsfähigkeit beteiligen sich möglicherweise eher an dieser Art von Suchprozess, was wiederum die Wahrscheinlichkeit erhöht, Atypikalitätsinferenzen abzuleiten und kontextuell zu integrieren. Im Gegensatz dazu stützen sichProzesse wie die Erkennung von Redundanz auf der Grundlage von Skriptwissen möglicherweise auf leicht zugängliches Hintergrundwissen und stellen daher relativ geringe Anforderungen an exekutive Ressourcen. Aus dieser Perspektive scheint die Ableitung von Atypikalitätsinferenzen von der Fähigkeit abzuhängen, kontextuell angemessene Interpretationen zu konstruieren und zu bewerten.

Das Ausbleiben eines Effekts von Theory of Mind bedarf weiterer Betrachtung. Eine mögliche Erklärung ist, dass sich Atypikalitätsinferenzen von anderen pragmatischen Inferenzen unterscheiden, für die Effekte von Mentalisierung berichtet wurden (Fairchild & Papafragou, 2021), da sie zumindest in frühen Verarbeitungsphasen ohne detailliertes Schlussfolgern über die Überzeugungen oder Intentionen des Sprechers abgeleitet werden können (Brown & Dell, 1987; Grigoroglou & Papafragou, 2019;

Lockridge & Brennan, 2002). Insbesondere könnten Zuhörer Informationsredundanz relativ zu skriptbasierten Erwartungen erkennen und als pragmatisch markiert interpretieren, ohne die mentalen Zustände des Sprechers explizit zu repräsentieren. Aus dieser Perspektive beruht die Ableitung von Atypikalitätsinferenzen stärker auf dem Zusammenspiel von Hintergrundwissen und Erwartungen hinsichtlich Informativität als auf Überlegungen zu kommunikativen Absichten. Gleichzeitig bleibt es möglich, dass spätere Phasen des Ableitungsprozesses, wie das Konstruieren und Bewerten von Erklärungen für die erschlossene Atypikalität, in stärkerem Maße Mentalisierung einbeziehen. Eine weitere Möglichkeit betrifft das zur Erfassung von Theory of Mind verwendete Maß, das möglicherweise nicht die Aspekte der Mentalisierung erfasst, die in diesem Kontext am relevantesten sind.

Diese Interpretation steht im Einklang mit aktueller laufender Forschung (Bila, 2026), die dieselben experimentellen Materialien und ein ähnliches Paradigma verwendet, jedoch zwischen verschiedenen Komponenten von Theory of Mind unterscheidet. Während bei Messgrößen, die auf affektive Aspekte der Theory of Mind abzielen, kein Effekt zu beobachten ist, zeigt sich ein Effekt bei Messgrößen, die das Schlussfolgern über die Überzeugungen und Absichten anderer betreffen. Diese Unterscheidung steht im Einklang mitder Auffassung, dass Theory of Mind mindestens zwei Subkomponenten umfasst: affektive Theory of Mind, die sich primär mit dem Schlussfolgern über emotionale Zustände anderer beschäftigt, und kognitive Theory of Mind, die das Schlussfolgern über Überzeugungen und kommunikative Absichten anderer beinhaltet (Shamay-Tsoory et al., 2006; Tager-Flusberg & Sullivan, 2000). Die in der vorliegenden Arbeit verwendete Messgröße zielt in erster Linie auf Ersteres ab, die für die an der Ableitung von Atypikalitätsinferenzen beteiligten Prozesse möglicherweise nicht direkt relevant ist. Im Gegensatz dazu könnte die kognitive Theory of Mind eine wichtigere Rolle bei der Erkennung pragmatischer Markierung oder der Konstruktion passender Erklärungen spielen. Insgesamt legt dies nahe, dass die Rolle von Theory of Mind bei der Ableitung von Atypikalitätsinferenzen davon abhängen könnte, wie sie operationalisiert wird.

Abschließend untersuche ich, wie Atypikalitätsinferenzen mit anderen Arten pragmatischer Implikaturen zusammenhängen. In Kapitel 10 vergleiche ich Atypikalitätsinferenzen mit mehreren Typen skalenbasierter Implikaturen bei denselbenTeilnehmenden, einschließlich skalarer Implikaturen, Numeralia und Disjunktionen (De Neys & Schaeken, 2007; Marty et al., 2013; Singh et al., 2016; Potts et al., 2016). Dieses Design ermöglicht es zu prüfen, ob Personen, die dazu neigen, eine Art von Implikatur abzuleiten, ähnliche Tendenzen auch bei anderen Arten zeigen. Die Ergebnisse zeigen nur schwache Korrelationen zwischen Atypikalitätsinferenzen und Implikaturen, die an lexikalische Skalen gebunden sind. Während einige skalenbasierte Implikaturen zusammen auftreten, weisen Atypikalitätsinferenzen ein eigenständiges Antwortprofil auf. Dies legt nahe, dass sie nicht auf dieselbe Weise abgeleitet werden wie Implikaturen, die auf lexikalischen Skalen basieren.

Zusammen genommen liefern diese Befunde eine Grundlage für das Verständnis der Ableitung und Akkommodation von Atypikalitätsinferenzen, ihrer Variabilität zwischen Rezipierenden, der Faktoren, die sie modulieren, sowie ihres Verhältnisses zu anderen Typen pragmatischer Inferenzen.

Acknowledgments

The journey to completing this PhD has been both challenging and rewarding, and I am deeply grateful to the many people who accompanied me along the way.

First and foremost, I would like to thank Prof. Dr. Vera Demberg for being such a thoughtful supervisor. I greatly appreciate your openness and all the discussions we had – they always helped me think things through. It was always easy to approach you, and I always felt encouraged to ask questions and explore ideas. I am continually impressed by how broad and deep your knowledge is, and I am very grateful for your collaborative supervision style, which gave space to my own ideas and encouraged me to think independently. Thank you for everything, Vera.

I would also like to thank my second reviewer, Prof. Dr. Stefania Degaetano-Ortlieb, for agreeing to review this thesis and for her time and support. Thank you, Stefania.

I am very grateful to my collaborators, Sasha, Jia, and Tony, for everything I learned from you, both scientifically and personally. Working with you was not only enriching, but also genuinely enjoyable. I would also like to thank my colleagues, Ekaterina Kravtchenko, Jack Duff, Emilia Ellsiepen, Merel Scholman, Katja Häuser, Marian Marchal, Guifu Liu, Nataliia Bila, Sebastian Schuster, and Pratik Bhandari, for their helpful feedback and for the many conversations that helped me refine my ideas. Many thanks to Charlotte Kurch, who worked with me as a Master's student, for her sharp mind and remarkable scientific curiosity. I am also grateful to all members of the PLEAD research group for the open exchange of ideas and for the joint input on the research, experimental designs, and analyses. I would like to thank Prof. Dr. Alexander Koller for the clarity and depth of his thinking. It was a pleasure to work with you on the same project.

This work also relied on the support of several people who contributed to the experimental preparations. I would like to thank Ahmed and Mansoor for maintaining the Lingoturk server and for helping me implement the experiments. I truly admire your dedication and attitude towards your work. I also thank Dave Howcroft for recording the audio materials and for providing the voice for the first dual-tasking experiment. I remember how much work it was, but it was a really enjoyable experience to do these recordings together.

I am very thankful to all participants who took part in my studies. Thank you for your time, your feedback, and your honest participation.

I was lucky to be part of a research group that created such a welcoming and warm environment. I truly appreciated our group meetings and the many conversations that helped shape this work, as well as all the fun activities beyond work – hikes, board game nights, Christmas events, and many more. You made this journey a lot more enjoyable. Thank you all. And thank you, Mayank, for our volleyball games. It was such a great summer. Thank you, Emilia and Frances, for our little conversations

along the way.

I would also like to express my sincere gratitude to Gabi and Khanda for their kindness and support. Your administrative help was invaluable, and I truly do not know how I would have managed all the bureaucracy without you. Khanda, it was always a pleasure to talk to you about so many different things. I really enjoyed our small chats whenever we had them.

I am also grateful to the psycholinguistic community and to all the colleagues I had the chance to meet over the years at conferences and workshops. I greatly appreciated the feedback and the openness that I encountered and I feel very fortunate to be part of such a supportive and encouraging community.

The SFB1102 supported this work by funding my research and providing many opportunities throughout my PhD, including workshops to develop both hard and soft skills, as well as the invited talk series, which helped me connect with other researchers. My sincere thanks go to the project coordinators, Marie-Ann, Sabine, and Heike. I always appreciated how readily you helped with all kinds of bureaucratic matters.

Without my friends and family, this PhD would not have been nearly as meaningful or enjoyable. To all my friends – thank you. We shared so much together: travels, long conversations about life, and so many great moments. I am really glad you were part of these years. Thanks to my mom for everything. Оля и Маша, спасибо вам. Не знаю что я бы делала без вас.

Most of all, I want to thank my partner. Thank you for supporting me in so many ways, for all the delicious food you cooked, for proofreading my texts and helping with so many everyday things, and for listening to all my stories about John and why he did or did not pay the cashier. Maria und Werner, vielen Dank für eure Freundlichkeit, eure Ermutigung und euren Glauben an mich. Ich bin so dankbar, euch kennengelernt zu haben. Ihr habt mir das Gefühl gegeben, hier ein Zuhause zu haben.

It was a really good time at Saarland University, and I am grateful to all the people who made it so.

Contents

Chapter 1

Introduction

Human communication relies on more than the literal meanings of words. In many situations, comprehenders must go beyond what is explicitly said in order to recover the speaker's intended meaning. This process is known as pragmatic inference: listeners integrate linguistic information with contextual knowledge, world knowledge, and assumptions about the speaker's communicative intent. Through this process, utterances that appear straightforward on the surface can give rise to interpretations that are not directly encoded in the sentence itself.

A large body of work in experimental pragmatics has investigated how pragmatic inferences are derived and what cognitive mechanisms support them. Much of this research has focused on implicatures based on lexical scales, most prominently scalar implicatures, such as the interpretation of *some* as implying *not all*. Experimental studies examining these phenomena have addressed a range of questions, including the time course of implicature derivation and the involved cognitive resources. This research has also revealed considerable variability in pragmatic interpretation. Listeners differ systematically in their tendency to derive pragmatic interpretations. Many studies have, therefore, investigated whether this variability may be related to differences in cognitive and socio-pragmatic characteristics, but the available evidence remains mixed (e.g., Nieuwland et al., 2010; Heyman & Schaeken, 2015; Antoniou et al., 2016; Yang et al., 2018; Fairchild & Papafragou, 2021). This variability raises one of the central questions for experimental pragmatics: **which factors determine whether a listener derives a pragmatic interpretation in a given context**.

The present dissertation addresses this question in the context of pragmatic inferences triggered by informational redundancy (Kravtchenko, 2022)[1]. Kravtchenko (2022) showed that utterances that appear redundant relative to script knowledge can lead listeners to derive additional interpretations. Importantly, these inferences are not tied to lexical scales or specific linguistic items but instead arise from the interaction between linguistic material and background knowledge about typical event

[1]See also Kravtchenko & Demberg (2015, 2022b,a).

sequences. In particular, when a speaker explicitly mentions an event that is normally expected to occur in a given situation, listeners may interpret this redundancy as communicatively meaningful. One possible outcome of this process is the derivation of an **atypicality inference**, whereby the listener concludes that the event in question is somehow unusual. For example, in a narrative about visiting a restaurant, stating that someone ate there may appear redundant because eating is typically part of the restaurant script. Rather than treating the utterance as merely restating an obvious fact, listeners may interpret the explicit mention as signaling that the event is not fully typical in the given context; for instance, that the person does not usually eat when going to restaurants and instead typically orders only drinks.

While Kravtchenko (2022) established that informational redundancy grounded in script knowledge can give rise to atypicality inferences, the factors that influence their derivation remain largely unexplored. In particular, it remains unclear why some listeners derive atypicality inferences in response to informationally redundant utterances, while others do not. Investigating the factors that modulate the derivation of such inferences provides an opportunity to better understand how pragmatic interpretation operates in situations where implicature arises from background knowledge rather than from lexical scales.

The central goal of this dissertation is therefore to investigate which factors influence the derivation of atypicality inferences. The dissertation approaches this question from two complementary perspectives. The first perspective examines **external constraints on cognitive resources** when encountering informationally redundant materials. A long-standing line of research in experimental pragmatics has investigated whether deriving pragmatic inferences requires cognitive effort and whether limiting available resources affects interpretation. Theoretical accounts of pragmatic processing make different predictions about the cognitive demands involved in deriving pragmatic interpretations (e.g., Levinson, 2000; Sperber & Wilson, 1996; Degen & Tanenhaus, 2019), and these predictions have often been tested by manipulating cognitive load during comprehension. One common approach introduces a secondary task that competes for attentional or working memory resources during pragmatic processing. Such paradigms have been widely used in studies of scalar implicatures (De Neys & Schaeken, 2007; Marty et al., 2013; Cho, 2020). However, it remains unclear whether similar predictions hold for atypicality inferences.

The second perspective considers **natural variability in cognitive and personality characteristics across individuals**. Another line of research in experimental pragmatics has examined how differences in cognitive abilities and personality traits influence pragmatic interpretation. Studies investigating scalar implicatures and related phenomena have linked interpretation patterns to factors such as working memory capacity or theory of mind abilities, although empirical evidence remains mixed (Antoniou et al., 2016; Fairchild & Papafragou, 2021; Yang et al., 2018; Nieuwland et al., 2010). Examining these factors makes it possible to investigate whether

differences observed in the processing of atypicality inferences can be explained by individual cognitive profiles.

Together, these two perspectives provide complementary approaches to investigating the factors that influence the derivation of atypicality inferences. Experimental manipulations of cognitive load allow testing whether the availability of cognitive resources constrains the derivation of such inferences. Analyses of individual variability, in turn, examine whether differences in cognitive and socio-pragmatic characteristics across individuals are associated with differences in the likelihood and strength of deriving atypicality interpretations.

The next section presents the research goal and research questions of the dissertation.

1.1 Research goal and questions

The present dissertation investigates comprehender-specific factors involved in the derivation of atypicality inferences. Its primary goal is to identify the cognitive and personality-based factors that modulate their derivation. To achieve this goal, I first propose a step-by-step framework of how atypicality inferences are derived, which provides the theoretical foundation for investigating potential influencing factors (RQ1). Before testing such influences, it is necessary to establish whether atypicality inferences are derived consistently within individuals, whether there is systematic variability across comprehenders, and whether these patterns remain stable over time (RQ2). I then examine the role of external constraints, asking whether limiting cognitive resources affects the derivation of atypicality inferences, as predicted by pragmatic processing accounts (RQ3). In addition, I explore whether individual differences in cognitive and personality-based factors can explain variability in how comprehenders derive atypicality inferences (RQ4). Finally, I compare atypicality inferences with other types of implicature to investigate whether they rely on shared mechanisms or whether they represent a distinct class of pragmatic reasoning (RQ5).

Research Question 1. How can the derivation of atypicality inferences be theoretically formalized into distinct cognitive steps, and how does this framework provide a basis for investigating the factors that influence their derivation?

While previous studies (e.g., Kravtchenko & Demberg, 2022a,b) have examined atypicality inferences and proposed general mechanisms for their interpretation, the specific steps by which such inferences are derived have not been explicitly formalized. Building on previous findings and related literature on informational redundancy and pragmatic processing, I propose a step-by-step framework that formalizes the reasoning path comprehenders may follow when deriving atypicality inferences (see Chapter 3). This framework includes detecting informational redundancy, recognizing a potential violation of conversational norms, inferring event atypicality, and accom-

modating this inference into the discourse representation. It serves as a theoretical foundation for my empirical studies, which investigate the factors that may influence the derivation of atypicality inferences.

Within this framework, two open issues emerge. The first issue concerns the final step of the proposed derivation process, namely contextual accommodation. This step requires comprehenders to integrate the inferred atypicality into the discourse representation (e.g., inferring that Mary habitually does not eat when going to restaurants, and then accommodating this inference by revising the discourse representation to include the assumption that Mary typically only orders drinks). Although such accommodation has been hypothesized, it has not been directly tested, and the rating-based methodology used by Kravtchenko (2022) cannot reveal whether and how it occurs. The second issue concerns the methodological interpretation of rating-based measures in a more general sense. While ratings have previously been used as a proxy for atypicality inferences, it is not clear to what extent they capture the underlying reasoning processes. On the one hand, ratings may not reflect atypicality inferences at all but instead tap into broader plausibility judgments or task-specific response strategies. On the other hand, because ratings provide only a final outcome measure, they may mask certain intermediate reasoning steps, such as the initial computation and subsequent rejection of an inference. In turn, if ratings do not adequately measure atypicality inferences, they cannot serve as a measure to investigate the factors that influence their derivation.

Taken together, this motivates the following sub-questions (investigated experimentally in Chapter 7):

(a) Does the derivation process of atypicality inferences involve contextual accommodation, i.e., integrating the inferred atypicality into the discourse representation?

(b) To what extent do the rating-based measures used by Kravtchenko (2022) adequately capture the derivation of atypicality inferences?

Research Question 2. Are atypicality inferences derived consistently within individuals and across time, and how much variability exists across individuals?

To investigate which cognitive and personality-based factors modulate the derivation of atypicality inferences, it is first necessary to establish whether such inferences show consistent patterns within individuals and whether there is systematic variability across comprehenders. Without this, any findings on individual differences could be dismissed as reflecting random variation or task-specific strategies rather than stable and meaningful tendencies. Previous studies on other types of pragmatic inferences, such as numerals or scalar implicatures, have demonstrated within-subject consistency and between-subject variability during a single session (Singh et al., 2016;

Marty et al., 2013; Tavano & Kaiser, 2010; Panizza et al., 2009). However, no such evidence exists for atypicality inferences. In addition, the question of whether response preferences remain stable over time has in general received little attention in the literature, particularly in adult populations (cf. Taguchi, 2012, but in the context of L2 population).

According to this, I identify the following sub-questions:

(a) Do individuals show consistent response preferences (e.g., pragmatic or literal) within a single session?

(b) Is there systematic variability across individuals, with some consistently pragmatic and others consistently literal in their responses?

(c) Do individuals' response preferences remain stable over repeated testing sessions?

Sub-questions (a) and (b) are investigated experimentally in Chapter 7, while sub-question (c) is addressed in a longitudinal study in Chapter 10.

Research Question 3. Do external constraints on cognitive resources affect the derivation of atypicality inferences?

One way to investigate the factors that influence the derivation of atypicality inferences is to impose an external load on a particular resource, thus reducing its availability for all participants. This creates a strong constraint on all comprehenders and allows us to test whether the resource is necessary for the derivation process in principle. If the resource in question genuinely modulates inference, then placing it under load should reduce the rate or strength of pragmatic interpretations – an effect that can be observed at the population level given a sufficiently large sample size.

In the literature on pragmatic processing, this approach has been widely used to test the predictions of theoretical accounts that adopt a processing cost perspective, claiming that different types of implicature require cognitive resources to varying degrees (e.g., the Default account proposed by Levinson, 1987; Relevance Theory, Sperber & Wilson, 1996; and the constraint-based approach of Degen & Tanenhaus, 2019). From this perspective, processing cost reflects the need for sufficient capacity in a specific resource in order to derive an implicature.

Experimentally, the presence or absence of processing cost has typically been investigated with dual-task designs, most often focusing on working memory load, in which participants complete a pragmatic task while simultaneously performing a secondary task that reduces available capacity. To date, such studies have, however, focused on scale-based implicatures (e.g., scalar terms or numerals; De Neys & Schaeken, 2007; Marty et al., 2013; Cho, 2020). Moreover, theoretical accounts do not provide direct predictions for atypicality inferences, instead also focusing on scale-based types. Consequently, two steps are required: first, the predictions of these accounts with respect

to atypicality inferences must be evaluated; and second, the dual-task method must be adapted to test atypicality inference derivation.

Accordingly, two additional sub-questions arise:

(a) What do current pragmatic processing accounts predict about the effect of resource constraints on atypicality inferences?

(b) How is the strength of atypicality inferences affected when working memory capacity is experimentally limited (e.g., under dual-task load)?

In my dissertation, I therefore evaluate the predictions of theoretical accounts (Chapter 4) and adapt the dual-task approach to atypicality inferences (Chapter 9), using two types of secondary task: one tapping attentional resources through a continuous dot-tracking task (not previously used in this domain) and another targeting verbal working memory through a reading span task. This design allows me to test whether working memory capacity modulates the derivation of atypicality inferences, thereby evaluating the predictions of pragmatic processing accounts regarding their processing cost.

Research Question 4. Which individual factors predict variability in the derivation of atypicality inferences?

In contrast to Research Question 3, which examines the effects of externally imposed resource constraints, the present question focuses on natural variation in comprehenders' cognitive characteristics. Rather than manipulating the availability of a resource for all participants, I adopt the individual differences approach, where I assess participants' cognitive and personality-related traits and examine whether variability in these traits predicts differences in atypicality inference derivation.

Given a sufficiently large sample, this approach allows for the investigation of whether natural variation in specific cognitive capacities is associated with stronger or weaker tendencies to derive atypicality inferences. Previous research on other types of implicature has shown that this variability may be linked to individual differences in cognitive and socio-cognitive abilities (see e.g., Antoniou et al., 2016; Yang et al., 2018; Heyman & Schaeken, 2015; Fairchild & Papafragou, 2021; Nieuwland et al., 2010; Feeney & Bonnefon, 2012).

Building on the formalization of the derivation process for atypicality inferences in Research Question 1 (see Chapter 3), as well as on the previous literature on other types of implicature, I identify a set of factors that are theoretically relevant to the proposed steps of atypicality inference derivation. These include working memory capacity, inhibition, memory updating, theory of mind, autistic traits, cognitive reflection, fluid intelligence, and print exposure. The theoretical motivation for selecting these factors is discussed in Chapter 5.

Accordingly, the following sub-question arises:

(a) Do cognitive and personality-related factors, such as working memory, inhibition, memory updating, theory of mind, autistic traits, cognitive reflection, fluid intelligence, and print exposure, predict variability in deriving atypicality inferences?

The experimental investigations addressing this research question are reported in Chapters 8 and 10.

Research Question 5. How do atypicality inferences compare to other implicature types?

While the previous research questions have examined atypicality inferences in isolation, the present question addresses how they relate to other types of pragmatic implicature. In the pragmatic literature, scalar implicatures (e.g., bare *some*, numerals, and disjunctions) are commonly treated as generalized conversational implicatures (GCIs; see e.g., Levinson, 2000; cf. Sperber & Wilson, 1996; Degen & Tanenhaus, 2019). In Chapter 4, I argue that atypicality inferences share properties with particularized conversational implicatures (PCIs), insofar as they are highly context-dependent and not anchored in lexical scales. However, it remains unclear whether this theoretical positioning is reflected in the cognitive factors that modulate them.

Experimental studies rarely examine multiple implicature types within the same sample of participants. Even when they do, response tendencies are typically analyzed separately rather than compared directly (e.g., Fairchild & Papafragou, 2021). Recent work suggests that performance across pragmatic phenomena may show only limited overlap and may not be reducible to a single general pragmatic ability (Çiftlikli & Demirel, 2022; Floyd et al., 2025). These findings motivate a comparison of atypicality inferences with well-studied scale-based implicatures within the same group of participants.

To address this research question, I therefore include several types of scalar implicatures (bare *some*, bare numerals, bare disjunctions, embedded variants in non-monotonic contexts, and politeness implicatures with *some*) alongside atypicality inferences.

This design allows assessment of whether response tendencies generalize across implicature types and whether atypicality inferences pattern together with scale-based implicatures or exhibit a distinct profile within the broader pragmatic landscape.

Accordingly, the following sub-questions arise:

(a) Do individuals' responses on atypicality inferences correlate with their responses on other implicature types (e.g., scalar implicatures, disjunctions, numerals)?

(b) Are the same cognitive or personality-based traits predictive of response tendencies across implicature types or do different factors play a role?

The experimental investigation addressing these questions is presented in Chapter 10.

1.2 Contributions of the research

Contribution 1. Formalization and empirical validation of the derivation process of atypicality inferences (RQ1).

This dissertation provides a step-by-step formalization of the derivation of atypicality inferences. Building on previous work, I propose that comprehenders (i) detect informational redundancy relative to script knowledge, (ii) recognize a potential violation of conversational norms, (iii) infer that the described event is atypical, and (iv) accommodate this inference within the discourse representation. This framework provides a theoretical account of the reasoning path comprehenders may follow when deriving atypicality inferences and establishes a basis for investigating the cognitive and contextual factors that may influence this process.

In addition, I empirically investigate the final step of this process, namely contextual accommodation. The results show that comprehenders not only assign lower typicality ratings, but actively integrate the inferred atypicality into the discourse, often by constructing context-specific explanations. At the same time, the findings show that previously used rating-based measures do not fully capture the underlying reasoning process. Ratings may mask intermediate stages of interpretation, such as uncertainty or rejection of the inference, and therefore provide only a partial reflection of the derivation process. This has important methodological implications for the study of atypicality inferences and pragmatic processing more generally.

Contribution 2. Evidence for stable but systematically variable response tendencies in atypicality inference derivation (RQ2).

The derivation of atypicality inferences is neither random nor uniform across comprehenders. Within a single experimental session, individuals show consistent response preferences, tending toward either pragmatic or literal interpretations. At the same time, there is systematic variability between individuals: some participants consistently adopt a pragmatic interpretation, whereas others consistently prefer a literal one.

Moreover, these response tendencies remain stable over time. A longitudinal study demonstrates that individual preferences persist across testing sessions separated by approximately two months. This stability suggests that pragmatic consistency appears to be a trait inherent to an individual, at least within the context of the same pragmatic task.

I also provide evidence on scale-based implicature types (e.g., scalar implicatures, numerals, and disjunctions), which allows comparison of response consistency across pragmatic phenomena. Compared to atypicality inferences, these implicature types may exhibit stronger learning effects, where participants may shift their interpretations from literal to more pragmatic responses, and vice versa. This pattern suggests that participants may adapt their interpretation strategies over the course of the experiment. Taken together, these findings suggest that pragmatic response patterns

are systematic overall and atypicality inferences tend to be even more stable over time.

Contribution 3. External constraints on working memory do not affect the derivation of atypicality inferences (RQ3).

Adopting a processing-cost perspective, I evaluate the predictions of prominent pragmatic processing accounts regarding the derivation of atypicality inferences. According to these accounts, the derivation of atypicality inferences should be cognitively costly and therefore sensitive to limitations in cognitive resources; for example through limitations in working memory capacity.

Across a series of three large sample experiments using dual-task paradigms, I investigate whether experimentally limiting cognitive resources affects the derivation of atypicality inferences. Two types of secondary task are employed: a novel visuomotor tracking task targeting attentional resources and a reading span task targeting verbal working memory capacity. Across both paradigms, no reduction in atypicality inferences is observed under cognitive load. These findings suggest that atypicality inference derivation does not rely critically on the executive resources targeted by the secondary tasks. The absence of robust dual-task effects challenges the assumption that all pragmatic inferences are uniformly resource-demanding and indicates that the processing profile of atypicality inferences differs from that often reported for scale-based implicatures.

Contribution 4. Reasoning ability, rather than executive function, is associated with variability in atypicality inference derivation (RQ4).

While measures of executive function (including working memory, inhibition, and updating) did not reliably predict individual variability in atypicality inference derivation, reasoning-related measures showed some evidence of being associated with stronger atypicality inferences. In particular, measures of fluid intelligence and cognitive reflection were related to participants' tendency to derive atypicality inferences, although this relationship was not observed consistently across all analyses.

Taken together, these results suggest that variability in atypicality inference derivation is not primarily driven by individual differences in executive functions but may instead reflect differences in higher-level reasoning abilities. In particular, reasoning ability may facilitate the accommodation step of the derivation process proposed in RQ1, which requires comprehenders to construct and evaluate context-specific explanations for atypical events.

Contribution 5. Atypicality inferences exhibit a distinct response profile relative to other implicature types (RQ5).

Finally, the dissertation situates atypicality inferences within the broader scope of pragmatic implicatures. Although scalar implicatures are commonly analyzed as generalized conversational implicatures and atypicality inferences have been argued to

resemble particularized implicatures, it has remained unclear whether such theoretical distinctions are reflected in comprehenders' response patterns.

The results show limited correlations between atypicality inferences and scale-based implicatures (e.g., scalar implicatures, disjunctions, and numerals). In contrast, stronger correlations are observed among implicature types that are conceptually closer to one another. This pattern suggests that atypicality inferences do not simply align with scale-based implicatures but instead exhibit a distinct response profile. At the same time, not all scale-based implicatures behave uniformly: numeral implicatures, for example, show weaker relationships with other scalar phenomena and display different patterns of responses across the experiment. Taken together, these findings contribute to a more nuanced understanding of implicature computation and suggest that different types of pragmatic inference may not form a uniform class with respect to how they are interpreted by comprehenders.

Overall conclusion. Taken together, the findings of this dissertation are interpreted within the proposed framework of atypicality inference derivation and show that these inferences exhibit stable individual response tendencies. Their derivation does not appear to rely critically on executive-function resources, but variability across comprehenders may be related to differences in higher-level reasoning abilities involved in contextual accommodation. Moreover, atypicality inferences show a response profile that differs from that of scale-based implicatures, suggesting that pragmatic inference types may not form a uniform class.

1.3 Previous publications of the materials in this dissertation

Some of the materials and results presented in this dissertation have already been published in papers I co-authored. This section outlines which parts are based on published work, what modifications were made, and what role I played in the original publications. It also highlights how the published materials have been integrated and contextualized within the broader aims of the dissertation.

Ryzhova, M., & Demberg, V. (2023). Processing cost effects of atypicality inferences in a dual-task setup. *Journal of Pragmatics*, 211, 47-80. doi: https://doi.org/10.1016/j.pragma.2023.04.005.

This publication forms the basis for Chapter 9, which reports a series of three experiments investigating the processing cost of atypicality inferences in a dual-task setup. The motivation, experimental design, analyses, and interpretation presented in that chapter are largely consistent with the published version, with minor changes introduced to align with the structure and format of the dissertation (e.g., adjustments in referencing and slight rewording).

Chapter 4, which presents background literature on processing cost accounts and dual-task experimental methodology, also draws directly on the background content of this publication. While most of the text was retained, I made minor edits for style and clarity, including slight rewording and adjustments to fit the overall tone of the dissertation. In addition, several expansions were made to provide additional context on the Gricean framework, clarify the place of atypicality inferences within the PCI/GCI classification, and explain related dual-task designs in the existing literature.

As the first author, I was responsible for designing and conducting the experiments, analyzing the data, and writing the manuscript. My co-author, who is also my doctoral supervisor, provided conceptual guidance, supervision, and editorial feedback throughout the research and publication process.

The findings in Chapter 9 were also published in the proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020) in (Ryzhova & Demberg, 2020).

Ryzhova, M., Mayn, A., & Demberg, V. (2023). What inferences do people actually make on encountering a redundant utterance? An individual differences study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45, 2631-2638. URL: https://escholarship.org/uc/item/88g7g5z0.

The experiments presented in Chapters 7 and 8 were first published in this article. From the publication, I incorporated parts of the original motivation (reformulated and extended for the purposes of the dissertation), as well as all experiment-related content: the annotation scheme, data analysis, and main conclusions.

Several components were substantially extended in the dissertation. Firstly, compared to the publication, the reporting of analysis decisions, predictions, participant demographics, and results was more elaborated in the dissertation. Secondly, the description of the annotation scheme was vastly expanded to better clarify how and why certain annotation decisions were made. Thirdly, in the analysis section, I added a qualitative analysis of participants' explanations, focusing on how individuals accommodated atypicality inferences in context. In addition, I introduced a second level of annotations that captured other types of inferences and common inferencing mistakes. These aspects were not discussed in the original article. Finally, both chapters also include additional visualizations and plots that were not part of the publication.

In addition, the conclusions in Chapter 7 were extended to incorporate these additional analyses and to summarize broader take-away points. In Chapter 8, I also expanded the discussion on the role of reasoning ability in derivation of atypicality inferences.

As the first author, I was primarily responsible for the data analysis and all visualizations, while the experimental idea and design were shaped in discussion with

the second and third authors. I was responsible for designing the annotation scheme, including defining the categories, setting the decision criteria, and writing up the guidelines, with valuable feedback from the second author during the development process. The paper was written collaboratively with the second author, under the guidance of the third author.

The second author and I jointly annotated the final dataset.

The findings in Chapters 7 and 8 were also presented at the 18th International Pragmatics Conference (IPRA 2023)[2].

Ryzhova, M., Loy, J., & Demberg, V. (2022a). Pragmatic comprehension in individuals is stable across time and some tasks. In F. Frau, L. Bischetti, C. Pompei, B. Scalingi, F. Domaneschi, & V. Bambini (Eds.), *Book of Abstracts – XPRAG 2022*. Italy: University School for Advanced Studies IUSS Pavia. doi: https://doi.org/10.17605/OSF.IO/C4KP2.

Chapter 10 builds on a study that was previously presented at a conference. Motivation was adapted with major changes to better align it with the broader aims of the dissertation. The analyses and conclusions follow the same structure as in the original version, but are presented here in greater detail, with additional comments on the general approach, experimental design, analysis decisions, and results.

As the first author, I was responsible for the experimental design, data collection, data analysis, and interpretation. The text of the abstract was jointly written with the second author, under the supervision of the third author.

The study in Chapter 10 was also presented at the CogSci 2022 conference (Ryzhova et al., 2022).

1.3.1 Additional related publications

Kurch, C., Ryzhova, M., & Demberg, V. (2024). Large language models fail to derive atypicality inferences in a human-like manner. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, & Y. Oseki (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 86-100). Bangkok, Thailand: Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2024.cmcl-1.8.

In collaborative work with a master's student whom I co-supervised, we examined whether large language models (LLMs) can derive atypicality inferences. While humans interpret informationally redundant utterances as signaling atypical behavior, LLMs consistently fail to derive atypicality inferences in zero-shot settings. Few-shot prompting produces partial improvements, but largely reflects shallow pattern

[2]Link to the program booklet: link

matching rather than pragmatic reasoning. Further analyses show that models possess script knowledge, but fail to treat informational redundancy as a violation of conversational norms required for pragmatic inference.

The derivation steps of atypicality inferences presented in Chapter 3 were first formalized as part of this publication. In the dissertation, I substantially extend the content by elaborating on each step and discussing the cognitive mechanisms that may be involved in the derivation process.

As co-advisor of the first author's master's thesis, I was solely responsible for the conceptual development, formalization, and write-up of the derivation steps of atypicality inferences presented in the publication.

Hong, X., Ryzhova, M., Biondi, D., & Demberg, V. (2024). Do large language models and humans have similar behaviours in causal inference with script knowledge? In D. Bollegala, & V. Shwartz (Eds.), Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024) (pp. 421–437). Mexico City, Mexico: Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2024.starsem-1.34.

In this study, human processing of causal relations in script-based narratives was compared with the behavior of LLMs. Results from a self-paced reading experiment show that humans stumble across causally incoherent text segments, exhibiting longer reading times in such cases. At the same time, readers can easily integrate script-predictable information, even when the explicit causal event is omitted from the story. When the same paradigm is applied to LLMs, only the newest models show similar behavior in detecting causal conflicts, while all tested models fail to replicate human behavior when the causal event is omitted, suggesting that script knowledge is not sufficiently represented in these models. As co-first author (equal contribution), I was responsible for the human experiment and statistical analyses, and jointly contributed to the paper write-up.

Part I

Background

Chapter 2

Atypicality inferences

In everyday communication, speakers often provide more information than strictly necessary. Empirical research in linguistics and psycholinguistics demonstrates that overinformativity can manifest in various forms. It occurs across multiple linguistic levels, including phonetics, syntax, and discourse.

Some studies suggest that overinformativity is often tolerated by listeners, typically more so than underinformativity (Walker, 1993; Pechmann, 1989; Engelhardt et al., 2006; Baker et al., 2008)[1], while others even report communicative advantages. For example, in reference game paradigms, redundant expressions have been shown to aid comprehension and alignment with the listener (Rubio-Fernández, 2016; Tourtouri et al., 2019; Vogels et al., 2020; Rohde et al., 2021; Li et al., 2023).

The work of Kravtchenko (2022) focuses on overinformativity at the discourse level, where redundancy arises not from linguistic repetition or referential overspecification, but from stating something already available through shared world knowledge. This results in the message being *informationally redundant*, in the sense that it carries low informational utility for the listener[2]. Kravtchenko (2022) investigates how comprehenders process such utterances and whether they engage in pragmatic reasoning to accommodate redundancy. An example of such an informationally redundant message is shown below.

(2.1) *Mary went to a restaurant.* ***She ate there!***

Semantically, this utterance provides no new information, as eating is an expected part of *going to the restaurant* activity. While some pragmatic theories (e.g., Grice, 1975) posit that speakers aim to be appropriately informative, they often leave unaddressed whether informationally redundant utterances like these are perceived as

[1] Cf. Davies & Katsos (2010, 2013). See also Veenstra & Katsos (2018), for an overview on over- and underinformativity in sentence judgment task.

[2] In this dissertation, I use the terms *overinformativity* and *informational redundancy* largely interchangeably (see Section 2.2, for further discussion).

infelicitous, irrelevant, or pragmatically marked. More importantly, they do not fully consider whether listeners engage in pragmatic repair when encountering such utterances, and if so, what the interpretive consequences are.

Kravtchenko (2022) shows that comprehenders not only recognize the redundancy but interpret such utterances via pragmatic enrichment. That is, they adjust their beliefs about the discourse common ground to accommodate the speaker's communicative intent. In Example 2.1, rather than treating "*She ate there!*" literally, listeners infer that Mary does not usually eat at restaurants, and that her eating there on this occasion is somehow atypical. This kind of belief revision transforms the utterance into a pragmatically informative one and gives rise to so-called **atypicality inferences**.

Kravtchenko's results, meanwhile, might also suggest that not all comprehenders consistently derive atypicality inferences. While this variability has not been systematically examined, the data might hint at individual differences in susceptibility to the inference, possibly due to differences in recognizing redundancy or engaging in pragmatic reasoning. This also resonates with previous findings on individual variability and processing cost in other types of pragmatic phenomena, such as scalar implicatures (e.g., Bott & Noveck, 2004; Dieussaert et al., 2011; Fairchild & Papafragou, 2021), and motivates the present work, which investigates comprehender-specific cognitive factors that might shape atypicality inferences.

In the following chapter, I begin with a detailed analysis of the findings of Kravtchenko (2022), which serve as the empirical foundation for my work (Section 2.1). I then turn to the broader theoretical concepts relevant to these results, focusing on informativeness and the role of overinformativity in conversation (Section 2.2). Then in Chapter 3, I introduce the derivation process of atypicality inferences, which provides the basis for predictions about individual variation and processing cost, which will be explored in more detail in later chapters.

2.1 Findings of Kravtchenko (2022)

Kravtchenko (2022) investigated whether excessive informational redundancy distorts the message received by the listener and can elicit pragmatic inferences (see also Kravtchenko & Demberg, 2015, 2022b,a). A series of experiments demonstrated that informationally redundant event descriptions can shift listeners' prior beliefs about the discourse common ground. When comprehenders encounter an utterance that redundantly states what is already strongly implied (as in Example 2.1), they accommodate the perceived redundancy through pragmatic inference, revising assumptions about what is typical or expected.

Kravtchenko (2022) modeled informational redundancy in terms of script-based knowledge (i.e., structured knowledge about common everyday activities and their sub-events; see e.g., Bower et al., 1979). According to major pragmatic theories (e.g., Grice, 1975), explicitly stating highly predictable information violates conversational norms, such as the Maxim of Quantity. Findings from Kravtchenko (2022) indicate that utterances describing events strongly expected within the script (*conventionally habitual* events) are interpreted pragmatically: rather than being taken at face value, they are understood as implying that the event may not be typical for the individual in question.

Importantly, the effect is context-sensitive. When the broader discourse context already suggests that the event is unlikely or atypical, the utterance no longer appears redundant, and no atypicality inference is drawn. This supports the view that such inferences arise only when there is a pragmatic mismatch between contextual expectations and the informativeness of the utterance.

The results further show that the effect is not triggered by all utterances. Descriptions of plausible but not script-predictable events (*non-habitual* events) do not lead to similar belief updates. This suggests that informational redundancy, rather than mere mention, is what drives the derivation of atypicality inferences. In the following section, I describe the results of Kravtchenko (2022) in detail.

2.1.1 Materials

Kravtchenko (2022) designed 24 stories describing common activities with a 2 (story context: *neutral* vs. *biasing*) x 3 (utterance: *conventionally habitual*, *non-habitual*, vs. *no utterance*, where the utterance was omitted and the story ended after the discourse setup) factorial design – see Table 2.1 (*going to a restaurant* activity) and Table 2.2 (*going shopping* activity), for two examples of experimental items across all conditions.

Each story centered on a stereotyped activity, selected for its clear and distinct sub-events, which typically followed a roughly consistent temporal order (e.g., going shopping, going to a restaurant, buying a subway ticket, or taking a flight) – see

Table 2.1: Experimental materials from Kravtchenko (2022); An example of ***going to a restaurant*** story by context condition (neutral vs. biasing) and utterance condition – conventionally habitual vs. non-habitual event is mentioned in the utterance (highlighted in gray). A **baseline** for both context conditions does not include an utterance text block.

Context	**neutral** Mary is a journalist who often goes to restaurants after her interviews.	**biasing** Mary is a journalist who often interviews restaurant waiters, but doesn't like eating out.
Discourse setup	Yesterday, she went to a popular Chinese place. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.	
Utterance	**conventionally habitual event** David said to Sally: "I ran into Mary leaving that Chinese place. She ate there! "	**non-habitual event** David said to Sally: "I ran into Mary leaving that Chinese place. She got to see their kitchen! "
Question habitual	How often do you think Mary usually eats, when going to a restaurant?	
Question non-habit	How often do you think Mary usually gets to see the kitchen, when going to a restaurant?	
Filler question 1	How often do you think Mary usually goes to restaurants?	
Filler question 2	How often do you think Sally and David usually run into each other?	

Chapter A, Table A.1, (Context version I), for a full list of experimental materials developed by Kravtchenko (2022).

The stories consisted of several blocks (but were presented to participants in plain continuous text without special formatting or highlights): context, discourse setup, and the utterance (which was omitted in the baseline *no utterance* condition). The **context** introduced the main activity featured in the story (e.g., going to a restaurant or shopping). Crucially, it established common ground regarding the main character's general habits related to a specific event within that activity – the event later targeted in the utterance and questions. Another function of context was to show that the main character of the story is often involved in the described activity (e.g., *Mary often goes to restaurants...* or *Mary is a journalist who often interviews restaurant waiters...*, in the *neutral* and *biasing* context conditions, respectively). Such framing licensed participants to make judgments about the typicality of the target event.

Table 2.2: Experimental materials from Kravtchenko (2022); An example of ***going shopping*** story by context condition (neutral vs. biasing) and utterance condition – conventionally habitual vs. non-habitual event is mentioned in the utterance (highlighted in gray). A **baseline** for both context conditions does not include an utterance text block.

Context	**neutral** John often goes to the grocery store around the corner from his apartment.	**biasing** John is typically broke, and doesn't usually pay when he goes to the grocery store.
Discourse setup	Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.	
Utterance	**conventionally habitual event** Susan said to Peter: "John just came back from the grocery store. He paid the cashier! "	**non-habitual event** Susan said to Peter: "John just came back from the grocery store. He got some apples! "
Question habitual	How often do you think John usually pays the cashier, when going shopping?	
Question non-habit	How often do you think John usually gets apples, when going shopping?	
Filler question 1	How often do you think John usually goes to the grocery store?	
Filler question 2	How often do you think Susan and Peter usually talk to each other?	

The **discourse setup** established the specific situational context of the story, introducing the relevant characters and their relationships. It served to create a shared common ground among the story characters, ensuring that it would be plausible for all of them – and, by extension, the reader – to be aware of the main character's typical habits regarding the critical event associated with the activity. Across the materials, half of the stories involved three discourse characters: the main character who engaged in the activity, a second character who learned about this from the main character, and a third character to whom this information was later passed. In the remaining stories, only two characters were involved: the main character and a second character who received the information directly.

The **utterance block** featured the critical experimental manipulation. The utterance described the engagement of the main character in a specific event, which was either (1) a *conventionally habitual* event – one strongly associated with the activity and likely to be inferred even without explicit mention[3]; or (2) a *non-habitual*

[3]As Kravtchenko (2022) points out, the term *conventionally habitual* is used *"to specify that the*

event that was plausible within the context but not strongly predictable based on general knowledge about the activity (for example, eating vs. interviewing people in a restaurant). While constructing the experimental materials, Kravtchenko (2022) writes about certain constraints when choosing target events, in particular conventionally habitual ones: "*It needed to be possible for the script to play out without the habitual event having taken place – otherwise, the discourse would be incoherent, or the inference would not be drawn. For example, one arguably cannot play tennis at all, without using a racket.*" (p. 58).

Depending on the number of characters in the story, the utterance was either produced by the main character (in two-character stories) or by another character (three-character stories). For instance, in the restaurant story, the utterance "*I ran into Mary leaving that Chinese place. She ate there!*" (three-character version) would become "*I just went to that Chinese place. I ate there!*" (two-character version).

An exclamation mark at the end of the utterance was used to signal that the utterance was produced as an intentional and marked communicative act. In this way, the exclamatory form served to disambiguate utterance's discourse function, distinguishing it from utterances that anchor a conversation, mark hesitation, or preface further information, and instead it highlighted the speaker's intent to draw attention to the event's occurrence. The interpretive role of the exclamation mark and its potential interaction with pragmatic reasoning are discussed in Section 2.1.4.

Stories in the *no utterance* condition did not include any utterance. This condition served as a baseline for estimating general beliefs about the event being a part of this activity when it was not explicitly mentioned.

Central to the experiment was the idea that utterances vary in their informativity, and that this variation can shape how listeners interpret event's typicality. Specifically, only an utterance that redundantly mentions an event already strongly implied by the context may trigger a pragmatic inference – namely, that the event is not as typical as expected. To systematically test this, Kravtchenko (2022) manipulated the informativity of the target utterance through variation in the **story context**, to check that atypicality inferences do not arise when the context does not prompt the *conventionally habitual* utterance to actually be informationally redundant.

To manipulate utterance's informativity, Kravtchenko (2022) introduced two types of story contexts: *neutral* and *biasing*. The *neutral* context was designed to be script-congruent. It did not present any information which was not consistent with the related script i.e., with the common assumptions about how the target activity is performed. In the *biasing* context the common ground was manipulated so that the typical assumptions about the activity (related to the event in the target utterance) may not hold (e.g., *"Mary often interviews restaurant waiters but* ***does not like eating out****."* or *"John is* ***typically broke and doesn't usually pay*** *when going*

event almost invariably occurs as part of the activity script (under normal conditions and for typical individuals)" (p. 56). In the present work, I share the same definition. However, I discuss the assumption of scripts' individual invariance in Section 12.1.

grocery shopping.").

Before the main experiments, three norming studies were conducted to verify the following criteria: (1) participants consistently judged target events as conventionally habitual or non-habitual, given the activity; (2) the contextual manipulation reliably shifted perceived typicality for both habitual and non-habitual events; and (3) the activity remained coherent and could plausibly proceed even in the absence of the habitual or non-habitual events. All items in the final set of materials satisfied the three norming criteria[4].

Questions

The questions were designed to assess how typical participants believed both the *conventionally habitual* and *non-habitual* events were for the main character of the story – see **Question habitual** and **Question non-habit** in Tables 2.1 and 2.2.

Each question was answered using a continuous scale ranging from 'Never' to 'Always', mapped onto a 0–100 range, with 50 labeled as 'Sometimes' – see Figure 2.1.

Figure 2.1: A slider used by experimental participants in Kravtchenko & Demberg (2022a). Source: Kravtchenko & Demberg (2022a).

2.1.2 Predictions

The predictions concern typicality ratings for the *conventionally habitual* and *non-habitual* events across all combinations of story context and utterance conditions. These predictions are based on the assumption that pragmatic reasoning is triggered when an utterance redundantly states information that could have been inferred from related world knowledge. The expected patterns of typicality ratings across conditions are summarized in Table 2.3 and are discussed in detail below.

Effects of conventionally habitual events. The following predictions concern typicality judgments made in response to the question about the *conventionally habitual event*. Only conditions in which this event was either not mentioned (baseline) or explicitly mentioned in the utterance are considered in the analysis. In the condition combining a *neutral* context with no target utterance (baseline), participants are expected to infer (based on commonsense knowledge about the activity) that the *conventionally habitual* event typically occurs. Accordingly, this event should receive high typicality ratings.

When the *neutral* context is followed by an utterance explicitly mentioning the *conventionally habitual* event, the utterance becomes *informationally redundant*, re-

[4]For details, see Kravtchenko (2022), p. 58.

Table 2.3: Expectations about the typicality of the target events across all combinations of story context and utterance conditions in Kravtchenko (2022).

Context → Utterance ↓	**Neutral**	**Biasing**
No utterance (baseline)	high	low
Conventionally habitual	lower than the baseline*	–
Non-habitual	–	–

∗ Indicates the derivation of an atypicality inference triggered by perceived redundancy.
– Indicates no expected difference in typicality relative to the corresponding baseline condition.

sulting in a violation of conversational norms – specifically, the expectation of informativity (Grice, 1975). In this case, participants are expected to infer that the target event is atypical, assuming that the main character does not usually engage in it. This reflects the derivation of an atypicality inference. In terms of ratings, participants are expected to assign lower typicality scores to the *conventionally habitual* event compared to the baseline.

In the condition where a *biasing* context is presented without a target utterance, the contextual cues are expected to lead participants to judge the *conventionally habitual* event as unlikely to occur. As a result, typicality ratings for the event should be low. However, when the same *biasing* context is followed by an utterance that explicitly mentions the *conventionally habitual* event, the utterance introduces new and relevant information. In this case, it is not perceived as redundant, and no atypicality inference is expected. For instance, in the *going to a restaurant* story, participants should not infer that Mary is even less likely to eat in restaurants than suggested by the biasing context alone.

Effects of non-habitual events. The following predictions concern typicality ratings provided in response to the question about the *non-habitual* event. These comparisons serve as a critical control: they are intended to show that the predicted pragmatic effect does not arise merely from the presence of an utterance. Rather, it is expected to occur only when an utterance redundantly mentions an event that would have been inferred even without explicit mention. Accordingly, only conditions in which the *non-habitual* event was either not mentioned (baseline) or explicitly mentioned in the utterance are included in the analysis. While *conventionally habitual* events can be inferred as part of the activity described in the stories, *non-habitual* events cannot be inferred automatically. Although plausible, such events occur only occasionally within the broader context of the activity. Therefore, in both baseline

context conditions (i.e., with no utterance), participants are expected to give average typicality ratings close to 50 on a 0-100 scale, corresponding to 'Sometimes'. In contrast to the effects expected for *conventionally habitual* events, the explicit mention of a *non-habitual* event is not predicted to significantly alter participants' beliefs about how typically the main character engages in that event. This is because mentioning a *non-habitual* event does not create informational redundancy and, therefore, does not trigger the pragmatic reasoning expected for the *conventionally habitual* event.

2.1.3 Results

A total of 700 native English speakers read the stories and answered questions about them. Each participant read 6 experimental stories and 4 filler stories. Each story topic and condition was presented to a given participant only once, resulting in a between-subjects design for typicality ratings[5]. The conditions with *no utterance* served as baselines for measuring participants' initial beliefs about event typicality. The remaining conditions, in which an event was explicitly mentioned in the utterance, measured participants' updated beliefs about event typicality.

A linear mixed-effects regression was used to analyze typicality ratings, with separate models fit for *conventionally habitual* and *non-habitual* events, based on respective question types.

In each model, the *utterance* factor had two levels: either *no utterance* (i.e., the baseline condition) or an explicit utterance referring to the relevant event type (*conventionally habitual* or *non-habitual*, depending on the model). All predictors were effect/sum coded.

Conventionally habitual event (*'He paid the cashier!'*). The results are illustrated in Figure 2.2. In the *neutral* story context, participants showed strong beliefs about the typicality of the conventionally habitual event when it was not explicitly mentioned (*no-utterance* baseline condition; $mean = 85.79$), confirming that such events are perceived as highly conventional and expected. As predicted, explicitly stating the event in the utterance led to lower typicality ratings ($mean = 72.37$), consistent with the derivation of an atypicality inference.

In the *biasing* story context, which framed the event as atypical for the main character (e.g., "*John is typically broke and does not usually pay*"), the same event was perceived as less typical in the baseline condition ($mean = 48.00$). When the event was explicitly mentioned in this context, ratings changed only slightly ($mean = 45.71$).

A regression analysis revealed a significant interaction between utterance type (no utterance vs. utterance with conventionally habitual event) and story context (neutral vs. biasing): $\beta = -10.77, p < .001$ (see Table 2.4). Post-hoc tests indicated that the

[5]The results were also replicated using a within-participants design (Kravtchenko & Demberg, 2015).

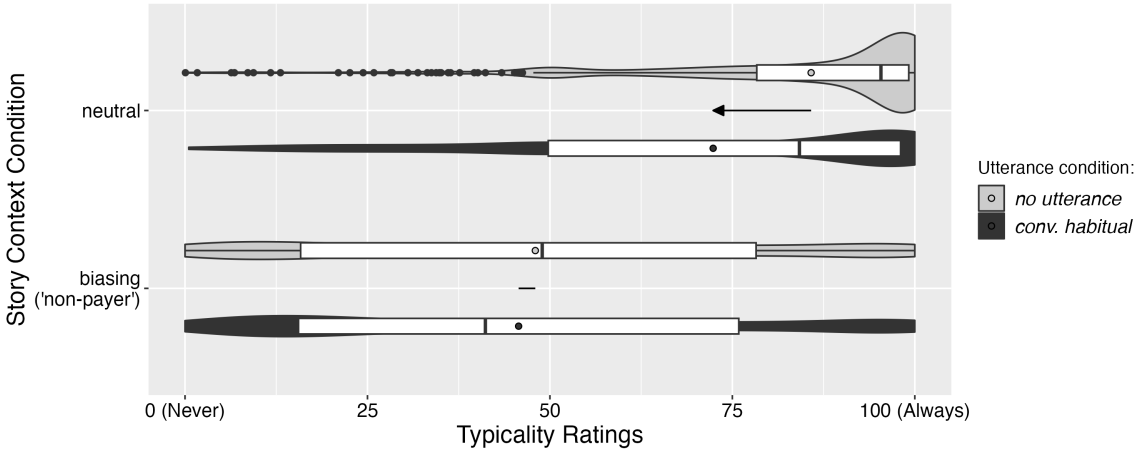


Figure 2.2: Results of Kravtchenko & Demberg (2022a). *Conventionally habitual* (paying the cashier) event analysis. The plot shows changes in typicality ratings depending on whether the *conventionally habitual* utterance is seen (*'He paid the cashier!'*), as well as whether the story context causes the utterance activity to be perceived as non-habitual. Violin plots, overlaid with box plots, show the distribution of typicality ratings. A violin plot is simply a smoothed and mirrored histogram: the fatter the distribution at a given point, the more instances there are of that particular event typicality estimate. Circles represent mean values. Arrows show statistically significant differences between ratings in *no utterance* and *conv. habitual utterance* conditions.

interaction was driven by significantly lower typicality ratings in the neutral context when the event was mentioned explicitly ($\beta = -13.21, p < .001$).

Table 2.4: Results of Kravtchenko & Demberg (2022a). *Conventionally habitual* event analysis.

	β	SE(β)	t-value	p
Intercept	63.03	1.84	34.32	**<.001**
Context: neutral	32.38	3.33	9.72	**<.001**
Utterance: conv. habit	-7.83	1.71	-4.58	**<.001**
Context * Utterance	-10.77	2.40	-4.50	**<.001**

Non-habitual event (*'He got some apples!'*). In both story contexts, only slight differences in typicality ratings were observed depending on the presence of the non-habitual utterance (*neutral* context (no-utterance vs. utterance): *mean* = 40.80 vs. 42.47; *biasing* context: 38.49 vs. 39.56) – see Figure 2.3.

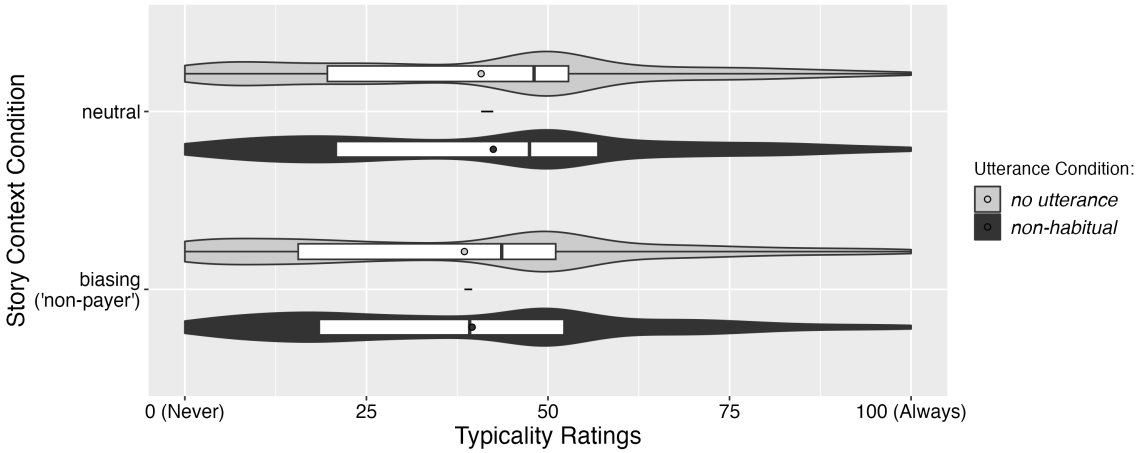


Figure 2.3: Results of Kravtchenko & Demberg (2022a). *Non-habitual* (buying apples) event analysis.

A regression analysis revealed no significant interaction between utterance and story context, nor significant main effects of either factor (see Table 2.5).

Table 2.5: Results of Kravtchenko & Demberg (2022a). *Non-habitual* event analysis.

	β	SE(β)	t-value	p
Intercept	40.29	1.86	21.69	**<.001**
Context: neutral	2.88	2.07	1.39	0.2
Utterance: non-habit	1.34	1.85	0.73	0.5
Context * Utterance	0.01	2.14	0.01	1

2.1.4 Discussion

The results indicate that participants perceive informationally redundant utterances as non-optimal at face value and engage in pragmatic reasoning to reconcile the apparent violation of conversational norms. Comprehenders seem to recognize redundancy based on their initial beliefs about activity typicality and to interpret the context accordingly, adjusting their expectations. Crucially, no inference is drawn when the otherwise predictable, conventionally habitual utterance is rendered atypical by the preceding context. Moreover, the effect cannot be attributed to the mere presence of an utterance, as no belief update is observed in the control condition involving non-habitual utterances.

The data also suggest that the inference process might be not uniform across comprehenders. Although Kravtchenko & Demberg (2022a) did not investigate the processing cost or individual variability, their results on belief updates provide motivation to explore these dimensions. The overall belief update observed in the utterance condition is relatively modest, but the distribution of responses reveals a long tail to the left. This pattern may indicate that some participants made substantial belief revisions, while others showed no meaningful change relative to the no-utterance baseline. Such variability might point to potential individual differences e.g., in sensitivity to informational redundancy and/or in the cognitive resources engaged during pragmatic reasoning, raising the questions about the mechanisms underlying atypicality inferences.

The role of exclamation mark. The exclamation mark in the critical utterance (highlighted in gray in Tables 2.1 and 2.2) was used to amplify the speaker's intent, i.e., to signal that the utterance is indeed intentionally conveyed and contains an important relevant piece of information. However, the exclamation mark may also act as a supporting cue for the derivation of atypicality inferences, as it can signal surprise or something unusual (Rett, 2011). As Kravtchenko & Demberg (2022a) discuss, even though it could be argued that exclamation forces atypical interpretation independently of the utterances' informativity, this cannot be an explanation of their findings, as there are no signs of such effects of exclamation in other experimental conditions. In addition, in the follow-up experiments, Kravtchenko & Demberg (2022a) examined utterances in which the speaker's intent was conveyed through discourse markers rather than exclamation, or where no such cues were present. These experiments showed that the effect remained. Although the size of the belief update was smaller than in the experiment including exclamation mark. Given that utterances paired with exclamation marks appear to introduce greater variability in responses across participants, it makes the exclamation condition particularly suitable for investigating how atypicality inferences interact with individual cognitive and personality profiles, which is the focus of the present research.

2.2 Concept of informativeness and overinformativity

According to the rational communication framework, interlocutors are expected to convey the appropriate amount of information to their conversational partners. This means they should (1) make their messages as informative as is required (for the current purpose of the exchange) and (2) **not** make their messages more informative than is required – see Grice (1975), for the so-called maxim of quantity. In a similar manner, Levinson (1987) formulates it in terms of the Q-principle (*do not provide a statement that is informationally weaker than your knowledge of the world allows, unless providing a stronger statement would contravene the I-principle*) and the I-principle (*say as little as necessary*). Findings in experimental pragmatics have demonstrated that violation of this maxim can result in pragmatic implicatures.

In the early stages of the field, greater attention was directed toward the lower bound of informativeness, as defined in (1). More specifically, a large portion of the studies focused on violations of (1) that arise from the interpretation of scalar terms and numerals. By definition, scalar terms can be arranged on a scale according to their degree of informativeness (e.g., {some, all}), where the usage of a weaker term is entailed in the stronger one (i.e., the stronger term 'all' includes the weaker term 'some' within its scope). In formal logic and semantics, terms like 'some' act as existential quantifiers – they assert the existence of something by establishing a lower bound ('at least one'), without specifying an upper limit (Horn, 1972, 1992). In this way, by saying "*I ate some candies.*", I can also describe the situation in which I ate all the candies. However, following Grice (1975) and the Rational Communication framework, if a speaker uses a weaker term, it can be assumed that they had a reason for not using a stronger one, provided that the speaker is cooperative and adheres to the conversational maxims. Given this, a listener is supposed to assume that once a weaker term is uttered, the stronger one does not hold; i.e., 'some' finds its upper bound and its meaning is pragmatically enriched to 'some but not all'. In this case, if I use the sentence "*I ate some candies.*" to describe a situation where I ate all the candies, my statement becomes underinformative. Moreover, such pragmatic enrichment can make utterances seem awkward, infelicitous, or even untruthful. The pioneering work of Noveck (2001) and Bott & Noveck (2004) serves as a good example here. They provide quantitative evidence that sentences like "*Some elephants have trunks*" are generally perceived as less truthful because pragmatic enrichment of 'some' to 'some but not all' contradicts common knowledge, i.e., the fact that all elephants should have trunks (see also Bambini & Domaneschi, 2024; Degen & Tanenhaus, 2015; Röhrig, 2010; Mazzaggio et al., 2022; Holtgraves & Kraus, 2018; Khorsheed et al., 2022; Panizza et al., 2009; Spector, 2013; Marty et al., 2013, for other studies and overviews of implicatures based on scales). Consequently, in Chapter 10, I examine scalar implicatures in more detail, focusing on their relation

to atypicality inferences and general pragmatic ability.

On the other hand, the second part of the Maxim of Quantity (*the interlocutors should not make their messages more informative than is required*; or in terms of the I-principle: *say as little as necessary*), postulates avoiding overinformative messages (Grice, 1975; Horn, 1984; Levinson, 2000). A particular subtype of overinformativity that is especially relevant to the present discussion is *informational redundancy*. Kravtchenko (2022) defines informational redundancy as a problem of excessive wordiness or precision, where speakers provide more detail than is strictly necessary, resulting in low informational utility of the message. Informational redundancy can be illustrated through cases of overinformative nominal modification (e.g., saying *a small fluffy dog* to identify the only dog available in a given context; Kravtchenko & Demberg, 2022a), or, as in the case of atypicality inferences, through saying something that is already strongly implied by shared world knowledge.

In this dissertation, I follow Kravtchenko (2022) in treating informational redundancy as a specific form of overinformativity. Consequently, I use the two terms interchangeably when referring to atypicality inferences.

Under pragmatic theories, overinformative messages are often considered suboptimal or infelicitous. However, in natural dialogue, overinformativity is surprisingly common. Comprehenders often tolerate it more easily than underinformativity, and redundant utterances are not necessarily judged as unacceptable[6]. For example, in a direction-giving experiment, Baker et al. (2008) found that a large proportion of utterances were redundant, particularly when speakers interacted with strangers or when the listener signaled a lack of understanding. This type of redundancy involved repetitions in which speakers reiterated information to ensure comprehension when they perceived that their interlocutor did not understand. Similarly, Walker (1993) analyzed the recordings of a call-in financial radio show and observed a high proportion of informationally redundant utterances. She argues that such utterances address limitations of human attention and working memory capacity, serve narrative functions, and make elements of the common ground salient again for the comprehender. A range of production studies, including those focusing on nominal modification, might initially suggest that speakers generally follow the principles of rational communication, preferring to omit typical or inferable information in favor of atypical aspects, such as the mention of non-conventional instruments (*to stab with an icepick vs. a knife;* Grigoroglou & Papafragou, 2016), materials (*wool vs. ceramic bowl;* Mitchell et al., 2013), shapes (*octagonal vs. round mug*; Mitchell et al., 2013), or colors (*pink vs. yellow banana*; Sedivy, 2003; Rubio-Fernández, 2016). However, when overinformativity does occur, several studies show that it can also serve a useful function in communication. Pragmatic enrichment enables listeners to efficiently narrow the search space in a visual scene and disambiguate the target object from potential competitors, as demonstrated in studies using a reference game setup (Rubio-Fernández, 2016; Tourtouri et al., 2019; Vogels et al., 2020; Rohde et al., 2021; Li et al., 2023).

[6]Cf. Davies & Katsos (2010, 2013).

A recent work by Rees & Rohde (2023) discusses another type of inference related to overinformativity, which does not require targeting specific classes of words or the "prototypicality" of object properties. Instead, it focuses on overinformativity that occurs at the level of events and their informational utility. They consider simple utterances like "*I saw that the library walls are blue.*" and claim that under certain conditions such utterances do not fulfill the requirements of informativity without pragmatic enrichment. Pragmatic enrichment, in turn, leads to the inference that what is informative in this utterance is not merely the color of the walls, but the event of the walls having been painted a new color, implying that the walls used to be different. Rees & Rohde (2023) identify the knowledgeability of the speaker (or, more precisely, the listener's understanding of the speaker's knowledge state) as the condition under which the above inference arises. For instance, when the speaker discusses a location they are familiar with (e.g., their school), pragmatic enrichment occurs because the listener assumes the speaker knows the usual state of the location. Conversely, if the speaker refers to a place they are not expected to visit regularly but maybe only once (e.g., the prime minister's house), pragmatic enrichment does not happen. To quantify pragmatic enrichment, Rees & Rohde (2023) asked participants if they thought the situation was the same before or not (e.g., whether the walls were blue a few months ago or not). The proportion of responses indicating that the walls used to be different (pragmatic enrichment) was higher for familiar locations (the school) compared to unfamiliar locations (the prime minister's house). They conclude that pragmatic enrichment happens when the utterance violates listener's expectations of informativity, which is tied to listener's awareness of the speaker's knowledgeability.

Indeed, the concept of informativeness is deeply connected with the **mutual knowledge** and beliefs of interlocutors about each other. This can manifest in various ways, such as through shared encyclopedic knowledge and assumptions about what is accessible to the interlocutor, or through more specific beliefs about the conversational partner's experiences or expertise, which help determine what they are likely to know or not know. Methodologically, this means that experimental materials should account for this factor by clearly specifying the speaker's and listener's knowledge, as well as their mutual awareness of each other's informational states, in order to accurately model overinformativity. In turn, at the level of pragmatic processing, the way informativeness is packaged for the interlocutors can influence the cognitive cost associated with inference derivation – a topic I explore in more detail in Chapters 3 and 4.

For example, in the study of Rees & Rohde (2023), which I described above, overinformativity of the utterance is achieved using the following setup. The target utterance "*I saw that the library walls are blue.*" is spoken by a schoolgirl named Suzy to her father. In the familiar location condition, the context is defined as follows:

> Suzy went back to school after the summer holidays and is telling her dad about her day at school.

In turn, in the unfamiliar location condition, the conversational context changes to the location that Suzy is not familiar with:

> Suzy went on a school trip to the prime minister's offices is telling her dad about her day at the prime minister's offices.

In both conversational context conditions, Suzy and her father share mutual knowledge regarding Suzy's awareness of the wall color. The context incorporates this mutual awareness either implicitly (since it is a conversation between two closely related people – a father and daughter) or explicitly in the text, through a description of the location Suzy visited. Importantly, all of this is also accessible to the experimental participants, who act as overhearers in the setup: Suzy is not addressing them, but rather speaking to her dad.

Rees & Rohde (2023) show that common ground between speakers and listeners is an important factor influencing the interpretation of utterances and their perceived overinformativity (see also Degen & Tanenhaus, 2019, for a review of factors influencing pragmatic processing; I also consider some of these factors in Chapter 4). Moreover, the informativeness of a message is shaped not only by shared knowledge, but also by the situational, conversational, and linguistic context[7].

Another factor that strongly guides the perceived informativity of an utterance is **background world knowledge** or prior beliefs, which people continuously rely on in language processing (Degen & Tanenhaus, 2019; Winograd, 1972). World knowledge influences not only language production (as discussed above, speakers tend to omit nominal modifiers when the information is shared; e.g., they typically avoid saying "*yellow*" when referring to a banana), but also language comprehension. Participants' world knowledge directly influences their judgments of sentence truth-values in scalar implicatures. For example, the pragmatic interpretation of the utterance "*Some elephants have trunks.*" contradicts world knowledge, and thereby renders such utterances underinformative[8].

In practice, however, it is often difficult to determine what individual comprehenders know and do not know, especially when working with a random sample of participants tested in the laboratory or online. In this sense, background world knowledge can be unsystematic and individual-specific, making it challenging to control for in the experimental setting (Kravtchenko, 2022). This matters because an inference might fail to arise not because participants reject it or fail to derive it due to cognitive limitations, but simply because they do not consider the initial message to be overinformative due to specifics of their world knowledge.

As described in Section 2.1, Kravtchenko (2022) addresses this bottleneck by using script knowledge as a proxy for background world knowledge. Script knowledge refers to a structured set of expectations about the typical sequence of events involved in

[7]See Horn (1991, 1993, 2014) for the discussion of how redundancy can also result from the restating of facts that are already presupposed.

[8]In addition, world knowledge has been shown to guide reference resolution (Kehler et al., 2008; Chambers et al., 2004) and the interpretation of declarative sentences (Hagoort et al., 2004).

familiar activities, such as going to a restaurant or grocery shopping, and is considered more stable and widely shared among individuals (Minsky, 1975; Schank & Abelson, 1975; Schank, 1980; Abelson, 1981)[9]. In the following section, I provide background on the concept of script knowledge.

2.2.1 Script knowledge

Script knowledge has long been a focus in the field of discourse coherence and causal inference, used to study coherence during narrative text comprehension and to investigate the understanding and prediction of events in text (Zwaan et al., 1995; van den Broek, 1990b). It has also been significant in the field of natural language processing, where it helps to improve machine understanding of narratives and is related to AI knowledge representation (Schank & Abelson, 1975; Barr, 1981; Modi, 2017; Wanzare, 2020; Ostermann, 2020; Hong et al., 2024).

Scripts can be viewed as specialized components of world knowledge that focus on routine daily activities, capturing temporally structured sequences of events (or actions) in stereotypical scenarios such as going to a restaurant or going shopping (Schank & Abelson, 1975; Bower et al., 1979; Wanzare, 2020; Kravtchenko, 2022). A script describes events related to the activity in a progressive temporal order from an activity's initiation to its completion. For example, Figure 2.4 illustrates the internal structure of the *going to a restaurant* script.

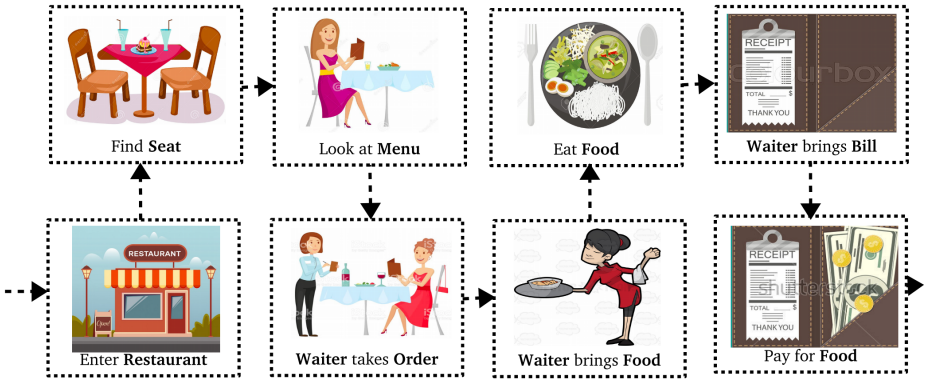


Figure 2.4: Example of a restaurant script showing part of internal script structure. Source: Wanzare (2020).

A wide body of research on script knowledge highlights several key properties of these structured knowledge representations, particularly in relation to human memory and language comprehension. It has been shown that once the script is invoked, the event sequence from this script is also activated – that is, comprehenders anticipate a certain course of events once they encounter the topic (Bower et al., 1979; Zwaan et al., 1995).

[9] Cf. Section 12.1.

Importantly, when an event is omitted from the description of such routine activities, comprehenders do not generally interpret this as indicating that the event did not occur; rather, they assume it was implied (see e.g., Hong et al., 2024). Consequently, comprehenders often fill in missing details during recall based on their script knowledge (Bower et al., 1979). These findings highlight the crystallized (or, in other words, deeply entrenched) nature of scripts: comprehenders may struggle to recall which events were explicitly mentioned and which were merely implied, underscoring how strongly scripts shape memory (Zwaan et al. 1995; see also Graesser et al., 1979).

Overall, these findings demonstrate that scripts strongly influence attention and memory, playing a pivotal role in information processing. Building on this, Kravtchenko (2022) showed that when script-related information is made explicit, it can, in certain contexts, be perceived as informationally redundant. This perceived redundancy may constitute a violation of the Maxim of Quantity, prompting listeners to derive atypicality inferences.

Chapter 3

Derivation process of atypicality inferences

The derivation scheme described in this chapter was first published in (Kurch et al., 2024):

Kurch, C., **Ryzhova, M.**, & Demberg, V. (2024). *Large language models fail to derive atypicality inferences in a human-like manner.* In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, & Y. Oseki (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 86–100). Bangkok, Thailand: Association for Computational Linguistics. doi: 10.18653/v1/2024.cmcl-1.8

Current theoretical accounts of pragmatic processing offer competing perspectives on how implicatures are derived and under which conditions they arise. Some theories suggest that certain implicature types are derived automatically and with minimal cognitive effort, while other implicatures are treated as inherently effortful (Levinson, 2000). Still other accounts adopt a more context-dependent perspective, viewing the cognitive effort associated with pragmatic inferences as a function of contextual support (Sperber & Wilson, 1996; Degen & Tanenhaus, 2019). While these accounts provide a good theoretical foundation for understanding pragmatic processing, none explicitly address atypicality inferences, nor do they specify in detail which comprehender-specific factors may influence the derivation process.

At the same time, empirical research provides a body of findings on methodological, contextual, and individual-level factors that can shape implicature derivation (see, e.g., De Neys & Schaeken, 2007; Feeney & Bonnefon, 2012; Marty et al., 2013; Cho, 2020; Fairchild & Papafragou, 2021). These findings suggest that variability in pragmatic processing may be related to differences in the cognitive and personality-related characteristics of comprehenders. However, such factors have not yet been investigated in relation to atypicality inferences.

In this chapter, I, therefore, specify the potential derivation steps of atypicality inferences in more detail and identify the cognitive and personality-related factors that may be involved in each step of this process. This approach is necessary in light of the theoretical and empirical gaps noted above: investigating why atypicality inferences may vary across individuals requires a structured basis for understanding what the derivation process may involve. In Chapters 4 and 5, I return to these steps, elaborating further on these factors, including their potential contribution to processing cost, and linking them to the theoretical accounts of pragmatic processing and to the empirical evidence on individual differences in pragmatic processing.

While Kravtchenko (2022) alludes to the reasoning process involved in deriving atypicality inferences, the steps of this process are not explicitly formalized. Building on Kravtchenko's findings and related literature on informational redundancy and pragmatic processing, I define the following steps that capture the reasoning path comprehenders may follow when deriving atypicality inferences:

1. identify redundancy based on script knowledge;
2. recognize that this redundancy is infelicitous, as it violates conversational norms;
3. infer that the event mentioned in the target utterance is atypical;
4. and explicitly accommodate this atypicality within the situational context.

The goal here is to identify which cognitive and personality-related factors may be involved in each step, under the assumption that variation in these factors may modulate the likelihood or strength of deriving atypicality inferences across individuals.

In what follows, I examine these steps in more detail. **The first step** is to gather all the necessary information required to identify redundancy. As noted in Section 2.2, whether a piece of information (fact or event) is redundant depends on both the situational and conversational contexts, as well as the interlocutors' knowledge. Consequently, collecting, filtering, and processing this relevant information may engage executive functions, particularly working memory. Working memory is responsible for temporary storage of information in a readily accessible form during tasks such as comprehension, reasoning, and problem solving (Cowan, 2014), and has been shown to modulate comprehension in a variety of language tasks. Moreover, as I will discuss in Chapter 5, executive functions, and in particular working memory, have been extensively studied in relation to pragmatic processing, especially for phenomena such as scale-based implicatures (De Neys & Schaeken, 2007; Dieussaert et al., 2011; Marty et al., 2013; Marty & Chemla, 2013; Antoniou et al., 2016). One point of contention concerns the nature of script knowledge, which underpins atypicality inferences. As discussed in Section 2.2.1, script knowledge consists of tightly interconnected mental representations that encode stereotypical everyday situations through event sequences. Because these script-related events are strongly causally

linked, as shown in production and comprehension studies (Bower et al., 1979; Zwaan et al., 1995; Wanzare, 2020; Hong et al., 2024), once a script is evoked in a narrative – such as *going shopping* – the entire sequence of relevant events tends to be activated immediately. Consequently, all necessary information may already be readily available for the comprehender to evaluate the redundancy of the target utterance, potentially reducing the involvement of additional memory resources at this stage.

The second step concerns the Gricean perspective on producing redundant information. It arises from the Maxim of Quantity under which redundancies should be avoided in rational communication (Grice, 1975; Levinson, 1987). In other words, the comprehender must recognize that the utterance identified as redundant violates conversational norms. Realizing such a violation may draw on socio-pragmatic abilities (e.g., theory of mind), as well as on individual differences in autistic traits and higher-level reasoning abilities such as cognitive reflection and fluid intelligence.

After recognizing that the utterance is redundant and therefore violates conversational norms, **the third step** involves drawing an atypicality inference as part of an accommodation process. In other words, the comprehender *repairs* the utterance that might otherwise be viewed as infelicitous by positing that the information is not truly redundant but, instead, signals something unusual or unexpected. By doing so, the comprehender reconciles the apparent violation of the Maxim of Quantity, assuming that the speaker had a reason for including the seemingly superfluous information – namely, that the event in question does not usually happen. The cognitive mechanisms involved in the second step may also be engaged in the third step, as the comprehender proceeds from recognizing redundancy to drawing an inference.

By **the fourth step**, I propose that the derivation process does not end once atypicality is identified. Because any conversation or narrative unfolds within a specific situational context, it becomes necessary to reconcile the newly inferred atypicality with the established script. In other words, once the inference disrupts the script, it must be accommodated rather than left in conflict with typical expectations. For example, in the case presented in Table 2.2, John might be judged as a habitual shoplifter or as someone who uses self-checkout machines instead of paying the cashier directly. I contend that this accommodation step is crucial for maintaining the inference derived in the third step; if no explanation for the atypical behavior is provided, the inference may ultimately be canceled (Garmendia, 2023). Finally, as I argue in Chapter 5, this process may be closely related to higher-level reasoning abilities, such as cognitive reflection and fluid intelligence, among other cognitive factors.

It is important to note that the investigations in this dissertation do not aim to confirm the precise nature of the derivation process for atypicality inferences. The proposed steps are intended to guide the selection of cognitive and personality-related factors that may contribute to variability in atypicality inference derivation. These factors are further discussed in Chapters 4 and 5, and are subsequently tested in the experimental chapters.

Chapter 4

Accounts of pragmatic processing and processing cost

This chapter was published as part of the following article: **Ryzhova, M.**, & Demberg, V. (2023). *Processing cost effects of atypicality inferences in a dual-task setup. Journal of Pragmatics*, *211*, 47–80. 10.1016/j.pragma.2023.04.005. The text has been included with minor modifications for formatting and consistency with the dissertation.

In this chapter, I adopt the perspective of theoretical accounts of pragmatic processing that investigate whether implicature derivation is cognitively costly or can proceed with little or no effort. The accounts considered include the Gricean framework (Grice, 1975), the Default account (Levinson, 2000), the Relevance theory (Sperber & Wilson, 1996), and the Constraint-Based account of Degen & Tanenhaus (2019). These accounts make different predictions about processing cost depending on implicature properties, and, for example, on the degree of contextual support available for implicature derivation.

I consider how their predictions can be extended to atypicality inferences, with particular focus on the properties of these inferences that may render their derivation effortful. This perspective connects to the main goal of the dissertation, namely to identify the factors that influence the derivation of atypicality inferences, and in particular to the Research Question 3, which asks whether their derivation is affected by externally constrained cognitive resources. I show that, while for scalar implicatures, these two groups (the Default account and the alternative contextualized accounts of Sperber & Wilson, 1996 and Degen & Tanenhaus, 2019) differ in their predictions, they agree on the cost for atypicality inferences.

In Section 4.1, I move to the methodological side and review the dual-tasking experimental paradigm, which has been widely used to test predictions about processing cost, primarily of scale-based implicatures. In Chapter 9, I adapt this methodology to atypicality inferences in order to test whether constraining cognitive resources affects

their derivation.

Under Grice's model of rational communication, the listener, in their search for the speaker's meaning, is guided by the Cooperative Principle, grounded in four maxims of conversation. In his work, Grice described the procedure that a listener should follow to grasp the speaker's intended meaning (Grice, 1975). For example, when an implicature is triggered by a scalar term such as *some*, as in "*Mary ate some apples*", according to Grice, the utterance chosen by the speaker is compared with alternative utterances that the speaker could have produced. In this case, a relevant lexical alternative to *some* is *all*, since *some* can be interpreted as *some or all*. The listener can then reason about these alternatives while taking into account the Cooperative Principle, and in particular the Maxim of Quantity. This maxim requires that the speaker be maximally informative and therefore use *all* if they believe it to be true. From the fact that the speaker did not choose the more informative alternative, the listener may infer that *all* is not true; the resulting implicature is that Mary ate *some but not all* apples. However, Grice himself did not consider the proposed steps as a psychologically grounded theory of implicature derivation, but rather a philosophical discussion of the steps involved (Zufferey et al., 2019; Mazzarella, 2014). Later theories were based on Grice's work and attempted to provide more cognitively plausible accounts of pragmatic processing.

The proponents of the Default account (Horn, 1972; Chierchia, 2004; Levinson, 2000) emphasized the distinction between generalized and particularized conversational implicatures (GCIs and PCIs, respectively – see Huang, 2012; Recanati, 2004, for an overview). GCIs were claimed to be widely independent of context, possessing a clearly defined set of lexical alternatives – as in the *"Mary ate some candies"* example, where the alternatives are tied to {some; all} scale. According to Levinson (2000), these inferences are assumed to occur frequently and are therefore precompiled and virtually free of cost. In other words, based on this account, people have more exposure to the scalars and their lexical scales, which at the very least reduces the derivational cost of GCIs compared to PCIs (since the set of alternatives in GCIs is available by default) and at the very most makes it automatic and completely cost-free (Chierchia, 2004). In turn, PCIs were featured as contextually dependent implicatures, where the set of alternatives is strongly tied to the conversational context. They were viewed as extremely costly, as they are different in each context, and hence the derivation cannot be trained or automatized.

According to the PCI/GCI classification, atypicality inferences are clearly particularized, given their high degree of contextual specificity. The alternatives depend heavily on the specific situation in which the actor is placed. For instance, consider the scenario in Example 4.2, in which Lisa brought her swimsuit when going swimming.

(4.2) *Today, Lisa went to the swimming pool [...]* ***She brought her swimsuit!***

The alternative in this case might be that she does not usually bring her swim-

suit, perhaps suggesting that she is absent-minded. In contrast, in a context where it is stated that Lisa paid the cashier when grocery shopping (as in Table 2.2), the alternative might be that she does not typically pay, for example because she steals food or uses self-checkout machines. While the core of the inference tends to follow a recurrent pattern (namely, that the actor is not typically involved in the target event), the subsequent accommodation step, in which the comprehender integrates the inferred atypicality into the situational context, is highly context-sensitive (as discussed in Chapter 3). Each scenario may support multiple plausible alternatives for how the atypicality is interpreted (e.g., being forgetful, or going to a naturist swimming pool where a swimsuit is not required). Given that, at the time of derivation, comprehenders may generate and evaluate several context-specific alternatives, the Default account predicts that atypicality inferences should be cognitively costly.

In contrast to the Default account, Relevance theory (Sperber & Wilson, 1996) abandons the PCI/GCI distinction and reconciles them in the following way: under this framework, Gricean maxims are reduced to one single principle of communication relevance. Every utterance raises the expectation of relevance which listeners seek to satisfy by picking up the best hypothesis about what the speaker could mean by saying this utterance. Further, the listener's search for the best hypothesis is determined in terms of cost and benefits. Benefits reflect arriving at the speaker's true meaning, while the cost is the processing effort that the listener meets while searching for the meaning. Thus, Relevance theorists emphasized the role of contextual support in both PCI and GCI. An utterance said in a neutral context is predicted to be more costly, as it requires greater effort in pragmatic enrichment. However, if a pragmatic interpretation is primed by the context, it can also be less costly than a literal interpretation. According to this line of reasoning, the findings that the scalar implicature similar to *"Mary ate some candies."* can be effortful (e.g. shown in Dieussaert et al., 2011; Antoniou et al., 2016; Bott & Noveck, 2004) is explained from the position that it is not sufficiently primed in the context to avoid the effortful search for relevance.

From the relevance-theoretic perspective (Sperber & Wilson, 1996), processing of atypicality inferences would be predicted to proceed as follows: during a conversation, listeners integrate their representations of the world with the linguistic signal. Their assumptions about the world have different strengths that can dynamically change during the conversation. Crucially, new assumptions can be formed when processing utterances. Before committing an assumption to memory, a cognitive system (they call it a deductive device) checks whether this assumption is already there (e.g., that people usually bring their swimsuits when going to the pool). If the strength of the current assumption is not very high, repetition has no effect other than strengthening it. However, when a speaker utters information that is already highly predictable from the listener's assumptions (which is the case in bringing one's swimsuit in the *going swimming* context), to preserve relevance, the listener applies the additional step of processing the repetition and deriving extra contextual effects (namely, that

Lisa does not usually bring her swimsuit for a particular reason). Pragmatic enrichment of informationally redundant utterances should hence be associated with some processing costs.

Recently, Degen & Tanenhaus (2019) proposed a new way of understanding pragmatic inferences called the "Constrained-based account". This approach aligns with Relevance theory, but additionally proposes a set of factors that influence how humans infer meaning. Their paper focuses on scalar inferences and proposes that the likelihood and speed of making a scalar inference depend on how much support it receives from the surrounding context. If the context supports the pragmatic meaning, the probability of making an inference increases, but if the context does not support it, the probability decreases.

In the first place, implicature computation depends on the listener's ability to distinguish between different states of the world that can be potentially reported by the speaker. For example, by saying *"Mary ate some candies"*, two prospective states of the world are that Mary ate *all* candies and that Mary ate *some but not all* candies. In conversation, the relevant states of the world can be highlighted to interlocutors via the conversational or situational context, which can give rise to a question under discussion (QUD). An example of a QUD in the context of the utterance "*Mary ate some apples.*" could for instance be "Did Mary eat *all* apples?". In this case, an alternative for *some* (namely *all*) is explicitly highlighted and the implicature that Mary ate some but not all apples is more likely to arise. Consider, on the other hand, the QUD "Did Mary eat *any* apples?" – here, the quantifier *any* makes the implicature irrelevant (shown experimentally, for example, in Yang et al., 2018; Kursat & Degen, 2020).

Furthermore, the computation depends on the properties of the target utterance and its alternatives (what could have been said instead of what the speaker said). According to Degen & Tanenhaus (2019), the target utterance can be characterized by its production cost and/or informativeness. When the listener has access to the alternative utterances (for example via the QUD), they can reason about the cost that the speaker underwent for the chosen formulation compared to the cost the speaker would have undergone if an alternative would have been uttered. Imagine, for instance, that the word *all* would be more costly to utter than the word *some*. In this case, the less informative utterance "Mary ate *some* candies." would be less likely to give rise to the implicature (shown for reference game setup in Rohde et al., 2012).

The other group of factors that form the conversational context is how the listener judges the speaker. For instance, if the listener knows that the speaker lacks a reliable source of information about exactly how many candies Mary ate, then the implicature is less likely to arise (see Geurts, 2010, for discussion). Next, assuming that the speaker is properly informed, the implicature computation further depends on whether the speaker ostensively tries to be less informative than they could. As a reminder, the use of less informative terms (e.g., *some* compared to *all*) is not consid-

ered a lie but only restricts the interpretation to be greater than zero. For example, in the context of Mary being a guest at someone's house where the candies were meant for the host's children, the speaker's intention behind saying *"Mary ate some candies."* can be to save Mary's face, i.e., to stay polite with respect to Mary who might feel embarrassed for taking something not clearly offered, rather than to informatively report that Mary ate all the candies. If the listener judges the speaker in this way, then the implicature is less likely to arise (shown experimentally, for example, in Bonnefon et al., 2009). Finally, even when the speaker is maximally knowledgeable and has no intention to be underinformative, the use of a less informative term *some* can be attributed rather to speaker's inability to be properly informative, than to the actual intent to convey an implicature that Mary ate some but not all candies (for example, if the speaker is not a native user of the language; see Fairchild et al., 2020).

Additionally, world knowledge in the form of listeners' prior beliefs can modulate implicature computation. Teresa Guasti et al. (2005) discuss that the experimental materials, where the domain of target world knowledge is not clearly defined, can mask the implicature. For example, in the statement "*Some giraffes have long necks.*" the implicature computation might depend on a subset of giraffes one has in mind. If the comprehender considers a population of giraffes compared to animals with shorter necks (e.g., lions), then the statement with "some" is underinformative and should be rejected because in fact all giraffes have long necks. However, if the comprehender divides the population of giraffes into baby and adult ones, the implicature in the statement "*Some giraffes have long necks.*" is masked: the statement is fairly informative since adult giraffes have longer necks compared to baby giraffes. Another example of the interplay of world knowledge and scalar implicature computation was presented by Degen et al. (2015). They show that scalar implicatures can cause revision of prior knowledge about the world. In their study, subjects were asked to judge how many of the objects (e.g., marbles or feathers) sank in water after hearing an utterance with *some*: "*Some objects sank in water.*". Given the fact that marbles are expected to sink in water more than feathers, one expects prior knowledge about marbles to dominate the implicature (meaning that subjects would answer that all marbles sank). However, as Degen et al. (2015) show, subjects rather tended to make the implicature about marbles (that some but not all marbles sank) – to the same extent as for feathers (e.g., by assuming special properties of the water or the material marbles were made of).

Finally, the implicature computations depends on what information the speaker and the listener share in common ground: does the speaker know that the listener knows their communicative intentions, implied QUD, do they share the information about how they judge each other, do they have any communication conventions, what is known about Mary or, for example, candies etc. For example, Rees & Rohde (2023) manipulated knowledgeability of the speaker and showed that the speaker's familiarity with the context can give rise to pragmatic inferences (see Section 2.2, for a detailed discussion of this study).

To summarize, when presented out of context, the scalar implicature triggered by "*Mary ate some candies.*" does not receive any contextual support, according to Degen & Tanenhaus (2019): both prospective meanings, *all candies* and *some candies*, are equally likely – neither the question under discussion, speaker's personality, potentially relevant world knowledge, nor common ground are made explicit. Thus, such implicatures should be effortful.

Next, I consider what role, if any, the factors discussed in Degen & Tanenhaus (2019) might play in atypicality inferences. In the experimental materials, the conversational context introduces all characters as knowing each other well. Other than that, the context describes an everyday activity and is written in a neutral manner such that none of the actors placed in this context (e.g., the person who speaks about Mary, the recipient of the utterance, or Mary herself) carry any specific properties in terms of their personality, cooperativity or reliability. Similarly, the context does not address the question under discussion – no contextual information is given that can reveal what the speaker wants to convey by uttering the informationally redundant utterance (see Section 2.1.1 and Chapter 6, for the description of experimental materials). Consequently, the set of possible alternatives is not explicit, contrary to scalar implicatures – in the materials used for investigating atypicality inferences, it is completely up to comprehenders' inferencing process. The atypicality inferences are, in fact, triggered by the redundancy of the utterance, but this is different from contextual support of the actual inference: claiming that the context supports the inference would amount to saying that the mention of going swimming supports the inference that the swimmer in the story does not usually bring their swimsuit, which seems like a crazy claim to make.

Moreover, one might expect even stronger effects of cognitive burden on pragmatic processing here than for scalar implicatures. It is crucial to keep in mind that, compared to atypicality inferences, scalar implicatures, nevertheless, are based on the lexical scales – independently of how the conversational context is defined. In neutral contexts, the set of alternatives is exhaustively formed based on lexical scales. However, in non-neutral contexts, the implicature is derived still based on the scale – albeit accessed in context (Foppolo et al., 2021). In turn, the note raised by Degen & Tanenhaus (2019) about the contextual factors is especially crucial for the particularized pragmatic inferences which arise very situatively, in unique contexts. The set of alternatives consequently might be hard to properly constrain in the listener's mind which can lead to a processing delay or reduced rate of the inferences. This note is one more reinforcement for examining implicatures that do not underlie lexical scales.

On the other hand, one should not expect that contextual support is a cure-all factor for successful pragmatic processing. Following the propositions of Relevance theory, anything that can be relevant for inferencing (according to Degen & Tanenhaus (2019), either coming from the properties of alternatives, speaker's personality or anything else) first needs to be retrieved and then assessed (Sperber & Wilson, 1996). Thus, the inference-relevant information might be ignored or go unnoticed

when comprehenders are placed in a situation where they do not have enough cognitive resources.

Taken together, the reviewed accounts support the prediction that atypicality inferences should incur processing cost. This follows from their contextual dependence and the need to construct and evaluate situation-specific alternatives, which are not lexically specified. In contrast to scalar implicatures, where the set of alternatives is constrained by lexical scales, atypicality inferences require comprehenders to rely more heavily on background knowledge and reasoning, which may increase processing demands. At the same time, these predictions remain largely theoretical, as the processing cost of atypicality inferences has not yet been investigated. It is therefore necessary to test whether atypicality inferences are, in fact, costly during comprehension. One approach in the previous literature on pragmatic processing is to examine the effects of externally constraining cognitive resources, most commonly through dual-tasking paradigms that target working memory. If atypicality inferences rely on cognitive resources, limiting these resources should affect their derivation.

In the next section, I review the dual-tasking paradigm, which has been widely used to test predictions about processing cost in the domain of scalar implicatures. This provides the methodological background for the experimental investigation in Chapter 9, where the dual-tasking approach is applied to test whether constraining cognitive resources affects the derivation of atypicality inferences.

4.1 Dual-tasking methodology in processing cost research

In the literature on memory, dual-tasking has been shown to effectively tap the executive and attentional resources. The deficit in corresponding sub-components of memory negatively affected performance of the tasks (Holding, 1989; Baddeley et al., 2009, 1991).

The idea is that the secondary task will compete for cognitive resources with the pragmatic processing task, for instance, in terms of working memory, which as I discuss in Chapter 3, might modulate the strength of atypicality inferences. If pragmatic processing is effortful in the sense that working memory resources are needed in order to retrieve relevant alternatives and combine the textual information with situational information and world knowledge, pragmatic inferences can be expected to be attenuated when fewer such resources are available during processing.

The idea stems from the Relevance theory claim (and was further broadened in the Constrained-based account of Degen & Tanenhaus (2019) by providing concrete factors forming the contextual support) that when any implicature is not primed in neutral contexts, pragmatic enrichment requires cost for analytical thinking in which working memory is involved (Sperber & Wilson, 1996). In addition, in the studies of individual differences in pragmatic processing, working memory capacity has been

found to modulate the success of GCI derivation (Dieussaert et al., 2011; Feeney et al., 2004). Thus, in the absence of sufficient working memory capacity, the rate or strength of inferences should become lower than in the condition with no cognitive burden. In the Default account of Levinson (2000) the same line of argumentation is applicable for PCIs, whereas GCIs are claimed automatic – so, no cost is predicted for them in the dual-tasking setup. I discuss the role of working memory capacity in pragmatic processing in more detail in Section 5.1.

One of the pioneering studies devoted to scalar implicatures by De Neys & Schaeken (2007) examined the rate of pragmatic interpretations in the absence of context. Subjects were presented with underinformative sentences ("*Some tuna are fish*") and were asked to decide if the sentence was true or false. False responses signified computation of a scalar implicature. As a secondary task, they used a classic spatial storage task, namely the dot memory task (Bethell-Fox & Shepard, 1988). Before encountering an underinformative sentence, subjects were asked to memorize a dot pattern presented in a 3x3 grid. After answering the target question, they had to reproduce the pattern. In the low load condition, the dots were vertically or horizontally aligned, which made them easier to memorize, see Figure 4.1. In the high load condition, the matrix contained a rather complex non-linear pattern. The results suggested that under a high load condition, the rate of pragmatic responses was lower than under a low load condition, supporting the claim about the implicatures' costliness in neutral contexts. This finding constitutes evidence against the Default account.

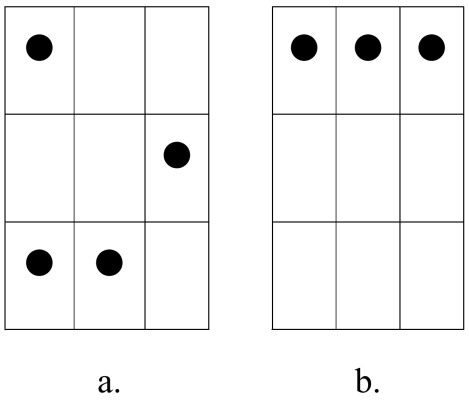


Figure 4.1: Example of the dot memory task setup used by De Neys & Schaeken (2007). Panel (a) shows the dot pattern used in the high load condition. Panel (b) shows the dot pattern used in the low load condition. Source: De Neys & Schaeken (2007).

In contrast, Fairchild & Papafragou (2021) did not find any effect of load on scalar implicatures, using the same memory dot secondary task. Their experiment contrasted effects of working memory with effects of theory of mind. They found that both working memory capacity and theory of mind correlated with judgments of scalar implicatures but only ToM had a unique effect, while the effect of WMC disappeared once ToM was included as a predictor. Fairchild & Papafragou (2021)

suggest that working memory modulates the implicature computation via ToM but not directly – one would need sufficient working memory capacity to carry out ToM computations, or would need sufficient working memory resources to hold results in memory. Similarly, Marty et al. (2013) showed that tapping participant's memory resources interferes with the derivation of scalar implicatures, which speaks against the Default account of Levinson (2000). Participants in their study were told to memorize the sequence of letters before the main task (four letters in the high load condition vs. two letters in the low load) and reproduce it afterwards. The study did however not find any cost associated to inferences about numerals. In a nutshell, the results of Marty et al. (2013) for scalars support contextualized accounts of pragmatic processing (Degen & Tanenhaus, 2019; Sperber & Wilson, 1996), while for numerals the Default account is supported (Levinson, 2000). This shows that even the implicatures that have been, for a long time, claimed conceptually similar – both scalar and numeral implicatures are based on lexical scales and they were classified as GCIs – can still behave differently.

To achieve a stronger interference between the main and secondary tasks, Cho (2020) used a linguistic secondary task (reading span task) in the study of processing cost during online comprehension of scalar implicatures. Their finding also supports the contextualized accounts of pragmatic processing. Based on the analysis of self-paced reading times measured at the final sentence region of underinformative sentences (e.g., *Some birds have wings and* ***beaks***.), there was a difference in reading times between a pragmatically felicitous and a pragmatically infelicitous condition in the no-load condition. This effect however disappeared under increased memory load, indicating that participants may be less likely to recognize the pragmatic problem when under load which in line with contextualized accounts (Degen & Tanenhaus, 2019; Sperber & Wilson, 1996).

To summarise, the results of the previous studies to a large extent disfavor the Default account of Levinson (2000) – under cognitive resources deficit, the implicatures were less likely to arise (as measured by the rate of pragmatic responses) or took significantly more time for processing (as measured in reaction times). In most studies, such resource limitations are implemented through tasks that tax working memory, suggesting that working memory resources are particularly relevant for implicature derivation. Importantly for my investigations, these results show that the dual-task paradigm is in principle well-suited for detecting cognitive load associated with pragmatic implicatures and also that many types of pragmatic implicatures remain underresearched, with most studies focusing on scalar implicatures. In particular, there are no previous dual-task studies investigating particularized implicatures.

The present dissertation extends this approach to atypicality inferences. In Chapter 9, I apply the dual-tasking paradigm to investigate whether their derivation is sensitive to constraints on working memory resources.

Chapter 5

Inter-individual variability in pragmatic processing

The previous chapter reviewed theoretical accounts of pragmatic processing and their predictions regarding the cognitive cost of implicature derivation. Extending these accounts to atypicality inferences, their derivation is expected to be cognitively demanding, given their particularized nature. At the same time, it was noted that these predictions remain largely theoretical and require empirical testing, which is addressed in Research Question 3.

Against this background, the present chapter turns to a complementary perspective on variability in pragmatic processing, as formulated in Research Question 4. Rather than manipulating cognitive resources, this approach examines whether natural variability in comprehenders' cognitive and personality traits can account for variation in the derivation of atypicality inferences. The theoretical framework outlined in Chapter 3 provides a basis for identifying which factors are likely to be relevant in this respect. Building on this framework, the present chapter reviews candidate factors that have been proposed in the literature to account for individual differences in pragmatic processing. These include components of executive function, such as working memory capacity, inhibition, and memory updating; socio-pragmatic abilities (such as theory of mind or autistic traits); higher-level reasoning abilities (fluid intelligence and cognitive reflection); as well as linguistic experience. While previous research has primarily focused on scalar implicatures, it remains an open question which factors modulate the derivation of atypicality inferences. The goal of this chapter is therefore to provide an overview of these factors and to motivate their inclusion in the empirical investigations reported in Chapters 8 and 10.

To the extent that individuals exhibit consistent biases toward a logical or pragmatic meaning, the question that follows is whether such tendencies can be attributed to traits specific to an individual. Research on the role of individual differences in

implicature processing has identified two broad accounts in relation to response preferences – cognitive factors and personality traits (Antoniou et al., 2016). The cognitive account is based on the assumption that implicature generation is effortful (Bott & Noveck, 2004); hence, variation in cognitive resources may influence the extent to which individuals are able to compute an implicature. In line with this, studies examining the interpretation of *some* using a dual-task methodology (e.g., with a secondary dot-tracking task) have found that implicature rates decrease as task complexity increases (De Neys & Schaeken, 2007; Marty et al., 2013). This suggests that cognitive factors such as working memory capacity (WMC) may influence the ability to generate an implicature. However, it has been noted that these studies were not designed as a direct test of the role of WMC, and hence do not discount the possibility that effects may have been due to a "third variable" unaccounted for. Antoniou et al. (2016) suggest that WMC may impact implicature processing via a Theory of Mind (ToM) component; that is, comprehenders have to track a speaker's epistemic state in order to reason that they intend to convey the implied (but unspoken) meaning – an inferential process that is slow and costly. This is supported by recent findings that higher ToM skills are related to stronger tendencies in comprehenders to draw a pragmatic implicature (Fairchild & Papafragou, 2021).

The personality account, on the other hand, emphasizes the role of personality traits, in particular real-world pragmatic skills, as a source of variation in implicature generation. Nieuwland et al. (2010) for instance showed that the magnitude of an EEG response to underinformative *some* (e.g., "Some people have lungs") is modulated by pragmatic ability (as measured by the Autism-Spectrum Quotient), with pragmatically skilled participants showing a larger N400 effect in response to underinformativeness. Personality-based factors such as self-perceived honesty have also been reported to vary with interpretation of the scalar term *or* (Feeney & Bonnefon, 2012).

In this chapter, I review cognitive and personality factors relevant to the derivation of implicatures, with a focus on atypicality inferences. This provides the motivation for the experiments on individual differences reported in Chapters 8 and 10.

5.1 Working memory capacity

In the early stages of research on pragmatic processing, working memory capacity (WMC) emerged as one of the key factors that could potentially influence the derivation of pragmatic implicatures, in particular those relying on lexical scales. Early studies on scalar implicatures reported longer response times in cases when subjects derived an implicature compared to when they did not (e.g., Noveck et al., 2001; Bott & Noveck, 2004; Bott et al., 2012; Noveck & Posada, 2003). These findings were among the first to suggest the presence of a processing cost associated with deriving scalar implicatures. Several cognitive processes have been proposed to contribute to

this cost: deriving a scalar implicature may require taking the speaker's knowledge state into account, overriding an initial literal interpretation, and coordinating linguistic and contextual information. Working memory capacity has therefore been proposed as one of the factors that may modulate the success of these processes and, consequently, the successful derivation of scalar implicatures. De Neys & Schaeken (2007) further used a dual-tasking methodology to directly manipulate the load imposed on working memory and found that scalar implicature processing draws on working memory resources. Under conditions of reduced memory resources, subjects computed fewer scalar implicatures, while the load did not affect the responses to semantically true and pragmatically felicitous sentences. Using a similar paradigm, Dieussaert et al. (2011) subsequently found that only individuals with low working memory capacity were affected by the cognitive load that comes from the secondary task and derived fewer pragmatic responses when the load was increased.

On the other hand, Marty et al. (2013); Marty & Chemla (2013) found that when an implicature is contextually supported, a secondary task does not affect its derivation (the rate of derived implicatures remains the same between load conditions). Based on these findings, Marty et al. (2013) propose that working memory may have limited involvement in the actual process of deriving implicatures. Instead, its primary role appears to lie in the decision-making process of whether to derive the implicature.

Antoniou et al. (2016), however, argue that contextual factors can facilitate not only the decision-making process but also the inferential processes involved in deriving implicatures: contextual support can make both processes easier or introduce complexities that impose additional cognitive load on working memory. Antoniou et al. (2016) were the first to test scalar implicatures in an individual differences paradigm, showing that participants with higher working memory capacity derived more implicatures than those with lower capacity. Further, Yang et al. (2018) found that working memory capacity predicts participants' sensitivity to context in implicature derivation, thus connecting the findings of Marty et al. (2013) and Antoniou et al. (2016).

There is, however, evidence that the role of WMC may be auxiliary, in that it supports perspective-taking processes required for successful implicature derivation (Fairchild & Papafragou, 2021; Lin et al., 2010). Fairchild & Papafragou (2021) tested scalar implicatures, metaphors, and indirect requests and found that, across all three phenomena, working memory capacity made no unique contribution to implicature derivation once theory of mind abilities were taken into account. Thus, working memory may be involved either in supporting theory of mind computations themselves or in holding and integrating their outcomes with other information during the interpretation of pragmatic meaning. The role of theory of mind in pragmatic processing is further discussed in Section 5.5.

Fairchild & Papafragou (2021) further discuss that, given their findings, it may be necessary to investigate the specific subcomponents of executive functions and their

role in pragmatic processing. They acknowledge that some accounts consider working memory as just one subcomponent, alongside, for example, inhibition and memory updating (Miyake et al., 2000; Cooper, 2010). Given this perspective, Fairchild & Papafragou (2021) propose that these subcomponents should be tested to gain a more comprehensive understanding of their roles in pragmatic inference. A potential role of inhibition and memory updating is discussed in Sections 5.2 and 5.3, respectively.

While recent evidence challenges the direct involvement of working memory in the implicature derivation, it remains a key component of executive function that supports a range of cognitive processes. Working memory has been shown to play a role in various language processing phenomena. Therefore, it is important to assess individual differences in working memory capacity. Moreover, certain aspects of atypicality inference derivation may require executive resources. For instance, activating relevant world and script knowledge (such as understanding that grocery shopping involves payment) and linking this knowledge to the narrative.

Reading span test (RSpan)

Previous studies on scalar implicatures and other types based on lexical scales have used a range of tasks to measure different aspects of working memory capacity, ranging from visuo-spatial to verbal working memory (see Kane et al., 2004; Conway et al., 2005, for discussion of different tasks). In line with Cho (2020); Antoniou et al. (2016), I use an RSpan test that focuses on the verbal component of working memory (Caplan & Waters, 1999; Just et al., 1996; Cooper, 2010). In this task, individuals are asked to judge the grammaticality of several sentences while holding the final words of the sentences in memory. They are then asked to recall the final words of each sentence in a series in the order in which they were presented. The task thus involves both language processing and storage components Scholman et al. (2020).

Materials. The materials were taken from Scholman et al. (2020), who designed the items to closely resemble those employed by Waters (1996) and Waters et al. (1987). The material consisted of 56 sentences (a range of $[8, 11]$ words). Half of the sentences contained a verb that required an animate object (e.g., *to scare*, *to confuse*), while the other half required an animate subject (e.g., *to escape*, *to donate*) – see examples in Section 9.3.3.

Experimental instructions. Participants were presented with a sequence of sentences and asked to judge whether they were acceptable. After evaluating two or more sentences, they were prompted to recall the final words of the sentences. Participants were informed that both tasks (acceptability judgments and word recall) were important, but that performance on the acceptability task was more critical. They were instructed to perform the task as accurately as possible and to do their best on the recall task.

Afterward, participants were provided with examples illustrating which sentences should be judged as acceptable and which should be rejected. Before the main experi-

ment, they completed a test trial involving only the acceptability task, which consisted of eight sentences presented sequentially. Following this, participants completed two practice trials.

Subjects were reminded of the instructions throughout the experiment.

Experimental procedure. The experiment started with an acceptability task, in which participants were presented with eight sentences sequentially. In this task, they were not required to recall the final words of the sentences. The purpose of this task was to measure the average time each participant took to judge the acceptability of a sentence, which then served as a time threshold (i.e., a cutoff) in the main experiment. If a participant exceeded their mean judgment time plus two standard deviations, a time-out message appeared and the next sentence was presented.

In the main experiment, participants completed 16 trials and saw a total of 56 sentences. The number of sentences per trial ranged from 2 to 5, with four trials for each set size. The order of trials, as well as the order of sentences within each trial, was randomized for each participant.

Measure of performance. Following Scholman et al. (2020), Conway et al. (2005), and Friedman & Miyake (2005), working memory capacity was assessed using the partial-credit unit (PCU) scoring method. PCU was calculated as the mean proportion of words within a set that were correctly recalled, with values ranging from 0 to 1. Minor typographical errors were permitted, allowing for a one-character difference between the correct answer and the provided response.

5.2 Inhibition

Inhibition refers to the ability to suppress the most immediate or dominant/ natural reaction to the stimulus in favor of a more appropriate behavior that is consistent with the current goal (MacLeod, 2007; Friedman & Miyake, 2004; Friedman et al., 2006). Research has demonstrated a correlation between inhibition and higher-level executive functions (Perret, 1974; Dimitrov et al., 2003; Kim et al., 2018), as well as expressive language abilities (Thomas et al., 2022).

Revisiting Fairchild & Papafragou (2021)'s assertion regarding the potentially critical role of executive functions' subcomponents for pragmatic processing, it's worth noting that there are indeed two prominent theoretical accounts that explore the role of inhibition within the broader framework of executive function (see Robert et al., 2009, for a comprehensive review).

The first, so-called, resource account posits that deficiencies in inhibition are primarily driven by a reduced working memory capacity. This perspective suggests that inhibition, by itself, consumes cognitive resources (Conway & Engle, 1994; Engle et al., 1995). Therefore, individual differences in inhibition may stem from variations in working memory capacity rather than differences in inhibitory control per se (Engle et al., 1995; Roberts et al., 1994).

On the other hand, the second account emphasizes the independent role of inhibition and connects individual differences in inhibitory control to the mechanisms of inhibition (Bjorklund & Harnishfeger, 1995; Harnishfeger, 1995; Hasher et al., 2007). According to this perspective, inhibition plays a significant role in estimating working memory capacity, in the sense that it may determine which processes are no longer goal-oriented and, therefore, should not access memory resources (May et al., 1999).

Taking both theoretical accounts into consideration, it is interesting to examine a direct impact of inhibition on processes related to pragmatic language use. It's worth noting, however, that experimental evidence supporting the role of inhibition in pragmatic processing is still relatively limited and is primarily derived from research conducted in neuroatypical populations and from studies on language development (for example, see Rints et al., 2015, who finds individual differences in inhibition predicting pragmatic language use in ADHD children). In particular, a portion of the evidence stems from studies on figurative speech, e.g., in the context of metaphors (Sana et al., 2021; Chiappe & Chiappe, 2007). In these studies, there is the hypothesis that inhibition may play a role in suppressing the literal interpretation of an implied metaphor. For instance, in the statement "lawyers are sharks," the literal interpretation would suggest that lawyers physically resemble fish (e.g., that they have fins and/or other fish-like physical qualities), while the metaphorical meaning implies that lawyers are aggressive. Thus, Chiappe & Chiappe (2007) discovered that both working memory and inhibition predicted the quality of metaphor interpretation and the time required to come up with such interpretations. Moreover, the group analysis revealed that the disparities in interpretation quality between different working memory groups diminished in significance when measures of inhibitory control were considered as covariates. This implies that individual differences in inhibitory control underlie differences between the working memory groups. Similarly, Carriedo et al. (2016) showed in children that both working memory and inhibitory control became increasingly demanding for comprehending more difficult metaphors and for individuals with limited reading experience and lower levels of semantic knowledge.

Taking into account the findings of the aforementioned studies, it is conceivable that inhibitory control may play a similar role in inhibiting the literal interpretation of informationally redundant utterances. Specifically, it may aid in rejecting the straightforward and commonsense course of events (such as "Lisa should always bring her swimsuit to the swimming pool because it's what people usually do") in favor of deriving atypicality inferences.

However, it is worth noting that the role of inhibition may not be crucial for deriving all types of pragmatic inferences. For instance, Antoniou et al. (2016) did not find evidence for the involvement of inhibition in deriving scalar implicatures in neurotypical adults, when participants were presented with underinformative statements with 'some'[1].

[1]In their study, Antoniou et al. (2016) employed a sentence-picture verification task that included statements like 'There are stars on some of the cards.' These statements were underinformative when

Stroop task

The phenomenon known as the Stroop effect was firstly demonstrated by the psychologist John Ridley Stroop (Stroop, 1935). It reflects the human tendency to experience challenges in naming the actual color of a word when that word spells out a different color's name. This discovered effect subsequently served as the foundation for developing a psychological test, known as the Stroop test (e.g., used to measure inhibitory control in Miyake et al., 2000; Kane & Engle, 2003; Friedman et al., 2008; Friedman & Miyake, 2004). This task has become widely employed for assessing inhibitory control, including the studies of pragmatic phenomena (e.g., in Chiappe & Chiappe, 2007; Antoniou et al., 2016)[2].

Materials. Materials and the experimental setup were adapted from Miyake et al. (2000); Kane & Engle (2003); Friedman et al. (2008); Friedman & Miyake (2004). Five colors (red, green, blue, yellow, purple) and corresponding words were used in the experiment. The experimental list comprised of 60 congruent trials (where the color matches the word; e.g., the word 'green' was presented in green ink) and 40 incongruent trials (where the color of the word is different to the word, e.g., the word 'purple' is presented in green ink). In the congruent trials, each color/word appeared 12 times, while in the incongruent trials, each non-matching color/word combination appeared 2 times. The order of trials was randomized for each subject.

Experimental instructions. Subjects were instructed to identify the colors of words by pressing the corresponding buttons. Further, they were familiarized with the keys (where the words were printed in the matching ink):

- r for red words
- g for green words
- b for blue words
- y for yellow words
- p for purple words

Two textual examples were provided to the participants, which were followed by a brief practice session (five words in each, congruent and incongruent, conditions). Subjects were explicitly instructed to provide their answers as fast as possible, as they were timed.

presented in the context where participants were shown 5 cards, and all 5 cards did have stars on them.

[2]Some other tasks commonly utilized in literature to assess inhibitory control are also built upon the fundamental observation that humans tend to exhibit delayed reactions when the stimulus and the required response are incongruent either in terms of differences in color (as in the Stroop test), spatial location (as in the Simon task; Simon & Rudell, 1967), or directional alignment (as in the Flanker task; Eriksen & Eriksen, 1974).

Before the study began, participants were asked to decline participation if they had been diagnosed with any form of color vision deficiency. Following the experiment, participants were once again asked to disclose any color vision deficiency diagnoses they might have had. They were informed that this disclosure would not affect their payment for participating in the experiment.

Experimental procedure. After the practice session, participants completed 100 trials of the main experiment.

Each trial consisted of a word displayed in colored ink against a black background, positioned at the center of the screen. Within each trial, there was an initial 500 ms period of a black screen, followed by the presentation of a white fixation cross (+) for 200 ms, and then another 100 ms of a black screen. The stimulus, which was the word, remained on the screen for a maximum of 2000 ms.

If a participant did not submit a response within this time frame, a "Wrong" feedback message was displayed for 500 ms, after which the next trial began. If a response was provided within the time limit, feedback (either "Correct" or "Wrong") was displayed for the same duration.

Then the new trial began.

Measure of performance. Performance was assessed by calculating the mean latency difference between incongruent and congruent conditions for each participant, reflecting the approach used in previous studies (Stroop, 1935; Antoniou et al., 2016; Friedman & Miyake, 2004). Consequently, more negative differences indicate weaker inhibitory control.

5.3 Memory updating

Memory updating is another construct of executive functions. It was shown to correlate with working memory in both verbal and visuo-spatial working memory tasks (Miyake et al., 2000; St Clair-Thompson & Gathercole, 2006; Ecker et al., 2010). However, it is argued to measure a different aspect of executive function and reflect subjects' ability to effectively monitor and update working memory representations (Morris & Jones, 1990). According to Miyake et al. (2000); Morris & Jones (1990), it is responsible for monitoring the incoming information in terms of its importance and suitability, and subsequently refreshing the information in the working memory (WM) by replacing old and irrelevant information with newer, more relevant information.

The current research in experimental pragmatics has not yielded a significant amount of studies regarding memory updating and its relation to pragmatic processing. As an exception, recent findings of Schuster et al. (2023) can be considered, where they find that the memory updating capacity correlates with the magnitude of semantic-pragmatic adaptation in usage of uncertainty expressions like "probably" and "might".

However, memory updating capacity has been shown to play an important role in reading comprehension for both adults and children (Butterfuss & Kendeou, 2018; Muijselaar & de Jong, 2015; Whitely & Colozzo, 2013; Linares & Pelegrina, 2023; Palladino et al., 2001; Carretti et al., 2009). Research in narrative comprehension and causality relations argues that mental models of situations described in the text must be updated as new information about the unfolding course of events is encountered (Morrow et al., 1987; van den Broek, 1990a; Graesser et al., 1997; Radvansky et al., 2014; Zwaan et al., 1995). Subsequently, Radvansky & Copeland (2001) found that memory updating was the strongest predictor (among other executive functions tested) of a successful situation model updating during reading. Similarly, the results of Palladino et al. (2001) suggest that poor reading comprehension might be associated with poorer memory updating capacity. They find that poor comprehenders produced lower recall and more errors in memory updating tasks than good comprehenders[3].

Although the above studies on reading comprehension do not specifically focus on implicature derivation, their results provide a good foundation to hypothesize that memory updating can, in fact, be a good predictor of individual differences in the computation of atypicality inferences. The present experimental materials consist of short narratives about everyday situations, so it is expected that individuals will integrate the forthcoming events into their situational model during reading. When the informationally-redundant utterance is encountered, subjects who noticed its over-informativity should update and substitute their assumptions regarding the target event typicality. In addition, the derived atypicality of the target event may necessitate a broader contextual shift in relation to the story characters and/or narrative circumstances, which should also be updated in the memory. Thus, it can be hypothesized that if memory updating capacity is insufficient, atypicality inferences may be rejected or not derived at all.

Keep track task

The keep track task is one of the tasks used to measure memory updating capacity (originally proposed by Yntema, 1963). The version of the task used in Chapter 10 is described below and follows the implementation reported in Miyake et al. (2000), adapted from Yntema (1963).

Materials. The vocabulary size was six categories (animals, colors, countries, distances, metals, and relatives) with six items per category:

- **animals:** bear, dog, horse, lion, pig, wolf
- **colors:** blue, green, orange, purple, red, yellow

[3]Poor and good comprehenders were selected according to their performance in the standardized test for assessing reading comprehension in Italian language (Cornoldi et al., 1991) that included reading short passages of text and answering multiple choice questions

- **countries:** Australia, Brazil, China, France, India, Japan
- **metals:** gold, silver, bronze, copper, zinc, iron
- **distances:** meter, mile, inch, kilometer, foot, yard
- **relatives:** mother, father, brother, sister, aunt, uncle

Experimental instructions. Before starting the task, participants were introduced to six categories along with their corresponding items to familiarize themselves with the materials. They were then given instructions specifying that in each trial they would be presented with several categories and shown 15 words from the list. Their task was to remember the last words presented in each of the specified categories and type them at the end of the trial. Participants were also given a textual example, as well as a practice trial with feedback. They were instructed not to use any external means to aid memorization and instead to rely solely on their own memory.

Experimental procedure. The task comprised a total of 12 trials. At the beginning of each trial, participants were presented with the names of either 2, 3, or 4 categories (four trials per set size; order randomized) that they were required to keep track of. These category names remained displayed at the bottom of the screen throughout the trial. Once participants were ready, they initiated the trial by pressing a "Start" button. Subsequently, 15 items were presented sequentially on the screen, with each item displayed for 1500 ms. The items were selected randomly but always included 2 or 3 items from each of six possible categories. After viewing all 15 items, participants entered the most recent words they recalled for each of the target categories. They then received feedback, which included the tracked categories, their responses, and the correct responses. Additionally, they were informed of the number of correctly recalled words. The next trial then began.

Measure of performance. The score was calculated as the number of correctly recalled words, with a maximum of 36. Higher scores indicate greater memory updating capacity.

5.4 Socio-pragmatic abilities

Socio-pragmatic abilities have been suggested to influence implicature derivation in a number of studies (see, for example Nieuwland et al., 2010; Yang et al., 2018; Feeney & Bonnefon, 2012; Katsos & Bishop, 2011; Mazzaggio & Surian, 2018). The idea is rooted in findings that individuals with autism spectrum disorder can exhibit pragmatic deficits, indicating difficulties in deriving pragmatic implicatures (Baron-Cohen, 1988). As suggested by Baron-Cohen (1997), autistic individuals may have a reduced ability to infer the mental states of others. In turn, the ability to mentalize (also known as 'theory of mind' – a factor discussed in Section 5.5) is often considered crucial for implicature derivation (see Grice, 1975, and the communicative principle).

In studies with neurotypical populations, socio-pragmatic abilities, particularly autistic traits, have been shown to correlate with pragmatic responses. Individuals with higher levels of autistic traits may be less likely to adopt the interlocutor's perspective or reason about why something was said, and may therefore respond more literally (Yang et al., 2018). For example, Nieuwland et al. (2010) showed that participants with higher levels of autistic traits were less sensitive to underinformativity. Similarly, Mazzaggio & Surian (2018) found that an autism-related cognitive phenotype is negatively associated with a tendency to spontaneously derive scalar implicatures [4].

Given the observed correlation between socio-pragmatic abilities and the derivation of scalar implicatures, it is reasonable to hypothesize that this factor may also be linked to the derivation of other types of inferences, such as atypicality inferences. For example, individuals with higher levels of autistic traits may be less likely to derive an atypicality inference, as they may have more difficulty recognizing the pragmatic markedness of an informationally redundant utterance.

The autism spectrum quotient

Socio-pragmatic abilities are often assessed using the Autism Spectrum Quotient (AQ) proposed by Baron-Cohen et al. (2001b). This self-administered test has been widely used not only to distinguish neuroatypical from neurotypical populations, but is also claimed to be sensitive to more subtle differences within neurotypical populations. In the pragmatic processing literature, this measure has been widely used to investigate individual differences in neurotypical adults (Nieuwland et al., 2010; Antoniou et al., 2016; Mazzaggio & Surian, 2018; Yang et al., 2018).

The AQ is a self-administered test designed to assess the extent to which adults with typical IQ exhibit what are commonly referred to as 'autistic traits'. These include reduced social and communication skills, limited imagination, exceptional attention to detail, and difficulties with attention switching or a strong focus of attention. The questionnaire has been validated in both neurotypical adults and individuals on the autism spectrum, providing not only binary classification but also a continuous measure of autistic traits.

Materials. The materials consisted of 50 statements that participants were asked to rate based on whether they applied to them, using a 4-point Likert scale ("definitely agree", "slightly agree", "slightly disagree", "definitely disagree"). These statements were designed to encompass five categories, each containing 10 items, corresponding to areas of cognitive differences associated with autism at the time the test was developed (American Psychiatric Association, 1994; Rutter, 1978; Wing & Gould, 1979):

[4] It is important to note, however, that these experimental findings are not consistent. Antoniou et al. (2016) and Heyman & Schaeken (2015) do not find effects of socio-pragmatic abilities on the derivation of scalar implicatures.

- social skill
- attention switching
- attention to detail
- communication
- imagination

Approximately half of the items were designed to elicit a "disagree" response, while the remaining items were intended to elicit an "agree" response in individuals with higher levels of autistic traits.

Experimental instructions. Participants were instructed to read 50 statements and indicate how much each statement applied to them. They were asked to read the statements carefully but not to spend too much time on each one.

Experimental procedure. Following the instructions, participants were presented with 50 statements displayed in a pseudo-random order, as specified by the original authors of the test (Baron-Cohen et al., 2001b). The task could only be completed once participants had rated all statements.

Measure of performance. The total Autism Quotient (AQ) score was calculated as the number of statements to which participants responded in a manner consistent with higher levels of autistic traits. The total score ranged from 0 to 50. In addition, subscale scores were calculated by summing such responses within each category, yielding scores from 0 to 10. Higher scores on both the total AQ and the subscales indicate a stronger presence of autistic traits (Baron-Cohen et al., 2001b).

5.5 Theory of mind

Theory of Mind (ToM) refers to the cognitive ability to attribute mental states to oneself and others. Having a model of mental states involves being able to represent both one's own experiences (beliefs, intentions, and knowledge) and those of others (Goldman, 2012). This ability allows individuals to explain and predict others' behavior, i.e., to perform 'mindreading'. A crucial aspect of this capacity is recognizing another individual as an intentional agent – in other words, being aware that one's own mental state is not identical to that of another person (Carlson et al., 2013).

The capacity to derive pragmatic implicatures has been extensively discussed in relation to ToM (Bosco et al., 2018). As pragmatic abilities involve an inferential process that bridges the gap between what a speaker literally says and what they actually mean, it is reasonable to assume that implicature derivation involves attributing mental states and using this information to predict and recognize the speaker's communicative intentions (Grice, 1975).

The role of ToM has been highlighted in a range of studies in neuroatypical populations; for example, in individuals with schizophrenia, where impairments in figurative language comprehension have been linked to ToM deficits (Mo et al., 2008; Champagne-Lavau & Stip, 2010; Bosco et al., 2012). However, the role of ToM remains controversial; see Bambini et al. (2016) and Parola et al. (2018), who report no evidence of ToM impairment in a substantial proportion of patients with schizophrenia who nevertheless exhibit pragmatic deficits.

In neurotypical populations, a similar pattern emerges: while some studies emphasize the role of mentalizing abilities in making pragmatic inferences (namely, scalar implicatures, see Nieuwland et al., 2010; Mazzaggio & Surian, 2018; Fairchild & Papafragou, 2021), others find no significant effects of ToM on pragmatic processing (Antoniou et al., 2016).

I hypothesize that deriving atypicality inferences, in particular accommodating informational redundancy and forming explanations, may involve assessing the speaker's knowledge state. For example, interpreting why Mary behaves in an unexpected way (e.g., she does not usually eat in restaurants because she goes there only for drinks or follows a lifestyle that limits her spending) may require the comprehender to consider the context and the knowledge states of the interlocutors. In particular, the speaker and the listener may be assumed to share knowledge about Mary's typical behavior, such as the fact that she does not usually eat at restaurants.

The reading the mind in the eyes test

The reading the mind in the eyes test (RMET) was developed by Baron-Cohen et al. (2001a) to measure the ability to represent the mental states of others. The test employs a forced-choice format. Participants are instructed to select the word that best describes the emotion conveyed by the eyes in a picture where the rest of the face is not visible. The test has been shown to be applicable to both neuroatypical individuals and neurotypical adults, and is therefore widely used in psycholinguistic research. For example, Fairchild & Papafragou (2021) used this test to investigate individual differences in the derivation of different implicature types, including scalar implicatures, metaphors, and indirect requests.

However, existing discussions of the test (see e.g., Quesque & Rossetti, 2020) suggest that it may be overly centered on a particular aspect of mind-reading, namely the assessment of theory of mind in the context of emotional judgments, as illustrated in Figure 5.1. Based on their discussion, it can be suggested that the test may only indirectly relate to the broader capacity to attribute mental states, while placing greater emphasis on the recognition of facial expressions.

Nonetheless, this test has two distinct advantages. First, it is sensitive not only to differences in neuroatypical populations but also to variation among neurotypical adults with typical IQ. This contrasts with many clinical ToM tasks, in which neurotypical adults often reach ceiling performance (e.g., Frith & Corcoran, 1996). The

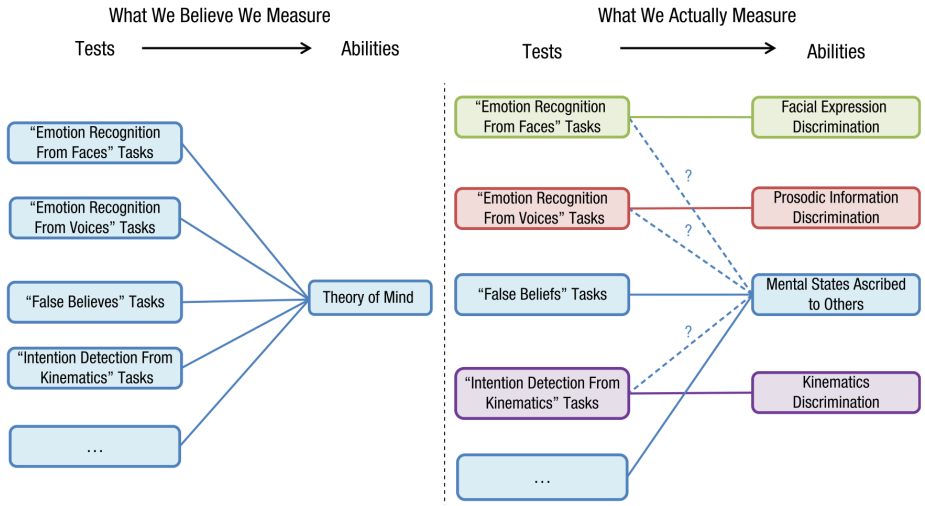


Figure 5.1: Illustration highlighting the possibility that many classic theory of mind tasks may rely on lower-level cognitive processes. Source: Quesque & Rossetti (2020).

RMET provides a range of scores, allowing a more nuanced assessment of individual differences in neurotypical populations.

In a similar vein, many other tests primarily focus on child development or are designed to assess theory of mind deficits in children on the autism spectrum. For example, the Strange Stories test proposed by Happé (1994) examines mental state inferences derived from narratives, in which participants are asked to provide context-appropriate explanations for a character's behavior. However, evaluations of this test suggest that although it is sensitive enough to distinguish between neuroatypical and neurotypical children and adults, neurotypical adults without clinical diagnoses typically achieve ceiling performance, making it difficult to investigate individual variability[5].

Fairchild & Papafragou (2021) used this test in combination with the RMET to assess theory of mind abilities in neurotypical adults and to examine their relation to pragmatic implicature derivation. However, the authors do not report the distribution of scores for the Strange Stories test in isolation. Instead, they provide a composite score combining the RMET and Strange Stories tests. As a result, it is not possible to determine the extent of variability attributable specifically to the Strange Stories test, or whether participants in their sample reached ceiling performance, thereby contributing most of the variability to the RMET.

Furthermore, it is worth noting that tests designed for children often incorporate scenarios and contexts that may not align with the social experiences of adults. This mismatch may also influence performance on such tasks.

[5]Another example is the Sally-Anne task developed by Baron-Cohen et al. (1985), which is designed to evaluate false belief attribution in children and is not suitable for adults.

Below, I provide a detailed description of the reading the mind in the eyes test, including its materials and experimental setup.

Materials. The materials used in the test were taken from Baron-Cohen et al. (2001a) and consisted of 37 images of the eye region, each associated with four emotion words. For each image, only one option was correct.

Experimental instructions. Participants were informed that the test assessed their ability to recognize emotions from images of the eyes. They were instructed that they would be presented with an image and should select the most appropriate emotion from four options. Before proceeding to the 36 experimental trials, participants completed one practice trial and received feedback on the correctness of their response. During the experimental trials, no feedback was provided, but at the end of the experiment participants were given their total score (i.e., the number of correctly answered items).

Experimental procedure. Each trial consisted of an image depicting the eye region, displayed on the right side of the screen. Four emotion options were shown around the image, one in each corner. On the left side, the same options were presented as clickable radio buttons, which participants used to select their responses – see Figure 5.2. After selecting an answer and clicking the 'Next' button, the next trial began.

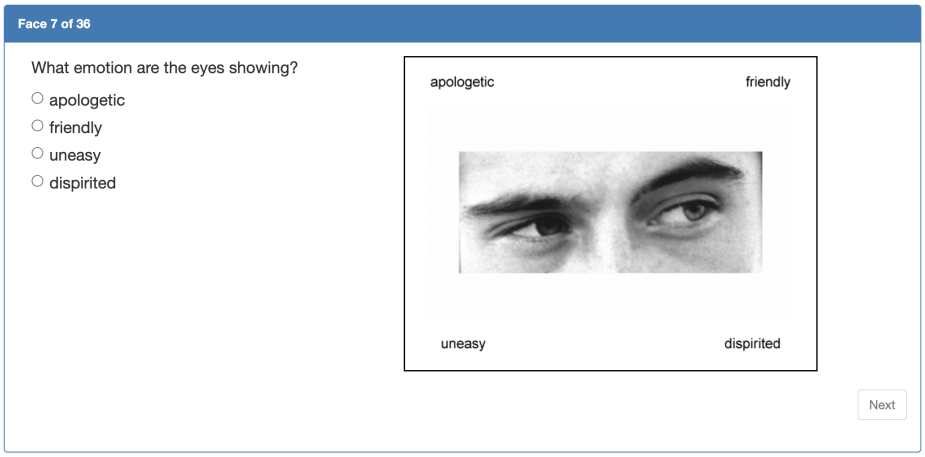


Figure 5.2: An experimental trial of the RMET.

Following the test, participants were asked a set of additional questions that were later used to exclude some participants from the analysis:

- Have you taken this test before?
- If you are not a native speaker of English, did you recognize all the words used to describe emotions?
- Did you encounter any technical difficulties or interruptions during the study?

- Did you use any external resources or in any way provide inaccurate information? (It is acceptable if you used a dictionary to look up some words.)
- Do you have any additional comments for the researcher (e.g., questions, suggestions, or concerns)?

Measure of performance. Following Baron-Cohen et al. (2001a), participants' performance in the test was assessed based on the number of correctly answered items, resulting in scores ranging from 0 to 36. Higher scores indicate greater theory of mind ability.

5.6 Cognitive reflection

In their work, Frederick (2005) defined cognitive reflection as '*the ability or disposition to resist reporting the response that first comes to mind*' (Frederick, 2005, p. 35). More broadly, Frederick (2005) situates cognitive reflection within dual-process theories of human cognition (Sloman, 1996; Stanovich & West, 2000; Kahneman & Frederick, 2002; see also De Neys, 2006).

In essence, these theories propose the existence of two cognitive systems: System 1, whose processes occur spontaneously and require fewer attentional resources, and System 2, which requires attention, effort, motivation, and concentration. Accordingly, System 1 is often characterized as 'automatic' and 'intuitive', whereas System 2 is described as 'reflective' and 'analytical', and is associated with cognitive reflection ability.

For example, consider the Bat and Ball problem provided below:

(5.3) *A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?* (Kahneman & Frederick, 2002, p. 58)

People often answer that the ball costs 10 cents – an intuitive and impulsive response driven by System 1. This answer is incorrect, as it would result in a total cost of $1.20 rather than the $1.10 specified in the problem. Frederick (2005) hypothesizes that arriving at the correct answer (5 cents) requires engagement of System 2 processes.

Toplak et al. (2014) further elaborate on dual-process theory by highlighting the trade-off between the two systems in terms of power and cost. System 2 allows for solving a wide range of problems, including novel ones, with a high degree of accuracy. However, this flexibility comes at a substantial cognitive cost. In contrast, System 1 has lower computational power, as it relies on heuristics that can lead to errors and biases, but requires fewer processing resources. Importantly, Toplak et al. (2014) discuss a hierarchical relationship between System 1 and System 2 computations. System 1 processes can run in parallel with those of System 2 without interference

and are often automatically triggered by certain types of tasks, such as the one in Example 5.3. System 2 processes, in contrast, can interfere with System 1, as they may need to override responses generated by System 1 (see also Evans & Stanovich, 2013). Drawing on this distinction, Toplak et al. (2014) suggest that cognitive reflection can be viewed as a thinking disposition that contrasts with 'miserly' processing. Individuals may differ in their tendency to invest mental effort in further reflection rather than relying on intuitive judgments.

To measure cognitive reflection ability, Frederick (2005) developed the Cognitive Reflection Test (CRT), which is described in more detail at the end of this section. The test consists of items that elicit an incorrect intuitive response, similar to the problem in Example 5.3. Arriving at the correct response requires resisting the intuitive answer and engaging in reflective thinking.

CRT scores have been shown to be good predictors of performance in rational thinking tasks. For example, Frederick (2005) originally evaluated the CRT in relation to risk and time preferences and found it to be one of the best predictors of performance in these domains, compared to other measures of cognitive ability[6]. Toplak et al. (2011) further conducted an extensive evaluation of the CRT across 15 tasks associated with rational thinking (including probabilistic and statistical reasoning tasks, as well as tasks involving cognitive and logical biases[7]). They found that CRT scores were the strongest predictor of performance across these tasks, surpassing measures of intelligence and executive function.

The findings of Toplak et al. (2011) suggest that cognitive reflection, as measured by the CRT, extends beyond differences in decision-making styles. This view is also reflected in related work (see, e.g., Campitelli & Labollita, 2010; Stanovich, 2012). Cognitive reflection appears to tap into a broader set of characteristics than originally proposed by Frederick (2005). For example, Stanovich (2012) propose a tripartite model that extends dual-process theories by distinguishing between intelligence, rationality, and reflective thinking as separate constructs. Accordingly, the CRT may capture variance beyond traditional measures of cognitive ability, as general intelligence alone may not fully account for the ability to initiate override and simulation processes (Stanovich, 2012).

It is worth noting, however, that more recent findings may challenge this view. For example, Welsh (2022) conducted an extensive correlation analysis of CRT scores with a range of individual difference measures, including intelligence (e.g., fluid and crystallized intelligence, memory abilities, quantitative and spatial abilities, processing

[6]The measures included the Scholastic Achievement Test and American College Test (as measures of academic achievement), the Wonderlic Personnel Test (general cognitive ability), and the Need for Cognition Scale (enjoyment of thinking); Cacioppo & Petty, 1982.

[7]The complete list of tasks included: Causal Base Rate, Sample Size: Hospital Problem, Sample Size: Squash Problem, Regression to the Mean, Gambler's Fallacy 1, Gambler's Fallacy 2, Conjunction Problem, Covariation Detection, Methodological Reasoning, Bayesian Reasoning, Framing Problem, Probabilistic Reasoning: Denominator Neglect, Probability Matching, Sunk Cost, Outcome Bias.

speed), personality (Big Five traits), decision-making styles (e.g., rationality, intuition, need for cognitive closure), and attentional abilities (e.g., focused, sustained, and divided attention). Welsh (2022) find that the CRT primarily reflects fluid intelligence, with additional associations with crystallized and quantitative abilities. However, it shows little unique variance beyond that shared with other cognitive measures.

In the domain of pragmatic processing, CRT scores have been shown to correlate with the rate of pragmatic responses in reference game tasks (Mayn & Demberg, 2022), while Heyman & Schaeken (2015) found that participants with higher CRT scores were more consistent in their responses to underinformative sentences. If atypicality inferences behave similarly, individuals with a greater cognitive reflection ability may be more likely to engage in the later stages of the derivation process, in particular inferring atypical behavior and accommodating it in context by generating explanations.

Cognitive reflection test (CRT)

Materials. The original version of the CRT proposed by Frederick (2005) consisted of three items. Subsequently, numerous extensions and revised versions of the test were developed (Primi et al., 2016; Baron et al., 2015; Sirota & Juanchich, 2018; Thomson & Oppenheimer, 2016; Toplak et al., 2014), as performance on the CRT has been shown to be sensitive to prior familiarity with the test (see also Stieger & Reips, 2016). A recent version was introduced by Mayn & Demberg (2022), which combines items from previously proposed CRT variants (Primi et al., 2016; Baron et al., 2015; Sirota & Juanchich, 2018; Thomson & Oppenheimer, 2016; Toplak et al., 2014). The revised test includes six critical items (three verbal and three numerical), along with four non-trick 'decoy' items. Examples of each question type are provided in Table 5.1.

Experimental instructions. Participants were instructed to respond to ten questions. They were encouraged to consider their answers carefully but not to spend too much time on each item.

Experimental procedure. Following the instructions, participants were presented with the questions in random order. To respond, they were required to type in their answers. Beneath each question, participants also indicated in an open-ended format whether they had encountered that question before.

Measure of performance. The CRT score was calculated as the number of correctly answered critical items out of those that participants reported not having seen before. Participants who indicated familiarity with more than three of the six critical items were excluded from the analysis.

Table 5.1: Examples of each question category in the revised version of the cognitive reflection test. Source: Mayn & Demberg (2022)

Question Type	Question	Correct answer	Distractor Answer
Verbal	If you have a match box with only one match in it and you walk into a dark room where there is an oil lamp, a newspaper and wood – which thing would you light first?	Match	Oil lamp
Numeric	Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?	29	30
Decoy	Sara, Emma, and Sophia embark on a river trip. Each of them brings one supply item for the trip: a kayak, a cooler of sandwiches, and a bag of apples. Sara brought the apples and Emma didn't bring anything edible. What did Sophia bring?	Sandwiches	NA

5.7 Fluid intelligence

In addition to the CRT, which has been argued to be closely related to fluid intelligence (Welsh, 2022), this section also considers fluid intelligence itself.

Fluid intelligence is a core component of human cognitive abilities, supporting deductive and inductive reasoning, particularly in novel situations (Kyllonen & Kell, 2017). According to the Cattell/Horn/Carroll theory of intelligence, it is contrasted with crystallized intelligence, which reflects declarative knowledge acquired through learning and experience (Cattell, 1963; Horn & Cattell, 1967; Carroll, 1993).

In studies on scalar implicatures, there is currently limited evidence that implicature derivation is modulated by fluid intelligence. For example, Spychalska et al. (2014) investigated the comprehension of scalar implicatures in a truth-value judgment task and identified two groups of comprehenders: "pragmatists" and "logicians". Importantly, these groups did not differ in non-verbal intelligence scores (see also Antoniou et al., 2016), who included fluid intelligence as a control factor to avoid potential confounds. A positive effect of reasoning ability on pragmatic responding has, however, been observed in a pragmatic reference game (Mayn & Demberg, 2022).

Given the strong relationship between fluid intelligence and cognitive reflection observed by Welsh (2022), I consider both factors as potentially influencing the derivation of atypicality inferences. Both can be viewed as contributing to a general capacity for reasoning (Stanovich, 2012). Deriving an atypicality inference requires reasoning about contexts in which an apparently redundant utterance would be informative (e.g., stating that Mary ate at a restaurant is not redundant if she usually only orders drinks). It is therefore hypothesized that abstract reasoning ability may modulate

the process of constructing a context that accommodates the inferred atypicality.

Raven's progressive matrices test

Raven's progressive matrices (Raven's IQ) is commonly used to measure non-verbal fluid intelligence and was originally developed by Raven (1936). Following Mayn & Demberg (2022), a shortened version consisting of 10 items is used, as it has been shown to be highly correlated with the full test (Bilker et al., 2012).

Materials. Ten items of increasing difficulty were selected from a full-length test, following Bilker et al. (2012). Each item consisted of a geometric pattern with a missing part. For each pattern, participants were presented with multiple options, only one of which correctly completed the pattern.

Experimental instructions. Participants were informed that they would be presented with an image of a pattern and that their task was to select the option that best completed it from a set of six to eight alternatives. In other words, they were to choose the option that best matched the pattern.

They were also provided with an example item, including the pattern, the response options, and the correct answer.

Subjects were also informed that the expected completion time for the test was 5 minutes and they were encouraged to aim to finish within this time frame. A clock was displayed at the top of the screen to help them keep track of time. However, they were assured that there would be no penalty for exceeding this time limit. The subjects were also informed that the questions were organized in order of increasing difficulty.

Experimental procedure. After reading the instructions, participants were presented with 10 items and were required to submit a response before proceeding to the next item.

Measure of performance. Performance was assessed as the proportion of correctly answered items, yielding scores ranging from 0 to 1, with higher values indicating higher fluid intelligence.

5.8 Exposure to print

Reading as a form of experience has been shown to support the development of a range of cognitive abilities, as demonstrated in a growing body of research (Liu et al., 2019; Stanovich et al., 1995; Kidd & Castano, 2013; Castano et al., 2021; Johnson & Arnold, 2021; Inohara et al., 2017; Ritchey, 2011). Despite this, research on language processing and comprehension has historically focused more on executive functions, particularly working memory capacity.

One possible explanation, discussed by Scholman et al. (2020), is that the literature often reports a strong correlation between working memory capacity and exposure to

print (see Peng et al., 2018, for a meta-analysis; see also Farmer et al., 2017; Stanovich et al., 1995). This has led to the view that variability in language experience may, at least in part, reflect differences in working memory capacity. However, this relationship has been argued to be potentially spurious, suggesting that additional factors may underlie the association between these constructs (Scholman et al., 2020; Farmer et al., 2017; Stanovich et al., 1995). From this perspective, variability in linguistic experience cannot be fully reduced to differences in working memory capacity. Therefore, as argued by Scholman et al. (2020), it is important to consider both constructs, working memory capacity and linguistic experience, to disentangle their respective contributions.

In pragmatic processing research, the role of print exposure has also been discussed in relation to metaphor production. For example, Chiappe & Chiappe (2007) found that children with greater exposure to print produced higher-quality metaphors. This finding has been linked to earlier work suggesting that print exposure supports vocabulary growth and the development of richer semantic networks, which are important for understanding figurative language (see also Stamenković et al., 2019).

In addition, Yang et al. (2018) investigated the role of language skills in the derivation of scalar implicatures with *some* in a story-sentence matching task[8]. They found no evidence of a relationship between language abilities and scalar implicature derivation. However, the authors hypothesized that language skills may still play a role in other types of pragmatic phenomenon, particularly those that arise in more linguistically rich contexts, where comprehenders are demanded to process more complex discourse and integrate larger amounts of textual input. From this perspective, atypicality inferences may provide a suitable test case to explore the effects of language skills.

In summary, there is currently limited evidence that language skills modulate pragmatic processing; the available literature is relatively sparse and tends to focus primarily on children and/or neuroatypical populations. However, a substantial body of research highlights the broader role of language experience in supporting various aspects of language and cognition. For example, linguistic experience has been shown to shape expectations about coherence relations (Scholman et al., 2020), support the understanding of semantic relationships and engagement in logical reasoning (Scribner & Cole, 1981; Stanovich & Cunningham, 1993), facilitate the comprehension of less frequent linguistic structures (Freed et al., 2017), and more generally enhance reading comprehension (Cipielewski & Stanovich, 1992) and verbal abilities (Stanovich & Cunningham, 1993; Stanovich et al., 1995). In this regard, individuals with greater linguistic experience may be more likely to possess richer semantic representations, which may facilitate the inference of atypicality and the generation of plausible explanations for atypical behavior during the accommodation step.

[8]Yang et al. (2018) used the Peabody Picture Vocabulary Test and the Author and Magazine Recognition Task to measure language skills.

Author recognition test

The author recognition test (ART), introduced by Stanovich & West (1989), is one of the commonly used measures of print exposure. In this task, participants are presented with a list of real and fictitious author names and must identify which correspond to actual authors. Importantly, the ART does not directly measure linguistic experience or reading habits, but rather serves as a proxy. As noted by Scholman et al. (2020), this implies that high levels of literacy may also be achieved through means other than extensive reading of fiction. Nevertheless, despite this limitation, the ART has been widely used in studies investigating a range of linguistic phenomena (Schuster et al., 2023; Johnson & Arnold, 2021; Freed et al., 2017).

Materials. The materials consisted of 130 names, half of which were real authors and half fictitious. Following Scholman et al. (2020), the real names were taken from the version of the test reported by Acheson et al. (2008), while the fictitious names were taken from the adaptation proposed by Martin-Chang & Gould (2008).

Experimental instructions. Participants were instructed to read each name and determine whether it belonged to a real author. They were explicitly encouraged not to guess, but to respond only when confident in their judgment. In addition, they were instructed to respond as quickly as possible without prolonged deliberation.

Experimental procedure. The names were presented in alphabetical order. For each name, participants had ten seconds to make a decision. No feedback was provided.

Measure of performance. The ART score was calculated by subtracting the number of falsely identified fictitious names from the number of correctly identified real authors. Higher scores indicate greater print exposure.

Part II

Experimental investigations

Chapter 6

Experimental materials

In the experiments reported in this dissertation, I used experimental materials originally developed by Kravtchenko (2022) (see Section 2.1.1), which comprised twenty-four stories about everyday activities such as *going shopping*, *going swimming*, *taking a train*, or *writing letters*. Although the core structure of the materials was retained, several modifications were introduced to better suit my research goals. In this chapter, I describe these modifications. A complete set of experimental materials is presented in Chapter A.

Selection of experimental conditions

One central difference between my experimental design and that of Kravtchenko (2022) is related to the selection of experimental conditions. Kravtchenko employed a 2 (story context: *neutral* vs. *biasing*) × 3 (utterance: *conventionally habitual*, *non-habitual*, *no utterance*) factorial design (see Section 2.1.1, Table 2.1, for an example of an experimental item under all conditions). In contrast, I include only those conditions in which atypicality inferences are expected to arise: specifically, the *neutral* story context paired either with a *conventionally habitual utterance* or with *no utterance* (baseline).

The *biasing* context and *non-habitual* utterance conditions were originally intended as controls to show that atypicality inferences do not arise when the story context marks the target event as atypical or when the utterance itself is non-habitual. Since my focus is on the comprehender-specific factors that modulate derivation of atypicality inferences, these conditions are not relevant for my research goals and are therefore omitted.

To better reflect the simplified structure of my design and to clarify the relevant experimental contrast, I re-label the two retained conditions in terms of informational redundancy (IR), see Table 6.1. Thus, in the **with-IR condition**, the utterance intro-

duces a conventionally habitual event within a neutral context, creating informational redundancy – for example, stating that someone ate in the restaurant is redundant because it is generally expected that people eat in restaurants. In the **without-IR** condition, the utterance block is omitted, and thus no redundancy should arise. This condition serves as a baseline against which the effect of informational redundancy is measured.

Table 6.1: An example of ***going to a restaurant*** story in with-IR (a target IR-utterance is highlighted in gray) vs. without-IR (no utterance block) conditions.

a. Context + Discourse Setup	
Mary is a journalist who often goes to restaurants after her interviews. Yesterday she went to a popular Chinese place where she ran into her friend David. Later that day David ran into Sally, a mutual friend of him and Mary.	
b. Utterance (by condition)	
with-IR	**without-IR**
David said to Sally: "I ran into Mary leaving that Chinese place. She ate there! "	NA

Following Kravtchenko (2022), in the without-IR condition, no utterance is present and therefore no informational redundancy is introduced. In this case, comprehenders are expected to believe that the target event (e.g., eating) occurs with high frequency in the given activity (e.g., going to a restaurant), due to strong script-based associations (Bower et al., 1979; Zwaan et al., 1995; van den Broek, 1990b, see also Section 2.2.1).

In the with-IR condition, the utterance explicitly states a conventionally habitual event (e.g., "She ate there!"), which may be perceived as informationally redundant and trigger atypicality inferences, leading to a belief update, that is, that the event is not typical for the main actor of the story.

To measure the belief update, I follow the approach used by Kravtchenko (2022), asking participants to rate the typicality of the target event using questions such as: "*How often do you think Mary usually eats when going to a restaurant?*". Responses are collected using a continuous scale ranging from 0 (Never) to 100 (Always), with 50 marked as 'Sometimes'. The key prediction is that typicality ratings will be significantly lower in the with-IR condition compared to the without-IR condition, reflecting the derivation of atypicality inferences.

Choice of the form for the target utterance

An additional consideration in selecting the final set of experimental materials was how to frame the target utterance in the stimuli, as this influences the strength of the

belief update (i.e., how strongly atypicality inferences are triggered). Kravtchenko (2022) tested three different forms of the utterance to determine whether informational redundancy alone is sufficient to elicit atypicality inferences or whether some degree of discourse or prosodic emphasis is necessary for their derivation. Specifically, they examined (1) a plain utterance lacking explicit speaker intent ("She ate there."), (2) an utterance with prosodic emphasis ("She ate there!"), and (3) an utterance with a discourse marker ("Oh yeah, and she ate there."). The results showed that the belief update (measured as the difference in typicality ratings between without-IR and with-IR conditions) was present across all three forms. However, the exclamation form produced the largest difference in ratings between the story conditions (see Table 6.2). To maximize statistical power in my experiments, I selected stimuli with the exclamation mark.

Table 6.2: Kravtchenko (2022). Mean typicality ratings and standard deviation in parentheses across story conditions (without-IR vs. with-IR) for the three utterance forms: Period ("She ate there."), Exclamation ("She ate there!"), and Discourse Marker ("Oh yeah, and she ate there."). The difference in typicality ratings between story conditions was significant for all three forms.

Utterance Form → **Story Condition ↓**	**Period**	**Exclamation**	**Discourse Marker**
without-IR	85.59 (19.42)	85.79 (19.52)	84.71 (19.89)
with-IR	80.30 (23.01)	72.37 (28.79)	73.84 (26.04)

Chapter 7

Exp. 1. What inferences do people actually make on encountering informational redundancy?

This study was published in Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2023; Ryzhova et al., 2023) and presented at the 18th International Pragmatics Conference (IPRA 2023)[1].

7.1 Motivation of the study

Kravtchenko (2022) showed that mentioning events that are highly predictable in a certain context overtly triggers pragmatic inferences, as in the example below:

(7.4) *Mary went to a restaurant.* ***She ate there!***

When comprehenders encounter such an informationally redundant (IR) utterance (in bold), they rate the probability that Mary usually eats in a restaurant lower than if it is not mentioned. Kravtchenko (2022) interprets such belief changes as a repair mechanism that accommodates the common ground to make the IR utterance informative with respect to the context (e.g., that Mary does **not** usually eat at restaurants). Consequently, the behavior that is typically entailed by the context (eating when going to a restaurant) becomes atypical for the utterance's referent.

As discussed in Chapter 3, derivation of atypicality inferences may consist of the following steps: 1) identify redundancy based on script knowledge; 2) recognize that this redundancy is infelicitous, as it violates conversational norms; 3) infer that the event mentioned in the target utterance is atypical; 4) and explicitly accommodate this atypicality within the situational context.

[1]Link to the program booklet: link

Steps 3 and 4 result in a representation in which the mentioned event is **not** predictable for the person in question (i.e., Mary does not usually eat in restaurants). This should be reflected in a lower typicality rating for the event, measured via responses to the question "*How often do you think Mary usually eats, when going to a restaurant?*".

In Step 4, it is hypothesized that, in order to obtain a coherent discourse representation, the comprehender may construct an explanation that renders the mentioned event informative (e.g., it is not redundant to say that Mary ate in the restaurant if she typically goes to restaurants only for drinks and does not order food). However, it has never been directly tested whether such an accommodation process actually occurs and whether low typicality ratings truly reflect this kind of contextual update.

If the assumptions linking typicality ratings to such contextual updates are incorrect, the theoretical consequences are non-trivial. A weak correspondence between the ratings and the interpretations they are assumed to reflect would call into question the validity of using such ratings as a proxy for measuring the pragmatic effect. This would, in turn, limit the ability to use rating-based data to evaluate which factors influence the derivation of atypicality inferences, as the measure may capture general plausibility judgments or task-specific strategies rather than the integration of an unexpected event into the discourse model. Moreover, substantial discrepancies between ratings and directly elicited explanations would suggest that the two measures tap into distinct underlying processes: one potentially reflecting automatic activation of world knowledge, the other explicit reasoning. In such a case, interpreting ratings as direct evidence for atypicality inference derivation could lead to misleading conclusions about both the presence of the inference and the factors assumed to modulate it, highlighting the need for other experimental approaches.

Another possibility is that typicality ratings may mask intermediate stages of interpretation. Research on scalar implicatures shows that comprehenders can initially compute an implicature but subsequently reject it. This process can be detected with time-sensitive methodologies – for example, by tracking the proportion of looks toward the literal vs. pragmatic interpretation in a Visual World eye-tracking paradigm (see e.g., Schwarz et al., 2016, 2015). In the rating-based design of Kravtchenko (2022), however, such intermediate computations would remain invisible, as the rating would capture only the comprehender's final decision. If some participants go through such a "compute–evaluate–reject" process, ratings may underestimate the proportion of comprehenders who initially generate the inference, potentially obscuring the effects of individual differences and misrepresenting the factors that influence the derivation of atypicality inferences.

In the present study, the above concerns are addressed by asking participants to provide an explanation for their typicality estimates of the target event (i.e., to ask not only how often they think Mary usually eats, when going to a restaurant, but also why they think so). Annotating these explanations will provide a qualitative and quantitative picture of how people accommodate redundancy and will allow to check

whether the atypicality interpretations are indeed identifiable based on the typicality ratings or to what extent the ratings can mask an implicature computation.

Another question is whether participants' biases in their interpretation of informational redundancy are consistent (i.e., whether individuals consistently favor a pragmatic or a literal interpretation) and if so, whether these biases are modulated by cognitive or personality traits. Given that previous studies on atypicality inferences have used a few-shot approach (Kravtchenko, 2022; Kravtchenko & Demberg, 2022a), it is unclear to what extent the drawing of the inference is consistent within a specific participant and whether someone's tendency to draw or not draw such inferences can be explained by their individual traits. In the present chapter, I test whether subjects consistently draw atypicality inferences, providing the basis for exploring its relationship to individual traits. The next Chapter 8 builds on this by investigating how the observed patterns relate to measures of individual differences.

7.2 Materials

Twenty-four short stories about everyday activities (e.g., going grocery shopping, taking a train, visiting a restaurant) were taken from Kravtchenko (2022) (see Chapter 6 for the description of the experimental manipulation, Section 2.1 for a detailed description of the original study, and Chapter A for a list of stories and questions). Each story consisted of two parts: (a) Context + Discourse Setup block, which introduced the characters and activated relevant script knowledge; (b) Utterance block, which varied depending on the condition. In the with-IR story condition, the utterance mentioned a conventionally habitual event that is highly predictable given the story topic (e.g., "She ate there!" in a *going to a restaurant* story), thus introducing informational redundancy (IR). In the without-IR condition, no utterance followed the setup, and no such redundancy was introduced. This condition served as the baseline. Table 7.1 shows an example story under both conditions.

Filler items followed the same format. However, instead of an IR-utterance, some fillers included an utterance with a non-predictable event (e.g., "She recently got a promotion!"), while others omitted the utterance block and aligned with the without-IR baseline condition (though they differed in terms of the associated typicality question – see below).

Questions

Two questions were associated with each story: a typicality question and an explanation question. The typicality question was answered on a scale from 0 (*Never*) to 100 (*Always*), with 50 marked as *Sometimes*.

The typicality questions for target manipulation were taken from Kravtchenko (2022). In the with- and without-IR story conditions, the typicality question always

Table 7.1: Experiment 1. An example of *going to a restaurant* story in with-IR, without-IR, and filler conditions.

a. Context + Discourse Setup		
Mary is a journalist who often goes to restaurants after her interviews. Yesterday she went to a popular Chinese place where she ran into her friend David. Later that day David ran into Sally, a mutual friend of him and Mary.		
b. Utterance (by condition)		
with-IR	**without-IR**	**filler**
David said to Sally: "I ran into Mary leaving that Chinese place. She ate there!"	NA	David said to Sally: "I ran into Mary leaving that Chinese place. She recently got a promotion!"
Typicality question		
with-IR / without-IR	**filler-rel**	**filler-irr**
How often do you think Mary usually eats, when going to a restaurant?	How often do you think Mary usually gets to see the kitchen, when going to a restaurant?	How often do you think Mary usually gets a promotion?
Explanation question		
Why did you place the slider in this particular position?		

referred to the conventionally habitual target event (e.g., *How often do you think Mary usually eats, when going to a restaurant?*).

Filler questions were designed to prevent participants from developing expectations about the relationship between the presence of an utterance and the typicality of the event. To achieve this, fillers varied both in the presence of an utterance and in the event to which the typicality question referred. In fillers without an utterance, the question always referred to a non-predictable event. In fillers with an utterance, the question either targeted the uttered event or a different non-predictable event that was not mentioned in the story. This variability helped prevent participants from developing strategies based on a fixed association, for example, that utterances always refer to highly predictable events or that questions following stories without utterances always concern predictable events.

A second explanation question was included in all conditions and asked participants to explain their typicality judgment in their own words (*Why did you put the slider in this particular position?*).

7.3 Experimental design and procedure

To reduce learning effects while increasing the number of data points per participant, the main experiment was conducted in two sessions, with at least two weeks in between. The 24 items from Kravtchenko (2022) were used to construct 8 balanced experimental lists with 3 items appearing in the target with-IR condition, 3 in the control without-IR condition, and 4 in the filler condition in each list. In the second session, the subjects saw lists of the same structure but consisting of items they had not seen in the first session. Across the two sessions, 6 observations per subject in the with-IR condition and 6 in the without-IR condition were obtained[2].

At the beginning of the experiment, participants were instructed to read several stories and answer questions about them. Before starting the main experiment, they completed a short set of practice questions designed to both familiarize them with the slider interface and assess their attention. Participants were not allowed to proceed unless they passed all practice questions. No additional attention checks were included in the main experiment. The practice phase included two tasks: (1) On the first screen, participants saw two sliders and were instructed to move one completely to the left and the other completely to the right; (2) In the second practice question, participants were trained on how to complete an actual experimental trial.

After the practice session, the main experiment began. Each story was presented on a separate screen, followed by a typicality question and the slider scale to provide an answer. Participants had to click on the scale for the slider to appear, to make sure that they did not just click through, leaving it in the initial position. When participants gave a rating and clicked on "Next question", the slider froze and a textbox appeared, along with the question "Why did you put the slider in this particular position?" – see Figure 7.1 for an example of the experimental trial. Participants could not proceed to the next story until both questions were answered.

The experiment concluded with a brief questionnaire in which participants reported their native language and their interpretation of the study goal. They were also invited to provide feedback and report any technical difficulties. Participants were informed that their responses to these questions would not affect their participation or compensation.

The experimental procedure was identical in both sessions. In addition, after the second session, five cognitive and personality measures were collected to examine whether individual differences could explain variability in participants' responses. The corresponding analysis is presented in Chapter 8.

[2]For 18 participants, the first session did not contain any items in the without-IR condition. These participants saw all 6 without-IR items in the second session instead.

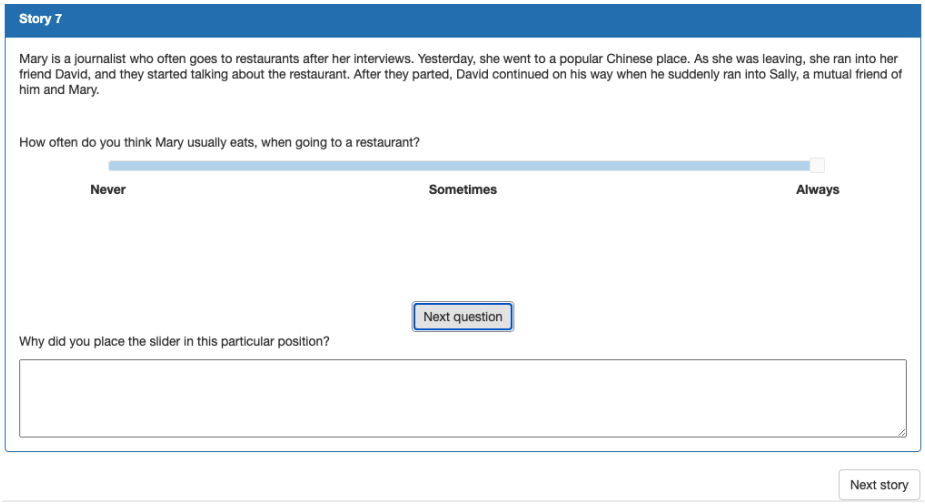


Figure 7.1: Experiment 1. Example of an experimental trial (as presented to participants) for a *going to a restaurant* story in the without-IR condition

7.4 Annotation scheme and annotation procedure

Prior to data collection, the following annotation scheme was designed to classify participants' explanations of their typicality ratings in the with-IR condition. In this condition, the story included an informationally redundant utterance, and participants were asked to express their belief about the typicality of a target event X that was explicitly mentioned in the story (e.g., "How often do you think Mary usually eats when going to a restaurant?"). In the accompanying explanation question, participants were asked to justify the rating they had provided.

The annotation scheme was structured in two levels. The main motivation for developing the scheme was to determine whether a participant had made an atypicality inference, based on their open-ended explanation of the rating they provided. As discussed in Section 7.1, the rating alone may be unreliable or noisy. Therefore, the highest level of the annotation scheme was divided into three categories: *no-atypicality* (no atypicality inference is derived), *atypicality* (an atypicality inference is derived), and *unclear* (i.e., it could not be determined whether an inference had been made). The complete scheme is presented in Table 7.2 and is discussed in detail below. Within each of the three categories, subcategories were introduced to capture more fine-grained differences in participants' responses to the informationally redundant materials.

7.4.1 Annotation scheme

Atypicality inference is not drawn (*no-atypicality* category)

No-atypicality category included responses in which participants did not show evidence of drawing an atypicality inference. This top-level category consisted of two subcategories: 1) *Non-cooperative* subcategory included responses where

participants noticed the informational redundancy but attributed it to the speaker's non-cooperativeness. In Gricean terms, these were cases in which participants recognized a violation of a conversational maxim but did not repair it by generating an inference (Grice, 1975; Huang, 2014). Additionally, since participants acted as overhearers in the present experiment, responses that commented on the overall "quirkiness" of the story (rather than specifically on the speaker) were also included in this subcategory. 2) *Normal* subcategory included responses in which participants stated, in various formulations, that the target event X was typical or expected given the story context.

Atypicality inference is drawn (*atypicality* category)

The primary criterion for assigning a response to this category was that the explanation indicated that the main character is not usually (or not always) involved in the target activity X. Importantly, the annotation scheme was not designed to capture the strength of the atypicality inference; participants' explanations could reflect varying degrees of (a)typicality. Consider the following examples:

I assigned this rating because...

(7.5) *...I think Mary* ***does not usually*** *eat in restaurants*

(7.6) *...I think she* ***sometimes*** *eats* ***but not often***

(7.7) *...She eats* ***on occasion***

(7.8) *...I* ***don't think*** *she* ***always*** *eats in restaurants*

(7.9) *...****there is a chance*** *that circumstances do not require her to actually eat in restaurants*

(7.10) *...it is* ***never impossible not to eat*** *in restaurants*

(7.11) *...she* ***never eats*** *in restaurants*

In examples (7.5)–(7.11), each response implies that Mary does not always eat in restaurants, which may signal an atypicality inference. However, the degree to which this inference is expressed varies across the formulations. Previous work on quantifier interpretation has shown that there is no consistent mapping between quantifiers and event probabilities, and that there is considerable variability between speakers in how such expressions are used (Ramotowska, 2022; Schuster, 2020; Clark, 1990; Wallsten et al., 1986; Pepper & Prytulak, 1974). Moreover, listeners have been shown to adapt their interpretation of quantifiers (or uncertainty expressions) based on contextual information, such as the speaker's personality or the valence of the event. For example, Bonnefon & Villejoubert (2006); Bonnefon et al. (2011) demonstrated that when an event has more severe consequences, people tend to interpret expressions like "possible" or "probable" as indicating a higher likelihood than when the same phrases are used in neutral or positive contexts (see also Holford et al., 2022; Schuster, 2020).

Experimental materials in the present study involve everyday situations that, on their own, do not imply high-risk outcomes for the story characters. That is, the

events described in the target utterances are not inherently emotionally valenced in the way that, for example, serious illness or winning the lottery might be. However, the ways in which participants **accommodate** the informational redundancy may result in explanations that invoke such, for example, socially undesirable scenarios. Consider, for instance, a *going shopping* scenario in which the target IR-utterance is "*John paid the cashier!*". One way to accommodate the redundancy is to infer that John sometimes steals. Another, less socially marked, interpretation is that John sometimes uses self-checkout machines. Even though both responses could involve the same quantifier (e.g., "sometimes"), their implications differ, and thus the interpretation of the quantifier may change accordingly. For these reasons, attempting to estimate the strength of the atypicality inference would require a high degree of guesswork on the part of the annotators (particularly since it is not expected that all participants will provide elaborative explanations for the atypical behavior). As such, this annotation category is intended to identify the presence of an atypicality inference, without attempting to quantify the degree of belief update.

Within the *atypicality* category, I distinguish between two factors that reflect the steps involved in drawing atypicality inferences. The first is whether participants mention the IR utterance in their explanation (e.g., *I think Mary does not usually eat in restaurants because David commented that she ate this time*). This signals that they noticed the informational redundancy in the story – a necessary condition for any further inferential processing to occur. The second is whether they provide a reason for why Mary does not usually eat in restaurants (e.g., she interviews people in restaurants, has limited financial means, or usually orders only drinks). This explicitly shows that they accommodated the atypicality within the situational context. Both features help to clarify how participants arrive at the inference. Mentioning the utterance connects the belief update directly to the linguistic input that triggered it, while providing a reason shows that the inference has been integrated into a mental model of the story. In this way, annotating these features makes the inference process more transparent. Based on the intersection of these two factors, the *atypicality* category was further divided into four subcategories: whether the IR utterance was mentioned or not (_*utt* vs. _*noutt*), and whether the context was extended or not (_*elaborative* vs. _*concise*) – see examples in Table 7.2.

It is unclear if atypicality inference was drawn (*unclear* category)

Some responses may not provide sufficient information for annotators to determine whether an atypicality inference was drawn. Several types of ambiguity fall under this top-level category, which are described as subcategories below.

First, participants might explicitly indicate that they did not know how to answer the typicality question or how to justify the rating they gave. If no further explanation is provided, such responses should be assigned to the *not-sure* subcategory.

In addition, some responses may express uncertainty in conjunction with a general statement about 'normal' behavior (e.g., *People usually eat in restaurants, but here*

I'm not sure). These cases are difficult to interpret, as they fall somewhere between inference absence, rejection, and acceptance. If no further information is provided, these responses cannot be confidently classified as either *no-atypicality* or *atypicality*, nor do they exhibit clear signs of implicature rejection. As such, they should also be annotated as *not-sure*.

BUT: Expressing uncertainty, however, does not necessarily justify assignment to the *unclear* category. For example, some participants may express hesitation while still clearly indicating that the event is atypical (e.g., *I think Mary doesn't usually eat in restaurants, but I'm not sure*). In such cases, it may be unclear whether the uncertainty reflects doubt about the inference itself or simply about the strength of the belief update. Nonetheless, because the atypicality inference is explicitly drawn, such responses are assigned to the *atypicality* category rather than the *unclear* category, if no further information is provided.

Second, it is possible that some participants notice the informational redundancy and even draw an atypicality inference, but ultimately choose not to accommodate it within the given context. In such cases, the inference is rejected after being initially computed. One example of this is shown in Table 7.2 (see the *atypicality-reject* label, under the *unclear* category). Below, I break this response down step by step:

It is strange that David mentioned this, so maybe she doesn't always eat? But after interviews Mary will be tired – she cannot just go to a restaurant for a drink after a long day. So she should eat.

1. **Noticing informational redundancy:** *It is strange that David mentioned this,*
2. **Deriving atypicality:** *so maybe she doesn't always eat?*
3. **Failed accommodation:** *But after interviews Mary will be tired – she cannot just go to a restaurant for a drink after a long day.*
4. **Inference rejected:** *So she should eat.*

Such cases are especially important because they cannot be identified based on ratings alone. Since the implicature is ultimately rejected, one would expect a high typicality rating accompanying such an answer. However, the rating reflects only the final outcome of the reasoning process and may not capture the earlier derivation of an atypicality inference. Using such ratings to classify participants as either pragmatic or literal comprehenders would therefore introduce bias. Individuals who go through such a "compute–evaluate–reject" sequence cannot be judged as purely literal comprehenders. These responses form a separate *atypicality-reject* subcategory within the higher-level *unclear* category.

Table 7.2: Experiment 1. Annotation categories for participants' answers to the explanation question "*Why did you place the slider in this particular position?*" in the with-IR story condition (e.g., *<...> Mary went to the restaurant. <...> She ate there!*, where *eating* is the target event)

Category	Description	Example
ATYPICALITY INFERENCE IS NOT DRAWN (*no-atypicality* category)		
Criterion: the response states that the target event X is normal and/or comments on the strangeness of the story.		
no-atypicality_ non-cooperative	The response mentions the informational redundancy of the utterance but attributes it to the speaker's non-cooperativeness.	Eating in restaurants is typical. It's a little odd that David mentioned it like this.
no-atypicality_ normal	The response states that the target event X is typical or expected given the story context.	Usually when you go to a restaurant, it is to eat.
ATYPICALITY INFERENCE IS DRAWN (*atypicality* category)		
Criterion: the response states or implies that the main character does not usually engage in the target event X. Two additional features are considered in subcategorization: • whether the IR utterance is mentioned in the response; • whether a reason for the atypical behavior is provided.		
atypicality_ utt_ elaborative	• The response links the atypicality of the target event X to the IR utterance (*_ utt*). • A reason for the atypical behavior is provided in the response (*_ elaborative*).	Since David mentioned it, it sounds like she doesn't always eat at restaurants. Maybe she also sometimes interviews people in restaurants.

atypicality_ utt_ concise	• The response links the atypicality of the target event X to the IR utterance (*_ utt*). • No reason for the atypical behavior is provided in the response (*_ concise*).	Since David mentioned it, it sounds like she doesn't always eat at restaurants.
atypicality_ noutt_ elaborative	• The response does not mention the IR utterance (*_ noutt*). • A reason for the atypical behavior is provided in the response (*_ elaborative*).	Maybe she also sometimes interviews people in restaurants.
atypicality_ noutt_ concise	• The response does not mention the IR utterance (*_ noutt*). • No reason for the atypical behavior is provided in the response (*_ concise*).	[because] She doesn't usually eats in restaurants. There is a chance that circumstances do not require her to actually eat there.
IT IS UNCLEAR IF THE INFERENCE WAS DRAWN (*unclear* category)		
Criterion: the response cannot be clearly assigned to either the *no-atypicality* or *atypicality* categories.		
unclear_ atypicality-reject	The response states or implies that the main character does not usually do the target event X, but this inference is later explicitly rejected.	It is strange that David mentioned this, so maybe she doesn't always eat? But after interviews Mary will be tired – she cannot just go to a restaurant for a drink after a long day. So she should eat.

unclear_ not-sure	The response expresses uncertainty, with no additional explanation provided. **Additional remarks:** • Responses that claim the target event is typical (i.e., belong to the *no-atypicality* category) but then express uncertainty still fall under the *not-sure* category. • Responses that state the atypicality inference but also express uncertainty about it still fall under the *atypicality* category.	Not sure. You can't tell from the passage how often Mary eats in restaurants. Usually when you go to a restaurant, it is to eat but I am not sure. **Counterexample:** Mary can sometimes just order some drinks but I am not sure. [here it should be *atypicality_ noutt_ elaborative*]
unclear_ other	All other responses that cannot be reliably assigned to a specific category.	

7.4.2 Annotation procedure

Two annotators labeled participants' explanations of their typicality ratings using the categories defined in Table 7.2. They had access only to the textual responses and did not see the typicality ratings from the first question, in order to avoid potential bias.

Before annotation, the dataset of participant explanations was divided into batches based on story topic. After annotating each batch, the annotators discussed any discrepancies and resolved them collaboratively.

7.5 Predictions of the study

The present study extends prior work on atypicality inferences in several ways. First, it replicates the key findings of Kravtchenko (2022); Kravtchenko & Demberg (2022a), while also explicitly linking participants' typicality ratings to their reasoning, as obtained from open-ended explanations. In addition, the study explores how participants justify atypical behavior, and whether individuals show consistent interpretation biases throughout the experiment.

Replication of Kravtchenko & Demberg (2022a). In line with Kravtchenko & Demberg (2022a), I predict that typicality ratings for the target event will be lower

in the with-IR condition (where the informationally redundant utterance "*She ate there!*" is present) than in the without-IR condition (where the utterance is omitted) – see Chapter 6 for more detail.

Typicality ratings and given explanations. In the with-IR condition, participants' explanations in response to the question "*Why did you place the slider in this particular position?*" are expected to reflect the reasoning behind their ratings. Specifically, I predict that lower typicality ratings will be associated with explanations that indicate an atypicality inference, compared to explanations that do not. However, I also hypothesize that typicality ratings alone may not always capture the underlying pragmatic processes. In particular, some participants may initially derive an atypicality inference but ultimately reject it; in such cases, the rating would reflect only the final 'rejection' outcome. As a result, ratings in the *atypicality-reject* category are expected to be closer to those in the *no-atypicality* category than to those in the *atypicality* category.

Accommodation within the situational context. I also take an exploratory perspective on how participants make sense of atypical behavior. As discussed in Chapter 3, inferring that Mary does not usually eat in restaurants does not necessarily mark the end of the derivation process. The atypicality must still be made sense of in context – it cannot exist in isolation, as it needs to be integrated into a coherent representation of the discourse. In other words, the listener must re-evaluate the situational context to accommodate the inferred atypicality. This final step, however, has not been directly tested in prior work. Although the present study does not directly prompt participants to explain the intended meaning of the IR utterance, and instead asks them to explain their rating, participants may still spontaneously reveal how they adjusted the context to support their claim about event atypicality. In this sense, it is worth examining the kinds of justifications participants offer, i.e., how they make sense of the atypical behavior.

Consistency within subjects. Previous work has shown stable individual differences in implicature derivation: some individuals consistently draw pragmatic inferences, while others tend to respond more literally (Noveck & Posada, 2003; Marty et al., 2013; Foppolo, 2007; Singh et al., 2016; Pagliarini et al., 2018). Building on this, I hypothesize that participants will show consistent biases in the derivation of atypicality inferences as well: some individuals will systematically draw atypicality inferences, while others will interpret informational redundancy more literally across trials.

It is important to note that the present study does not test belief change within participants – that is, how their typicality ratings for a given story might shift de-

pending on whether they see the IR utterance or not.[3][4] Due to the between-subjects design, each participant sees each story in only one version (either with or without the IR utterance). As a result, it is not possible to directly assess how participant's beliefs would compare across conditions; for instance, how high or low their typicality rating would have been had they seen the alternative version of the same story.

Instead, participants' explanations in the with-IR condition may offer an indirect window into their prior beliefs, through a contrast effect (Plous, 1993). I expect that participants who draw an atypicality inference do so by comparing the IR utterance to what they consider the normal course of events – for example, responses such as "*usually, people eat in restaurants, but in this case...*" suggest that the inferred atypicality is grounded in a comparison to prior expectations.

7.6 Analysis

For the analysis of participants' answers to the typicality question (*How often do you think Mary usually eats when going to a restaurant?*), typicality ratings were transformed from the original $[0, 100]$ scale to the open unit interval $(0, 1)$. This transformation involved dividing the original ratings by 100 and replacing exact 0 and 1 values with 0.001 and 0.999, respectively.

The transformed ratings were analyzed using a beta mixed-effects regression model, with a logit link for the location parameter μ and an identity link for the precision (dispersion) parameter φ, as implemented in the `glmmTMB` package (Brooks et al., 2017) in R (R version 4.2.1; `glmmTMB` version 1.1.4).

The choice of a beta distribution is justified by the nature of the data: typicality ratings are bounded by the slider endpoints and exhibit a strong negative skew – see Figure 7.2 for histograms of the typicality ratings in both experimental conditions. The beta distribution is well-suited to modeling such proportion-like data, as its shape, determined by the parameters μ and φ, allows it to flexibly model skewed and bounded distributions (Ferrari & Cribari-Neto, 2004).

As fixed effects, the model included the story condition (with-IR vs. without-IR). This variable was sum-coded as $+0.5/-0.5$ for the with-IR and without-IR conditions, respectively. To investigate potential learning effects over time, trial order was also included as a fixed effect. For the analysis, the original trial order from both experimental sessions was recoded into a continuous sequence: presentation orders of items from the two sessions were concatenated so that the first half of the sequence

[3]Both the present study and the original study by Kravtchenko & Demberg (2022a) estimate between-subject effects, rather than within-subject belief updates. Kravtchenko & Demberg (2015) used a within-subject design and found the results comparable to Kravtchenko & Demberg (2022a). In Chapter 10, I also adopt a within-subject design, in which participants are exposed to both versions of each story across two sessions, separated by a significant time delay.

[4]However, unlike Kravtchenko (2022); Kravtchenko & Demberg (2022a), where each participant saw only one item per condition, the present study includes six items per condition per participant.

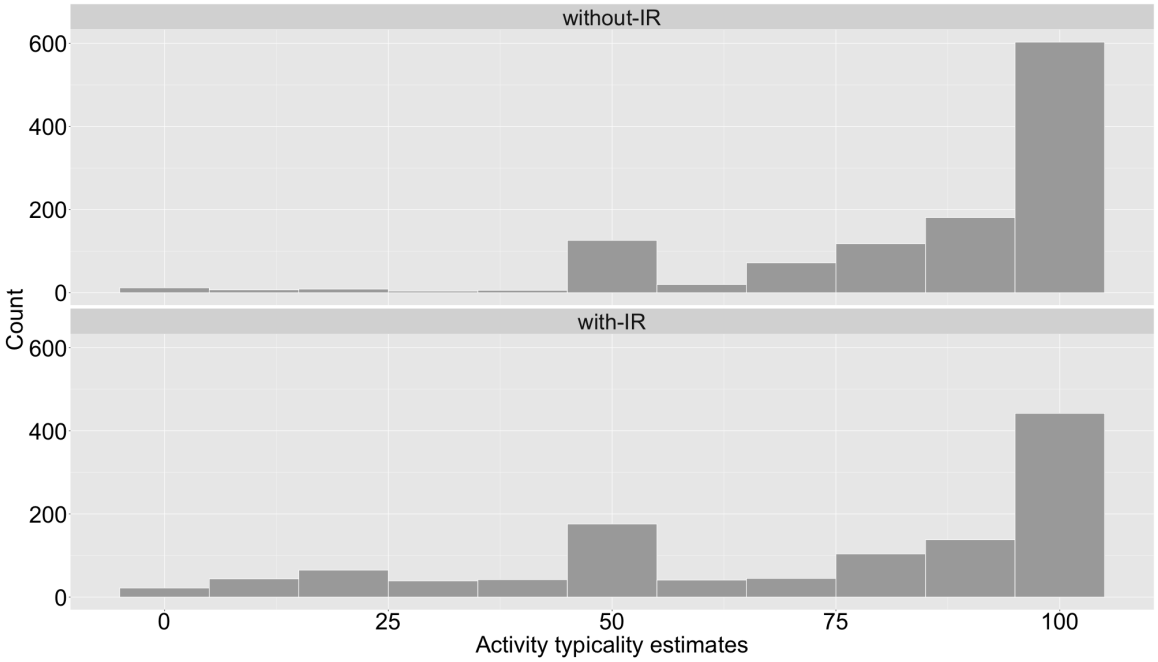


Figure 7.2: Experiment 1. Distribution of the non-transformed typicality ratings by story condition (with-IR vs. without-IR).

corresponded to session 1, and the second half to session 2.

The model's maximal random-effects structure included by-subject and by-item random intercepts, as well as random slopes for story condition. In cases of non-convergence, the random-effects structure was progressively simplified until convergence was achieved, following recommendations from Barr et al. (2013). Any such model simplifications are reported in the Results section.

7.7 Participants

Two groups of participants were recruited via the online crowdsourcing platform Prolific[5]: 160 in the first group and 176 in the second. All participants were native English speakers and had an approval rate of at least 95% based on a minimum of 50 prior submissions.

The first group was not informed that the study involved two experimental sessions, which resulted in a low return rate for session two: only 80 out of 160 participants (50% return rate) completed the second session. In contrast, the second group was explicitly informed about the two-session structure[6], leading to a higher return rate of 84% (148 out of 176 participants returned).

In total, 228 out of 336 participants returned for the second session. No partic-

[5] https://www.prolific.com/

[6] Apart from this, the experimental instructions were identical for both groups.

ipants were excluded based on their responses to the demographics questionnaire. However, 11 participants were excluded due to data loss on the server. An additional 24 participants were excluded according to predefined criteria for individual difference measures: 19 participants were excluded because they were familiar with three or more critical items on the CRT, and 5 participants were excluded because they responded 'No' to nearly all items on the ART – see Chapter 8. After these exclusions, the final sample included 193 participants (mean age = 36.5 yrs (sd = 12.2), age range = [18; 69]; 71% female).

7.8 Results

7.8.1 Replication of the main effect

Participants gave lower typicality ratings in the with-IR condition ($mean = 72.63$) compared to the without-IR condition ($mean = 85.71$). As shown in Table 7.3, the mean ratings in both conditions are comparable to those reported in Kravtchenko & Demberg (2022a).

Table 7.3: Experiment 1. Mean subjects' typicality ratings with standard deviation in parentheses, in both the present study and the original study of Kravtchenko & Demberg (2022a).

Study/Story condition	with-IR	without-IR
Present study	72.63 (29.82)	85.71 (20.61)
Kravtchenko & Demberg (2022a)	72.37 (28.79)	85.79 (19.52)

A generalized beta mixed-effects regression analysis showed a significant effect of story condition ($\beta = -0.45, SE = 0.06, z = -7.96, p < .001$) indicating that participants gave lower typicality ratings in the with-IR condition compared to the without-IR condition – see Table 7.4. This replicates the key finding of Kravtchenko (2022), namely that the overt mention of a highly predictable event leads to lower typicality ratings; see also Figure 7.3.

Next, no significant effect of trial order was found ($\beta = -0.004, SE = 0.004, z = -0.9, p = .37$), meaning that participants did not show any learning effects in the study.

7.8.2 Relationship between annotations and ratings

The participants' answers to the explanation question (*Why did you place the slider in this particular position?*) were annotated by two annotators according to the scheme described in Table 7.2. In total, 1158 responses from the with-IR condition were annotated. (6 responses per participant). Annotations were first performed

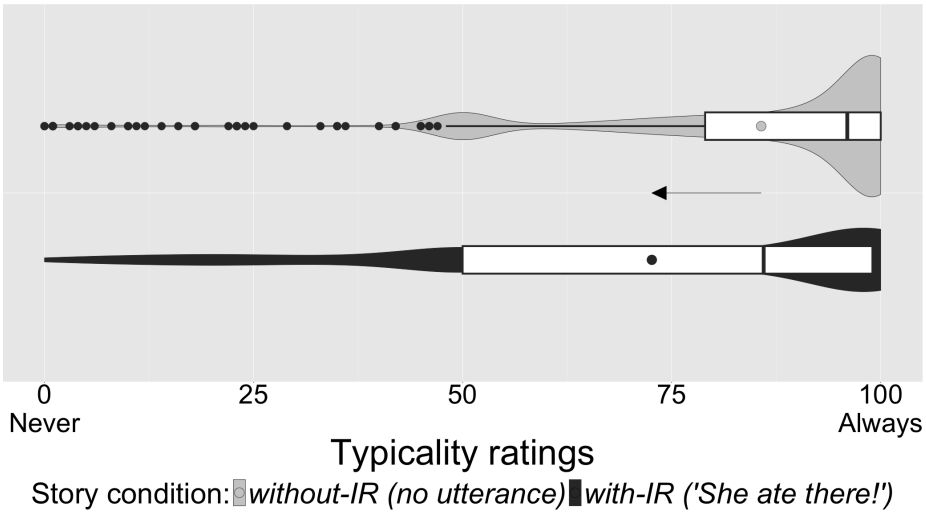


Figure 7.3: Experiment 1. Two violin plots overlaid with box plots showing the distribution of typicality ratings depending on the story condition: without-IR (no utterance) vs. with-IR ('*She ate there!*'). Circles represent mean values. The arrow indicates a statistically significant difference between the ratings in two story conditions.

independently. Inter-annotator agreement was substantial (Cohen's $\kappa = 0.74$, $p < .0001$, 95% CI [0.70, 0.77]). All disagreements were then resolved jointly.

Deriving atypicality inference vs. not. Figure 7.4 shows the mean ratings associated with each annotation category (top panel) and the frequency of each tag (bottom panel). *atypicality* is the most frequent tag ($N = 528$), and it indeed corresponds to a much lower average typicality rating ($mean = 51.84$, $sd = 28.06$) than the ratings given by people whose responses indicated that they made no atypicality inference (*no-atypicality* category: $N = 457$, $mean = 93.82$, $sd = 11.38$).

Table 7.4: Experiment 1. Replication of the main effect. This table shows the effect sizes (β), standard errors (SE), z-values, and p-values for the beta model of participants' transformed typicality ratings. The dispersion parameter φ was estimated at 1.86.

	β	*SE*	z	p
Intercept	1.38	0.08	16.6	<.001
Condition (with-IR)	-0.45	0.06	-7.96	**<.001**
Trial order	-0.004	0.004	-0.9	.37
Random effects	Variance			
Subject	0.17			
Item	0.08			
Condition | Item	0.02			

Interestingly, the *atypicality-reject* cases were found to correspond to similar ratings as in the *no-atypicality* cases where the pragmatic inference was not made ($N = 71$, $mean = 95.46$, $sd = 8.97$). This shows that ratings indeed correspond to participants' final decision.

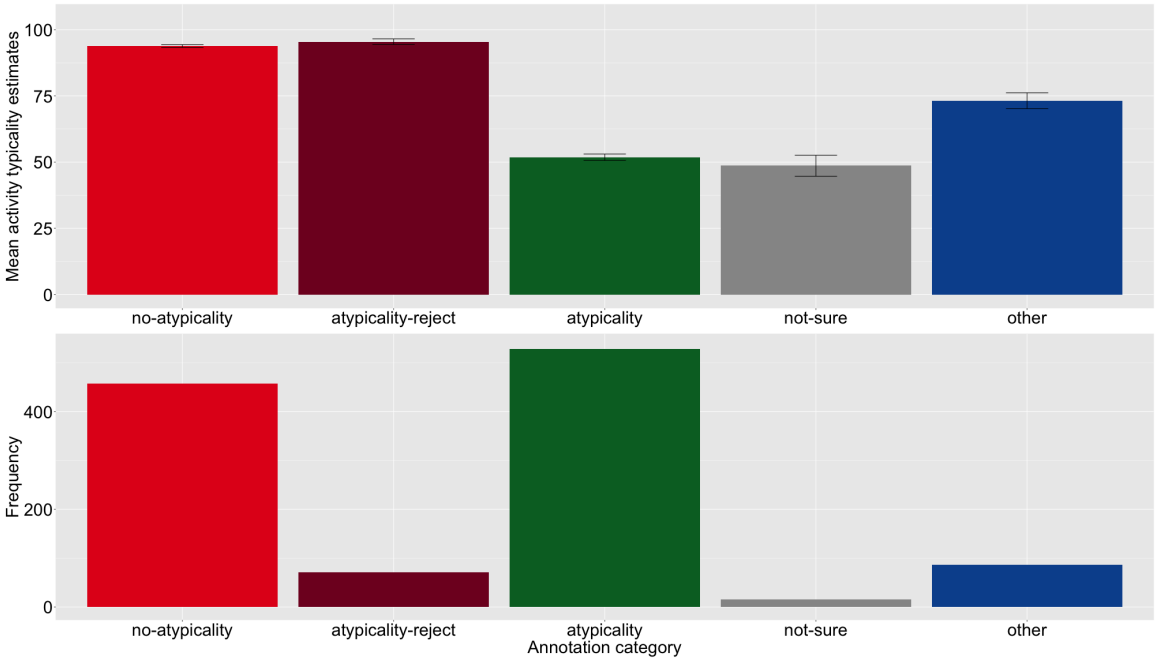


Figure 7.4: Experiment 1 (with-IR condition). Mean typicality ratings ($\pm SE$; upper panel) and frequency of responses (bottom panel) for each annotation category. The subcategories *atypicality-reject*, *not-sure*, and *other* are shown separately for clarity but correspond to the higher-level *unclear* category as defined in Table 7.2.

Explanations within *no-atypicality* category. A closer examination of each annotation group is provided below. Within the *no-atypicality* group, participants occasionally made explicit reference to the speaker's lack of cooperativity (annotated as *non-cooperative*; see Table 7.2 for details). This occurred in only 8 out of 457 trials, with a mean typicality rating of 87 ($sd = 14.79$). This value did not differ numerically from the rest of the responses in the *no-atypicality* group. Next, during the annotation process, it was observed that in some trials participants justified the **typicality** of the target activity by referring to the utterance (e.g., *She usually eats in restaurants because David said so*). Given the unexpected nature of such justifications, these responses were annotated separately for further analysis. However, only 16 such cases were identified, with a mean rating of 94.31 ($sd = 7.52$), indicating that most participants did not interpret IR-utterances as reinforcing claims of normality or typicality.

Explanations within *atypicality* category. Responses within the *atypicality* group were further analyzed based on two factors: whether the utterance was mentioned in the response (_ *utt* vs. _ *noutt*) and whether a more elaborative explanation of the atypical behavior was provided (_ *elaborative* vs. _ *concise*). Figure 7.5 shows the mean typicality ratings and frequencies for each subgroup. In the majority of trials, participants referred to the utterance as the source of the atypical behavior (375 out of 528 responses; _ *utt* annotations). Additionally, in nearly half of the responses within the *atypicality* group, participants provided an explicit explanation of atypical behavior (231 out of 528 trials; _ *elaborative* annotations). Notably, more elaborative explanations were associated with higher mean typicality ratings compared to responses without such explanations (67.90, $sd = 24.48$ for _ *elaborative* vs. 39.36, $sd = 24.05$ for _ *concise*).

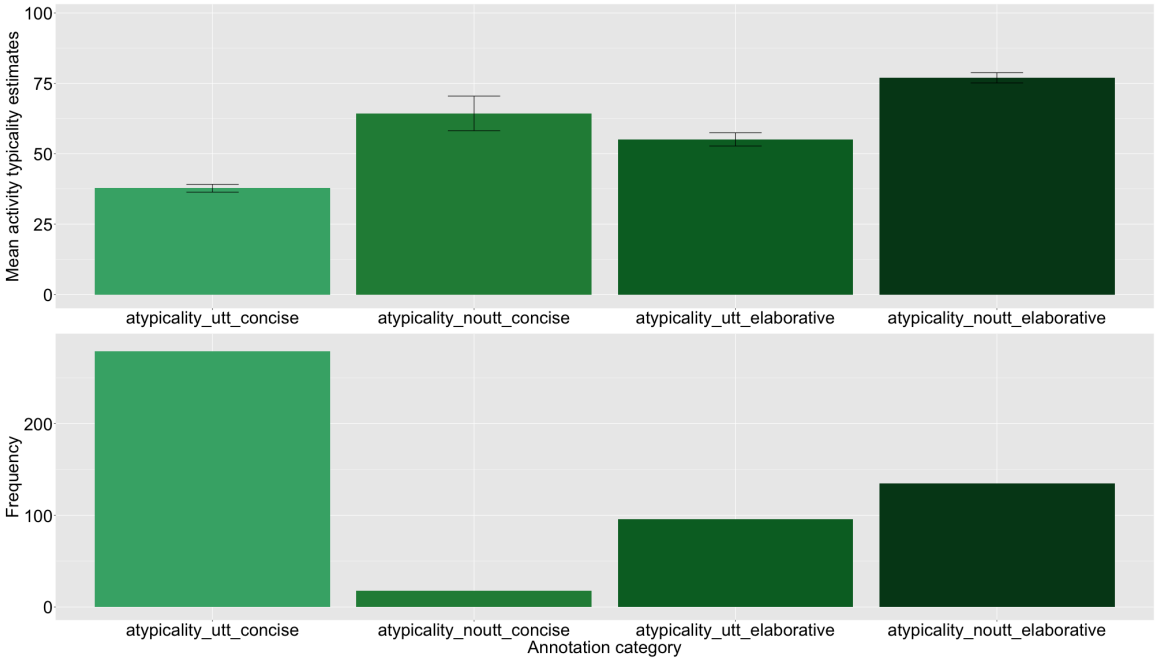


Figure 7.5: Experiment 1 (with-IR condition). Mean typicality ratings ($\pm SE$; upper panel) and frequency of occurrence (lower panel) within the *atypicality* category.

Explanations within *unclear* category. The *unclear* category comprises cases in which annotators could not determine with certainty whether an atypicality inference had been drawn (see Figure 7.4; *atypicality-reject*, *not-sure*, and *other* tags). In total, this category includes 173 trials, with a mean typicality rating of 80.06 ($sd = 25.53$). Cases in which participants drew but subsequently rejected an atypicality inference (*atypicality-reject*) have already been discussed above. In an additional 16 trials, participants were unable to explain their assigned rating (*not-sure*). The mean typicality rating for this subset (48.62, $sd = 15.90$) suggests that participants tended

to use the midpoint of the scale to express uncertainty. Finally, responses labeled as *other* (86 cases) were examined in greater detail to better understand the difficulties faced by the annotators.

In most cases, participants referred either to the story context alone (e.g., [because] *She goes to restaurants after her interviews*) or to the utterance (e.g., *As David said, she ate there*). Such responses are problematic because both the context and the utterance can be used in support of either *atypicality* or *no-atypicality* interpretation, particularly in certain story contexts. As illustrated by more elaborate explanations based on the story context, some participants appeared to assume that the target event should always occur when the script is invoked. However, the conditional nature of the question (*How often do you think Mary usually eats**, when going to a restaurant?***) was not always handled correctly, and in some cases the probability of the target event was equated with the probability of the subject participating in the corresponding script (e.g., *People go to restaurants for eating. The story says "Mary often goes to restaurants," which means almost always; therefore, she almost always eats there*). Although this type of reasoning was rare, it indicates that responses relying solely on the story context can give rise to multiple interpretations, including misreadings of the question or probabilistic errors. Similarly, responses that cite only the utterance can be used either to reinforce a claim of normality or typicality (as in the *no-atypicality* group; e.g., *because David said that Mary ate there, she usually eats in restaurants*) or to support an atypicality inference (as in the *atypicality* group; e.g., *because David said that Mary ate there, she does not usually eat in restaurants*).

Additionally, 6 responses within the *other* tag indicated that participants had made an error when placing the slider (e.g., *I wanted to move the slider 75% to the right* or *Actually, on reading the question more carefully, I think [revision of the belief update]*). Although such responses could, in principle, be classified as either *atypicality* or *no-atypicality*, they introduce an artificial distortion in the relationship between ratings and annotation categories and were therefore classified as *unclear*.

Other inferences. Finally, during the annotation process, raters observed that in some trials participants provided explanations that appeared to involve inferences about the story characters but were not directly related to the (a)typicality of the target event. These responses were marked in the secondary annotation for further analysis. In total, 30 such cases were identified, the majority of which were primarily classified within the *no-atypicality* and *atypicality-reject* categories. See Table 7.5 for the distribution of these inferences across the primary annotation categories.

Both categories involve the claim that the target event typically occurs. However, in the *atypicality-reject* category, participants primarily identified informational redundancy and, having rejected the atypicality inference, related it to other aspects of the story (e.g., inferences about a particular restaurant: *Why would anyone go to a restaurant without eating? I read David's "She ate there" comment as "She ate THERE" rather than "She ATE there". So I assume Mary is like the rest of us and*

Table 7.5: Experiment 1 (with-IR condition). Frequencies of occurrence for other inferences not related to the (a)typicality of the target event across primary annotation categories.

Primary annotation category	Number of answers
no-atypicality	17
atypicality-reject	4
atypicality	5
other	4

goes to a restaurant to eat.). Similarly, some participants in the *no-atypicality* group also identified informational redundancy but related it to alternative aspects of the story without drawing an atypicality inference (e.g., *it sounds like this is something Mary does, but other people don't approve (maybe the restaurant is too expensive?), so they mention it when she does it*). When an explanation supported an atypicality inference but still allowed for an alternative interpretation, the response was nevertheless assigned to the *atypicality* group (e.g., *I don't think David would have mentioned this if Mary always eats in restaurants, unless he was being sarcastic*). In such cases, the presence of an alternative explanation primarily affected the strength of the belief update rather than whether an atypicality inference was drawn.

7.8.3 Subject-specific strategies

The trial-level analysis of annotation groups and categories showed that participants produced a wide range of responses, from drawing an atypicality inference to rejecting any belief update. The next step is to examine whether consistent patterns or strategies can be identified in how participants process redundant information. Because mean typicality ratings computed per participant may result in some loss of information, a more informative approach is to analyze the distribution of annotation tags for each participant.

Figure 7.6 illustrates individual variability in participants' response strategies. Each bar in the plot represents a participant and is divided into six segments, corresponding to the six items presented in the with-IR story condition (and thus six explanation responses). Each segment is colored according to the annotation tag assigned to the corresponding response. For example, the first four participants never drew an atypicality inference (their bars are entirely red), whereas the fifth participant drew an atypicality inference once (one out of six segments is colored green). Overall, the plot indicates that participants adopted a range of strategies in accommodating informationally redundant utterances, with some participants showing consistent patterns across trials.

Subjects were further grouped into three classes: *literal* and *pragmatic*, if four or more of their six explanations in the with-IR condition were classified as *no-*

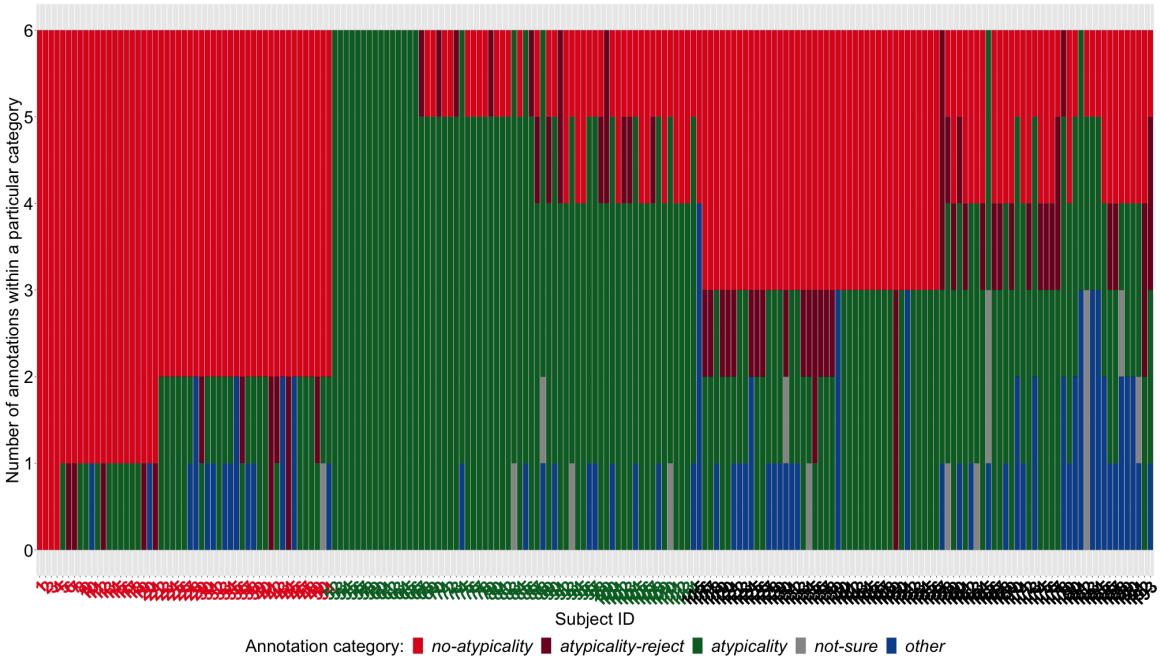


Figure 7.6: Experiment 1 (with-IR condition). Variability of annotation tags across participants. Each bar represents a participant and corresponds to six annotations (y-axis), as each participant saw six items in the with-IR condition. Each segment of a bar represents a single trial and is colored according to the annotation assigned to that response. Participant IDs on the x-axis are colored according to the dominant annotation in their responses (**red**: majority of annotations belong to the *no-atypicality* group; **green**: majority belong to the *atypicality* group; **black**: neither category dominates).

atypicality/*atypicality-reject* or *atypicality*, respectively, and *inconsistent* otherwise. Figure 7.7 shows the mean typicality ratings per story condition for each group.

The majority of participants showed consistent behavior, falling into either the pragmatic class ($N = 63$) or the literal class ($N = 51$). For the pragmatic class, there is a substantial difference in ratings between conditions (84.57 (21.28) in the without-IR condition vs. 55.51 (30.49) in the with-IR condition), whereas for the literal class there is little difference in typicality ratings (87.63 (18.76) in the without-IR condition vs. 88.75 (20.20) in the with-IR condition). The average rating for the inconsistent class ($N = 79$) in the with-IR condition (75.86 (27.54)) falls between those of the literal and pragmatic classes, whereas the mean rating in the without-IR condition (85.37 (21.16)) is comparable to that of the other two classes.

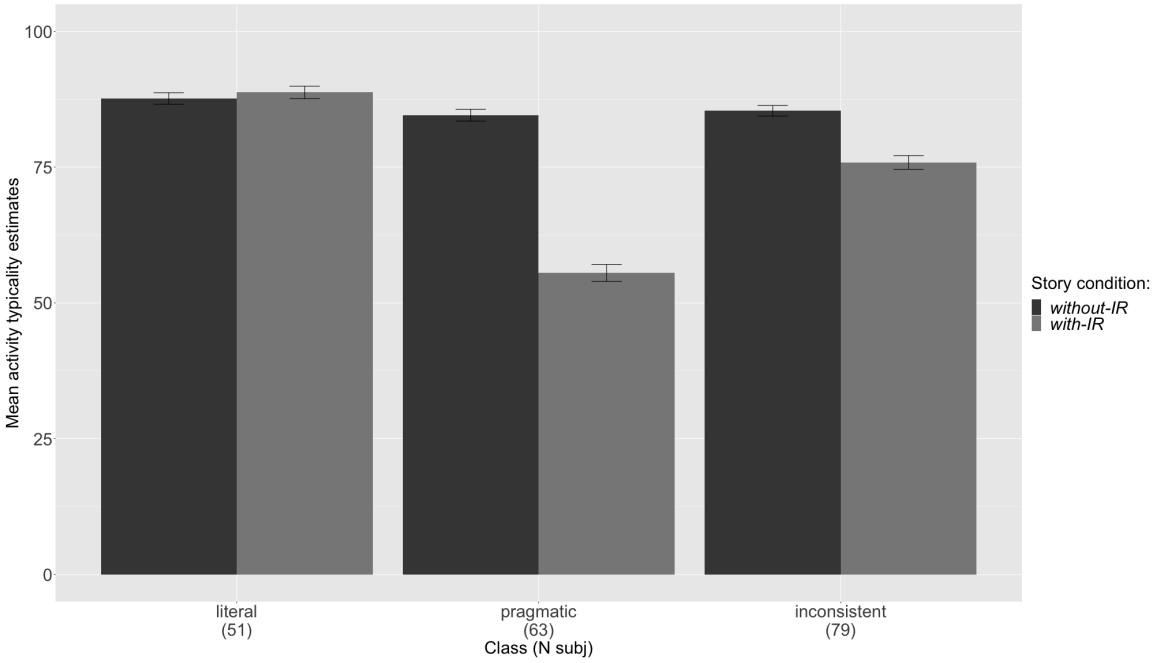


Figure 7.7: Experiment 1. Mean typicality ratings ($\pm SE$) in both story conditions (with-IR vs. without-IR) across three groups of subjects (literal, pragmatic, and inconsistent respondents).

7.9 Discussion and Conclusions

In this study, the previous assumption that the reduction in typicality ratings for the redundant event corresponds to an atypicality inference is largely confirmed: participants' explanations of their belief updates indicate that they interpret the target event (e.g., eating in the context of going to a restaurant) as not typical for the main character.

At the same time, the combined analysis of ratings and explanations provides a more detailed picture of what typicality ratings capture. While ratings clearly reflect participants' final interpretations, they do not necessarily reveal the full reasoning process that leads to these interpretations. In particular, cases in the *atypicality-reject* category show that participants may initially derive an atypicality inference but ultimately reject it, resulting in ratings that pattern with the *no-atypicality* category. Although such cases constitute only a small proportion of the dataset (6%), they demonstrate that ratings may fail to capture intermediate stages of inference derivation. A similar issue arises in *not-sure* responses, where participants appear to use the midpoint of the scale to express uncertainty, as well as in cases involving reasoning errors or misinterpretations of the task. Taken together, these findings suggest that typicality ratings correspond well to participants' final interpretations in most cases and can serve as a useful proxy for atypicality inferences, while at the same time capturing only a simplified version of the underlying reasoning process.

Although such cases constitute only a small proportion of the data in the present study, their frequency is not guaranteed to always be similarly low, and future work is needed to examine these cases more closely.

The analysis of participants' explanations further shows substantial variability in how informational redundancy is accommodated. Among responses classified as *atypicality*, 44% included explicit extensions of the context that justified why the target event is not typical for the person in question. These explanations varied widely across items, often involving inferences about the character's habits, preferences, or circumstances. For example, in the *going shopping* scenario, participants who concluded that the main character does not usually pay the cashier explained this either by assuming socially undesirable behavior (e.g., shoplifting) or by appealing to more neutral alternatives (e.g., the use of self-checkout). Similarly, in the restaurant scenarios, participants proposed explanations involving dietary preferences, financial constraints, or the nature of the character's activities. This variability highlights the flexibility of the accommodation process and shows that deriving an atypicality inference often involves constructing a coherent extension of the discourse model.

Crucially, these findings provide empirical support for the assumption that the derivation of atypicality inferences involves a step of contextual accommodation. A substantial proportion of participants not only indicated that the target event is not typical, but also grounded this inference in the story context by providing concrete explanations of the character's behavior. Moreover, participants' responses suggest that when they fail to arrive at a satisfactory explanation, the inference may be rejected altogether. This suggests that atypicality inferences are not abstract belief updates, but are integrated into a richer discourse representation that maintains coherence with the narrative.

Interestingly, responses that included more elaborative explanations were associated with higher typicality ratings than concise responses (mean = 67.90, sd = 24.48 vs. mean = 39.36, sd = 24.05). One possible explanation is that participants who did not provide detailed justifications were more confident that the event is not typical and, therefore, did not feel the need to explain their answer, as the inference may have seemed obvious to them. Overall, this observation suggests that the absence of elaboration does not necessarily mean that the inference is weaker.

In some cases, participants also drew inferences that were not directly related to the (a)typicality of the target event, for example by commenting on relationships between the characters or interpreting the utterance as part of a broader conversational move (e.g., sarcasm or emphasis). Although such responses were relatively rare, they indicate that informationally redundant utterances can give rise to a broader range of interpretations beyond atypicality.

In addition to trial-level variation, the results also show systematic differences between participants in how informational redundancy is interpreted. Although some participants exhibited variable response patterns across trials, many others showed consistent behavior, tending toward either a more pragmatic or a more literal in-

terpretation. This pattern suggests that the derivation of atypicality inferences is not uniform across comprehenders, but may reflect stable individual tendencies. The present results therefore provide evidence for both within-subject consistency and between-subject variability in the interpretation of informational redundancy.

Several methodological considerations should be taken into account when interpreting these findings. First, participants were asked to explain their slider placement rather than directly explain the speaker's utterance. This design was intended to avoid drawing too much attention to the utterance. However, it also means that participants' explanations may not fully reflect their interpretation of the utterance itself, but rather their reasoning about the rating task. While many responses still showed clear evidence of atypicality inference and contextual accommodation, future work could explore more direct ways of asking about interpretation, while being careful not to encourage participants to overthink their answers or try to guess the purpose of the study.

In addition, not all participants who derived an atypicality inference provided an explanation, which is not surprising. The effort required to type a response may have discouraged some participants from giving detailed explanations. At the same time, the absence of an explicit explanation does not necessarily mean that participants did not consider such reasoning or did not have a clear interpretation. This issue should be addressed in future work.

Finally, the annotation of participants' explanations provides useful insights into how they accommodate redundancy. However, since these explanations are given after the fact, they may not always reflect the reasoning that took place during comprehension (thinking in the moment). Participants may revise or refine their judgments while responding, which makes it difficult to fully reconstruct their initial interpretation. This is also supported by a small number of responses in which participants explicitly indicated that they had changed or reconsidered their answer.

In the next chapter, subjects' preferences to either draw or ignore the atypicality inferences are related to their cognitive profiles.

Chapter 8

Exp. 2. Atypicality inferences across individuals: The role of cognitive and personality-related traits in their derivation

This study was published in Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2023; Ryzhova et al., 2023) and presented at the 18th International Pragmatics Conference (IPRA 2023)[1].

8.1 Motivation of the study

Previous research on pragmatic processing has shown that responses to implicatures based on lexical scales, such as scalar implicatures, can be highly consistent within individuals (Heyman & Schaeken, 2015). At the same time, a growing body of work has provided evidence for systematic individual variability in the processing of such inferences, linking differences in response tendencies to cognitive and socio-pragmatic characteristics (e.g., Antoniou et al., 2016; Fairchild & Papafragou, 2021; Yang et al., 2018). However, this line of research has so far focused primarily on inferences tied to lexical scales, while it remains largely unclear whether similar patterns of individual variability can be observed for pragmatic inferences that arise from informational redundancy and script knowledge, such as atypicality inferences.

In Chapter 7, responses to informationally redundant materials were shown to be consistent within individuals, suggesting that participants can adopt stable response tendencies when interpreting such utterances. The present study builds on this find-

[1]Link to the program booklet: link

ing and asks whether these tendencies can be related to individual cognitive and personality-related factors discussed in Chapter 5.

Chapter 5 identified several classes of factors that may contribute to variability in pragmatic processing, including cognitive resources involved in maintaining and integrating information, individual differences in reasoning style, and socio-pragmatic traits. The present study focuses on a subset of these factors and tests whether they are associated with participants' response tendencies in the atypicality inference task.

8.2 Cognitive test battery

For participants who took part in the experiment described in Chapter 7, the following individual difference measures were collected to investigate whether they predict participants' responding tendencies in the atypicality inference task (listed in the order of completion): Reading Span test (RSpan), Cognitive Reflection Test (CRT), Author Recognition Test (ART), Raven's Progressive Matrices Test (Raven's IQ), and Autism Spectrum Quotient (AQ) – see Chapter 5, for the description of the test procedures.

8.3 Predictions

The following predictions are derived from previous work on individual differences in pragmatic processing and the factors discussed in Chapter 5. For each factor, the expected relationship with the derivation of atypicality inferences is outlined below.

Verbal working memory capacity. It has been argued that implicature derivation is effortful and therefore requires sufficient cognitive resources (Antoniou et al., 2016; De Neys & Schaeken, 2007; Fairchild & Papafragou, 2021). For example, Yang et al. (2018) found that individuals with higher working memory capacity showed higher context sensitivity when deriving scalar implicatures. It is hypothesized that atypicality inferences may also be costly and draw on executive function resources – meaning that individuals with higher working memory capacity would be more likely to derive them.

Cognitive reflection. It could be that derivation of some inferences requires overriding the literal interpretation to arrive at the pragmatic one. The Cognitive Reflection test (Frederick, 2005) taps into reflexivity and the tendency to override the intuitive but wrong response (Pennycook et al., 2016; Welsh, 2022). The rate of pragmatic responding in a reference game was shown to be modulated by participants' performance on the Cognitive Reflection Test (Mayn & Demberg, 2022), while Heyman & Schaeken (2015) found that participants with higher CRT scores were more consistent in their responses to underinformative sentences. If atypicality inferences behave similarly, individuals with higher ability to override the intuitive response would be more likely to derive them.

Exposure to print and language experience. It has been shown that individuals with higher print exposure are more sensitive to certain context cues (Arnold et al., 2018; Scholman et al., 2020). Thus, individuals with higher print exposure might notice and react to informational redundancy more easily, which may increase the likelihood of deriving atypicality inferences.

Non-verbal intelligence. The derivation of an atypicality inference requires reasoning about possible contexts in which the apparently redundant utterance may not be redundant anymore (e.g., stating that Mary ate at a restaurant is not redundant if she usually only orders drinks). It is thus hypothesized that abstract reasoning ability may modulate the process of coming up with a context that would accommodate the atypicality inference.

Socio-pragmatic abilities. As measured by the Autism Spectrum Quotient, socio-pragmatic abilities have been found to correlate with pragmatic responding (Yang et al., 2018): individuals with higher AQ scores may be less likely to put themselves in the interlocutor's position or reason about why the interlocutor said what they did, therefore responding more literally. Similarly, Nieuwland et al. (2010) found that underinformative sentences elicited an N400-effect only in pragmatically skilled participants, as indicated by low scores on the AQ Communication Subscale. Adults diagnosed with ASD have been shown to perform significantly worse on tests of Theory of Mind (Baron-Cohen et al., 2001a; Happé, 1994). Thus, people with higher scores on the AQ (indicating more autistic tendencies) are hypothesized to be less likely to draw an atypicality inference as they might have more trouble recognizing the redundancy and inferring its non-literal meaning.

8.4 Analysis

The obtained scores of the cognitive and personality tests were analysed with a principal component analysis technique (PCA) as implemented in the `psych` package (Revelle, 2022, version 2.2.5) in R and following the steps described by Tanner (2019) – see Section 8.5.3 for details. The resulted latent components were used in all further analyses, instead of the original measures.

For the analysis of subjects' answers to the typicality question (*How often do you think Mary usually eats, when going to a restaurant?*), a beta mixed effects regression was used. The procedure was identical to the one described in Section 7.6 with the following changes. As fixed effects, the model included a story condition story (with-IR vs. without-IR story), its interaction with the obtained latent components of the individual differences measures, and the trial order.

For analysis of subjects' explanations (*Why did you put the slider in this particular position?*), a generalized mixed effects binomial logistic model of the likelihood to compute an atypicality inference was built, as implemented in `lme4` package, version 1.1-30 (Bates et al., 2015). P-values were obtained using the Satterthwaite approx-

imation for degrees of freedom, as implemented in the `lmerTest` package, version 3.1 – 3 (Kuznetsova et al., 2017). The model contained the same set of fixed effects and an identical random effects structure.

In both types of analysis (modeling of the typicality ratings and modeling of the likelihood to draw an atypicality inference), a backward model selection was used to leave in the models only those predictors that significantly contributed to the model fit. For the model of typicality ratings, a model selection procedure was manually implemented. To obtain nested models and the corresponding values of the Akaike information criterion (AIC), the `MuMIn` package (version 1.46.0) in R was used (Bartoń, 2024). The comparison of each two nested models was performed as implemented in `stats` package (version 4.2.1), analysis of deviance (R Core Team, 2022). In the model of likelihood of drawing an atypicality inferences, the backwards model selection was conducted as implemented in `buildmer` package (version 2.8) in R (Voeten, 2023).

8.5 Results: scores on cognitive and personality tests

8.5.1 Descriptive statistics

The summary statistics are presented in Table 8.1. There was no indication of floor or ceiling effects in any of the scores.

Table 8.1: Experiment 2. Descriptive statistics for cognitive and personality measures collected for subjects participated in the experiment ($N = 193$).

Task	Possible range	Observed Range	Mean (SD)	Skewness	Kurtosis
ART	-65 – 65	-9 – 58	19.3 (14.28)	0.54	2.69
AQ	0 – 50	2 – 49	20.52 (8.37)	0.41	3.25
CRT	0 – 1	0 – 1	.31 (.26)	0.67	2.71
Raven's IQ	0 – 10	1 – 10	5.38 (2.12)	-0.09	2.08
RSpan	0 – 1	.01 – 1	.76 (.2)	-1.28	4.49

Note: ART = Author Recognition Test; AQ = Autism-Spectrum Quotient Test (overall score); CRT = Cognitive Reflection Test; Raven's IQ = Raven's Progressive Matrices Test; RSpan = Reading Span Test

ART scores were calculated by subtracting the number of incorrectly identified fictitious authors from the number of correctly identified real authors. Higher scores indicated greater exposure to print. The observed range aligns with previous literature

(e.g., Scholman et al., 2020, where scores ranged from 2 to 56 ($mean = 19.64$, $SD = 10.88$)).[2]

The overall AQ scale was used as a measure of autistic traits, with higher scores indicating a greater presence of autistic-like characteristics. Scores on this scale ranged from 2 to 49, with a mean of 20.52 ($SD = 8.37$). These results compare favorably with those reported by the test creators (Baron-Cohen et al., 2001b), where the neurotypical control group had a mean of 16.4 and $SD = 6.3$, with scores ranging from 0–5 to 41–45. However, the current study showed a higher proportion of extreme values, which may be explained by differences in the sampled populations. For example, participants in the present study were not asked whether they had an ASD-related diagnosis, which may have influenced the distribution of scores. Overall, these findings are consistent with those of a recent study by Lodi-Smith et al. (2021), which examined autistic traits in a non-clinical population ($mean = 20.21$, $SD = 8.53$, $[1; 48]$ *range*). That study included 1,139 participants with explicit variation in age and ethnicity, of whom 8.52% self-reported an ASD diagnosis.

CRT scores were calculated as the proportion of correctly answered critical questions, with higher scores indicating greater cognitive reflection ability. The mean proportion of correct responses on the CRT was .31 ($SD = .26$), with scores ranging from 0 to 1. These results are consistent with previous findings for this version of the test (see Mayn & Demberg, 2022). Welsh (2022) used a 7-item CRT consisting of three questions from Frederick (2005) and four questions from Thomson & Oppenheimer (2016). In their study, the mean number of correctly answered questions was 4.1 ($SD = 2.0$), corresponding to 59% correct responses. The higher performance observed by Welsh (2022) may, at least in part, be explained by differences in data cleaning procedures. In the present study, participants were asked for each item whether they had seen the question before (and thus might have already learned the correct answer). Participants who reported prior exposure to more than half of the items were excluded from the analysis.

The Raven's IQ score was calculated as the number of correctly answered questions, with higher scores indicating greater non-verbal intelligence. The mean score was 5.38 ($SD = 2.21$), which is consistent with the findings reported by Mayn & Demberg (2022), who used the same version of the test and observed scores ranging from 1 to 9, with a mean of 5.66 and $SD = 2.05$.

Finally, the RSpan score was calculated as the mean proportion of elements within a set that were correctly recalled, with higher scores indicating greater verbal working memory capacity. Scores ranged from .01 to 1, with a mean of .76 ($SD = .20$). These results are comparable to those reported by Scholman et al. (2020), where scores ranged between .12 and .94, with a mean of .75. However, compared to Friedman & Miyake (2005), the present results are more extreme: they reported a range between .46 and .90, with a mean of .68 ($SD = .09$). In line with Scholman et al. (2020),

[2]Note that Scholman et al. (2020) applied stricter exclusion criteria; for example, they excluded participants with negative scores (i.e., those who selected more non-authors than real authors).

the increased variability in the current sample may be attributable to differences in participant characteristics, such as age. Unlike Friedman & Miyake (2005), the present sample did not consist exclusively of undergraduate students and included a broader age distribution.

8.5.2 Pairwise correlation analysis

Next, a pairwise correlation analysis was conducted to examine the relationships between the obtained scores and assess the potential for dimensionality reduction. The results are presented in Figure 8.1 and are discussed below.

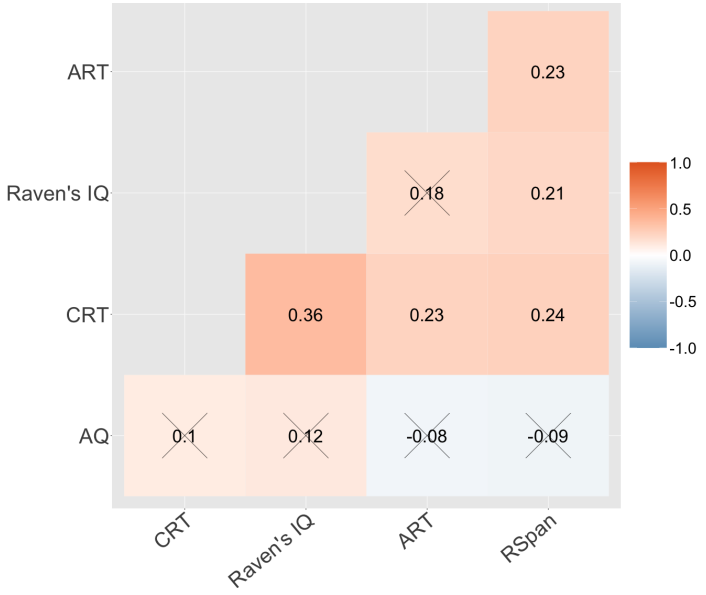


Figure 8.1: Experiment 2. Pairwise correlations ($N = 193$) of cognitive and personality measures collected for subjects participated in the experiment (p-values calculated with Holm correction; non-significant correlations are crossed out). **Note:** ART = Author Recognition Test; AQ = Autism-Spectrum Quotient Test; CRT = Cognitive Reflection Test; Raven's IQ = Raven's Progressive Matrices Test; RSpan = Reading Span Test

Firstly, there was a significant correlation of $r = 0.23$ ($p_{adjusted} < .01$) between RSpan and ART scores which is in line with the previous studies (e.g., Farmer et al., 2017; Payne et al., 2012; Scholman et al., 2020). RSpan scores also significantly correlated with Raven's IQ ($r = 0.21$, $p_{adjusted} = .02$) and CRT ($r = 0.24$, $p_{adjusted} < .01$). The correlation of verbal working memory capacity with intelligence and cognitive reflection abilities is also supported by previous studies (see Toplak et al., 2011; Engle et al., 1999).

Further, ART significantly correlated with CRT ($r = 0.23$, $p_{adjusted} = .01$). This aligns well with previous studies about the relationship between print exposure and

performance on (fluid) intelligence tests (e.g., Stanovich & Cunningham, 1992; Fleva et al., 2017).

Finally, the strongest correlation was observed between CRT and Raven's IQ scores ($r = 0.36$, $p_{adjusted} < .0001$, 95% CI $[0.23, 0.47]$). This finding is consistent with previous work reporting a positive association between cognitive reflection and fluid intelligence. For example, Hanaki et al. (2016) report a moderate correlation between CRT and Raven's Progressive Matrices (Spearman's $\rho = 0.306$, $p < .001$), and further suggest that CRT and Raven's IQ may share a common source, while not capturing entirely identical cognitive processes. This interpretation is also compatible with Welsh (2022), who found that CRT was strongly related to a broader fluid ability factor ($r = .626$), and also showed substantial associations with other cognitive ability factors (e.g., crystallized and quantitative ability). Importantly, in that study Raven's Advanced Progressive Matrices test was included as one of the indicators of the fluid ability factor, and not analyzed as a separate measure. Thus, Welsh (2022) does not provide a direct correlation between Raven and CRT, but it does support the broader conclusion that CRT is closely related to fluid intelligence.

The observed correlations suggest overlapping variance among several measures, particularly between CRT and Raven's IQ. To account for this shared structure in subsequent analyses, a dimensionality reduction approach is adopted.

8.5.3 Dimensionality reduction

Following Tanner (2019), I perform a principal component analysis (PCA) to examine whether the set of cognitive and personality measures can be reduced to a smaller number of underlying dimensions. All measures were centered and scaled prior to analysis.

An exploratory PCA extracting five components showed that four components accounted for 87% of the variance (see Figure 8.2). Based on this result, a four-component solution was retained. The components were rotated using the Varimax method to obtain orthogonal (uncorrelated) and more interpretable dimensions (Kaiser, 1958).

Inspection of the output (Table 8.2) showed that CRT and Raven's IQ load strongly on the same component (Component 1; loadings = 0.76 and 0.86, respectively). Given that both measures are associated with abstract reasoning and fluid intelligence, this component can be interpreted as a *Reasoning* factor. This interpretation is consistent with previous findings showing some association between CRT and Raven's IQ (Hanaki et al., 2016; Welsh, 2022), suggesting shared variance, as discussed in the previous section. The remaining measures each formed distinct components.

The obtained PCA components were consequently used as predictors in the subsequent analyses.

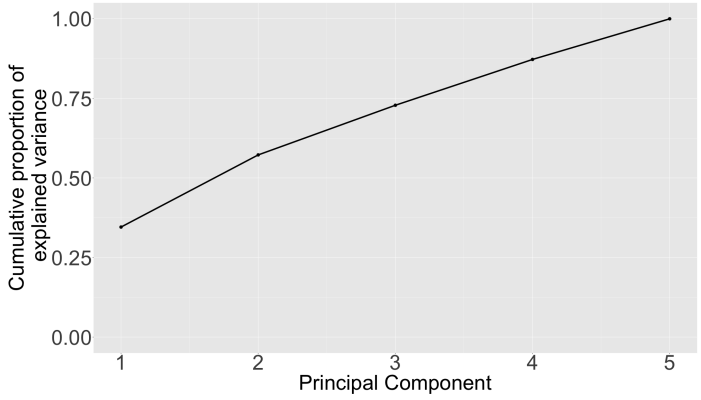


Figure 8.2: Experiment 2. A scree plot of the cumulative proportion of explained variance in the exploratory PCA.

Table 8.2: Experiment 2. Principal Components Analysis (varimax rotation) of cognitive and personality measures (N = 193). Standardized component loadings are shown.

Measure	Component 1	Component 2	Component 3	Component 4
ART	0.14	-0.05	**0.98**	0.10
AQ	0.08	**0.99**	-0.04	-0.05
CRT	**0.76**	0.07	0.18	0.13
Raven's IQ	**0.86**	0.03	0.01	0.06
RSpan	0.15	0.10	-0.05	**0.98**

Note: ART = Author Recognition Test; AQ = Autism-Spectrum Quotient Test; CRT = Cognitive Reflection Test; Raven's IQ = Raven's Progressive Matrices Test; RSpan = Reading Span Test

8.6 Results: analysis of individual differences

8.6.1 Individual differences in typicality ratings

Next, a beta mixed-effects regression model of the $(0, 1)$-transformed typicality ratings was fitted. The full model included story condition (with-IR vs. without-IR) and its interaction with the latent components derived from the individual differences measures (see Section 8.5.3). Trial order was also included as a fixed effect. The random-effects structure was simplified to achieve convergence and initially included only by-subject and by-item random intercepts.

A backward model selection procedure was then performed to obtain the minimal model, which is presented in Table 8.3. After removing non-significant predictors, it became possible to extend the model's random-effects structure to include by-item

random slopes for story condition without convergence issues.

Table 8.3: Experiment 2. Ratings model. Effect sizes (b), standard errors (SE), z-values, and p-values for the minimal mixed effects beta regression model of participants' ratings of the target activity typicality (the ratings were transformed to fit a beta distribution). The dispersion parameter is 1.87.

	b	*SE*	*z*	*p*
Intercept	1.35	0.07	18.97	<.001
Story condition (with-IR)	-0.45	0.06	-7.88	**<.001**
Reasoning	-0.02	0.04	-0.46	.65
AQ	-0.07	0.04	-1.75	.08
ART	0.08	0.04	2.20	**.03**
Story condition (with-IR) : Reasoning	-0.11	0.05	-2.40	**.02**
Story condition (with-IR) : AQ	-0.09	0.05	-1.99	**.05**
Random effects	Variance			
Subject	0.16			
Item	0.08			
Story condition | Item	0.03			

A significant effect of story condition was observed ($\beta = -0.45, SE = 0.06, z = -7.88, p < .001$), indicating that participants, on average, made atypicality inferences.

A main effect of ART was also found ($\beta = 0.08, SE = 0.04, z = 2.2, p = .03$), suggesting that participants with greater reading experience tended to give slightly higher ratings across both conditions. See also Figure 8.3 (top and bottom panels) for a visualization of the model's effect sizes.

Next, an interaction between AQ and story condition was observed ($\beta = -0.09$, $z = -1.99$, $p = .05$), indicating that participants with higher AQ scores gave lower ratings in the with-IR condition.

The top-left panel in Figure 8.3 shows that, in the with-IR story condition, participants with less pronounced socio-pragmatic abilities exhibited stronger atypicality inferences (i.e., a larger difference in mean typicality ratings between the story conditions) compared to those with more developed socio-pragmatic abilities (i.e., a smaller difference in mean typicality ratings between the story conditions).

Finally, an interaction between story condition and the reasoning component was observed ($\beta = -0.11, z = -2.4, p = .02$), indicating that participants with higher reasoning ability exhibited lower ratings in the with-IR condition compared to those with lower reasoning ability, while ratings in the without-IR condition were comparable (Figure 8.3, top-right panel). This, in turn, suggests that participants with higher reasoning ability were more likely to derive atypicality inferences. None of the other effects reached significance and were therefore excluded during backward selection.

The model of typicality ratings presented in Table 8.3, however, includes experimental trials in which the relationship between typicality ratings and the presence

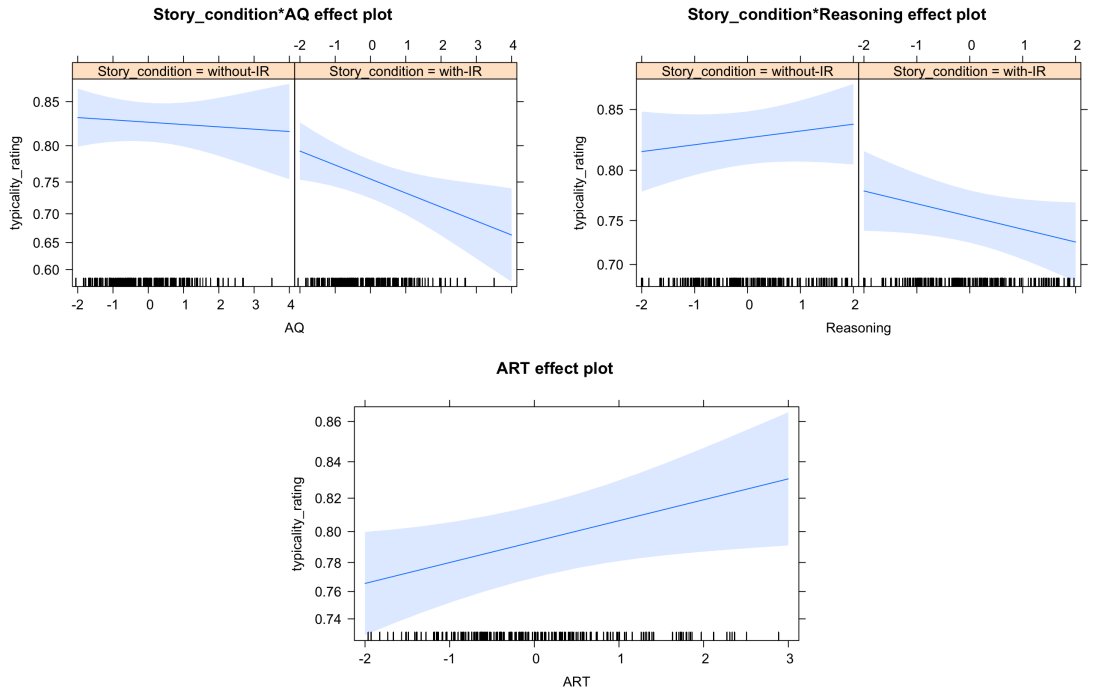


Figure 8.3: Experiment 2. Ratings model. Visualization of the effect sizes of the beta mixed effects regression model in Table 8.3. **Top row, left**: main effect of ART; **Top row, right**: interaction of story condition with Reasoning; **Bottom row**: interaction of story condition with AQ. **Note:** the measures of individual differences are the result of the principal component analysis conducted in Section 8.5.3

(or absence) of inference computation is noisy. As shown in Section 7.8.2 of the previous experiment, in some cases participants who assigned higher typicality ratings in the with-IR story condition nevertheless computed an atypicality inference (*atypicality-reject* tag). Conversely, some participants who assigned lower ratings did not compute the inference but instead indicated uncertainty in responding to the typicality question (*not-sure* annotation tag). Finally, in some trials, participants may have revised their beliefs after assigning the typicality rating (annotation tag *error* within the *unclear* annotation group).

Post-hoc investigations. For a follow-up analysis of typicality ratings and their relation to the cognitive and personality profiles of participants, trials for which it was unclear (based on the annotations) whether an atypicality inference had been computed were removed from the model.

The resulting data set included 1158 observations in the without-IR condition and 985 observations in the with-IR condition, reflecting the removal of 173 trials belonging to the *unclear* category. The total number of participants remained unchanged ($N = 193$). However, the trial exclusion procedure resulted in two participants having

only two observations in the target condition and 13 participants having only three observations in the target condition. The model is presented in Table 8.4.

Table 8.4: Experiment 2. Ratings model. Exclusion of *unclear* experimental trials in the with-IR story condition. Effect sizes (b), standard errors (SE), z-values, and p-values for the minimal mixed effects beta regression model of participants' ratings of the target activity typicality (the ratings were transformed to fit a beta distribution). The dispersion parameter is 1.88.

	b	*SE*	*z*	*p*
Intercept	1.33	0.07	18.62	<.001
Story condition (with-IR)	-0.50	0.06	-7.87	**<.001**
Reasoning	-0.02	0.04	-0.53	.60
AQ	-0.07	0.04	-1.96	.0504
ART	0.08	0.04	2.1	**.04**
Story condition (with-IR) : Reasoning	-0.11	0.05	-2.35	**.02**
Story condition (with-IR) : AQ	-0.11	0.05	-2.25	**.03**
Random effects	Variance			
Subject	0.16			
Item	0.08			
Story condition | Item	0.04			

Overall, the same set of parameters was significant, with an identical interpretation as in the model presented in Table 8.3. However, the interaction effect between AQ and story condition became stronger ($\beta = -0.11$, $p = .03$ vs. $\beta = -0.09$, $p = .05$).

8.6.2 Individual differences in annotations

Next, the relationship between responses to the explanation question ("*Why did you put the slider in this particular position?*") and participants' cognitive characteristics was examined. Only those experimental trials in which participants clearly either did or did not draw an atypicality inference (based on their explanations) were retained for analysis. Annotation tags belonging to the *unclear* category were therefore excluded, resulting in the removal of 15% of trials. Consequently, the likelihood of a participant providing a pragmatic response can be estimated and related to their cognitive characteristics.

A mixed-effects logistic regression analysis was conducted in which the dependent variable was whether a participant's explanation in the with-IR condition reflected an atypicality inference, as indicated by the *atypicality* tag (1 = atypicality inference, 0 = no atypicality inference). The dependent variable was regressed onto the individual difference measures, as well as trial order. The model also included by-participant and by-item random intercepts. A backward model selection procedure was then performed.

The results of the minimal model are presented in Table 8.5. The model revealed a significant effect of reasoning ($\beta = 0.48$, $SE = 0.13$, $p < .001$), indicating that participants with higher reasoning scores (a composite component of IQ and CRT) were more likely to provide pragmatic responses. A significant effect of AQ was also observed ($\beta = 0.29$, $SE = 0.12$, $p = .02$), likewise indicating a higher likelihood of pragmatic responses among participants with higher AQ scores.

Table 8.5: Experiment 2. Annotations model. Effect sizes (b), standard error (SE), z-values, and p-values for the minimal logistic regression model of the annotations of participants' explanations (atypicality vs. no-atypicality) in the with-IR condition.

	b	*SE*	*z*	*p*
Intercept	0.17	0.19	0.90	.37
Reasoning	0.48	0.13	3.81	**<.001**
AQ	0.29	0.12	2.32	**.02**
Random effects	Variance			
Subject	1.72			
Item	0.50			

Post-hoc investigations. In contrast to the ratings model, an important modeling decision in the logistic regression concerns whether to classify *atypicality-reject* responses ($N = 71$) as instances of atypicality inference or to exclude them from the analysis (as it was done in Table 8.5). On the one hand, participants who provided an *atypicality-reject* response did not accept the atypicality inference. On the other hand, they generated and considered it. It is therefore plausible that, in terms of cognitive processing, *atypicality-reject* responses may pattern more closely with *atypicality* responses than with *no-atypicality*.

To assess whether classifying *atypicality-reject* responses as instances of atypicality inference affects the logistic mixed-effects model, these responses were alternatively coded as 1 (i.e., inference drawn) and as 0 (i.e., no inference drawn). This post-hoc analysis yielded the same pattern of significant effects as reported in Table 8.5 across both specifications. When *atypicality-reject* responses were coded as indicating inference, the estimate for reasoning was $\beta = 0.46$ ($SE = 0.11$, $z = 4.06$, $p < .001$), compared to $\beta = 0.39$ ($SE = 0.12$, $z = 3.37$, $p < .001$) when they were coded as not indicating inference. For AQ, the estimates were highly similar between codings ($\beta = 0.27$, $SE = 0.11$, $z = 2.47$, $p = .01$ vs. $\beta = 0.27$, $SE = 0.12$, $z = 2.37$, $p = .02$).

These results suggest that the effect of reasoning is somewhat sensitive to how *atypicality-reject* responses are classified, with more precise and slightly larger estimates when they are treated as instances of inference. In contrast, the effect of AQ appears robust to this coding decision. Overall, *atypicality-reject* responses may, to some extent, pattern with *atypicality* responses rather than with *no-atypicality* responses, although the variation across specifications indicates that they may occupy

an intermediate status. Given the post-hoc nature of this analysis, this interpretation should be treated with caution. The observed pattern highlights variability in how participants evaluate and report inferred interpretations, which in turn points to a potential direction for future research.

8.7 Discussion and Conclusions

In the present study, the previously observed variability in whether participants derive atypicality inferences was examined in relation to their cognitive and personality traits. This variability was investigated using two complementary approaches: modeling participants' typicality ratings and modeling the annotations of their explanations.

Reasoning ability as a predictor of atypicality inferences. Across both analyses, the clearest cognitive predictor of pragmatic responding was the Reasoning component, which loaded primarily on Raven's Progressive Matrices and the Cognitive Reflection Test. Participants with higher scores on this component gave lower typicality ratings in the with-IR condition and were also more likely to produce explanations annotated as atypicality inferences.

A plausible interpretation is that reasoning ability primarily affects the *repair* process triggered by recognizing informational redundancy. In the present materials, the critical utterance states an event that is already highly expected from script knowledge. To arrive at an atypicality inference, comprehenders must go beyond this default interpretation and search for an alternative context in which the utterance becomes informative again. This requires, at minimum, recognizing a mismatch between what is said and what would normally need to be said, and then constructing a non-obvious explanation that reconciles the utterance with the discourse context. Participants with higher reasoning ability may be better at exactly this kind of search over possible interpretations.

This interpretation fits naturally with the two tests that define the reasoning component. Raven's matrices and the CRT are well known to correlate and are commonly taken to reflect, to a large extent, a shared component of fluid intelligence (e.g., Welsh, 2022). At the same time, they are often argued to place somewhat different functional demands on this shared capacity. Raven's matrices can be taken to index abstract pattern detection, efficient attention allocation within the problem space, and the comprehender's persistence in exploring the solution space before giving up (see Stocco et al., 2021; for discussion, see Liu & Demberg, 2026). By contrast, CRT – beyond capturing the tendency to resist an immediately available response (i.e. miserly processing) and to continue searching for a less obvious one (Frederick, 2005) – has also been argued to reflect, in addition, aspects of thinking disposition, such as the tendency to engage in reflective, open-minded and effortful thought (Toplak et al., 2011, 2014).

In the present task, these aspects of reasoning appear to be highly complementary. One possibility is that the same underlying capacity for fluid reasoning may support the search for an alternative interpretation after noticing that the IR utterance is not optimally aligned with the discourse context. From this perspective, the CRT-like ability may be more closely related to engagement in deeper pragmatic reasoning while Raven-like ability may support sustained exploration of the hypothesis space, helping individuals avoid giving up too early and more efficiently disengage from non-satisfiable hypotheses. Importantly, this should not be taken to imply that the two measures tap completely independent mechanisms. Instead, they likely represent different aspects of the same reasoning process.

The present findings are consistent with earlier work on individual differences in pragmatic reasoning in reference games. Mayn & Demberg (2022) report that better performance on pragmatic inference tasks is associated with both CRT and, in their confirmatory study, Raven's IQ. Importantly, they argue that these effects are more readily explained by reasoning demands of the task than by socio-pragmatic skill alone. The current results extend that pattern to atypicality inferences. This strengthens the idea that at least some pragmatic variability is linked to general reasoning resources, especially in tasks that require participants to compute a context-sensitive repair.

The reasoning effect observed here is also captured by the ACT-R model proposed by Liu & Demberg (2026), which was developed on the same dataset. When using reasoning score to estimate model parameters of exploration tendency, their model can better predict whether a participant can derive atypicality inference. On this view, higher Raven's performance does not reflect greater pragmatic knowledge as such, but rather a greater ability to search through candidate interpretations: persisting when an initial interpretation fails, abandoning unproductive strategies, and shifting to alternatives when needed. This idea builds on earlier work linking fluid reasoning to reinforcement-learning mechanisms that support shifting attention away from unsuccessful hypotheses (Stocco et al., 2021), as well as related ACT-R accounts of pragmatic reasoning in reference games (Duff et al., 2025). In this sense, the model shows how the relationship observed in the present study could arise.

Autistic traits as a predictor of atypicality inferences. In addition to reasoning ability, the analyses revealed an effect of participants' Autism-Spectrum Quotient (AQ) scores on the derivation of atypicality inferences. Specifically, higher AQ scores were associated with a greater likelihood of producing explanations annotated as atypicality inferences, as well as with lower typicality ratings in the with-IR condition. In other words, participants with more autistic traits appeared more likely to derive atypicality inferences.

This finding is unexpected. A large body of work suggests that autistic traits are typically associated with differences, and often difficulties, in pragmatic processing. While results in this area are mixed, there is a general tendency in the literature to

associate lower AQ scores with better performance on tasks that require integrating contextual information and inferring speaker meaning (see e.g., Nieuwland et al., 2010; Heyman & Schaeken, 2015; Antoniou et al., 2016; Mazzaggio & Surian, 2018; Yang et al., 2018). Against this background, a positive relationship between AQ and the derivation of pragmatic inferences is not straightforwardly predicted.

One important consideration is the nature of the AQ measure itself. The autism spectrum quotient test was originally designed as a self-report instrument to quantify autistic traits not only in clinical populations, but also in the general population (Baron-Cohen et al., 2001b). It comprises five theoretically motivated subscales: (i) *social skills*, capturing preferences for and ease of social interaction; (ii) *communication*, reflecting difficulties in conversational exchange and understanding others; (iii) *imagination*, relating to the ability to construct and engage with hypothetical scenarios; (iv) *attention switching*, indexing flexibility of attention and resistance to change; and (v) *attention to detail*, capturing a tendency to focus on fine-grained perceptual information. These domains reflect a broad and multifaceted construct rather than a single underlying ability.

Given this structure, it is tempting to relate the present effect to the *communication* (or mindreading) component of the AQ, as the task involves interpreting why a speaker would produce an apparently redundant utterance. At the same time, deriving atypicality inferences may also depend on sensitivity to subtle mismatches between the utterance and the discourse context, which might be related to the *attention to detail* component. Importantly, these two components would lead to different, and potentially opposing, predictions: while differences in communication-related abilities are often associated with reduced sensitivity to speaker intentions, increased attention to detail might facilitate the detection of informational redundancy. The fact that the present effect is positive makes it difficult to straightforwardly attribute it to either component.

Moreover, as emphasized by Kloosterman et al. (2011), the theoretically derived subscales of the AQ are not well supported by empirical factor-analytic evidence. In particular, a range of studies have proposed substantially different factor structures for the instrument (see, e.g., Austin, 2005; Hurst et al., 2007; Stewart & Austin, 2009 for alternative models; for an overview of these studies, see Kloosterman et al., 2011). In addition, Kloosterman et al. (2011) shows that several subscales exhibit relatively low internal consistency, further complicating their interpretation as coherent constructs. This means that isolating and interpreting individual subscales is not straightforward. As a result, attributing the present effect to any specific subscale would require a more principled psychometric approach, for example, deriving components at the item level using exploratory or confirmatory factor analysis.

An additional complication arises when considering how the AQ effect could be further explored within the present dataset given these psychometric limitations. One possible approach would be to decompose the AQ into its subscales or to test alternative factor structures, in order to identify which components drive the observed

effect. However, such analyses would necessarily be post-hoc and would involve a substantial increase in the number of statistical comparisons. This, in turn, would inflate the risk of Type I error. At the same time, applying appropriate corrections for multiple testing would substantially reduce statistical power, making it difficult to detect any true effects. As a result, these approaches would either increase the likelihood of false positives or render the analyses largely inconclusive. In this sense, further probing of the AQ effect within the current dataset is unlikely to yield reliable insights. Instead, what is needed is a study specifically designed to test the contribution of different components of autistic traits, with sufficient power and prior hypotheses. Until such evidence is available, the present finding should be treated as exploratory and interpreted with caution.

Taken together, the observed effect of AQ on atypicality inference is both surprising and difficult to interpret within the current theoretical and methodological framework. While it raises the possibility that certain aspects of autistic traits may be related to how comprehenders engage with informational redundancy, the present data do not allow for a clear account of the underlying mechanism. Given the psychometric complexity of the AQ , this finding should be treated as exploratory. Most importantly, it requires direct replication and more targeted investigation before any strong conclusions can be drawn.

Other effects. In addition to Reasoning and AQ, the rating-based analyses also revealed an effect of ART. However, this effect did not extend to the annotation-based analysis. This limits its interpretation, as it does not appear to generalize to participants' explicit explanations. This result should be interpreted with caution and may be another direction for future work.

By contrast, no effect of working memory capacity was observed. This is in line with findings reported by Fairchild & Papafragou (2017), and contrasts with earlier work showing an effect of working memory on pragmatic processing (De Neys & Schaeken, 2007). However, as argued by Fairchild & Papafragou (2021), effects of working memory may be reduced or disappear once other cognitive factors are taken into account. More generally, these differences may be due to differences in the tasks and measures used across studies.

Taken together, the results indicate that individual differences in reasoning ability and socio-pragmatic traits might be associated with variability in responses in the atypicality inference task. At the same time, an important limitation concerns the offline nature of the measures used here. Both ratings and written explanations were collected after participants had time to reflect on the stimuli, and therefore may not directly reflect the "thinking in the moment". As a result, the observed relationships between individual differences and pragmatic responses should be interpreted with some caution, as they may capture not only how inferences are derived, but also how they are evaluated and reported after the fact. Future work using more fine-grained, online measures (e.g., eye-tracking) would be needed to better under-

stand how differences in cognitive and personality traits relate to the derivation of atypicality inferences.

Chapter 9

Exp. 3. Atypicality inferences under external cognitive constraints: The effects of cognitive load on their derivation

This chapter was published as part of the following article: **Ryzhova, M.**, & Demberg, V. (2023). *Processing cost effects of atypicality inferences in a dual-task setup. Journal of Pragmatics*, *211*, 47–80. 10.1016/j.pragma.2023.04.005. The text has been included with minor modifications for formatting and consistency with the dissertation.

9.1 Motivation of the study

The previous chapter approached variability in pragmatic processing from the perspective of individual differences (Research Question 4). The present chapter adopts a complementary perspective by examining how externally imposed constraints on cognitive resources affect the derivation of atypicality inferences (Research Question 3). As discussed in Chapter 4, theoretical accounts of pragmatic processing (Levinson, 2000; Sperber & Wilson, 1996; Degen & Tanenhaus, 2019) differ in their predictions regarding the effortfulness of generalized implicatures, but converge in predicting that particularized implicatures, such as atypicality inferences, should be cognitively effortful. Some empirical evidence is consistent with this prediction. For example, Rees & Rohde (2023) investigated inferences that are similar to atypicality inferences in that they rely on informativity and background knowledge. Their results suggest that such inferences are not derived automatically and may involve processing costs, as reflected in response time differences. However, experimental work on the processing

cost of particularized implicatures remains limited.

Most previous research has instead focused on scalar implicatures, often using dual-task paradigms to test whether their derivation depends on the availability of cognitive resources (De Neys & Schaeken, 2007; Fairchild & Papafragou, 2021; Marty et al., 2013; Cho, 2020). From the perspective adopted in this dissertation (see Research Question 3), the key issue is not only whether pragmatic inferences are costly, but whether their derivation depends on the availability of cognitive resources. If so, constraining these resources should affect the likelihood or strength of the resulting inferences.

Atypicality inferences triggered by informationally redundant utterances (e.g., "*She brought her swimsuit!*" in the context of going swimming) provide a particularly suitable test case. Unlike scalar implicatures, they are purely particularized: they are highly context-sensitive, are not tied to lexical scales, and rely on world knowledge about typical event sequences (*scripts*; Schank & Abelson, 1975). Their derivation involves multiple processing steps, including the detection of redundancy, the recognition of pragmatic markedness, the inference of atypicality, and its accommodation within the discourse model. If these processes depend on cognitive resources, then limiting these resources should reduce or weaken the resulting inferences.

To test this prediction, the present series of studies adopts a dual-task paradigm (De Neys & Schaeken, 2007; Fairchild & Papafragou, 2021; Marty et al., 2013; Cho, 2020); see Section 4.1. Across three experiments, different cognitive resources are targeted. In Section 9.2 (Experiment 3.1), stimuli are presented auditorily and a visuo-motor tracking task is used to primarily tax attentional resources. In Section 9.3 (Experiment 3.2) and Section 9.4 (Experiment 3.3), participants perform a reading span task that more directly taxes verbal working memory. The use of these two types of secondary tasks allows the present study to examine the role of attention and verbal working memory in the derivation of atypicality inferences under externally imposed constraints.

9.2 Experiment 3.1: Atypicality inferences while performing a visuo-motor tracking task

9.2.1 Materials

The materials for this experiment were taken from Kravtchenko (2022) as described in Chapter 6. Among twenty-four stories designed by Kravtchenko (2022) the 20 items were selected with a highest pragmatic effect (defined as the difference in ratings between story conditions) in that study. Thus, the stories about baking, fueling, laundry, and pasta were not included in the experiment. Stories and experimental questions are presented in Chapter A; the example of an experimental item is also presented in Table 9.1.

The stories were read out aloud by a native speaker of American English. The informationally redundant (IR) utterances were recorded with exclamatory intonation.

Table 9.1: Experiment 3.1. Example of the "Going swimming" story and related questions

a. Context
Lisa likes to go swimming at a nearby pool after work. A couple days ago she was at the pool when she saw Harvey, another regular member, and they stopped to chat. After Harvey changed and went out into the pool area, he ran into Jen, another swimmer and a friend of Lisa's.
b. Optionally mentioned IR activity description (in bold)
Harvey said to Jen: "Lisa's here to swim, too. **She brought her swimsuit!**"
Questions
How often do you think Lisa usually brings her swimsuit, when going swimming?
How often do you think Lisa usually brings her children, when going swimming?
What does Lisa like to do after work?

Each story was followed by three story-related questions. The target typicality question aimed to assess participants' judgments about the target activity typicality mentioned in the informationally redundant utterance (*How often do you think Lisa usually brings her swimsuit, when going swimming?*). The control question addressed an activity that is generally non-predictable from the story (*How often do you think Lisa usually brings her children, when going swimming?*). In addition, to the experimental questions above that were designed by Kravtchenko (2022), here I introduce a third comprehension question that was about the content of the story (*What does Lisa like to do after work?*). The answers to this question were used for checking whether participants paid attention to the story content.

Filler items were also about everyday activities which had similar properties in terms of structure and length, but did not contain any informationally redundant utterances. Instead, filler stories contained utterances that were either a question ("Hey, do you know what time it is?", "So, what are you up to?", "Have you heard the news today yet?") or an event-unrelated statement ("You know, I'm really tired."). Filler stories were accompanied with the same experimental questions, as the target items.

9.2.2 Secondary task: Visuo-motor tracking

To manipulate the participants' available cognitive resources, a dot tracking task was used, where subjects were instructed to keep the dot inside of a box, while the dot continuously and randomly moved on a computer screen. Subjects could control the box via their computer mouse – see Figure 9.1 for an example of participants' screen while performing a tracking task.

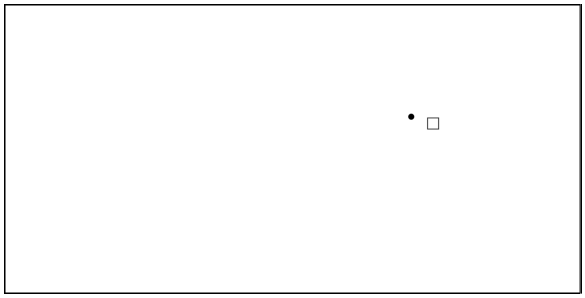


Figure 9.1: Experiment 3.1. Example of participants' screen while performing a tracking task

The dot tracking task is a type of a visuo-motor tracking task. In the literature on working memory, visuo-motor tracking tasks have been shown to trigger the constructs of working memory, thus affecting the amount of available cognitive and attentional resources (see Baddeley et al., 2009, Chapter 3, for an overview; Pype et al., 2010).

The present dot tracking task is also very similar in its essence to a continuous tracking and reaction task (the "ConTRe" task, see Mahr et al., 2012). The ConTRe task simulates the driving environment where subjects have to steer the wheel to keep a constantly moving yellow bar in-between of two blue bars. This task was repeatedly and successfully used as a secondary task to elevate cognitive processing load in studies of different linguistic complexity phenomena (see Vogels et al., 2020, for referential processing; Vogels et al., 2018, for semantic surprisal; Engonopulos et al., 2013 for relative clauses).

The dot tracking task and the ConTRe task are equivalent in that they both require constant attention by the participant and provide a continuous measure outcome. However its important advantage over the ConTRe task is that the dot tracking task can be easily used in remote testing. In the present study, it is expected that the dot tracking task will primarily tax attentional resources, such that in the dual tasking condition, less attention can be devoted to processing the linguistic stimuli. Alternatively, if participants prioritize the linguistic task, it is expected to be able to see that they perform less well on the tracking task during the time of computing the pragmatic inference.

The version of the dot tracking task used in this experiment was implemented according to the example of a dual-task from the website of Cognition Laboratory

Experiments, designed by John H. Krantz[1].

Following their design, the dot was controlled via three parameters: maximum angle variation (this describes how much the dot can change direction from moment to moment), speed, and size of the dot. The size of the box was set up equally to the size of the dot. Based on preliminary testing, the parameters were balanced such that the task was challenging but not impossible (size of the dot = 30, dot speed = 600, maximum angle variation = 180). The sampling rate for the dot and the box coordinates was set to 20 Hz (which amounted to taking a measurement every 50 ms).

9.2.3 Procedure

The experiment started with the instructions where both tasks were explained. Participants were told to listen to the stories carefully and consider their answers to the story-related questions.

Each experiment consisted of only eight trials (four critical trials and four fillers), such that each condition (with-IR vs. without-IR by high vs. no load) was only encountered once by the participants, and no story was encountered in both versions (with and without-IR) by the same participant. This experimental design hence made it impossible for participants to form expectations about what kinds of pragmatic inferences are required in the experiment or develop processing strategies.

In half of the trials, participants thus listened to the stories while tracking the dot with their mouse (high load condition). In the other half of the trials, they only needed to listen to the story. In these trials a fixation cross was displayed in the middle of the screen, and participants were asked to look at it during the duration of the trial (no load condition).

Each new trial in the high load condition started with the dot appearing in the middle of the screen. The dot began to move only after the participant hovered the cursor to the dot. They were instructed to follow the dot carefully with their mouse throughout the trial and keep the the dot inside of the box-cursor. After the dot started moving, participants performed 5 seconds of single-task tracking, before the audio began to play. For the analysis, the onset of the story was annotated, as well as the onset of the pragmatic utterance (*She brought her swimsuit!*). The time course of a trial in the high load condition is also illustrated in Figure 9.2.

Once the story ended, participants were redirected to a page with target and control questions. To answer these questions, participants had to indicate their estimates using a slider that ranged from 0 ('*Never*') to 100 ('*Always*'). Then, on a separate screen participants were shown a comprehension question which they had to answer in an open form. After all three story-related questions were answered, the next trial began.

[1] https://psych.hanover.edu/JavaTest/CLE/Cognition$_$js/exp/dualTask.html

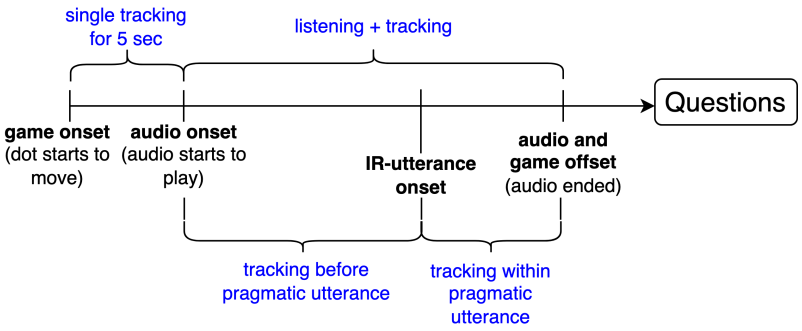


Figure 9.2: Experiment 3.1. The timecourse of a trial in the high load condition

On average, the experiment lasted 15 minutes.

9.2.4 Data collection

In total, there were 20 different stories. Each story had two different versions, one with the informationally redundant utterance, and one without that utterance. A two (with-IR vs. without-IR) by two (high vs. no cognitive load) design was used. Stories were randomized across twenty experimental lists such that each subject saw each condition only once, with no repetitions of a story topic. Each list thus contained four critical items. Four filler trials were also added to each list. For each participant, the order of items was randomized, as well as the order of target and control questions in each trial.

The data collection was conducted to ensure an approximately equal distribution of participants to each list.

9.2.5 Participants

382 eligible participants (mean age = 34 yrs; 60% female) were recruited online through the crowdsourcing platform Prolific. The task was open only to workers who stated English as their native language, and who had an approval rating of $> 95\%$. All participants reported no hearing problems and had normal or corrected-to-normal vision.

9.2.6 Analysis

Typicality ratings

Similarly to experiments in Chapters 7 and 8, subjects' ratings in the typicality question exhibited a strong negative skew – histograms of the typicality ratings for each experimental condition are displayed in Figure 9.3.

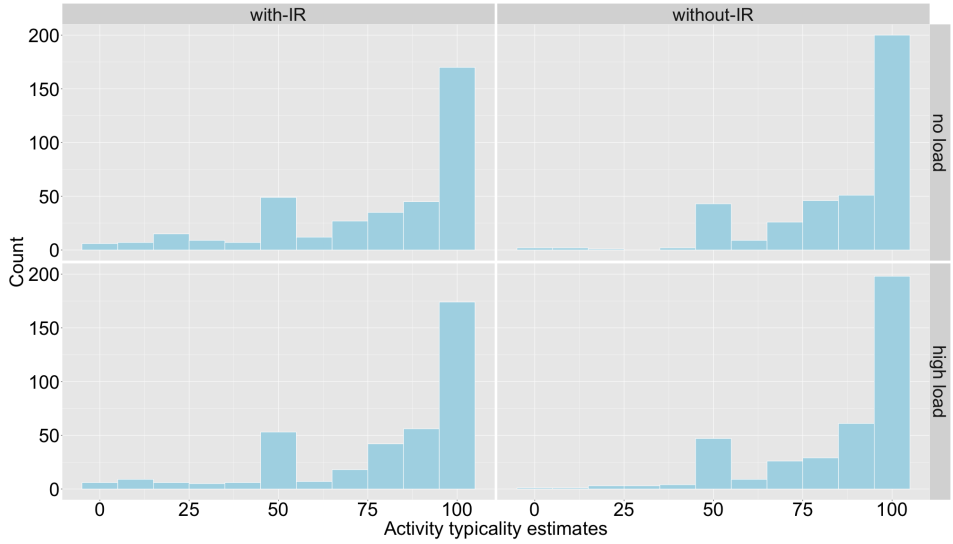


Figure 9.3: Experiment 3.1. Distribution of the non-transformed ratings in the target typicality question by story (with-IR vs. without-IR) and cognitive load (high vs. no)

Likewise, participants' ratings of target activity typicality were transformed from their original scale to the open unit interval $(0, 1)$. The transformation consisted of dividing the original ratings by 100 and substituting zero and one values with 0.001 and 0.999 respectively. Further the transformed ratings were analyzed using a Bayesian mixed effects beta regression model (with a logit link for location parameter μ and an identity link for precision parameter φ), as implemented in the `brms` package (Bürkner, 2017) in R (R version 4.0.3; `brms` version 2.15.0). As fixed effects, story (with-IR vs. without-IR story), load (high vs. no cognitive load), and their interaction were included. For all parameters, the default prior were used – see Table 9.2, for details. The number of iterations was set up to 16000 with a warm-up of 3000 iterations. Each parameter estimate converged with $Rhat = 1$.

Tracking deviations

Performance on the dot-tracking task was analyzed using by-subject mean tracking deviations as a response variable. The deviations were calculated as the euclidean distance (in pixels) between the dot and the cursor in each timestamp of a trial. For each subject, the mean tracking deviations per each interval were calculated, see Figure 9.2. The scaled by-subject mean tracking deviations were analyzed using Gamma mixed effects regression models (with an inverse link, as implemented in `lme4`, version 1.1.23; Bates et al., 2015). The choice of gamma family was justified by the skewed distribution of the tracking deviations – see Figure 9.4.

In each model, one binary predictor representing two different trial intervals was included (prior to pragmatic target sentence and pragmatic target sentence) as fixed

Table 9.2: Experiment 3.1. Prior specification for the Bayesian model of target activity typicality ratings

prior	class	coef	group
student_t(3, 0, 2.5)	Intercept		
(flat)	b	story	
(flat)	b	load	
(flat)	b	story:load	
student_t(3, 0, 2.5)	sd	Intercept	item
student_t(3, 0, 2.5)	sd	story	item
student_t(3, 0, 2.5)	sd	Intercept	subject
student_t(3, 0, 2.5)	sd	story	subject
student_t(3, 0, 2.5)	sd	load	subject
lkj_corr_cholesky(1)	L		item
lkj_corr_cholesky(1)	L		subject
gamma(0.01, 0.01)	phi		

effects – see Section 9.2.8 for more details. P-values were obtained using the Satterthwaite approximation for degrees of freedom, as implemented in the `lmerTest` package, version 3.1.2 (Kuznetsova et al., 2017).

Comprehension questions

To analyze answers to the comprehension questions, a logistic generalized mixed effects regression model of the proportion of correct responses was built (as implemented in `lme4`; Bates et al., 2015). Load, story condition and their interaction were used as fixed effects.

In all models, all factors were $+0.5/-0.5$ sum coded.

Models with the maximal random effects structure justified by the design were always fit at first. Thus, for ratings and logistic regression models, by-subject random intercepts and slopes for story and load conditions as well as by-item random intercepts and slopes for both factors and their interaction were included in the model. By-subject random slopes for the interaction were not included in the model, because there were no repeated measures for the interaction (each subject saw each condition only once). For tracking deviation models, by-subject and by-item random intercepts and random slopes for the interval were included in the model. In the case of non-convergence, the random effect structure was simplified progressively until convergence was achieved (Barr et al., 2013); any model simplifications are stated in the result section.

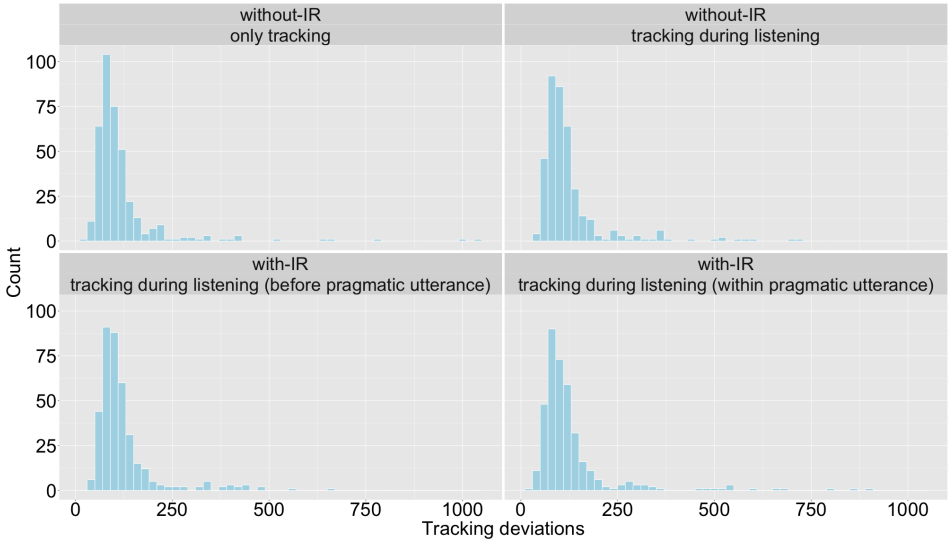


Figure 9.4: Experiment 3.1. Distribution of the non-transformed by-subject mean tracking deviations in the four intervals of interest

9.2.7 Results: pragmatic inferences

First off, it is important to test whether the pragmatic effect reported in Kravtchenko (2022) could be replicated in auditory settings: the activity typicality rating for stories including the informationally redundant utterance is expected to be lower than for the stories without informationally redundant utterance. A Bayesian mixed effects beta regression analysis was conducted with by-subject random intercepts and slopes for story and load and by-item random intercepts and slopes for story. Model results are shown in Table 9.3. There was found an evidence of a negative effect of informational redundancy in the story ($CI_{95} = [-0.42, -0.14]$), replicating the effect reported earlier.

Table 9.3: Experiment 3.1. Mean estimates, 95% credible intervals, and posterior coefficient probabilities (probability that coefficient > 0), for main effects in the Bayesian mixed effects beta model of linguistic judgements. The mean estimate of the precision parameter φ was equal to 2.53 with $CI_{95} = [2.28, 2.79]$

Parameter	Mean estimate	95% credible interval	P($\beta > 0$)
Intercept	1.56	[1.38; 1.75]	1
Story: with-IR	-0.28	[-0.42; -0.14]	0.00025
Load: high	0.02	[-0.09;0.13]	0.65
Story*Load	0.08	[-0.14; 0.29]	0.76

Of special interest for this experiment is the interaction between informational

redundancy and cognitive load. Here, the pragmatic effect is expected to be attenuated, i.e., there should be less of a difference in typicality ratings between with-IR and without-IR conditions in the high load condition. However, the experiment did not show any evidence of such an interaction ($CI_{95} = [-0.14, 0.29]$). The histograms of the posterior distributions for the population-level effects are presented in Figure 9.5. Figure 9.6 displays participants' non-transformed mean activity typicality ratings.

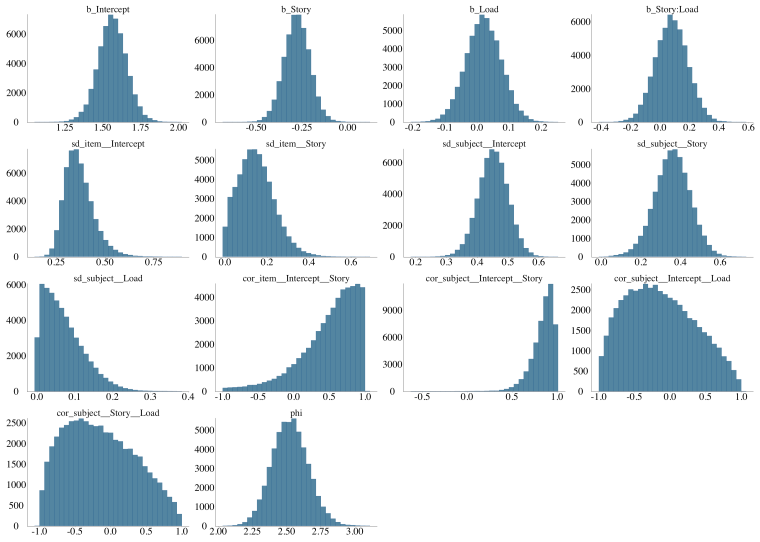


Figure 9.5: Experiment 3.1. Posterior distributions of parameter estimates in the hierarchical beta regression model of transformed typicality ratings

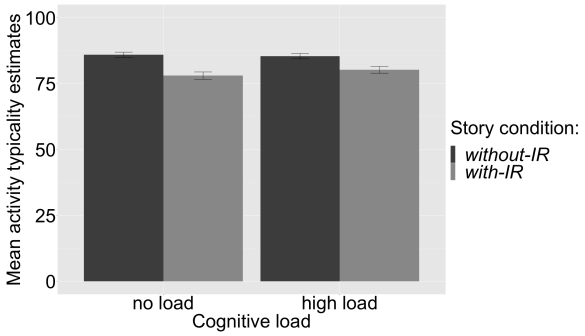


Figure 9.6: Experiment 3.1. Non-transformed mean participants' ratings of the target activity typicality (± SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (no vs. high) conditions.

In addition, subjects' answers to the comprehension questions were analysed. The answers were coded as 1, if the response was correct, and as 0 otherwise. To analyse the proportion of correct responses, a logistic mixed effects regression model was used. As fixed effects, the model included story, load, and their interaction. The random effects structure consisted of by-subject and by-item intercepts and by-subject slopes for load and by-item slopes for story, see Table 9.4.

Table 9.4: Experiment 3.1. Effect sizes (b), standard errors (SE), z-values, and p-values for the logistic model of the proportion of correct responses to the comprehension questions. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*z*	*p*
Intercept	2.08	0.29	7.28	***
Load: high	-0.32	0.16	-2.04	0.04 *
Story: with IR	0.04	0.23	0.17	ns
Story*Load	0.16	0.3	0.54	ns
Random Effects		*Variance*		
Subject		0.018		
load|Subject		1.36		
Item		0.1		
story|Item		0.3		

The model shows a significant effect of load, suggesting that subjects in the no load condition made fewer mistakes than subjects under high load. This might be considered as an evidence that the dot tracking task indeed did interfere with with language processing.

9.2.8 Results: tracking deviations

If two tasks, here visuo-motor tracking and language comprehension, compete for attentional resources, effects may be observed on both tasks. If participants prioritize the language comprehension task over tracking, then it is expected to observe (a) reduced tracking performance when comparing single task tracking to tracking during language comprehension and (b) reduced tracking performance when language processing is particularly effortful, i.e., when the pragmatic target utterance is processed compared to non-target utterances.

To address point (a), tracking deviations in single task tracking should be compared to tracking deviations during language comprehension[2]. If tracking deviations in dual tasking are higher than tracking deviations in single task mode, this provides evidence that the tasks indeed interfered with one another, and potentially also that participants prioritized the language task over the tracking task. To exclude possible effects of pragmatic processing, the comparison was made on the subset of data in the without-IR story condition. For each interval, by-subject mean tracking deviations were calculated and a Gamma mixed effects analysis was conducted. The model showed a significant main effect of task condition ($\beta = 0.04$, $SE = 0.01$, $t = 3$, $p < .003$ ***), showing that people performed less well in tracking while they listened to language at the same time.

[2]To reduce the noise, the first two seconds of tracking were excluded

Second, to investigate point (b), whether the pragmatic inference was difficult in particular, an analysis which compared tracking deviations during the non-critical region to tracking deviations in the critical region was conducted (tracking during listening before vs. within the pragmatic utterance). The comparison was performed on a subset of data including only the with-IR story condition in the high load condition. The Gamma mixed effects regression model of by-subject mean tracking deviations before and within the pragmatic utterance included by-subject random intercepts[3] and is shown in Table 9.5. The main effect of interval was significant at $p < .001$, suggesting that participants' tracking deviations were significantly higher in the interval of the pragmatic target utterance than in the interval preceding the onset of the pragmatic utterance (see the aggregated by-subject mean tracking deviations in Figure 9.7).

Table 9.5: Experiment 3.1. Effect sizes (b), standard errors (SE), t-values, and p-values in the Gamma mixed effects model (with inverse link) of tracking deviations in dual tracking intervals before vs. within pragmatic utterance. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*t*	*p*
Intercept	1.78	0.04	45.43	***
Interval: before	0.05	0.01	3.57	***
Random Effects		*Variance*		
Subject		0.24		

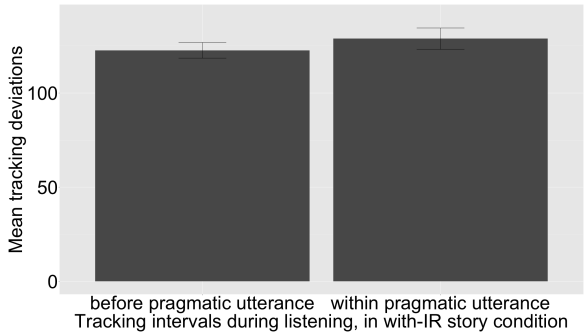


Figure 9.7: Experiment 3.1. Non-transformed by-subject mean tracking deviations (± SEM) aggregated by the interval (before vs. within the pragmatic utterance)

[3]Inclusion of by-subjects slopes is not warranted by the design, as there is only one data point available for each subject in each interval.

9.2.9 Discussion

In Experiment 3.1, the processing cost of atypicality inferences was investigated, using a dual-task design where subjects were instructed to listen to the stories with or without an informationally redundant utterance, while performing a visuo-motor tracking task.

The overall finding reported in Kravtchenko (2022) is replicated in auditory settings and in a dual-tasking setup using Bayesian mixed effects analysis. There was found evidence that participants' activity typicality ratings were lower when the predictable activity was mentioned explicitly in the story, in comparison to when it was not. Hence, participants accommodated informational redundancy by lowering their beliefs about the target activity typicality.

A key finding of Experiment 3.1 is a statistically significant effect of informational redundancy on tracking deviations. In the dual-tasking mode, tracking deviations were significantly higher during the pragmatic target utterance than during literal language processing. This indicates that subjects might have recruited extra cognitive resources in order to process material that elicits a pragmatic inference, directing attention away from the tracking task and hence leading to larger tracking deviations. However, as was pointed out by an anonymous reviewer for Ryzhova & Demberg (2023), an alternative explanation for the differences between the tracking deviations could be the exclamatory intonation in the region of the pragmatic utterance. To disentangle possible effects of prosody from effects of pragmatic processing on tracking deviations, a future follow-up study could repeat the experiment with a condition where stimuli are recorded in the neutral intonation – see also General Discussion on the role of the utterance's framing in atypicality inferences.

In contrast with studies of scalar implicatures (De Neys & Schaeken, 2007; Marty et al., 2013; Cho, 2020), no strong effect of cognitive workload on the derivation strength of atypicality inferences was found. This might be because participants potentially prioritized the language comprehension task over the tracking task. As there were differences in terms of task choice compared to earlier work that used working memory tasks as a secondary task, a second experiment was conducted using a linguistic working memory task, namely the reading span task, similar to Cho (2020). It is thus expected that the linguistic nature of both tasks will avoid prioritization of the pragmatic task over the secondary task, and allow observation of reduced levels of pragmatic inferences under load.

A possible limitation of the present experiment is also that the sample size might have been insufficient to detect a significant interaction (a small tendency in the predicted direction was observed, with 75% of posterior samples lying in the predicted > 0 direction). In the next Experiment 3.2, this is addressed by conducting a power analysis based on the observed main effect size of Experiment 3.1 (Gelman, 2018, March 15). Based on the power analysis, it is decided to collect approximately 1800 subjects – see details in Section 9.3.1.

9.3 Experiment 3.2: low load vs. high load

9.3.1 Power analysis

The results of Experiment 3.1 provide a strong basis for conducting a meaningful power analysis for the second experiment. In particular, the focus is on how many participants need to be collected in order to have a good chance of detecting an interaction effect. For the power analysis, the target power level was set to $\beta = 80\%$ and the target significance level to $\alpha = 0.05$. In order to determine the number of participants, it is also necessary to set a target effect size that should be detected. However, previous studies largely do not report effect size. It was therefore decided to proceed as follows: the target effect size for the interaction was set to half of the size of the main effect. This means that the aim is to ensure detection of a pragmatic effect reduction such that the typicality rating in the high load with-IR condition is reduced by half as much as the typicality rating in the low load with-IR condition compared to the baseline without-IR condition.

Then a two-step approach was taken. As a first rough approximation, the order of magnitude for the number of subjects was estimated based on Experiment 3.1. As there are some differences with respect to the experimental design between Experiment 3.1 and Experiment 3.2, the resulting sample is then used for a more precise power calculation. As power calculation libraries don't currently exist for linear mixed effects models with the beta family, the power calculation was performed using a normal distribution instead. Note that the results of the power analysis should be treated as conservative estimates of the lower bound of power, since the typicality ratings are strongly negatively skewed and the beta family gives a better fit to the data, and has more power to detect an effect than regression using a Gaussian distribution.

Recommendations by Gelman (2018, March 15) were followed in running a power analysis for the main effect of story (with-IR vs. without-IR) based on Experiment 3.1. The model included story and load as fixed effects, but used a gaussian distribution. The random structure included by-subject and by-item intercepts and by-item random slope for story – see model output in Table 9.6. Power analysis was performed in R 3.6.3 using `simr` 1.0.5 package (Green & MacLeod, 2016). The analysis showed that a power of 80% for the main effect of story can be reached with 200 subjects ([80.42; 85.18] 95% confidence interval for β) which, based on Gelman (2018, March 15)'s connection between main and interaction effects, suggested that for an interaction that is half the size of the main effect, one would need more than 1000 subjects. The power curve for a story predictor and details of power analysis are shown in Figure 9.8.

Subsequently, a more exact power estimation was calculated based on the data from 770 subjects collected in the Experiment 3.2 setup. A linear mixed effects regression model which included story, load, and their interaction as fixed effects was run again. The random structure included by-subject and by-item intercepts and

by-item random slopes for story. The main effect of story was equal to $\beta = -7.93$ (t-value $= -4$, p-value $< .001$ – see full model output in Table 9.7). In Experiment 3.2, the plan was to recruit no fewer than 1400 subjects. Assuming this number of subjects, a power analysis in `simr` package was run to estimate the power to detect an effect of interaction that is half of the size of the main effect of story. Based on 2000 simulations, it was found that the power to detect such an effect was estimated as 100% with a 95% confidence interval of $[99.82\%, 100\%]$ given 1400 subjects. For the actual study, it was decided to collect approximately 30% participants more.

Table 9.6: Experiment 3.2. Effect sizes (b), standard errors (SE), t-values, and p-values in the linear mixed effects regression model of the target activity typicality ratings. The model was used in the power estimation of the main effect of story – see Figure 9.8. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*t*	*p*
Intercept	82.48	1.66	49.72	***
Story: with-IR	-6.4	1.46	-4.39	***
Load: high	0.8	1.	0.8.	ns
Random Effects		*Variance*		
Subject		105.44		
Item		44.58		
Story|Item		22.91		

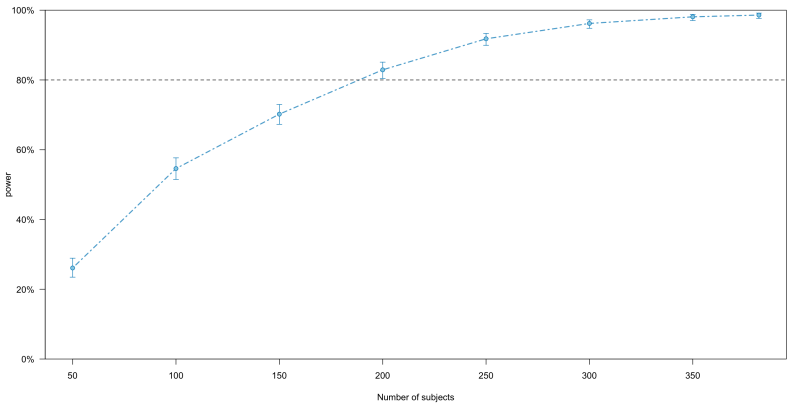


Figure 9.8: Experiment 3.2. Observed power (±95% CI) to detect a fixed effect of story with size '-6.4' calculated over a range of sample sizes. For a power estimation, a likelihood ratio test was used with parameter $\alpha = 0.05$ (1000 iterations)

Table 9.7: Experiment 3.2. Effect sizes (b), standard errors (SE), t-values, and p-values in the linear mixed effects regression model of the target activity typicality ratings. The model was used to estimate the number of subjects needed to detect a significant effect of interaction twice smaller than the main effect of load with the power of 80%. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*t*	*p*
Intercept	80.51	1.8	44.82	***
Story: with-IR	-7.93	1.98	-4	***
Load: high	0.53	0.77	-0.69	ns
Story*Load: with-IR	-0.44	1.53	-0.28	ns
Random Effects		*Variance*		
Subject		83.92		
Item		59.46		
Story|Item		67.09		

9.3.2 Materials

Materials were taken from Experiment 3.1. The target and control questions were identical to those used in Experiment 3.1. For consistency, the comprehension questions in the present experiment were rewritten such that the answer could also be given on a continuous scale (see Chapter A). The fourth filler question about the story characters was also added (*How often do you think Jen and Harvey usually see each other at the pool?*) to each story.

To answer the story-related questions, participants had to put a slider at a position on a scale that best reflected their response. Identically to Experiment 3.1, the scale ranged from zero ('Never') to one hundred ('Always').

In order for the stories to contain a clear answer to the new comprehension questions, the context of each story was modified. In Table 9.8, one can see an updated "Going swimming" story with a new comprehension question "*How often do you think the swimming pool nearby Lisa's office is open, when she finishes working?*". The number of correct "always" and "never" answers was balanced.

9.3.3 Secondary task: reading span task

The secondary task was the reading span task. It consists of reading a number of sentences and memorizing their last words for later recall. Sentences in this task are presented in sets of one to four sentences in a row. Each sentence was shown on a separate screen. In the four-sentence-condition, the subjects hence needed to answer four acceptability judgment questions and remember four words. The manipulation of having sets of 1-4 sentences allowed manipulation of the level of memory load on

Table 9.8: Experiment 3.2. Example of the "Going swimming" story and related questions

a. Context
Lisa likes to go swimming at a nearby pool after work, as they are always open when her working day is over. A couple days ago she was at the pool when she saw Harvey, another regular member, and they stopped to chat. After Harvey changed and went out into the pool area, he ran into Jen, another swimmer and a friend of Lisa's.
b. Optionally mentioned IR activity description (in bold)
Harvey said to Jen: "Lisa's here to swim, too. **She brought her swimsuit!**"
Questions
How often do you think Lisa usually brings her swimsuit, when going swimming?
How often do you think Lisa usually brings her children, when going swimming?
How often do you think the swimming pool nearby Lisa's office is open, when she finishes working?
How often do you think Jen and Harvey usually see each other at the pool?

subjects. To ensure subjects read the whole sentence and not just the last word, they also had to judge whether each sentence was acceptable or not.

The materials for the study were taken from Scholman et al. (2020) and consisted of one hundred sentences. Sentence length varied from eight to thirteen words. Forty-eight sentences contained a verb that required an animate subject (e.g., *escape*, *forget*) and fifty-two sentences contained a verb that required an animate object (e.g., *fascinate*, *impress*). Fifty-two sentences were acceptable. Unacceptable sentences were formed by inverting the animacy of the subject and object noun phrases. Below, one can see an example of an acceptable sentence with an animate subject verb (1), acceptable sentence with an animate object verb (2), and unacceptable sentence with an animate subject verb (3).

(1) It was the elephant that escaped from the zoo.

(2) It was the building that impressed the architect.

(3) It was the lawyer that disturbed the phone.

In total in each experimental list, subjects saw 20 different sentences[4] (see Section 9.3.5 for more details about the lists). For each subject, the mean last word

[4]The sentences in each list were equally assigned to each of the above conditions. However, the study did not aim to report any differences in the recall rate or the acceptability rate based on these conditions.

recall score (hereinafter, the LW-recall score – the number of words recalled correctly overall in the experiment) and the acceptability score (the number of sentences judged correctly overall in the experiment) were calculated.

The reading span task is a popular verbal working memory measure. In a dual task setting in pragmatic processing, it has previously also been used in another study (Cho, 2020). However, the format of the task was different from the present version: it asked the participants to remember the last words of sentences including the critical pragmatic items in the high load condition; the low load condition did not ask them to remember any words. The sentence encoding stage was separated from word recall by a simple arithmetic question. In contrast to present study, participants in Cho (2020) were given feedback for their answers, and there was no time limit for reading / encoding the sentences.

9.3.4 Procedure

Experiment 3.2 consisted of reading the stories with or without the informationally redundant utterance and answering story-related questions. The reading span task was used as a secondary task in this experiment to manipulate the amount of available cognitive resources.

Prior to the beginning of the experiment, participants were instructed to avoid using any tools to help them remembering the stories or words. In addition, the sentences were shown on the screen for a limited amount of time which was calculated based on the performance in the training session where subjects were asked to judge only the acceptability of sentences. The mean decision time was used in the main experiment as a threshold after which a 'timeout' message appeared. The timeout decisions were treated as incorrect in the later analysis.

In each experimental trial, participants first read one of the sets of one to four sentences and judged their acceptability. After judging all sentences in a set, participants were shown a story from the primary task, which they had to read carefully. Next, subjects answered the four story-related questions. After participants answered all the questions, they were asked to write down the last words of the sentences they saw before the story. In recall, participants were instructed to follow the original order in which the sentences were presented. The time course of a trial in the high load condition is illustrated in Figure 9.9.

9.3.5 Data collection

Again 20 experimental lists were constructed, where each contained one item from each of the four experimental conditions (story type: with-IR vs. without-IR by load conditions: high vs. low load). Experimental lists were designed such that each subject never saw the same item in different conditions, just like in Experiment 3.1.

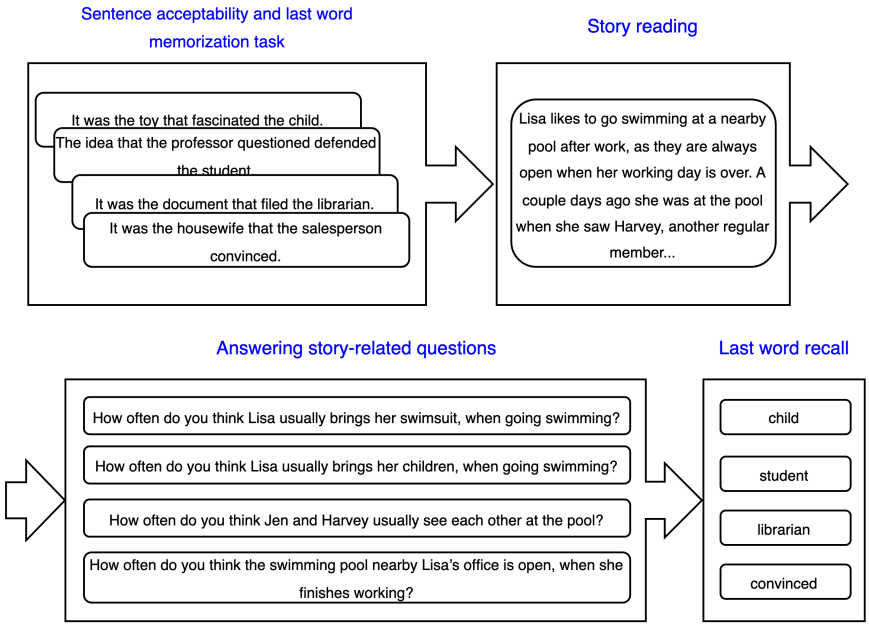


Figure 9.9: Experiment 3.2. The timecourse of a trial in the high load condition

Four filler stories that did not contain an IR-manipulation to each of the lists were also included.

In the low load condition, the set size in the secondary reading span task was one, while in the high-load condition, set size was four. For half of the filler stories, the secondary task set consisted of two sentences, while for another half it consisted of three sentences. Thus, each participant in total saw twenty different sentences from the secondary task and read eight stories in the main linguistic task.

The order of stories and the order of story-related questions was randomized for each subject.

On average, the experiment lasted 20 minutes.

9.3.6 Participants

Participation in Experiment 3.2 was open on the crowdsourcing platform Prolific to those workers who stated English as their native language, who had not taken part in a previous data collection with these items, and who had an approval rating of $> 95\%$.

The data collection happened in several stages in the period of August 2020 – August 2021. For the main analysis, the data from 1941 participants were available.

As a part of the data cleaning procedure, 154 subjects who recalled less than 3 words correctly were removed from any further analysis (to avoid analysing misunderstanding of the task or possible technical difficulties). 7 subjects were removed based on their answers to the questionnaire after the study. They had reported English as

their non-native language or used other means to help them remembering last words of the sentences.

After data cleaning, the dataset comprised data from 1780 subjects. This data forms the basis for the analyses of Experiment 9.3 (mean age = 34.3 yrs, sd = 12.4; 71% female; mean by-subject acceptability score = 18.2, sd = 1.9; mean by-subject last word recall score = 14.3, sd = 4.4).

9.3.7 Analysis

For the analysis of linguistic judgements in the target question, a Bayesian mixed effects beta regression was conducted. The procedure was identical to Experiment 3.1 as described in Section 9.2.6 with the following changes. In Experiment 3.2, the number of iterations was increased to 50000. This was done to ensure computational stability of the Bayes factor that was calculated for comparing a model with vs. without the story:load interaction. For the same reason, the random structure of the models of linguistic judgements was simplified. By-subject random slopes for story and load were removed. By the experimental design, there were only 2 data points per these parameters, which turned out to negatively influence the number of iterations needed to obtain a stable Bayes factor in model comparison.

The answers to the comprehension questions that implied a negative correct response were transformed to a positive scale by subtracting the original rating from the maximum rating of one hundred. Thus, higher ratings in transformed comprehension questions signified higher correctness of the responses. Figure 9.10 show the distributions of subjects' ratings in the comprehension question across the experimental conditions.

Further, transformed ratings were mapped to the open unit interval $(0, 1)$ as described in Section 9.2.6. To analyse the transformed ratings in comprehension questions, a generalized mixed effects beta regression model as implemented in glmmTMB package (version 1.0.2.1) in R (Brooks et al., 2017) was built. As fixed effects, the model included the load, story, and their interaction. Maximal random effects models (Barr et al., 2013) were used; in case of non-convergence, models were progressively simplified as described in Section 9.2.6.

For the analysis of performance in the secondary task, a generalized mixed effects binomial model as specified in `lme4` package (version 1.1.23) in R was built. It was tested whether the proportion of correctly recalled words in a trial was affected by the presence or absence of the IR-utterance or whether there was a significant effect of story redundancy and load interaction. See details in Section 9.3.9.

In all models, all factors were $+0.5/-0.5$ sum coded.

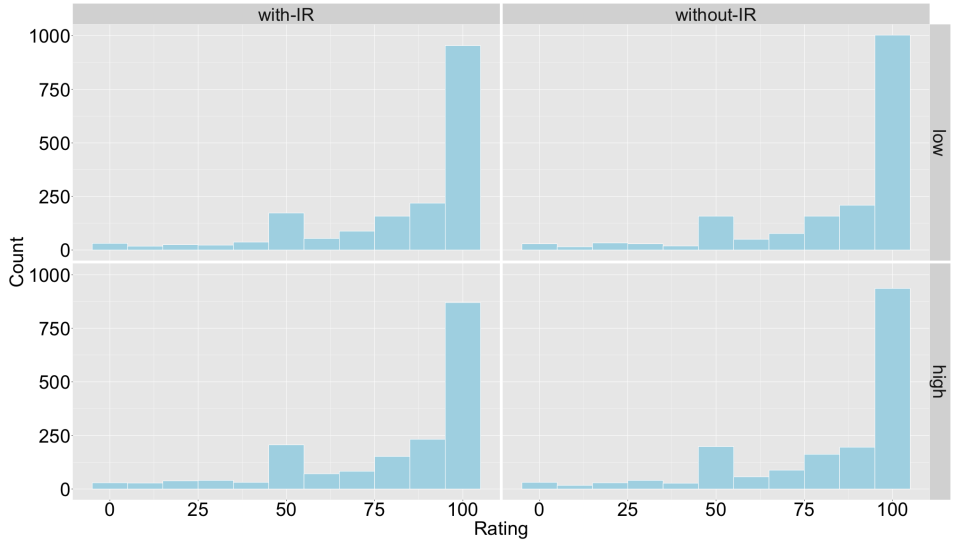


Figure 9.10: Experiment 3.2. Distribution of the ratings in the comprehension question (the higher the rating, the more correct the response is) by story (with-IR vs. without-IR) and cognitive load (high vs. low) conditions

9.3.8 Results: typicality ratings

A Bayesian mixed effects beta regression model with by-subject random intercepts and slopes for story and load and by-item random intercepts and slopes for story was built for participants' ratings of target activity typicality – see Table 9.9. There was evidence for a negative effect of informational redundancy – subjects showed lower ratings of target activity typicality when the IR-utterance was present in the story compared to stories without informational redundancy (Figure 9.11). Most of the posterior samples differed from zero in the predicted direction ($P(\beta < 0) = 0.998$). This replicates the main atypicality inference effect.

Table 9.9: Experiment 3.2. Mean estimates, 95% credible intervals, and posterior coefficient probabilities (probability that coefficient > 0), for main effects in the Bayesian mixed effects beta model of linguistic judgements. The mean estimate of the precision parameter φ was equal to 2.02 with $CI_{95} = [1.94, 2.11]$

Parameter	Mean estimate	95% credible interval	P($\beta > 0$)
Intercept	1.36	[1.2; 1.53]	1
Story: with-IR	-0.28	[-0.46; -0.1]	0.002
Load: high	0.02	[-0.04; 0.07]	0.72
Story*Load	0.00	[-0.1; 0.1]	0.5

No evidence for an effect of the interaction was found again ($\beta = 0, CI^{95} = [-0.1, 0.1]$). The histograms of the posterior distributions for the population-level

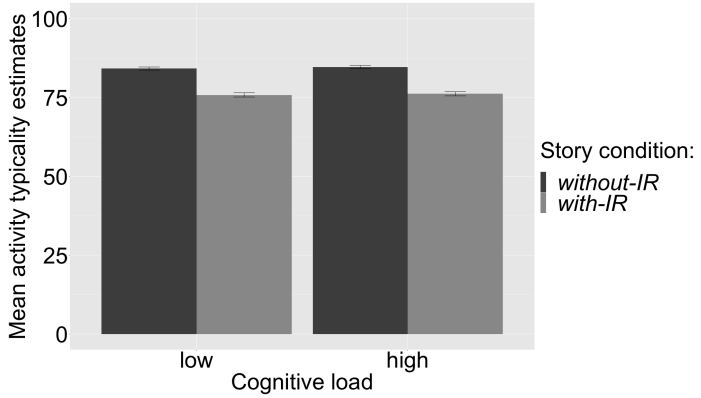


Figure 9.11: Experiment 3.2. Non-transformed mean participants' ratings of the target activity typicality (± SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (low vs. high) conditions

effects are presented in Figure 9.12.

The analysis of participants' ratings in the comprehension questions showed significant effects of Story and Load predictors. A generalized mixed effects beta regression model is displayed in Table 9.10. Subjects' ratings in the high load condition were significantly lower than in the low load condition suggesting that under high load, subjects were less attentive to the stories than under low load ($\beta = -0.07$, $p = .01$). In addition, a marginally significant effect of Story ($\beta = -0.05$, $p = .06$) might suggest that subjects' ratings were less correct when they read stories with IR-utterances compared to stories without the IR utterance – see Figure 9.13.

Table 9.10: Experiment 3.2. Effect sizes (b), standard errors (SE), z-values, and p-values for the generalized mixed effects beta regression model (with logit link) of the transformed ratings in comprehension questions. The dispersion parameter for beta family was estimated as 1.71. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*z*	*p*
Intercept	1.43	0.05	28.06	***
Story: with IR	-0.05	0.03	-1.85	.
Load: high	-0.07	0.03	-2.56	*
Story*Load	-0.03	0.05	-0.58	ns
Random Effects		*Variance*		
Item		0.05		
Load|Item		0.0009		

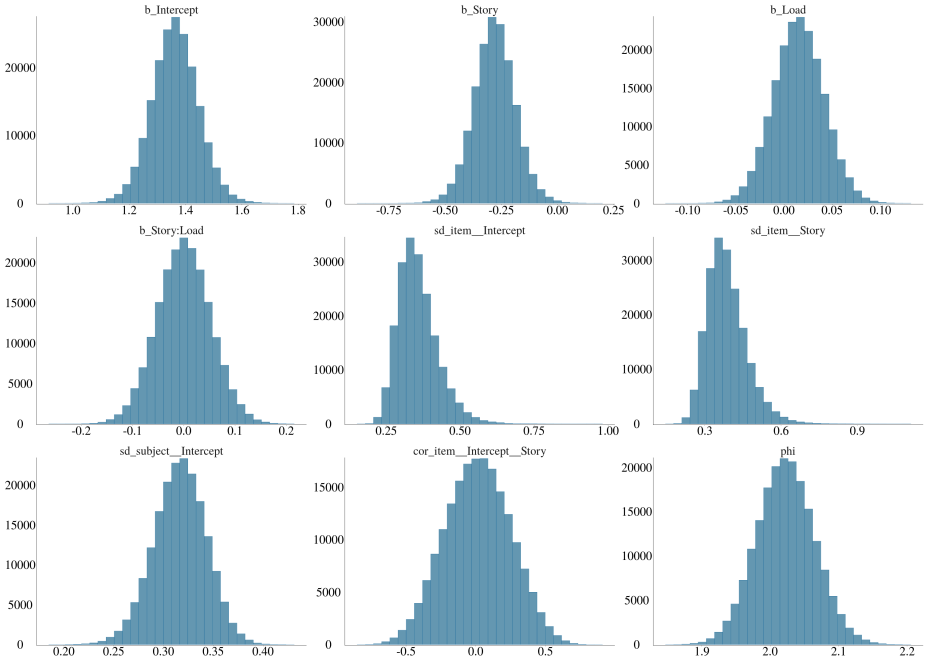


Figure 9.12: Experiment 3.2. Posterior distributions of parameter estimates in the hierarchical beta regression model of transformed typicality ratings

9.3.9 Results: recalled words

The mean proportion of correctly recalled words in the low load condition, when one word had to be recalled, was higher (86%) than in the high load condition (66%), when four words needed to be remembered, see Figure 9.14 and Table 9.11 ($b = -0.08$, $p < .05$). It was of particular interest whether the proportion of correctly recalled words would differ between the story redundancy conditions, as a main effect of story redundancy, or an interaction between load and story. Following the trade-off hypothesis, the proportion of correctly recalled words was analysed. A generalized mixed effects binomial regression model included story and load conditions as well as their interaction as fixed effects. The random structure included by-subject and by-item random intercepts. However, no main effect of story redundancy was found, and the interaction was also not statistically significant. The results are presented in the Table 9.11.

9.3.10 Bayes factor analysis

In Experiment 3.2, no evidence towards the processing cost of atypicality inferences was found, as well as no trade-off effects with the secondary task (in terms of the proportion of recalled words). In this section, the aim is to quantify the amount of evidence in favour of the null hypothesis, i.e., a model that does not include the story:load interaction to the alternative hypothesis where load is assumed to influence the strength of pragmatic responses, and thus the interaction is present in the model.

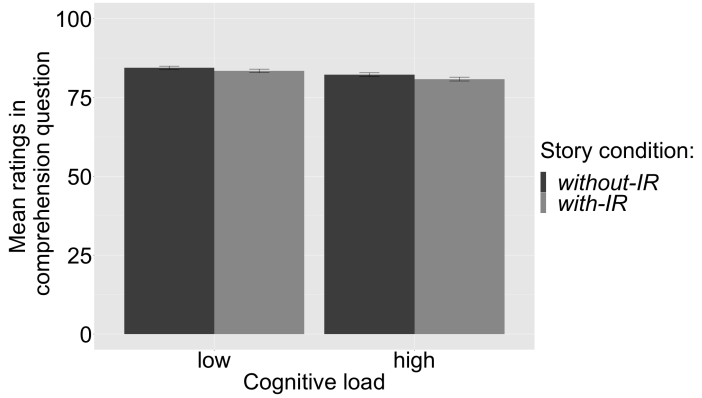


Figure 9.13: Experiment 3.2. Transformed mean participants' ratings in comprehension questions (± SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (low vs. high) conditions

Table 9.11: Experiment 3.2. Effect sizes (b), standard errors (SE), z-values, and p-values in the binomial mixed effects model (with logit link) of the proportion of correctly recalled words. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*z*	*p*
Intercept	1.58	0.07	21.51	***
Story: with-IR	-0.08	0.06	-1.48	ns
Load: high	-1.63	0.06	-28.28	***
Story*Load	0.13	0.11	1.23	ns
Random Effects		*Variance*		
Subject		1.92		
Item		0.08		

Following a Bayesian analysis workflow of Schad et al. (2023), firstly, four different priors (default, non-informative, weakly informative, and more informative) were set up to check for the stability of the Bayes factor and whether it critically differs depending on the strength of assumed interaction effect. As a default prior, a flat prior coming from a 'brm' function in R was chosen. For the other three priors, I specified only the prior for the fixed effects (intercept, story, load, and their interaction) and for the precision parameter ϕ. A full set of priors used in the present experiment is shown in Table 9.12.

The weakly informative and more informative priors were based on the posterior estimates from Experiment 3.1 (see Table 9.3). After a visual inspection of posterior distributions, a normal distribution was chosen to represent the prior for each of the above parameters (see Figure 9.5). The μs for Intercept and Story type were set to the mean estimates obtained in Experiment 3.1, as they were the most stable effects that were repeatedly obtained in pretest studies. The distributions for load and

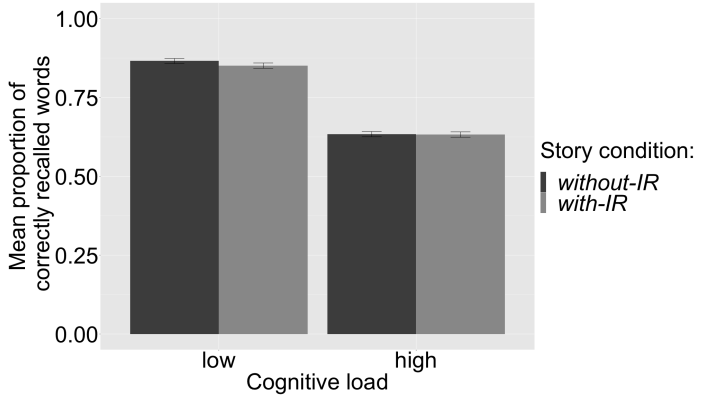


Figure 9.14: Experiment 3.2. Mean proportion of correctly recalled words in a trial (± SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (low vs. high) conditions

Table 9.12: Experiment 3.2. Prior specifications for the Bayesian model of target activity typicality ratings

default	non-informative	weakly informative	more informative	class	coef	group
student_t(3, 0, 2.5)	normal(0, 10)	normal(1.56, 1)	normal(1.56, 0.5)	Intercept		
(flat)	normal(0, 10)	normal(-0.28, 0.5)	normal(-0.28, 0.28)	b	story	
(flat)	normal(0, 10)	normal(0, 2)	normal(0, 0.2)	b	load	
(flat)	normal(0, 10)	normal(0, 2)	normal(0, 0.4)	b	story:load	
student_t(3, 0, 2.5)	as default	as default	as default	sd	Intercept	item
student_t(3, 0, 2.5)	as default	as default	as default	sd	story	item
student_t(3, 0, 2.5)	as default	as default	as default	sd	Intercept	subject
lkj_corr_cholesky(1)	as default	as default	as default	L		item
gamma(0.01, 0.01)	normal(1, 10)	normal(2.5, 2)	normal(2.5, 2)	phi		

story:load interaction were centered around zero. The σ estimates for each of the fixed effects in both weakly informative and more informative priors was set close to the corresponding confidence intervals in Experiment 3.1 but such that some variability still would be allowed (less variability in more informative prior and more variability in weakly informative prior). The parametrization in the non-informative prior for each of the fixed effects was set up to $N(0, 10)$ thus allowing for a wide range of different alternative hypotheses. The parametrization of the prior for ϕ was chosen to take into account that the data were expected to the skewed to the right to different degrees.

Note, that posterior estimates did not differ with different priors.

The full model had the same predictors as the model in Table 9.9. The null model differed from the Bayesian mixed effects beta regression model presented in Experiment 3.2 only by the absence of the interaction term. For each pair of null and full models the Bayes factor estimation was repeated seven times, to check for stability. The most important observation is that the Bayes factor is always in favour of the null hypothesis, independent of the exact settings of the prior. However, the strength of the estimate in favour of the null hypothesis does depend on the settings

of the prior. The observations are consistent with Schad et al. (2023) and Rouder et al. (2009): priors that favour the null hypothesis or smaller effects of the interaction exhibit a smaller odds ratio in favour of the null hypothesis compared to priors that allow for larger interaction effects. Thus, for the weakly informative prior the mean BF01 (evidence in favour of the null over the full model) was equal to $m = 26.87$, $sd = 9$, while for the more informative prior it was $m = 7.25$, $sd = 1.9$. When the prior allowed for a variety of of alternative hypotheses for the interaction (which were not supported by the data) the evidence in favour of the null model increased drastically: $m = 217.4$, $sd = 39.4$.

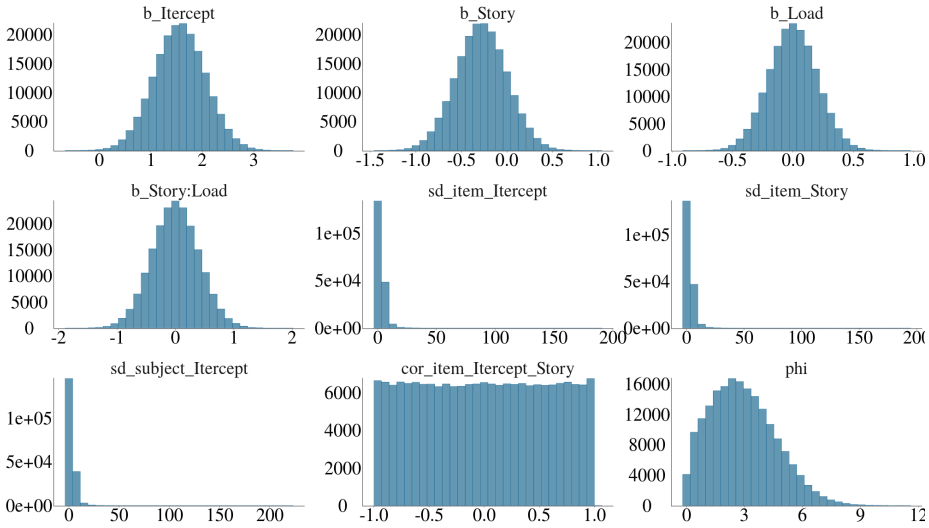


Figure 9.15: Experiment 3.2. Distribution of more informative prior for a set of fixed effect; Prior for other parameters were assigned by default

9.3.11 Discussion

In Experiment 3.2, the cost of atypicality inferences was tested by modulating the load via a reading span task. Compared to Experiment 3.1, using a secondary task of linguistic nature was supposed to increase the interference with the main pragmatic task.

However, there was found no direct evidence for an effect of load on the strength of pragmatic inferences. A Bayesian hierarchical Beta regression analysis of ratings to the target typicality questions included an estimate for a story:load term that was symmetric around zero ($\beta = 0, CI_{95} = [-0.1, 0.1]$). As a follow-up, a Bayes factor analysis was conducted in which the models with and without the interaction term were compared. The Bayes factor consistently suggested evidence in favour of the null hypothesis. Notably, even the models which in their prior assume already that the effect will be very small favoured the null.

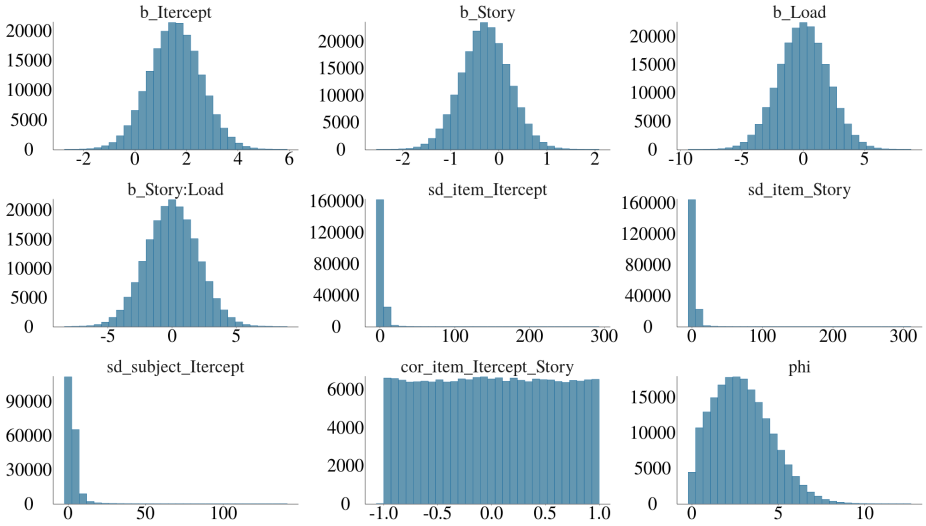


Figure 9.16: Experiment 3.2. Distribution of weakly informative prior for a set of fixed effect; Prior for other parameters were assigned by default

On the other hand, there was found no reduced performance in the secondary task. In general, subjects' mean proportion of recalled words was lower under high load than under low load which justified that conditions of the secondary task varied in difficulty for subjects. However, the number of correctly recalled words was not affected by the presence or absence of the informational redundancy in stories. Thus, under high load, subject did not seem to compensate the same level of performance in pragmatic task by loosing on words' recall.

Finally, consistently with Experiment 3.1, the load coming from the secondary task influenced ratings to the comprehension questions – subjects' answers were less correct under high load than under low load ($\beta = -0.07$, z-value $= 0.03$, p-value $=$ .01) but there was no significant effect of story redundancy or story:load interaction.

Considering the linguistic nature of the secondary task in Experiment 3.2, if the processing of atypicality inferences would be costly, the trade-off between the tasks here should have been more pronounced than in Experiment 3.1. However, neither an effect on pragmatic inferences, nor a compensatory effect in the secondary task performance were found, in contrast to Experiment 3.1. One possible explanation for the failure to find an effect in Experiment 3.2 could be that even the low load manipulation was already sufficiently taxing and distracting, that no difference between the conditions could be found. This is in contrast to Experiment 3.1, where the easy condition involved no secondary task. It was therefore decided to follow-up with a third experiment, which contrasts the high load condition with a no-load condition, where subjects did not have to remember any of the words prior to story reading and question answering.

9.4 Experiment 3.3: low load vs. no load

9.4.1 Materials and experimental setup

The experimental materials were taken from Experiment 3.2. The experimental design and procedure were identical to Experiment 3.2, with the following changes: Contrary to Experiment 3.2, the present experiment consisted of two blocks, a single task block and a dual task block. The dual task block contained the high load condition from Experiment 3.2, i.e., the trials were identical to Experiment 3.2, and each story was always preceded by four sentences. Thus, each subject saw 16 sentences for which the last word had to be memorized in the study. In the single task block, no preceding sentences were shown, and no memory task was required – subjects were only instructed to read the stories and answer story-related questions. The order of blocks was randomly assigned to each participant as they entered the study. Each block consisted of four stories: one story in with-IR condition, one story in without-IR condition, and two fillers. The order of stories within each block was randomized.

On average, the experiment lasted 15 minutes.

9.4.2 Participants

842 subjects completed the study (mean age = 42.2, sd = 13.8; 63.9% female, one subject preferred not to report; mean by-subject acceptability score = 14.4, sd = 1.5; mean by-subject last word recall score = 8.7, sd = 5). The requirements for participation were the same as in Experiment 3.2: eligible participants stated English as their native language, did not take part in the previous studies with the same materials and had an approval rating of 95%.

The data cleaning procedure was the same as in Experiment 3.2. 133 subjects, who recalled less than three words in the whole study (out of 16 possible words), were removed from any further analyses. The proportion of removed subjects based on this criterion is higher than in Experiment 3.2. The analysis of their performance and answers to the questionnaire after the study showed that excluded subjects did not report any technical difficulties; in the recall task, they either did not write any words or they wrote random words from the sentences, but not their last words. One possible explanation for such a high number of low performers might be that in Experiment 3.3, subjects were always faced with four words per trial, while in Experiment 3.2 they had also trials with one, two or three words. This might have made it easier to cross the threshold in Experiment 3.2 compared to Experiment 3.3.

It was decided to nevertheless maintain this criterion for excluding participants, as they might have focused only on the story reading task, and in that case would be less likely to exhibit any effect of load on pragmatic inferences. It is noted, however, that a post-hoc analysis of excluded participants showed no difference in their ratings

between the no load and the high load condition; adding these participants to the sample would hence not change results (see Section 9.4.3).

After data cleaning, 709 subjects (mean age = 42, sd = 13.8; 64.2% female, one person preferred not to report; mean by-subject acceptability score = 14.5, sd = 1.5; mean by-subject last word recall score = 10.2, sd =3.9) were kept for the analyses.

9.4.3 Results: typicality ratings

A generalized mixed effects beta regression model with by-subject and by-item random intercepts and by-item random slopes for story condition was built to analyze subjects' ratings of target activity typicality – see Table 9.13 and Figure 9.17. As in the previous experiments, a significant effect of informational redundancy was found ($\beta = -0.24$, $p < .01$). A significant effect of load was also found ($\beta = -0.11$, $p < .01$), meaning that subjects' typicality ratings were on average lower under high load than under no load. An interaction between story and load was not significant[5].

Table 9.13: Experiment 3.3. Effect sizes (b), standard errors (SE), z-values, and p-values for the generalized mixed effects beta regression model (with logit link) of linguistic judgements. The dispersion parameter for beta family was equal to 2.25. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*z*	*p*
Intercept	1.5	0.09	17.5	***
Story: with IR	-0.24	0.08	-3.1	**
Load: high	-0.11	0.04	-2.7	**
Story*Load	0.12	0.08	-1.5	ns
Random Effects		*Variance*		
Item		0.13		
Story|Item		0.09		
Subject		0.06		

The analysis of participants' ratings in the comprehension questions also did not show a significant effect story and load interaction – see Table 9.14. Only an effect of load was found, meaning that subjects were less attentive to a semantic content of the stories under high load, compared to no load – see Figure 9.18.

[5]Considering that the order of blocks could potentially influence the overall cognitive load of subjects throughout the experiment, the block order and its interactions were also included in the model. The idea behind was that if subjects were firstly exposed to the block with no secondary task, they might have had more resources compared to facing the the same block after the dual task. However, there was no significant 3-way interaction.

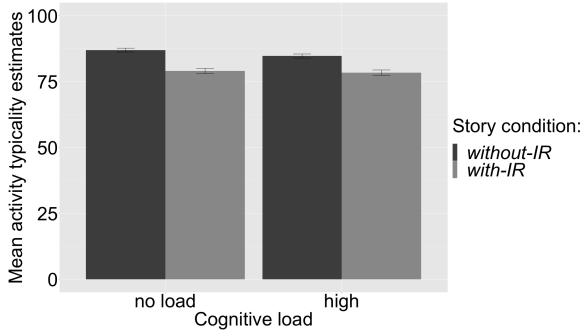


Figure 9.17: Experiment 3.3. Non-transformed mean participants' ratings of the target activity typicality (± SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (no load vs. high) conditions

Table 9.14: Experiment 3.3. Effect sizes (b), standard errors (SE), z-values, and p-values for the generalized mixed effects beta regression model (with logit link) of the transformed ratings in comprehension questions. The dispersion parameter for beta family was estimated as 1.71. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*z*	*p*
Intercept	1.67	0.06	29.84	***
Story: with IR	-0.07	0.04	-1.62	ns
Load: high	-0.22	0.04	-5.31	***
Story*Load	-0.05	0.08	0.62	ns
Random Effects		*Variance*		
Item		0.05		

9.4.4 Results: recalled words

The mean proportion of correctly recalled words was equal to 64.6% which was comparable with the recall rate under the high load in Experiment 9.3 (66%).

A generalized mixed effects binomial regression model of the proportion of correctly recalled words included only story condition. The random structure included by-subject and by-item random intercepts and by-item random slopes for story. There was found no main effect of informational redundancy. The results are presented in the Table 9.15 and Figure 9.19.

9.4.5 Discussion

Experiment 3.3 followed up on Experiment 3.2 by making the manipulation more extreme – instead of comparing low load to high load, the comparison here was between no load and the high load condition. This allows testing whether the lack of

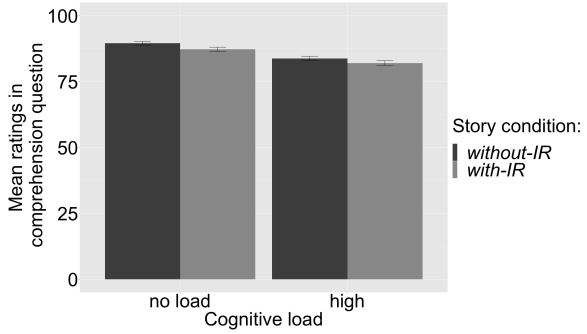


Figure 9.18: Experiment 3.3. Transformed mean participants' ratings in comprehension questions (± SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (no load vs. high) conditions

Table 9.15: Experiment 3.3. Effect sizes (b), standard errors (SE), z-values, and p-values in the binomial mixed effects model (with logit link) of the proportion of correctly recalled words. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*z*	*p*
Intercept	0.8	0.07	11.8	***
Story: with-IR	-0.03	0.1	-0.2	ns
Random Effects		*Variance*		
Subject		1.66		
Item		0.02		
Story|Item		0.32		

interaction between load and informational redundancy in Experiment 3.2 might be due to both the high load and low load condition both inducing cognitive load which affects pragmatic inferences, making it impossible to find a difference between these conditions.

However, no significant interaction between story and load was found. There was only a significant effect of load, showing that subjects' ratings of the target activity typicality, independently of the presence of an IR-utterance, were on average lower under high load, compared to no load. This means that subjects in general tended to pay less attention to the content of the story when they also had to remember the words. Subjects' ratings in the comprehension questions were also significantly lower under high load than under no load.

There were also no differences found related to the order in which the experimental blocks (single task and dual task) were presented. Similarly to Experiment 3.2, no trade-off effects in subjects' performance in the secondary task were observed: in the high load condition, there was no effect of informational redundancy on the proportion of correctly recalled words. Experiment 3.3 thus fully confirms the findings of

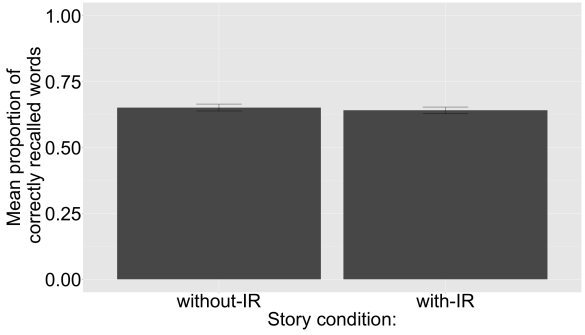


Figure 9.19: Experiment 3.3. Mean proportion of correctly recalled words in a trial (± SEM) aggregated by story condition (without-IR vs. with-IR)

Experiment 3.2.

9.5 General Discussion and conclusions

The experiments reported here were aimed at testing whether a specific type of particularized pragmatic inference, namely atypicality inferences triggered by informationally redundant utterances, are sensitive to externally imposed constraints on cognitive resources, in particular attention and verbal working memory. The results show no robust effects of this type of cognitive load on the derivation of atypicality inferences. This complements the findings reported in Chapter 8, where natural variability in verbal working memory capacity was examined and no effect on the derivation of atypicality inferences was observed. Taken together, these results suggest that both externally imposed constraints and inter-individual differences in these resources do not substantially modulate the derivation of atypicality inferences. Below, the implications of this finding are discussed with respect to processing cost, theoretical accounts of pragmatic processing, and the mechanisms involved in the derivation of atypicality inferences.

Kravtchenko (2022) previously showed that subjects tend to lower their initially high beliefs about the typicality of a highly predictable event (e.g., bringing one's swimsuit at the swimming pool) when this activity is mentioned explicitly. Atypicality inferences are highly context-dependent – in contrast to scalar implicatures, there is no context-independent lexically encoded alternative. Therefore, such inferences should be cognitively costly, according to theoretical accounts of pragmatic processing (Degen & Tanenhaus, 2015; Sedivy, 2007; Sperber & Wilson, 1996).

Following the methods previously used in the studies investigating the costliness of scalar implicatures, a dual-task design was used to manipulate cognitive load. Subjects had to perform two tasks in parallel, thus leaving less resources available for each task compared to when each of the tasks would be addressed separately. One of the predictions was that in a situation of cognitive burden, subjects potentially would

not recognize the redundancy and/or not draw any atypicality inferences based on the redundancy. In both cases, the strength of pragmatic inferences would be expected to be lower on average than in the low or no load conditions. In Experiment 3.1, the load was imposed via a visuo-motor tracking task, where subjects had to continuously track a dot while listening to the stories. In Experiment 3.2 and Experiment 3.3, the load came from a reading span task. In Experiment 3.2, subjects had to memorize 1-4 words and recall them after performing the main pragmatic task. In Experiment 3.3, subjects had to memorize either 4 words (high load) or no words (no load), thus making this experiment more directly comparable to Experiment 3.1.

None of the experiments showed direct evidence for processing cost associated with atypicality inferences – the strength of the atypicality inferences was equal in the high and low/no load conditions which contradicts the prominent existing accounts of the pragmatic processing. Only in Experiment 3.1, there was a small effect on the tracking deviation in the pragmatic condition, indicating that subjects might recruit additional resources for processing the pragmatic utterance, and then might compensate the same level of performance in pragmatic processing by sacrificing performance in the secondary task. It is noted, however, that it is at the point unclear whether the observed effect is due to pragmatic processing, or due to more low-level processes related to processing the exclamatory prosody.

In Experiment 3.2, no compensatory trade-off between the two tasks was found. The lack of effect is not considered to be attributable to power issues, as a large number of participants was hired in Experiment 3.2, enough to detect an interaction effect that is half the size of the main effect of story type with a probability $> 80\%$. The Bayes factor analysis consistently showed evidence in favor of the null hypothesis, i.e., that there is no interaction between cognitive load and pragmatic inference. Considering that in Experiment 3.2 the low load could have been potentially still strong enough to influence pragmatic processing, Experiment 3.3 was conducted, with no and high load conditions. The results of Experiment 3.3 were fully consistent with Experiment 3.2: no effects of load on pragmatic processing were found, nor any effects of informational redundancy on secondary task performance.

A first issue to consider is whether the load manipulations were successful. This is considered to be the case, as a reduction in comprehension answer accuracy was observed in all three experiments when the cognitive load induced by the secondary task was high. Comprehension questions were constructed such that they had clear Yes/No responses based on the text. Participants did not have to, and actually could not, infer the answer to these questions from world or script knowledge.

Secondly, the effect found in Experiment 3.1 could potentially be an artifact of the exclamatory intonation and in fact might not reflect difficulties in pragmatic processing. In this case, the higher tracking deviations found in the region within pragmatic utterance, compared to the region before, would reflect subjects' surprise to the exclamatory intonation itself. Since in Experiment 3.2 the stimuli were presented textually, the exclamation mark simply might not have had a comparable influence

as the auditory intonational prominence, and this could thus be the reason why no effect on secondary task performance was found in Experiment 3.2. To rule out this possible explanation, future work could repeat Experiment 3.1 in a setting where both conditions are using exclamatory prosody, or with an additional condition where the utterances are spoken in neutral intonation. It is noted, however, that the effect sizes for atypicality inferences without exclamatory intonation are expected to be lower in that condition, according to previous work by Kravtchenko (2022) (see also Chapter 6, in particular Table 6.2 comparing mean differences between the story conditions with and without exclamation mark in their work). The experiment would hence require several thousand participants to be well-powered. Finally, an important difference between Experiment 3.1 and Experiment 3.2 can be the modality in which the stories were presented to the participants. While there are no reported differences in memory recall between listening and reading (see e.g., Rogowsky et al., 2016), it still can differently affect online processing in the situation of cognitive overload.

Next, the connection between script knowledge and atypicality inferences in the context of retrieving information from memory will be discussed. As discussed in Section 2.2.1, atypicality inferences are based on script knowledge, which refers to the sequence of events related to everyday activities that are stored in our memory, such as swimming, buying food, and getting hair washed. This type of knowledge is highly crystallized in memory and can be immediately accessed and integrated with context. The literature on world and script knowledge suggests that interlocutors come into a conversation with a set of assumptions about the world, the conversational partner, and other encyclopedic facts of different granularity. These assumptions are progressively updated as the conversation evolves (Hagoort et al., 2004) and new information is introduced (Chwilla & Kolk, 2005). However, listeners cannot hold all assumptions in their working memory, so a search process becomes necessary. Each new piece of information can trigger different sets of assumptions from diverse sources, including perception, short-term memory, and long-term memory. The organization of encyclopedic memory plays a crucial role in the pursuit of relevance, and how information is chunked can either facilitate or hinder the search for accommodating inferences. According to Sperber & Wilson (1996), retrieving relevant information from memory can require more effort than interpreting an utterance, but some pieces of information might be more readily available and easier to retrieve than others.

Thus, one can hypothesize that overinformativity can be recognized very easily: the script is activated by the story topic and induces a high expectation of the target event; this target event is hence easy to retrieve, and its redundancy can be easily recognized. But could the absence of cost for atypicality inferences be explained this way? In a previous experiment reported in Chapter 8, explanations of the given ratings were elicited from participants. The results show that those participants who lowered the typicality rating in the informationally redundant condition often provided richer explanations, indicating engagement in the accommodation step of the derivation process. However, the present studies are not able to determine when

exactly such explanatory inferences are formed – whether immediately upon encountering the utterance, at the end of the sentence, or only at the point of making a judgment or providing a justification. In future work, a new experimental design should be used to investigate the relationship between forming alternative explanations and the effects of cognitive load.

The findings are now discussed in light of two other recent studies, Foppolo et al. (2021) and Fairchild & Papafragou (2021). Foppolo et al. (2021) shows that inference-relevant information, under some conditions, can be obtained while avoiding cost. Using a Picture Selection task, they compared the rate of pragmatic enrichment that children made in classic scalar implicatures (target sentence: *On my birthday cake, some of the candles are burning*; visual scene: a cake with 3/5 burning candles, 0/5, 5/5, and no burning candles) vs. particularized implicatures where to make an inference, children had to judge intentions of the interlocutor based on context (target sentence: *On my bed there is a teddy bear*; visual scene: the bed with a teddy bear and a penguin, only a teddy bear, only a penguin, and an empty bed). The results of this study suggest that contextual cues can be utilized fast enough to avoid the cost of deriving alternatives under some conditions. In their study, contrary to predictions of original accounts, the derivation rate of particularized implicatures was higher than that of scalar implicatures (86% vs. 74% respectively). In addition, the success rate of both implicature types correlated with morphosyntactic competence, while the theory of mind ability correlated only with scalar implicatures but not with particularized implicatures. Foppolo et al. (2021) hypothesize that, under the assumption that children have not yet mastered lexical scales at a level of adults, particularized implicatures, in their study, served contextual cues that helped to grasp the speaker's intentions and made alternatives more visible. While in scalar implicatures, children had to infer the alternative sets based on their mental state reasoning abilities.

It should also be considered whether the hypotheses regarding the processes being involved in particularized inferences are correct, or whether a different set of dual tasks, not involving attention or working memory, might have interfered more strongly with drawing atypicality inferences, and thereby would have elicited a dual-task effect. In a recent study, Fairchild & Papafragou (2021) found that individuals with better theory of mind abilities (ToM) were more likely to compute scalar implicatures. More importantly to the results, no evidence of a unique contribution of executive functions (including working memory capacity) was found. It is noted that in this concurrent study, Fairchild & Papafragou (2021) also failed to find an effect of cognitive load on pragmatic inferences. Fairchild & Papafragou (2021) hypothesize that the executive function might be involved in pragmatic processing only to the extent to which it is recruited in theory of mind reasoning (as such computations might be resources-demanding) and its further consolidation with the rest of contextual information. There might be a possibility that the results are due to atypicality inferences not requiring theory of mind reasoning – namely, the participant

(who acts as an overhearer in the experimental setup) might compute an atypicality inference irrespective of the speaker's or listener's (mutual) knowledge states. In support of this hypothesis, recent findings in language production are considered, which indicate that speakers mention atypical events because they cannot be inferred by a generic listener rather than due to the demands of a specific listener. That is why for computing an atypicality inference, the overhearer's mentalizing skills might not be involved in judging the mental states of the story's interlocutors (Brown & Dell, 1987; Grigoroglou & Papafragou, 2019; Lockridge & Brennan, 2002).

However, the hypothesis that atypicality inferences are entirely cost-free and do not involve theory of mind should be treated with caution. As argued above, recognizing redundancy may proceed without substantial effort due to the activation of script knowledge. However, the accommodation step, in which an explanation for the atypical event is constructed, may require reasoning about the knowledge states of the interlocutors. For example, to explain why Lisa brought her swimsuit, the comprehender may need to consider what is known about Lisa's typical behavior within the shared context of the story. To process explanations about Lisa (e.g., that she is forgetful or that she visits a nudist swimming pool and therefore usually does not bring her swimsuit), the comprehender of the story may need to consider the context and the knowledge states of the story characters. E.g., the speaker and the listener mutually know about Lisa's habits that she often forgets to bring her swimsuit to the swimming pool. This fact about Lisa (or any person who goes swimming) is not in world knowledge per se (only the knowledge that people bring their swimsuits is). The overhearer should compute that this information is in the characters' common ground and that mentioning the normally typical course of actions (that she brought her swimsuit) is informative. In fact, the story contexts are carefully set up to make sure that the characters in the stories are not strangers to each other but that they know each other well. The role of theory of mind in atypicality inference derivation is further investigated in Chapter 10.

Moreover, as shown in Chapter 8, reasoning ability predicted the derivation of atypicality inferences. This suggests that, unlike attention and working memory, which do not appear to constrain these inferences, higher-level reasoning processes may play a role in their derivation. In particular, reasoning may be involved in integrating contextual information and constructing coherent interpretations of the utterance, although the precise stage at which this contribution arises remains to be determined. It remains an open question, however, how cognitive load on reasoning processes could be experimentally manipulated within a dual-tasking paradigm.

Finally, it should be noted that, while the results show no effects of load on the pragmatic processing of a particular type of particularized conversational inferences, this conclusion cannot necessarily be generalized to other PCIs. Future studies should investigate the effect of cognitive load in the processing of other particularized implicatures, specifically those that are less dependent on script knowledge.

Chapter 10

Exp. 4. Atypicality inferences in comparison with other implicature types: Evidence from cross-task responses and individual differences

10.1 Motivation of the study

Understanding an utterance often requires going beyond the surface meaning of the words to making an inference about the speaker's intended meaning. This process, known as a pragmatic implicature, refers to a broad range of implicit meanings that listeners can infer based on pragmatic reasoning. A typical example is scalar implicatures where the quantifier *some* (e.g. "I saw some of your children today") is often interpreted as *not all*, despite the fact that *some* is logically compatible with the meaning *all* (Grice, 1975). In the literature, considerable debate revolves around the mechanism underlying implicature calculation, in particular whether the pragmatic interpretation arises by default, or only with enrichment of the logical meaning. Results reveal a lack of consensus across studies, with some researchers showing that the pragmatic meaning of *some* takes longer and more effort to derive (Bott & Noveck, 2004; Huang & Snedeker, 2009), while others provide support for rapid, immediate implicature (Feeney et al., 2004; Grodner et al., 2010). Variability in implicature sensitivity has also been reported with other scalar forms such as numerals (Marty et al., 2013; Panizza et al., 2009) or embedded disjunctions (e.g., "Every elephant caught a big butterfly or a small butterfly"; Singh et al., 2016; Pagliarini et al., 2018), as well as across different linguistic forms (e.g., quantifiers, modals, adjectives, etc.; van Tiel et al., 2016). Together, these findings paint an inconsistent picture with respect to whether and when listeners draw a pragmatic implicature.

One possible reason for this variability is that most experimental studies investigate a single pragmatic phenomenon in isolation. As a result, relatively little is known about how response tendencies relate across different implicature tasks within the same individuals. In particular, it remains unclear whether individuals who tend to derive pragmatic interpretations in one type of task show similar tendencies in other types, or whether response patterns vary depending on the specific implicature being tested. A small number of recent studies have begun to address this question by examining multiple pragmatic phenomena within the same group of participants. For example, Floyd et al. (2025) administered a large battery of pragmatic tasks and investigated the relationships between different aspects of pragmatic language use. Their results suggest that pragmatic behavior may involve several partially independent components rather than a single uniform pattern across tasks. By testing multiple implicature types within the same experiment, the present study enables investigation into whether response tendencies generalize across different pragmatic phenomena or whether atypicality inferences exhibit a distinct profile. Additionally, the design enables investigation into whether participants exhibit consistent response preferences across time and whether variability in interpretation is linked to individual differences in cognitive and personality traits (see Chapter 5).

Specifically, this study aims to answer three questions:

1. For a given individual, do they respond consistently (e.g., always literal vs. always pragmatic) to different types of implicatures?
2. Are individuals consistent in their responses across time?
3. Are participants' response tendencies driven by individual differences in cognitive and personality traits?

The types of implicature tested here allow a direct comparison between those commonly considered generalized (e.g. bare scalars with *some*, disjunctions, and numerals) and those viewed as more context-dependent (e.g. atypicality inferences). In addition, within the class of generalized implicatures, a range of less frequently studied syntactic forms was included (e.g. embedded *some* in non-monotonic contexts and embedded disjunctions in downward-entailing contexts), allowing a more fine-grained comparison among implicatures that are conceptually closer.

10.2 Materials

The main experiment consisted of two sessions separated by a two-month interval. In both sessions, participants were presented with stimuli from several pragmatic implicature tasks in random order. After each stimulus, they answered questions designed to probe their interpretation using a continuous scale ranging from 0 to 100. The labels on the scale varied depending on the type of item[1]. Across tasks, the

[1]Either *Never-Sometimes-Always*, *Definitely No-Definitely Yes*, or *Definitely False-Definitely True*.

left end of the scale corresponded to rejection and the right end to acceptance of the statement.

Each implicature type and the original studies from which the materials were drawn are described below.

Atypicality inferences: The materials for atypicality inferences were identical to those originally developed by Kravtchenko (2022). The stories describe everyday activities (e.g., going to a restaurant or going shopping). In the target condition (*with-IR*), the story includes an informationally-redundant utterance (e.g., "She ate there!") because the event is normally expected to occur in the described activity without explicit mentioning. In the control condition (*without-IR*), the utterance is omitted and thus no informational redundancy is introduced. In the present experiment, a subset of 20 stories from the original materials was used as experimental items. The remaining four stories (fueling, pasta, baking, and laundry) were included as fillers and were identical for all participants. The filler stories did not include an IR manipulation.

Each story was accompanied by four questions. One target question asked about the typicality of the conventionally habitual event (e.g., "How often do you think Mary usually eats when going to a restaurant?") using a continuous scale from 0 (*Never*) to 100 (*Always*), with the middle of the scale marked *Sometimes*. The remaining three questions served as fillers (see Chapter A).

Under literal interpretation, the typicality ratings in the target question are expected to be similarly high under both conditions. A pragmatic interpretation predicts lower typicality ratings in the target (*with-IR*) condition compared to the control (*without-IR*) condition.

Scalar implicatures with *some* (Bare *some*; De Neys & Schaeken, 2007): Participants judged underinformative sentences with *some* (e.g., "Some oaks are trees") on a scale from "Definitely False" to "Definitely True". Pragmatic interpretations here are evidenced by a "False" rating. In the control condition, the sentence was literally true (e.g., "Some trees are oaks" or "All oaks are trees"). Filler items were constructed so that the question had a clear negative response (e.g., "All tuna are birds" or "Some insects are pigeons").

The materials covered five semantic categories (birds, trees, flowers, fish, and insects), with four instances per category in the target condition – see Section B.1.1 for more details. The permutation of instances, categories, and usage of *some* / *all* quantifiers was balances according to Bott & Noveck (2004).[2]

Bare numerals (Marty et al., 2013): Participants saw pictures of four dice with dots colored either red or green and judged sentences containing numerals for which both a weak reading ("at least") and a strong ("exactly") reading were possible (e.g., "4 dots are red"). Responses were given on a scale from "Definitely False" to "Defi-

[2]All materials were pretested to ensure that sentences in each condition (target, control, and filler) yielded the expected ratings.

nitely True". Pragmatic interpretations are evidenced by "False" ratings in the target condition. In the filler condition, the sentence was false with respect to the picture.

Stimuli were generated following the procedure described in Marty et al. (2013); see Section B.1.2 for details.

Scalar implicatures with disjunction (Bare disjunctions; Singh et al., 2016)[3]: In the target condition, participants saw pictures of a character wearing two clothing items and judged sentences with a disjunction for which both a weak reading (inclusive or) and a strong reading (conjunctive or) were possible (e.g. "The man is wearing a hat or a scarf"). Under a literal interpretation, the sentence is true whenever at least one of the items is worn, whereas a pragmatic interpretation requires that exactly one of the items be worn. In the control condition, the character in the picture was wearing only one of two items mentioned in the sentence.

In the filler condition, two types of sentence were used ("The man is wearing a hat and a bowtie." or "The man is wearing a hat."), which were false with respect to the picture – see Section B.1.3 for more details.

Disjunctions in downward-entailing contexts (Embedded disjunctions; Singh et al., 2016): In the target condition participants saw pictures of three characters each wearing two clothing items and judged sentences with a disjunction (e.g. "Every man is wearing a hat or a scarf"), on a scale from "Definitely False" to "Definitely True". Pragmatic interpretations in the target condition are evidenced by "False" ratings in the condition when the characters are all wearing both items, indicating the strong disjunctive reading of *or*. In the control condition, the characters in the picture were wearing either one item (but not both) mentioned in the sentence.

In the filler condition, similarly to bare disjunctions, two types of sentences were used ("Every man is wearing glasses." or "Every snowman is wearing a belt and a scarf."), which were false with respect to the picture. See details in Section B.1.4.

Scalar implicatures with *some* in non-monotonic contexts (Embedded *some*; Potts et al., 2016). Participants saw pictures of three basketball players performing a number of shots. For each player, the outcomes of the shots were indicated by colored balls, with green balls representing successful shots and red balls representing misses. Participants judged whether a sentence matched the picture. The critical sentences contained *some* embedded under *exactly one* (i.e., *Exactly one player hit some of his balls*). Following Potts et al. (2016), pictures were constructed using three outcome types for each player: none (N), some but not all (S), and all (A) successful shots. The present study included two picture configurations. In the **NNA** condition, two players hit none of their balls, and one player hit all of his balls. In the **NSA** condition, the players hit none, some but not all, and all of their balls, respectively. These configurations differ in whether exactly one player satisfies the predicate *hit some of his balls* depending on how *some* is interpreted, and thus

[3]The visual materials were adapted from (Frank et al., 2016).

allow testing whether participants derive strengthened interpretations of *some* in this embedded environment. In the NNA condition, the sentence is true under a literal interpretation of *some* ('at least one') but false if *some* is interpreted as *some but not all*. In the NSA condition, the pattern reverses: the sentence is false under the literal interpretation but true if *some* is locally enriched to *some but not all*. No additional control condition was included.

In the filler items, the sentences were false with respect to the picture. More details on the generation of target and filler items are provided in Section B.1.5.

***Some* in face-threatening contexts (Politeness implicatures; Bonnefon et al., 2009):** Participants read short stories describing their own performance of a public activity. Each scenario ended with an underinformative statement of the form "Some people X-ed your [activity]", where X either boosted the participant's face (e.g., *loved*; target condition, where the scalar inference is expected) or threatened it (e.g., *hated*; control condition, where the inference is expected to be attenuated). Participants then judged whether an interpretation compatible with *all* remained possible (e.g., "Given what Denise told you, do you think that it is possible that everybody loved your speech?"), on a scale from "Definitely no" to "Definitely yes". Following Bonnefon et al. (2009), ratings are expected to be higher in the control (*hate*) condition than in the target (*love*) condition, as participants tend to cancel the implicature *some but not all* in face-threatening contexts. These items were excluded from further analyses due to strong order effects: participants showed a pattern of results similar to the original study only when they encountered the target (*love*) condition first. More details are provided in Section B.1.6.

10.3 Design and experimental procedure

Experimental design

The main experiment consisted of two experimental sessions separated by approximately two months. In each session, participants completed one of ten experimental lists. Each list contained 42 items drawn from eight implicature tasks included in the study. The breakdown of target, control, and filler items per implicature type for one session is shown in Table 10.1.

The experimental lists were organized into five paired sets (1–6, 2–7, 3–8, 4–9, and 5–10). Each participant received one list from a pair in Session 1 and the corresponding paired list in Session 2. This pairing ensured that items requiring condition reversal across sessions were presented in the opposite condition in the second session. In particular, for atypicality inferences and politeness implicatures, items appearing in one condition in Session 1 appeared in the complementary condition in Session 2.

For atypicality inferences, each list contained two target items in the with-IR condition, two control items in the without-IR condition, and four filler items. Across

Table 10.1: Experiment 4. Number of trials each subject saw per session of the main experiment by target/control/filler conditions.

Type	Number of trials
Atypicality	2/2/4
Bare *some*	4/2/2
Bare numerals	4/2/2
Bare disjunction	2/1/1
Embedded some NNA	2/0/2
Embedded some NSA	2/0/2
Embedded disjunction	2/1/1
Face-threatening *some*	1/1/0

the two experimental sessions, the lists were arranged such that participants saw the opposite condition for the same stories in Session 2: items presented in the with-IR condition in Session 1 appeared in the without-IR condition in Session 2, and vice versa. This allowed comparison of ratings within the same story across sessions.

The politeness implicature task contributed two items per session, corresponding to two story topics. Each participant encountered each story topic only once per session. Across the two sessions, the assignment of topics to conditions was reversed: topics presented in target condition in Session 1 appeared in the control condition in Session 2, and vice versa. In addition, the ordering of the two politeness items within a session was constrained so that one item appeared among the first half of the trials and the other among the second half.

For bare numerals, bare disjunctions, embedded disjunctions, and embedded *some*, stimuli were sampled for each participant from a larger pool of pre-generated items matching the relevant condition. For atypicality inferences, bare *some*, and politeness implicatures, items were instead distributed across the experimental lists with additional balancing constraints – see additional details in Section B.1.

Within each session, items from all implicature types were presented in randomized order.

Experimental procedure

Participants completed the experiment online on the Prolific crowdsourcing platform. On Prolific, participants were informed that the study was part of a series of seven related studies conducted over approximately two months (sessions 1, five cognitive tests as described below, and session 2)[4], and that completing all studies would result in a bonus payment. A schedule of upcoming studies was provided to participants in advance, which helped to ensure a high return rate between sessions.

[4]Participants were asked to complete the three additional cognitive tests as a follow-up after the study ended (see Section "Cognitive measures").

Upon starting the study, an instruction screen was presented explaining that they would be asked to answer questions about short stories, pictures, and sentences. The instructions emphasized that participants should respond by adjusting a slider on a continuous scale and that they should read each item carefully before answering.

Before the main experiment began, participants completed a short practice block designed to familiarize them with the response interface and the types of tasks used in the study. The practice block consisted of five items. The first item ensured that participants understood how to use the slider scale: participants were asked to move the slider completely to the right and then completely to the left. The remaining practice items illustrated the types of tasks used in the experiment. In one practice trial, participants viewed a picture of several objects and judged the truth of a sentence ("All Christmas trees are decorated with a star.") with respect to the picture using a scale ranging from "Definitely False" to "Definitely True" (correct answer: definitely true). In another practice trial, participants read a short sentence describing an event (e.g., acquiring a pet) and rated how likely several possible interpretations were by adjusting separate sliders for each option (e.g., dog, whale, cat, alligator) on a scale ranging from "definitely not" to "definitely yes" (correct responses: higher ratings for *cat* and *dog*, lower ratings for *whale* and *alligator*; this item was taken from Kravtchenko (2022)). The fourth practice item presented a picture of three boys accompanied by the sentence "Two boys are holding a banana." The sentence was false with respect to the picture, in which two boys were holding an apple and only one boy was holding a banana (Singh et al., 2016). Finally, a practice item introduced the basketball-shot display used in the embedded *some* task (Potts et al., 2016). Participants saw three players with green and red balls indicating successful and missed shots, respectively, and judged the truth of a sentence describing the situation (correct answer: false).

Participants were required to provide correct responses to the practice trials before continuing with the main experiment. After the practice block, participants were shown a final instruction screen informing them that the experiment would now begin and reminding them to read each item carefully and use the full range of the response scale.

The main experiment consisted of 42 items per session and took approximately 15 minutes to complete. Items appeared in a randomized order. In each trial, participants read a short story, sentence, or description accompanying a picture and responded by adjusting a slider on a continuous scale ranging from 0 to 100, with labels depending on the specific task (e.g., "Definitely False" to "Definitely True", "Definitely No" to "Definitely Yes", or "Never" to "Always").

After completing the experiment, participants answered a short set of demographic and debriefing questions, including their native language, their interpretation of the purpose of the study, and whether they experienced any technical problems during the experiment. Participants were also given the opportunity to leave additional comments.

Cognitive measures

In addition to the main experimental task, participants completed a set of cognitive and personality measures.

Between the two sessions of the main experiment (approximately one week apart), participants completed five tests in the following order: Autism-Spectrum Quotient (AQ; autistic traits), Keep Track Task (KTT; memory updating), Stroop task (Stroop; inhibition), Reading Span Test (RSpan; verbal working memory capacity) and Reading the Mind in the Eyes Test (RMET; theory of mind). If subjects could not join the study on time, it was reopened upon request.

At a later stage of the project, participants additionally completed three additional tests: Author Recognition Test (ART; print exposure), Cognitive Reflection Test (CRT; cognitive processing – the ability to suppress intuitive but incorrect responses) and Raven's Progressive Matrices (Raven's IQ; nonverbal intelligence).

A detailed description of each test is presented in Chapter 5.

10.4 Analysis

Three analyses were conducted, corresponding to the three research questions of the experiment. To address Research Question 1, correlations between participants' responses across the different implicature types were computed. Research Question 2 concerned the consistency of responses across the two experimental sessions. For this analysis, different approaches were used depending on the structure of the materials. For the implicature types involving quantifiers and disjunction, intraclass correlation coefficients (ICC) were computed between participants' mean target ratings in Session 1 and Session 2. For atypicality inferences, where responses varied substantially across stories and each participant evaluated only a subset of items, consistency across sessions was assessed using a Bayesian beta mixed-effects regression model that explicitly accounted for item-level variability. Finally, to address Research Question 3, mixed-effects regression models were fitted for each implicature type to test whether cognitive and personality measures predicted participants' interpretations (pragmatic vs. literal). A detailed description of each analysis is provided in the following subsections.

For all analyses, ratings in the NSA condition of embedded *some* were re-coded as $100 - rating$ so that lower ratings corresponded to pragmatic interpretations. This recoding was introduced to ensure that the interpretation of the rating scale was consistent across conditions: in all analyses, lower scores indicate a more pragmatic response.

Cognitive and personality measures. For the analyses, the following outcome variables were derived. RSpan performance was calculated as the mean proportion of words correctly recalled within each set. Stroop performance was operationalized as

the mean latency difference (in ms) between incongruent and congruent trials. KTT performance was measured as the proportion of correctly recalled words. AQ was represented by the total AQ score. RMET performance was calculated as the proportion of correctly answered items. CRT performance was defined as the proportion of correctly answered critical questions (out of those participants indicated they had not previously encountered). Raven's IQ was calculated as the number of correctly answered items. ART score was computed by subtracting the number of incorrectly identified fictitious authors from the number of correctly identified real authors.

The Pearson's pairwise correlations between the measures (Section 10.6.4) were calculated using `RcmdrMisc` package, (version 2.7-2; Fox, 2022), with p-values adjusted using the Holm method.

Implicature correlation analysis (Section 10.6.2). To investigate whether subjects' answers are consistent across different implicature types, a correlation analysis was performed. For the atypicality inferences, firstly the difference scores between target and control conditions were calculated for each subject and item, resulting in four data points per subject. The average of the four data points for each subject was used in the correlation analysis. For the rest of implicatures the mean ratings per subject in the target condition were calculated. The Pearson's pairwise correlations were calculated using `RcmdrMisc` package, (version 2.7-2; Fox, 2022), with p-values adjusted using the Holm method.

Consistency of responses across experimental sessions (Section 10.6.3).

To assess whether participants' responses remained stable across the two experimental sessions, separate analyses were conducted for atypicality inferences and for the remaining implicature types.

For bare *some*, bare numerals, bare disjunctions, embedded disjunctions, and embedded *some*, item-level variability was negligible and the target items were comparable within each implicature type. Consistency across sessions could therefore be assessed by aggregating responses at the participant level. For each participant and implicature type, mean target ratings were calculated separately for Session 1 and Session 2. The intraclass correlation coefficients (ICC) were then computed between these session-level means to estimate the reliability of participants' responses across sessions. Intraclass correlation coefficients were computed using a two-way mixed-effects model for absolute agreement (ICC(3,1)), using the `psych` package (version 2.3.6; Revelle, 2022).

For atypicality inferences, an aggregation-based reliability analysis was not appropriate. Each participant evaluated only a small subset of the available stories, and stories differed substantially in their baseline ratings. In addition, the same stories appeared across the two sessions with their experimental conditions reversed. As a result, computing session-level agreement measures would confound participants' behavior with story-specific variability. To account for this item variability, responses to atypicality inferences were analyzed using a Bayesian beta mixed-effects regres-

sion model. Prior to entering the analysis, the ratings to the target question were transformed from the original $[0, 100]$ scale to the open unit interval $(0, 1)$. This transformation involved dividing the original ratings by 100 and replacing exact 0 and 1 values with 0.001 and 0.999, respectively. The choice of a beta distribution is justified by the nature of the data: the ratings are bounded by the slider endpoints and exhibit a strong negative skew – the beta distribution is well-suited to modeling such proportion-like data, as its shape, determined by the parameters μ and φ, allows it to flexibly model skewed and bounded distributions (Ferrari & Cribari-Neto, 2004) – see Table 10.6 for histograms of the ratings in the experimental conditions across the tested implicature types. The model included story condition (with-IR vs. without-IR), experimental session (Session 1 vs. Session 2), and their interaction as fixed effects. Both predictors were sum-coded (-0.5, $+0.5$). Random effects structure included by-subject and by-item random intercepts as well as by-item random slopes for condition. The model was fitted in R using the `brms` package (version 2.20.4; Bürkner, 2017). Posterior summaries and parameter estimates were obtained using the `bayestestR` package (version 0.15.0; Makowski et al., 2019). The model was estimated with four Markov chain Monte Carlo chains and 4000 iterations per chain (including warm-up iterations), using the default priors implemented in `brms`. Posterior distributions were summarized using medians and 95% credible intervals. In addition, the proportion of the posterior distribution falling within a region of practical equivalence (ROPE) around zero was calculated using the `bayestestR` package.

Individual difference analysis (Section 10.6.5). Each inference type was analyzed in a separate beta mixed-effects regression model, with a logit link for the location parameter μ and an identity link for the precision (dispersion) parameter φ, as implemented in the `glmmTMB` package (Brooks et al., 2017) in R (R version 4.3.0; `glmmTMB` version 1.1.7).

For each inference type the model included the eight collected cognitive and personality scores as fixed effects. To account for potential learning effects over time, trial order was also included as a fixed effect. For the analysis, the original trial order from both experimental sessions was recoded into a continuous sequence: presentation orders of items from the two sessions were concatenated so that the first half of the sequence corresponded to session 1, and the second half to session 2.

All continuous predictors in the models were centered and scaled prior to the analysis. The binary variables were $+0.5/-0.5$ sum-coded, as described below.

In the models for bare *some*, bare numerals, bare disjunctions, embedded disjunctions, and embedded *some* only the target condition (where pragmatic inference is supposed to arise) was modeled. The control condition was omitted due to strong ceiling effects – see Figure 10.1 for the distribution of the ratings in the control condition, as well as the description in Section 10.6.1.

Further, based on the correlational analysis presented in Section 10.6.2, the ratings in bare disjunctions and embedded disjunctions were analyzed in the same model.

This model in addition included the implicature type as a fixed effect (sum-coded as $+0.5/-0.5$ for the bare disjunctions and embedded disjunctions types, respectively), as well as the interactions of the implicature type with eight measures of individual differences. Similarly, the data in the two conditions of embedded *some* (NNA and NSA) were analyzed in the same model. The variable was sum-coded as $+0.5/-0.5$ for NSA and NNA conditions, respectively. These two models additionally included the interaction of condition with trial order.

The models for atypicality inferences included ratings in both target and control story conditions due to observed variation in the control without-IR condition. This variable was sum-coded as $+0.5/-0.5$ for the with-IR and without-IR conditions, respectively. The model also included interactions of story conditions with eight measures of individual differences and the interaction of story condition with trial order.

For atypicality inferences, the model's maximal random-effects structure included by-subject and by-item random intercepts, as well as by-subject and by-item random slopes for story condition. For bare *some*, the model's maximal random-effects structure included by-subject and by-item random intercepts. For bare numerals, the model's maximal random-effects structure included by-subject random intercepts. For bare disjunctions, embedded disjunctions, and embedded *some* the model's maximal random-effects structure included by-subject random intercepts as well as by-subject random slopes for condition. In cases of non-convergence, the random-effects structure was progressively simplified until convergence was achieved, following recommendations from Barr et al. (2013). Any such model simplifications are reported in Section 10.6.5.

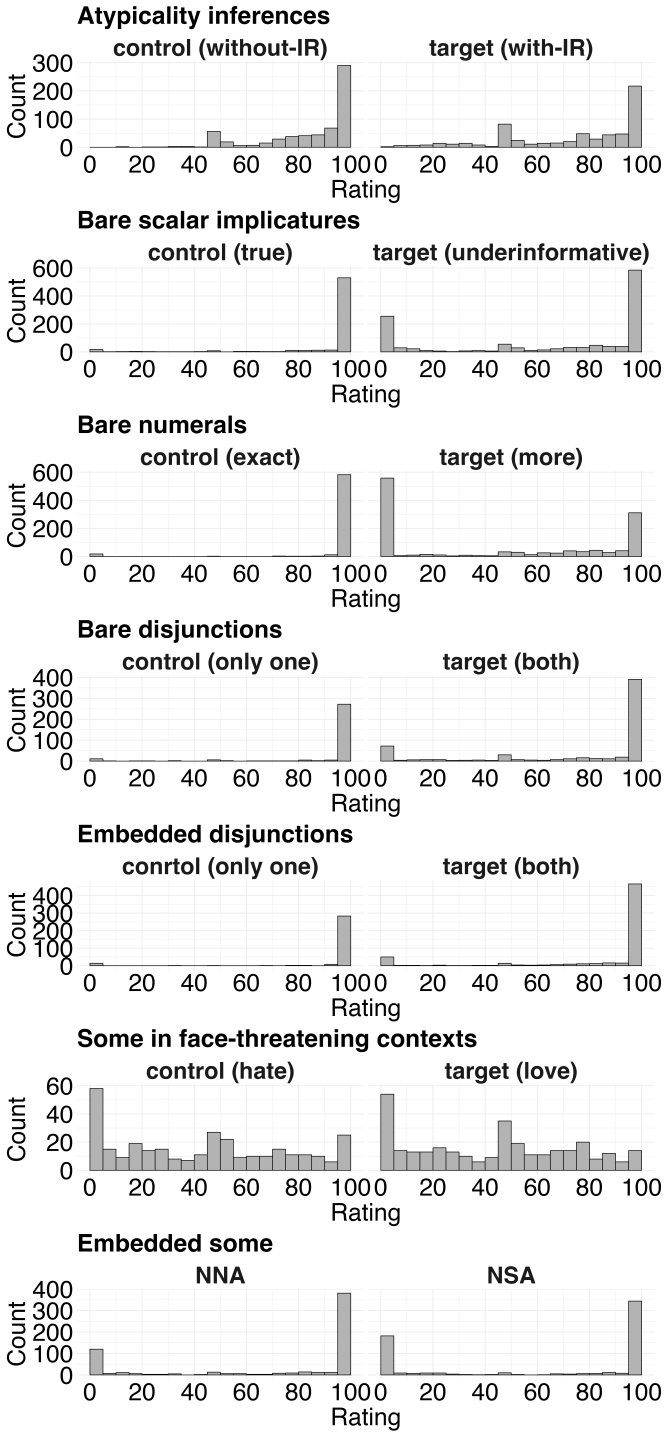


Figure 10.1: Experiment 4. Histograms of the ratings for each implicature type, shown separately for each condition. Ratings in the NSA condition for embedded *some* were reversed (100 − rating) so that, across all target conditions and implicature types, lower ratings correspond to pragmatic responses.

10.5 Participants

A total of 400 participants were recruited via the Prolific crowdsourcing platform. The study was open to native English speakers currently residing in English-speaking countries and reporting no vision problems.

Forty participants were excluded because they did not complete the full set of cognitive tests administered between the two experimental sessions (Autism-Spectrum Quotient, Keep Track Task, Stroop task, Reading Span Test, and Reading the Mind in the Eyes Test). Nineteen additional participants were excluded prior to analysis because they reported technical issues during the experiment or indicated in the post-experiment questionnaire that English was not their native language. Another 23 participants were excluded because their completion time for the experiment was substantially longer than the average ($>$ 32 minutes; mean completion time = 17 minutes). The study also included several control questions with objectively correct yes (100 rating) / no (0 rating) answers to assess attention and comprehension. No participants performed below the chance level and, therefore, no additional exclusions were made based on these checks.

After these exclusions, data from 318 participants were available. At a later stage of the project, these participants were re-invited to complete three additional cognitive measures: the Author Recognition Test (ART; print exposure), the Cognitive Reflection Test (CRT; cognitive processing – the ability to suppress intuitive but incorrect responses), and Raven's Progressive Matrices (Raven's IQ; nonverbal intelligence). A total of 156 participants completed these follow-up measures.

The final dataset therefore consisted of 156 participants who completed all cognitive measures and both experimental sessions with implicature tasks (89 female, 67 male; mean age = 41.4 years, SD = 15.1, range = $[19 - 74]$).

10.6 Results

10.6.1 Replication of pragmatic effects

As a first step, mean ratings in the target and control conditions are reported to verify that the experimental manipulations produced the expected response patterns for each implicature type. The descriptive statistics of participants' responses are presented in Table 10.2, and the distributions of ratings are shown in Figure 10.1.

Ratings for atypicality inferences in both conditions are comparable to those observed in the previous experiments reported in this dissertation.

For the remaining implicature tasks, the results generally follow the expected pattern: in tasks involving control conditions, mean ratings in the target condition were consistently lower than in the control condition, reflecting the intended pragmatic interpretation.

Table 10.2: Experiment 4. Mean participants responses (SD) in target and control conditions for each implicature type.

Implicature	Target (SD)	Control (SD)
Atypicality inferences	75.43 (25.86)	85.45 (18.27)
Bare some	66.46 (40.62)	93.08 (20.42)
Bare numerals	44.50 (43.65)	96.19 (17.01)
Bare disjunctions	77.79 (35.24)	92.80 (21.98)
Embedded disjunctions	86.22 (29.49)	94.31 (21.16)
Embedded some NSA	63.14 (45.38)	–
Embedded some NNA	72.19 (40.72)	–
Politeness implicatures	43.20 (30.88)	43.08 (32.50)

Direct numerical comparison with the original studies is difficult because they used various response scales. Nevertheless, the overall response patterns are comparable. For bare *some* and bare numerals, the difference between target and control conditions aligns with the effects reported in the original studies. Similarly, the patterns observed for bare and embedded disjunctions correspond to those reported by Singh et al. (2016) for adults, with higher ratings in the control than in the target condition. For embedded *some*, the relative pattern of responses in the NSA and NNA conditions is also consistent with the results reported by Potts et al. (2016), despite differences in response scales.

The only implicature type for which the expected pattern was not replicated concerns politeness implicatures.

Politeness implicatures. According to Bonnefon et al. (2009), ratings in the control *hate* condition should be higher than in the target *love* condition because listeners tend to cancel the implicature "*some but not all*" in face-threatening contexts. When someone says "some people hated your speech", the statement is already face-threatening, and there is no politeness reason to understate the number of people who disliked the activity. As a result, listeners may consider it possible that *all* people hated the activity. In contrast, saying "some people loved your speech" may be interpreted as a polite understatement, encouraging the inference "*some but not all*".

In the present study, this effect was generally not replicated. Mean ratings were very similar across conditions (*hate*: $M = 43.1$, $SD = 32.5$; *love*: $M = 43.2$, $SD = 30.9$). Further inspection revealed that the expected pattern emerged only when participants encountered the *love* condition first in the experimental session (*hate*: $M = 41.1$, $SD = 33.4$; *love*: $M = 35.0$, $SD = 29.4$). When the *hate* condition appeared first, the pattern was reversed (*hate*: $M = 44.7$, $SD = 31.7$; *love*: $M = 49.9$, $SD = 30.5$) – see Figure 10.2. Because the expected politeness implicature effect was not replicated and responses showed strong order effects, this inference type was excluded from the subsequent analyses.

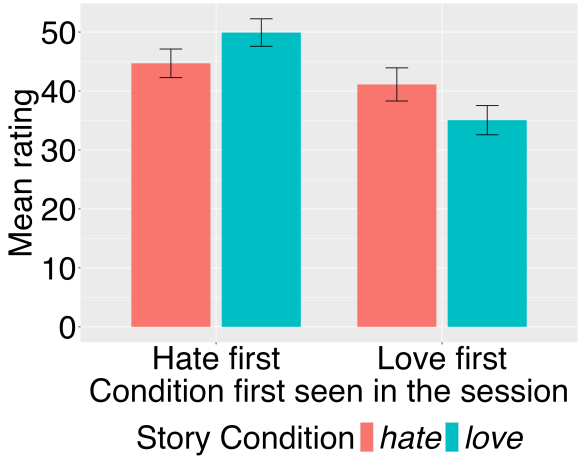


Figure 10.2: Experiment 4. Politeness implicatures. Mean ratings ($\pm SE$) by story condition and order of presentation.

10.6.2 Implicature correlations

To address the first research question, namely whether participants' responses are consistent across different implicature types, pairwise correlations were computed between participants' mean responses in each implicature task. For each participant, mean ratings were calculated separately for atypicality inferences, bare *some*, bare numerals, bare disjunctions, embedded disjunctions, and the two embedded *some* conditions (NNA and NSA). All variables were centered and scaled prior to the analysis. Pearson correlations were computed, and p-values were adjusted using Holm correction for multiple comparisons.

The correlation matrix is shown in Figure 10.3. The strongest association was observed between bare disjunctions and embedded disjunctions ($r = 0.86$, $p_{adjusted} < .0001$). A strong correlation was also found between the two embedded *some* conditions (NNA and NSA; $r = 0.69$, $p_{adjusted} < .0001$). In addition, bare *some* correlated with several implicature types, including bare disjunctions ($r = 0.55$, $p_{adjusted} < .0001$), embedded disjunctions ($r = 0.49$, $p_{adjusted} < .0001$), and embedded *some* in the NNA condition ($r = 0.36$, $p_{adjusted} < .0001$). A weaker correlation was observed between embedded disjunctions and the NSA condition ($r = 0.25$, $p_{adjusted} = .02$).

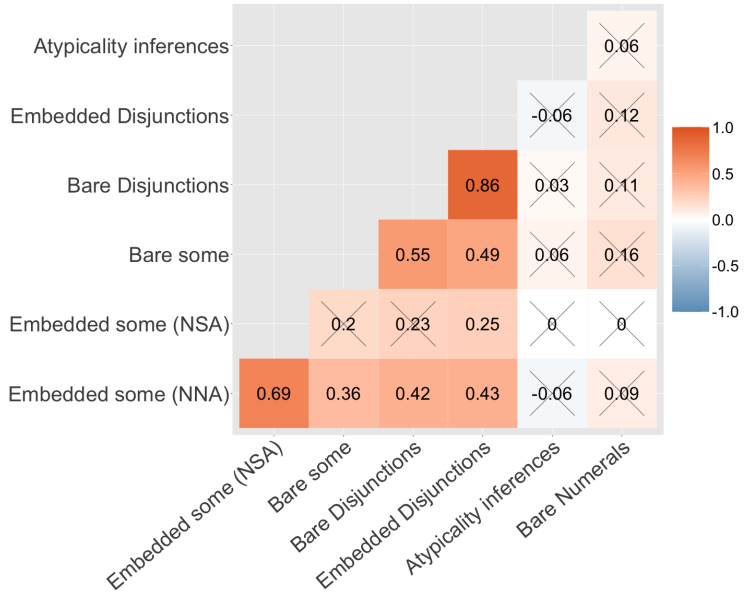


Figure 10.3: Experiment 4. Pairwise correlations ($N = 156$) of by-subject mean answers to different implicature types (p-values calculated with Holm correction; non-significant correlations are crossed out).

No significant correlations were observed for atypicality inferences or bare numerals after Holm correction. In particular, responses in the atypicality task were not reliably associated with responses in the other implicature tasks.

To further examine the structure underlying these correlations, exploratory and confirmatory factor analyses were conducted (see Section B.2 for full details). The results were consistent with the observed correlations. In particular, the analyses suggested two dimensions: one primarily associated with the disjunction tasks (with a weaker contribution from bare *some*) and another defined by the embedded *some* conditions. In line with the correlation analysis, atypicality inferences and bare numerals were not strongly captured by either factor.

10.6.3 Consistency of responses across experimental sessions

The second research question pertained to the consistency of responses across experimental sessions. For atypicality inferences, a Bayesian beta mixed-effects regression model was fitted with condition, experimental session, and their interaction as predictors. For the other implicature types, intraclass correlation coefficients (ICC) were computed between participants' session-level mean ratings in the target condition. The details of the analysis are described in Section 10.4.

Atypicality inferences

Posterior summaries of the model are presented in Table 10.3. The model revealed a clear effect of story condition. Ratings were lower in the with-IR condition than in the without-IR condition, indicating that the manipulation successfully triggered atypicality inferences.

Table 10.3: Experiment 4. Posterior summaries of the Bayesian beta regression model predicting ratings for atypicality inferences. Estimates are shown on the logit scale. The precision parameter was $\phi = 3.10$ with 95% CI $[2.83, 3.39]$.

Predictor	Median	95% CI
Intercept	1.56	[1.34, 1.78]
Story condition (with-IR)	-0.42	[-0.61, -0.25]
Session (Session 2)	-0.06	[-0.17, 0.06]
Story condition : Session	0.01	[-0.22, 0.24]
Group-level effects (SDs)		
Item intercept	0.41	[0.27, 0.62]
Item slope (story condition)	0.29	[0.08, 0.52]
Subject intercept	0.62	[0.52, 0.72]

Crucially for the present research question, there was no evidence that the size of the IR effect differed between sessions. The interaction between condition and session, which tests whether the manipulation effect changed across sessions, was centered around zero ($median = 0.01$, 95% credible interval $[-0.22, 0.24]$). The posterior distribution showed no directional evidence for a change in the effect (probability of direction = 54%), and the majority of the posterior mass (92%) fell within the region of practical equivalence around zero. This indicates that the size of the IR effect remained essentially constant across the two experimental sessions.

There was also no evidence for a general shift in ratings between sessions. The main effect of session was small ($median = -0.06$, 95% credible interval $[-0.17, 0.06]$), and the entire posterior distribution fell within the region of practical equivalence. This suggests that participants did not systematically change their overall rating behavior in the second session.

Taken together, these results provide no evidence that participants' responses to atypicality inferences changed across the two experimental sessions. The IR manipulation reduced ratings, and the size of this effect remained stable across sessions.

Other types of implicature

Table 10.4 presents the ICC estimates and their 95% confidence intervals for each implicature type. ICC values ranged from .47 to .57 across implicature types. Bare *some* yielded an ICC of .57 (95% CI [.46, .67]), while bare numerals showed a similar

level of reliability (ICC = .54, 95% CI [.42, .64]). Bare disjunctions and embedded disjunctions produced comparable values (ICC = .55 and ICC = .54, respectively). For embedded *some*, the NNA condition showed an ICC of .54 (95% CI [.42, .65]), whereas the NSA condition yielded a somewhat lower value (ICC = .47, 95% CI [.34, .59]).

Overall, the results indicate moderate consistency of pragmatic judgments across the two experimental sessions. Participants' responses were hence not perfectly stable across sessions, but individuals who tended to provide more pragmatic (or more literal) responses in one session generally showed similar tendencies in the other session.

The observed level of reliability suggests that pragmatic interpretations exhibit a degree of variability across sessions, while still reflecting stable individual differences in response tendencies.

Table 10.4: Experiment 4. Test-retest reliability across experimental sessions by implicature type.

Implicature type	ICC(3,1)	95% CI
Bare *some*	0.57	[0.46, 0.67]
Bare numerals	0.54	[0.42, 0.64]
Bare disjunctions	0.55	[0.43, 0.65]
Embedded disjunctions	0.54	[0.42, 0.64]
Embedded *some* (NNA)	0.54	[0.42, 0.65]
Embedded *some* (NSA)	0.47	[0.34, 0.59]

10.6.4 Cognitive and personality measures

Descriptive statistics

The descriptive statistics for each test is presented in Table 10.5, as well as the histograms of the distributions are shown in Figure 10.4.

In the experiment presented in Chapter 8 (Section 8.5), a subset of the cognitive and personality measures reported here was collected. Notably, the distributions of these overlapping measures were comparable across the two studies. Specifically, AQ scores were very similar (Chapter 8: $M = 20.52$, $SD = 8.37$; present study: $M = 20.48$, $SD = 7.17$). A similar pattern was observed for ART (Chapter 8: $M = 19.3$, $SD = 14.28$; present study: $M = 21.78$, $SD = 11.16$), CRT (Chapter 8: $M = 0.31$, $SD = 0.26$; present study: $M = 0.39$, $SD = 0.26$), Raven's Progressive Matrices (Chapter 8: $M = 5.38$, $SD = 2.12$; present study: $M = 5.54$, $SD = 1.93$), and RSpan (Chapter 8: $M = 0.76$, $SD = 0.2$; present study: $M = 0.75$, $SD = 0.17$).

Table 10.5: Experiment 4. Descriptive statistics for cognitive and personality measures collected for subjects participated in the experiment ($N = 156$).

Task	Possible range	Observed Range	Mean (SD)	Skewness	Kurtosis
AQ (overall)	0 – 50	4.00 – 47.00	20.48 (7.17)	0.63	3.65
RMET	0 – 1	0.47 – 0.97	0.76 (0.10)	-0.50	2.88
Stroop	–	-21.36 – 489.70	221.07 (95.25)	0.19	2.63
KTT	0 – 1	0.50 – 0.97	0.77 (0.11)	-0.61	3.28
RSpan	0 – 1	0.04 – 1.00	0.75 (0.17)	-1.26	5.07
ART	-65 – 65	2.00 – 51.00	21.78 (11.16)	0.50	2.55
CRT	0 – 1	0.00 – 1.00	0.39 (0.26)	0.43	2.62
Raven's IQ	0 – 10	1.00 – 10.00	5.54 (1.93)	-0.14	2.14

Note: **AQ** = Autism-Spectrum Quotient (overall score; autistic traits); **RMET** = Reading the Mind in the Eyes Test (theory of mind ability); **Stroop** = Stroop Task (inhibition); **KTT** = Keep Track Task (memory updating); **RSpan** = Reading Span Task (verbal working memory span); **ART** = Author Recognition Test (exposure to print); **CRT** = Cognitive Reflection Test (cognitive processing – the ability to suppress intuitive but incorrect response); **Raven's IQ** = Raven's Progressive Matrices (non-verbal intelligence).

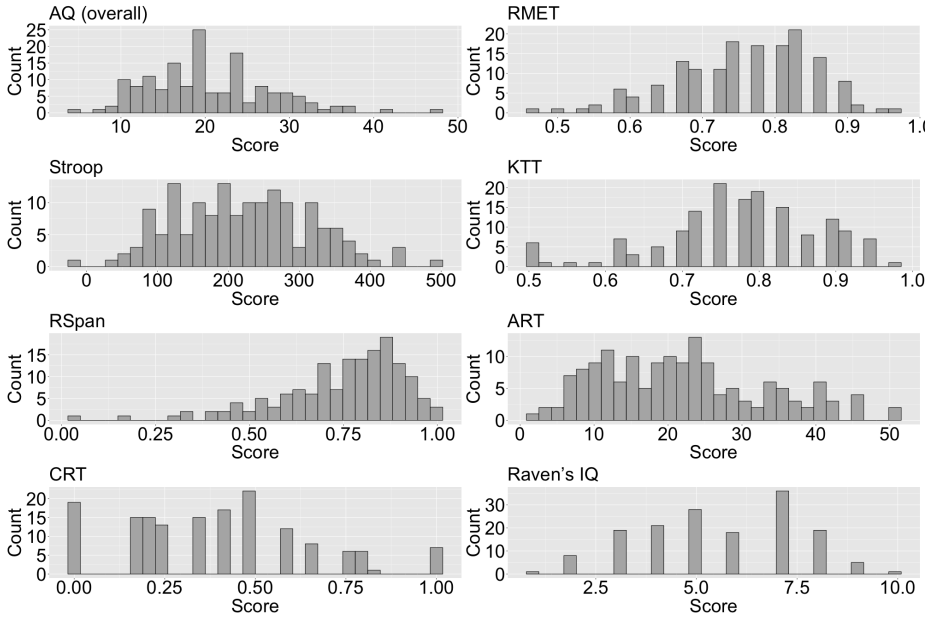


Figure 10.4: Experiment 4. Histograms showing the distribution of cognitive and personality measures collected from participants in the experiment ($N = 156$). Note: **AQ** = Autism-Spectrum Quotient (overall score; autistic traits); **RMET** = Reading the Mind in the Eyes Test (theory of mind ability); **Stroop** = Stroop Task (inhibition); **KTT** = Keep Track Task (memory updating); **RSpan** = Reading Span Task (verbal working memory span); **ART** = Author Recognition Test (exposure to print); **CRT** = Cognitive Reflection Test (cognitive processing – the ability to suppress intuitive but incorrect response); **Raven's IQ** = Raven's Progressive Matrices (non-verbal intelligence).

Pairwise correlations

Next, a pairwise correlation analysis was conducted to investigate the relationships between the obtained measures of individual differences. Figure 10.5 shows several positive correlations among the cognitive measures. After Holm correction for multiple comparisons, Raven's IQ was positively correlated with KTT performance ($r = 0.41$, $p_{adjusted} < .0001$). In addition, Raven's IQ was positively correlated with RSpan ($r = 0.28$, $p_{adjusted} = .01$) and with CRT ($r = 0.36$, $p_{adjusted} = .0001$). Finally, RSpan was positively correlated with CRT ($r = 0.32$, $p_{adjusted} = .002$). No other pairwise correlations were significant after Holm correction.

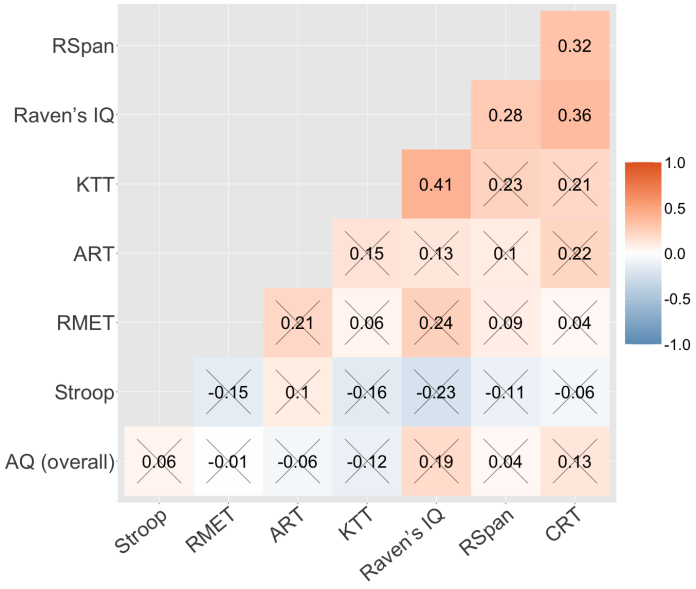


Figure 10.5: Experiment 4. Pairwise correlations ($N = 156$) of cognitive and personality measures collected for subjects participated in the experiment (p-values calculated with Holm correction; non-significant correlations are crossed out). Note: **AQ (overall)** = Autism-Spectrum Quotient (overall score; autistic traits); **RMET** = Reading the Mind in the Eyes Test (theory of mind ability); **Stroop** = Stroop Task (inhibition); **KTT** = Keep Track Task (memory updating); **RSpan** = Reading Span Task (verbal working memory span); **ART** = Author Recognition Test (exposure to print); **CRT** = Cognitive Reflection Test (cognitive processing – the ability to suppress intuitive but incorrect response); **Raven's IQ** = Raven's Progressive Matrices (non-verbal intelligence).

Compared to the experiment reported in Chapter 8 (Section 8.5), the overall pattern of correlations in the present study partially replicates earlier findings. In the earlier experiment, RSpan significantly correlated with ART, CRT, and Raven's IQ, and ART additionally correlated with CRT. The strongest correlation was observed between CRT and Raven's IQ. In the present study, the association between CRT

and Raven's IQ was replicated with an identical effect size ($r = 0.36$), suggesting a stable relationship between cognitive reflection and nonverbal intelligence across two samples of participants. Moreover, RSpan again showed positive correlations with both CRT and Raven's IQ, consistent with the earlier findings linking verbal working memory capacity with reasoning-related measures. Overall, the correlations among CRT, Raven's IQ, and RSpan were observed in both experiments.

At the same time, some differences emerged. In contrast to Chapter 8, ART did not show significant correlations with the other reasoning measures after Holm correction in the present study. Instead, a strong positive correlation was observed between Raven's IQ and KTT performance ($r = 0.41$, $p_{adjusted} < .0001$) – a measure that was not included in the earlier experiment. Importantly, the present study involved a larger set of individual-differences measures (8 vs. 5), resulting in a substantially greater number of pairwise comparisons and, consequently, a stricter Holm correction. This difference in correction procedure should be taken into account when comparing the two sets of results.

Overall, the findings are comparable with the previous literature, (Miyake et al., 2000) report no significant correlation between Stroop and keep-track tasks ($r(135) = 0.07, p = n.s$). Next, they found a significant correlation between working memory capacity and Stroop ($r = 0.2, p < .05$). Note, however, that in their study the working memory capacity was measured in an OSpan task.

In the present sample, total AQ scores were not correlated with RMET performance ($r = -0.01$). While Baron-Cohen et al. (2001a) reported a strong negative AQ-RMET correlation ($r = -0.53$) when combining neurotypical participants with individuals diagnosed with Asperger syndrome or high-functioning autism, subsequent work with more finely stratified samples has shown that this association is not robust between the groups. In particular, Baron-Cohen et al. (2015) observed a significant AQ-RMET correlation only in autistic females ($r = -.32$), but not in autistic males or in neurotypical control groups ($r = -0.13/-0.1, p = n.s$, in male/female controls, respectively).

It is worth noting that the observed correlation between non-verbal intelligence and memory updating (Raven's IQ and keep track task performance, respectively; $r = 0.41$, $p < .0001$) is not well documented in the literature. Although some studies report some correlation, it is usually lower – see e.g., Burgoyne et al. (2025), who found a correlation of $r = 0.29, p < .001$. At the same time they find correlation between RSpan and Raven's IQ ($r = 0.25$, $p < .001$), which is comparable with the results observed in the present study ($r = 0.28$, $p_{adjusted} = .01$).

Principal component analysis (PCA)

To examine whether the cognitive and personality measures reflected a smaller set of underlying dimensions, I conducted an exploratory principal component analysis (PCA), following the procedure described in Chapter 8 (Section 8.5).

An exploratory PCA was first fitted extracting eight components, corresponding to the total number of measures, in order to inspect the proportion of variance explained by successive components. Inspection of the cumulative variance plot indicated that six components accounted for 87% of the variance in the data (see Figure 10.6).

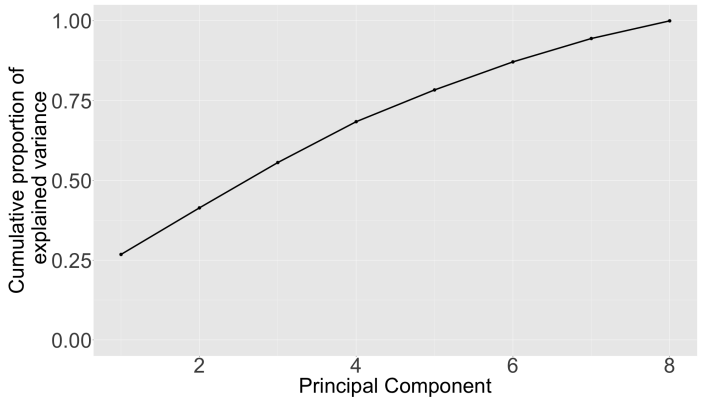


Figure 10.6: Experiment 4. A scree plot of the cumulative proportion of explained variance in the exploratory PCA.

Based on this inspection, a PCA with six components was fitted. The resulting components were rotated using Varimax rotation to obtain orthogonal components that maximize the variance of squared loadings and facilitate interpretation (Kaiser, 1958). The standardized component loadings are shown in Table 10.6.

Inspection of the loadings showed that the extracted components did not map well onto theoretically motivated cognitive constructs. Several components were dominated by a single measure (e.g. AQ, RMET, RSpan and Stroop), while two components reflected pairings of measures: KTT with Raven's IQ and ART with CRT. These groupings were data-driven and did not correspond straightforwardly to the constructs relevant to the present study. Notably, CRT and Raven's IQ, which might be expected to cluster as a reasoning-related dimension, did not primarily load on the same component. For this reason, the analysis of individual differences (Section 10.6.5) was conducted using the original cognitive and personality measures rather than PCA-derived components.

10.6.5 Individual Differences Analysis

The final research question was whether the consistency that is observed between implicature types and across sessions can be attributed to specific individual properties in terms of cognition or personality.

A separate beta mixed effects regression analysis was conducted for each type of inference. The main question here is whether pragmatic responses of different implicature types can be predicted by the profile of the participant (based on the

Table 10.6: Experiment 4. Principal Components Analysis (varimax rotation) of cognitive and personality measures ($N = 156$). Standardized component loadings are shown.

Measure	Component 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
AQ (overall)	-0.03	**0.94**	-0.02	0.01	0.00	0.08
RMET	0.05	0.02	0.09	**0.94**	0.07	-0.11
Stroop	-0.11	0.07	0.06	-0.10	-0.02	**0.95**
KTT	**0.91**	-0.18	0.07	-0.04	0.10	-0.02
RSpan	0.15	0.00	0.07	0.08	**0.96**	-0.02
ART	0.09	-0.16	**0.81**	0.29	-0.08	0.22
CRT	0.15	0.25	**0.70**	-0.22	0.37	-0.23
Raven's IQ	**0.68**	0.38	0.17	0.21	0.16	-0.23

Note: Primary loadings (absolute value $\geq .50$) are shown in **bold**. **AQ (overall)** = Autism-Spectrum Quotient (overall score; autistic traits); **RMET** = Reading the Mind in the Eyes Test (theory of mind ability); **Stroop** = Stroop Task (inhibition); **KTT** = Keep Track Task (memory updating); **RSpan** = Reading Span Task (verbal working memory span); **ART** = Author Recognition Test (exposure to print); **CRT** = Cognitive Reflection Test (cognitive processing – the ability to suppress intuitive but incorrect response); **Raven's IQ** = Raven's Progressive Matrices (non-verbal intelligence).

eight cognitive and personality measures that were collected). For detailed model specifications, see Section 10.4.

As discussed in Section 10.6.4, a principal component analysis (PCA) was conducted to examine whether the cognitive and personality measures could be reduced to a smaller set of interpretable components. However, the resulting components primarily reflected data-driven groupings and did not correspond to theoretically motivated cognitive constructs. For this reason, the models reported in this section include the centered and scaled raw cognitive and personality measures rather than PCA-derived components. Because several of these measures show moderate correlations (see Figure 10.5), potential multicollinearity should be taken into account when interpreting the model estimates. In particular, significant effects involving correlated predictors should be interpreted cautiously.

The results of atypicality inferences are presented first in two ways: 1) the model including all eight cognitive and personality measures collected in the present study, and 2) a direct replication of the analysis reported in Chapter 8. I then report the model results for the remaining implicature types.

Atypicality inferences: present study

A beta mixed effects regression model of the $(0, 1)$-transformed typicality ratings was fitted. The complete model included the story condition (with-IR vs. without-IR) and

its interaction with the eight cognitive and personality measures; see Section 10.4 for more details. A trial order was also included as a fixed effect. The random effect structure was simplified to reach convergence and included by-subject and by-item random intercepts and random slopes for story condition.

The results are presented in Table 10.7. There was a main effect of story condition, meaning that participants gave on average lower ratings in the with-IR condition compared to without-IR condition ($b = -0.43, SE = 0.08, z = -5.1, p < .001$), signifying the derivation of atypicality inferences. No other effects were significant. Trial order did not matter. Neither Raven's IQ nor CRT interacted with story condition, which is somewhat contrary to experiment presented in Chapter 7 where the PCA Reasoning component that incorporated both scores modulated the derivation of atypicality inferences.

Table 10.7: Experiment 4. Atypicality inferences: present study. Effect sizes (b), standard errors (SE), z-values, and p-values for the mixed-effects beta regression model of participants' ratings of target activity typicality (ratings transformed to fit a beta distribution). The dispersion parameter is 3.15.

	b	SE	z	p
Intercept	1.57	0.10	15.55	<.001
Story condition (with-IR)	-0.43	0.08	-5.10	**<.001**
Trial order	-0.04	0.03	-1.34	.18
AQ	-0.07	0.06	-1.11	.27
RMET	0.08	0.06	1.32	.19
Stroop	-0.03	0.06	-0.58	.56
KTT	0.07	0.06	1.14	.25
RSpan	-0.08	0.06	-1.29	.20
ART	-0.03	0.06	-0.58	.56
CRT	0.04	0.06	0.55	.58
Raven's IQ	0.02	0.07	0.27	.79
Story condition (with-IR) : Trial order	-0.00	0.06	-0.06	.96
Story condition (with-IR) : AQ	-0.09	0.06	-1.49	.14
Story condition (with-IR) : RMET	0.05	0.06	0.81	.42
Story condition (with-IR) : Stroop	0.06	0.06	0.94	.35
Story condition (with-IR) : KTT	0.07	0.07	0.99	.32
Story condition (with-IR) : RSpan	-0.11	0.06	-1.71	.09
Story condition (with-IR) : ART	-0.11	0.06	-1.68	.09
Story condition (with-IR) : CRT	0.07	0.07	1.02	.31
Story condition (with-IR) : Raven's IQ	-0.08	0.07	-1.14	.25
Random effects	Variance			
Subject	0.35			
Item	0.13			
Story condition | Item	0.07			

Atypicality inferences: a direct replication of the results in Chapter 8

In the model reported in the previous section, no significant effects of Raven's IQ or CRT were observed. This was somewhat unexpected, as in Chapter 8 the derivation of atypicality inferences was significantly modulated by a PCA-derived *Reasoning* component, which combined Raven's IQ and CRT scores.

To allow a more direct comparison with the earlier experiment, I conducted an additional analysis that closely follows the modeling procedure used in Chapter 8. In this analysis, cognitive measures that were not collected in the earlier study were omitted and a PCA was conducted on the remaining measures following the procedure described in that chapter. Details of this PCA are provided in Section B.3. The PCA yielded four components. Raven's IQ and CRT loaded on the same component, which I refer to as *Reasoning*, while the remaining measures each formed separate components, similar to the PCA structure observed in Chapter 8. These PCA component scores, together with their interactions with story condition, were then entered into the regression model, along with trial order and its interaction with story condition. The model results are presented in Table 10.8.

Table 10.8: Experiment 4. Atypicality inferences: replication of Chapter 8 results. Effect sizes (b), standard errors (SE), z-values, and p-values for the mixed-effects beta regression model of participants' ratings of target activity typicality (ratings transformed to fit a beta distribution). The dispersion parameter is 3.13. **Note:** AQ, RSpan, ART, and Reasoning refer to standardized PCA-derived component scores; see Section B.3 for details.

	b	SE	z	p
Intercept	1.57	0.10	15.44	**<.001**
Story condition (with-IR)	-0.43	0.09	-4.99	**<.001**
Trial order	-0.04	0.03	-1.23	.22
AQ	-0.08	0.06	-1.38	.17
RSpan	-0.05	0.06	-0.84	.40
ART	-0.01	0.06	-0.20	.84
Reasoning	0.07	0.06	1.24	.22
Story condition (with-IR) : Trial order	-0.00	0.06	-0.03	.97
Story condition (with-IR) : AQ	-0.09	0.06	-1.60	.11
Story condition (with-IR) : RSpan	-0.08	0.06	-1.34	.18
Story condition (with-IR) : ART	-0.06	0.06	-0.99	.32
Story condition (with-IR) : Reasoning	-0.05	0.06	-0.82	.41
Random effects	Variance			
Subject	0.36			
Item	0.13			
Story condition | Item	0.08			

Contrary to the findings reported in Chapter 8, the Reasoning component did not

significantly interact with the story condition. No other significant effects of cognitive or personality measures were observed. The implications of this inconsistency are discussed in the Discussion section.

Other implicature types

For each type of implicature, separate beta mixed-effects regression was fit as described in Section 10.4. The complete model outputs for each implicature type are presented in Section B.4.

Across all models, the only significant individual differences effect was observed for ART (exposure to print) in the model for embedded *some* (see Table B.13). Irrespective of condition (NNA vs. NSA), participants with higher exposure to print tended to interpret sentences with embedded *some* more literally ($b = 0.28, SE = 0.07, z = 3.99, p < .001$; see Figure 10.7).

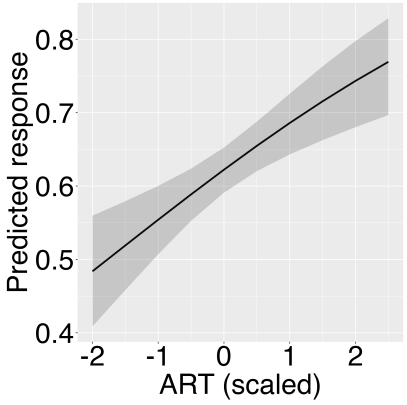


Figure 10.7: Experiment 4. Embedded *some*. Predicted response as a function of ART (higher scores signify stronger exposure to print) based on the fitted model. The solid line represents the model-predicted values, and the shaded area indicates the 95% confidence interval.

In the model for embedded *some*, there was additionally a main effect of condition (NNA vs. NSA; $b = 0.2, SE = 0.07, z = 2.82, p = .005$), indicating that participants gave lower scores to sentences in the NSA condition than in the NNA condition (NSA: M = 63.14, SD = 45.38; NNA: M = 72.19, SD = 40.72). This suggests that participants judged sentences in the NSA condition more pragmatically than those in the NNA condition.

Similarly, in the model for disjunctions there was a main effect of inference type (bare vs. embedded disjunctions; $b = 0.27, SE = 0.07, z = 4.17, p < .001$), indicating that participants gave lower scores to bare disjunctions than to embedded disjunctions (bare: M = 77.79, SD = 35.24; embedded: M = 86.22, SD = 29.49). This suggests that participants judged bare disjunctions more pragmatically than embedded disjunctions.

Next, a trial order effect was observed in the model for bare *some* ($b = -0.12, SE = 0.03, z = -3.36, p < .001$), indicating that as participants progressed through the experiment they gave more pragmatic responses (see Figure B.4).

In contrast, a significant trial order effect in the model for bare numerals ($b = 0.17, SE = 0.04, z = 4.91, p < .001$) showed that participants gave more literal responses as the experiment progressed (see Figure B.5).

10.7 Discussion and conclusion

This study is one of the first to investigate pragmatic comprehension on the same participant sample across different implicature types and throughout time. Overall, the results show that participants derived the targeted pragmatic interpretations across the tasks, including atypicality inferences and several types of scalar implicatures[5]. At the same time, participants' responses showed moderate consistency across sessions for the scale-based implicature tasks, while the effect of informational redundancy in the atypicality inference task remained stable across sessions. Further, the analyses revealed little evidence that the variability in pragmatic interpretations could be explained by the cognitive and personality traits measured in this study. The following sections discuss these findings in relation to the research questions.

10.7.1 Relationships between implicature types

The results of the correlation analysis provide partial support for distinguishing between particularized and generalized conversational implicatures (Grice, 1975). Responses to several scalar implicature tasks were positively correlated, suggesting that individuals who tended to derive pragmatic interpretations in one task were somewhat more likely to do so in others. In contrast, responses in the atypicality inference task, which fall under particularized implicatures, did not correlate with scalar implicatures, which are generally considered to be generalized conversational implicatures. This suggests that the mechanisms underlying the derivation of atypicality inferences differ from those involved in the derivation of scalar implicatures.

Within the group of generalized implicatures, the correlations were not uniform. Conceptually closer implicature types tended to correlate more strongly with each other. In particular, the two conditions of embedded *some* showed stronger correlations, as did bare and embedded disjunctions. These implicature types also correlated with bare *some*, although the correlations were generally weaker. Overall, this pattern suggests a degree of clustering among scalar implicatures based on their structural and conceptual similarity.

[5]Except for *Some* in face-threatening contexts (Politeness implicatures; Bonnefon et al., 2009), where the item order effects were observed.

Politeness implicatures (*Some* in face-threatening contexts; Bonnefon et al., 2009) were originally included in the design as an additional example of a more context-dependent implicature type, potentially comparable to atypicality inferences. However, this condition showed strong item-order effects and did not replicate the expected pragmatic effect, making the interpretation of the responses difficult. For this reason, politeness implicatures were excluded from the analysis reported here.

An interesting finding concerns numeral implicatures. Although numerals are often considered standard scalar terms with ambiguity between a literal ("at least") and a pragmatic ("exactly") meaning (e.g., Schulz & van Rooij, 2006), responses to numerals did not correlate with any other generalized implicature type. In addition, trial-order effects revealed an increasing acceptance of the literal interpretation for numerals, whereas the opposite tendency was observed for bare *some*. This suggests that the interpretation of bare numerals may be distinct from other implicature phenomena. Notably, this result is consistent with earlier work showing that numeral implicatures often pattern differently to other scalar terms such as *some*, both in acquisition (Noveck, 2001) and in experiments manipulating cognitive load (Marty et al., 2013).

Together, these results highlight the limitations of generalizing processing accounts of implicature derivation from one implicature type (in particular, inferences with *some*) to others.

10.7.2 Consistency of responses across time

The present study provides new data on implicature preferences within individuals. Previous studies have demonstrated a tendency for participants to have a distinct preference for the logical or pragmatic interpretation within a single experiment session (Noveck & Posada, 2003; Foppolo, 2007; Heyman & Schaeken, 2015); building upon this, the present study shows that these response preferences are stable within individuals over time. Moreover, this stability extends beyond classic scalar implicatures with *some* to other types of implicature.

For atypicality inferences, the effect of informational redundancy was comparable across the two sessions. No evidence was found that the size of the IR effect changed between sessions, and no systematic shift in ratings across experimental sessions was observed.

For the scale-based implicature tasks, stability was assessed using intraclass correlation coefficients (ICC) computed between participants' session-level mean ratings (ICC values ranged approximately from .47 to .57 depending on implicature type). The observed correlations suggest that participants who tended to give more pragmatic responses in one session generally showed similar tendencies in the other session, although the stability was far from perfect. This is consistent with the trial order effects observed in some of the tasks during the analysis of individual differences.

Taken together, the results suggest that pragmatic consistency appears to be a trait inherent to an individual, at least within the context of the same pragmatic tasks. However, the present data do not rule out the possibility that this consistency may vary with the nature of the task, since participants in the present study were presented with the same experiments for each implicature type across the two sessions. Thus, individual preferences may reflect task-related strategies specific to an experimental paradigm (cf. Tavano & Kaiser, 2010). A question that merits further research then is whether such internal stability holds across different tasks commonly used to test implicatures, such as sentence verification (Bott & Noveck, 2004) vs. picture verification (van Tiel et al., 2019).

One important methodological consideration when interpreting the consistency of responses across experimental sessions concerns the use of different approaches for different implicature types. For most implicatures, consistency was quantified using intraclass correlation coefficients (ICC) computed over participant-level aggregated responses, whereas atypicality inferences were analyzed using a Bayesian beta mixed-effects regression model that explicitly accounted for item-level variability.

These approaches capture different aspects of consistency. ICC summarizes test-retest reliability based on aggregated scores, collapsing across item-level variability and reflecting the stability of individual differences across sessions. In contrast, the Bayesian mixed-effects model operates at the trial level and accounts for variation across participants and items, allowing it to assess the stability of the IR manipulation across sessions. As a result, the two approaches are not directly comparable.

The use of different methods was motivated by differences in data structure. While items in most implicature tasks were assumed to be comparable and could be aggregated, atypicality inferences showed substantial item variability, involved only a subset of items per participant, and included a reversal of conditions across sessions. Under these conditions, aggregation would confound participant effects with item-related variability, making ICC-based estimates unsuitable.

A further consideration concerns differences in the number of observations across tasks. Participant-level means for bare *some* and numerals were based on more trials per session than in other tasks, including atypicality inferences, and are therefore likely to be more stable. On the other hand, this difference in the number of trials per participant may partly contribute to the learning effects observed for bare *some* and numerals, which were not found for other types of implicature in the study.

10.7.3 Role of individual differences

Lastly, the answers to pragmatic inferences across different types were related to the cognitive and personality profile of subjects. However, individual measures did not correlate with most of the implicature types.

In particular, the findings here can be seen contrary to those in De Neys & Schaeken (2007) and Fairchild & Papafragou (2021). In the study about scalar implicatures

comprehension, De Neys & Schaeken (2007) found a significant effect of working memory capacity (WMC) on the rate of pragmatic responses. Subjects in the low WMC group showed lower rates of pragmatic interpretations compared to the group of high WMC, under high cognitive load setup. Further, Fairchild & Papafragou (2021) found that when considering theory of mind ability (ToM) and WMC together, subjects with better ToM were more likely to compute implicatures. However, the measures targeting WMC did not make any additional contributions, after accounting for ToM. Contradictory results can be possibly explained by different tests that were used in Fairchild & Papafragou (2021) compared to the present study. For ToM, Fairchild & Papafragou (2021) derived a composite score from an abridged 12-trial version of the mind in the eyes task and strange stories test, while here only a single task was used for measuring ToM, namely a full version of mind in the eyes task.

Next, in the present study no effects of autistic traits were found. Previously, Nieuwland et al. (2010) found a difference in how individuals reacted to the under-informative sentences with *some*, based on their AQ-scores. Only the group of subjects with less autistic traits showed N400 effects in the critical regions, signifying pragmatic enrichment. While these results appear at odds with the present study, the inconsistency could reflect differences in comprehenders' on-line processing of implicatures and their eventual off-line interpretation. Future research should consider how individual differences may mediate both on-line and off-line comprehension in individuals for a more complete understanding of implicature phenomena.

The individual measures analysis in the present study only showed a significant effect of exposure to print (measured through the Author Recognition Test; ART) on embedded *some*. Participants with greater exposure to print gave responses consistent with the literal "some and possibly all" interpretation more often. One possible explanation for the ART effect is that print exposure did not facilitate pragmatic enrichment per se, but instead influenced how participants approached the embedded *some* sentence-picture verification task. Unlike bare *some*, where the scalar term is evaluated directly, the embedded version required interpreting *some* within the scope of the quantifier "exactly one", i.e., as part of a larger quantified construction. The ART effect did not interact with condition, suggesting that print exposure was related to a general response tendency across the embedded-some task rather than to one specific configuration (NNA vs. NSA). A cautious interpretation, then, is that greater print exposure may support a more stable or more conservative truth-conditional strategy when participants evaluate quantified sentences of this kind, rather than enhancing scalar implicature derivation more generally. This interpretation requires further investigation – for example, by testing additional configurations from Potts et al. (2016) while measuring participants' exposure to print – but it is consistent with Cowley et al. (2025), who showed that print exposure predicts how readers resolve ambiguity in quantified sentences and argued that greater print exposure may strengthen expectations for a dominant interpretation.

Finally, no cognitive or personality measures were found to influence the deriva-

tion of atypicality inferences in the present study. This contrasts with the findings reported in Chapter 8, where reasoning ability (operationalized as a composite measure of Raven's Progressive Matrices and the Cognitive Reflection Test) modulated participants' responses in the atypicality inference task. The implications of this discrepancy are discussed in the following section.

10.7.4 Effects of reasoning ability and autistic traits in atypicality inferences derivation

An earlier experiment in this dissertation (Chapter 8) investigated the relationship between individual differences in cognitive traits and the derivation of atypicality inferences. In that study, reasoning ability – operationalized as a latent component combining performance on Raven's Progressive Matrices and the Cognitive Reflection Test – was found to modulate participants' responses in the atypicality inference task. In the explanation model, higher reasoning ability increased the likelihood that participants explicitly produced an atypicality inference ($b = 0.48$, $SE = 0.13$, $z = 3.81$, $p < .001$). In the rating model, reasoning ability interacted with story condition, such that participants with higher reasoning ability showed a stronger effect of the IR-manipulation ($b = -0.11$, $SE = 0.05$, $z = -2.35$, $p = .02$). These results were interpreted as suggesting that reasoning ability may facilitate the process of accommodating informational redundancy by constructing a context in which the apparently redundant utterance becomes informative.

The present experiment did not replicate this effect. In the atypicality inference task, no reliable relationship was observed between reasoning ability and participants' typicality ratings (Story condition × Reasoning: $b = -0.05$, $SE = 0.06$, $z = -0.82$, $p = .41$).

One possible explanation for this discrepancy could be differences between the participant samples. However, the distributions of the cognitive measures collected in the two experiments were highly comparable. The mean scores on the overlapping measures (AQ, ART, CRT, Raven's IQ, and RSpan) were very similar across the two studies, and the pattern of correlations among the reasoning-related measures was largely replicated too. These similarities suggest that the cognitive profiles of the two samples were broadly comparable. This makes it unlikely that the absence of the reasoning effect in the present experiment can be attributed to differences in participant characteristics.

Several methodological differences between the experiments may help explain the observed discrepancy. First, the earlier study assessed inference computation using both typicality ratings and participants' written explanations, whereas the present experiment relied solely on rating responses. The analysis of explanations in the earlier study showed that ratings can sometimes mask the underlying reasoning processes involved in inference derivation. In particular, cases such as *atypicality-reject* responses indicate that participants may consider an atypicality inference but ultimately not

reflect it in their ratings. A similar mismatch between reasoning and ratings may arise in *not-sure* responses, where participants use the midpoint of the scale to express uncertainty, as well as in cases involving reasoning errors or post-hoc revisions of their judgments. Although such cases constituted a relatively small proportion of the data in the earlier experiment, their frequency is not guaranteed to be stable across experimental settings. In particular, because the present study did not include an explanation task, it is possible that instances in which participants entertained, rejected, or were uncertain about an atypicality inference remained undetected in the rating data. As a result, variability in how participants map their underlying reasoning onto rating responses may attenuate observable relationships between inference derivation and cognitive predictors.

Second, the two experiments differed in their overall task structure. In Chapter 8, participants completed only the atypicality inference task, whereas in the present study atypicality inference trials were embedded within a larger battery of implicature tasks. This broader context may influence how participants engage with the task, potentially encouraging more uniform or less elaborative response strategies and reducing the likelihood of actively accommodating atypical interpretations. Such differences in task demands may further contribute to variability in how inference-related reasoning is reflected in rating responses.

Finally, while the sample sizes of the two studies were comparable (156 vs. 193 participants in Chapter 8), the atypicality inference task in the present experiment involved fewer observations per participant (four items per participant and story condition vs. six items in Chapter 8). This reduction in the number of observations may limit the ability to detect interactions between cognitive predictors and experimental condition. However, the present design included a within-story reversal of experimental conditions across sessions, meaning that each participant encountered the same stories in both conditions. This feature helps to control for story-specific baseline differences. Nevertheless, the smaller number of observations per participant, together with reliance on rating responses alone and embedding atypicality inference trials within a broader battery of implicature tasks, may make individual differences more difficult to detect in the present design.

Overall, the results suggest that the relationship between reasoning ability and the derivation of atypicality inferences is not consistently observed across the two experiments. While reasoning ability was found to significantly modulate inference derivation in Chapter 8, no such relationship was detected in the present study. This discrepancy may be related to differences in the experimental design. Specifically, the earlier experiment assessed inference computation using both ratings and written explanations, focusing exclusively on the atypicality inference task. In contrast, the present study measured inference computation using ratings only, with atypicality inference trials embedded within a broader set of implicature tasks. However, it remains possible that the earlier effect was less robust than initially assumed.

Two directions for future research follow from these findings. First, a direct replica-

tion of the experiment in Chapter 8 using the same task structure and measures would help determine whether the relationship between reasoning ability and atypicality inference derivation can be reliably reproduced. Second, studies that systematically compare different task formats – for example, rating-based tasks and explanation-based measures, or isolated vs. mixed pragmatic task batteries – may help clarify under which conditions reasoning ability influences atypicality inference derivation.

A similar discrepancy can be observed for autistic traits (AQ). In an earlier experiment, AQ scores were associated with participants' responses in the atypicality inference task. However, no such relationship was observed in the present study. As with the reasoning effect, this difference may be due to methodological variations between the experiments or the earlier finding's limited robustness. Further research is needed to establish whether AQ reliably modulates the derivation of atypicality inferences.

Limitations

One potential limitation concerns the use of a continuous response scale across all implicature tasks. Although this choice provided a consistent response format and enabled fine-grained assessment of participants' judgments, it diverges from the original paradigms, which generally employed binary or Likert-type response options. This discrepancy may influence how participants engage with the task, possibly promoting more graded or uncertain judgments instead of clear-cut categorical choices. Consequently, the reliance on continuous scales may shape both overall response tendencies and the sensitivity to pragmatic inferences, and may constrain the extent to which the current findings can be directly compared with those of the original studies.

Another limitation concerns the reliability of measures across implicature types. In several tasks, participant-level measures were derived from relatively few trials per condition, which may reduce internal consistency and add noise. Consequently, individual differences may be estimated less accurately, potentially weakening correlations between tasks and obscuring consistency across sessions.

A further limitation concerns the range of implicature types included in the study. Although several well-studied types were tested, the set of tasks does not cover the full range of pragmatic inferences. Consequently, the results may not extend to other types of implicature.

Part III

Conclusions

Chapter 11

Discussion of the main findings

This chapter summarizes the main findings of the dissertation and relates them to the existing literature on pragmatic processing. The main goal of the dissertation was to investigate the comprehender-specific factors that can modulate the derivation of atypicality inferences. Across the studies, atypicality inferences were shown to be modulated by both reasoning-related factors and socio-pragmatic characteristics, in particular autistic traits. At the same time, they did not pattern with implicatures based on lexical scales, but instead showed a distinct profile. This suggests that atypicality inferences may rely on partly different processes than scale-based implicatures. The following sections discuss the findings in detail, starting with the derivation process of atypicality inferences.

11.1 Derivation process of atypicality inferences

Atypicality inferences constitute a type of pragmatic inference that arises in situations of informational redundancy or overinformativity. In such cases, a speaker explicitly mentions an event that is already strongly expected given background knowledge about everyday activities. This background knowledge can be described in terms of scripts, that is, structured representations of familiar situations and their typical event sequences (Schank & Abelson, 1975; Bower et al., 1979). When a script is activated during comprehension, its associated events become readily available in memory (Bower et al., 1979; Zwaan et al., 1995). As a result, explicitly mentioning one of these events, given the context, may be perceived as redundant. Rather than treating such utterances as infelicitous, comprehenders may interpret them pragmatically, and consequently revise their beliefs about how typical the event is for the individual in question.

This phenomenon was first investigated by Kravtchenko (2022), who showed that informational redundancy can lead to such belief updates. In particular, listeners

interpreted redundant utterances as signaling that the mentioned event is not typical for the individual in question. While Kravtchenko (2022) established the phenomenon and alluded to the reasoning process involved in deriving atypicality inferences, the steps of this process were not explicitly formalized. In particular, it was not specified in detail how such inferences are derived, nor which components of the derivation process give rise to variability across comprehenders.

To address this, I proposed a theoretical account of the derivation process of atypicality inferences (Chapter 3). While broader theories of pragmatic processing discuss what can underlie it (Grice, 1975; Levinson, 2000; Sperber & Wilson, 1996; Degen & Tanenhaus, 2019), they do not directly address atypicality inferences or provide a structured account of the steps involved in their derivation. Specifying such a process was necessary in order to formulate predictions about which factors may influence inference derivation and how variability across comprehenders may arise.

The proposed account treats atypicality inference derivation as a multi-step derivation process. The first step involves identifying redundancy relative to script-based expectations. When a script is activated, its associated events are readily accessible, and the comprehender can evaluate whether the explicitly mentioned event contributes new information. In cases where the event is already strongly expected, it is identified as redundant. The second step involves recognizing that this redundancy is pragmatically infelicitous, as it violates the cooperative principle and the expectation of informativity (Grice, 1975). This step requires the comprehender to consider that informational redundancy is pragmatically marked.

The third step involves deriving an atypicality inference as part of a pragmatic repair process. Instead of treating the utterance as infelicitous, the comprehender assumes that the speaker had a reason for including the redundant information. This leads to the inference that the event is not typical for the individual in question. The fourth step involves accommodating this inference within the discourse representation. Because inferred atypicality may conflict with script-based expectations, it must be integrated into the situation model, often by constructing an explanation that reconciles atypicality with the broader context.

This decomposition emphasizes that atypicality inferences arise from the interaction of several components, including script-based knowledge, pragmatic reasoning, and discourse-level integration. It also provides a basis for linking different types of cognitive factors to specific parts of the derivation process. Identifying redundancy may be influenced by factors involved in the activation of script-based information, such as working memory capacity. Recognizing the pragmatic markedness of redundancy may be related to socio-pragmatic abilities, such as theory of mind or autistic traits. Accommodating the inferred atypicality within the discourse representation may, in turn, depend more strongly on reasoning-related factors, such as cognitive reflection and fluid intelligence. The derivation scheme proposed here served as a framework for the empirical investigations of the dissertation and guided the selection of factors tested in the studies.

More broadly, the proposed step-wise account of atypicality inference derivation speaks to a general issue in pragmatics, namely how to relate theoretical descriptions of pragmatic meaning to the processes that give rise to such interpretations during comprehension. Classical and contemporary accounts emphasize that pragmatic interpretation depends on the communicative principle, conversational norms, contextual information, background knowledge, and consideration of alternatives (Grice, 1975; Levinson, 2000; Sperber & Wilson, 1996; Degen & Tanenhaus, 2019). However, these accounts typically characterize pragmatic inference at a relatively abstract level and do not specify in detail how different sources of information are coordinated during interpretation. Much experimental work has focused on phenomena such as scalar implicatures, where the relevant alternatives are lexically defined (e.g., *some* vs. *all*) and are assumed to be available to the comprehender as part of the linguistic system. In such cases, the empirical focus has often been on whether an inference is derived, how it is affected by processing cost, and how it varies with cognitive or contextual factors (De Neys & Schaeken, 2007; Marty et al., 2013; Fairchild & Papafragou, 2021; Degen & Tanenhaus, 2015). These approaches have been highly informative, but they rely on settings in which the space of alternatives is relatively well-defined. In contrast, atypicality inferences arise in situations where the alternatives are not given by lexical scales, but must be constructed from script-based knowledge about typical event sequences. This requires the comprehender to integrate background knowledge, detect informational redundancy, evaluate that it is marked pragmatically, and incorporate the resulting inference into the discourse representation. Making these components explicit provides a more fine-grained perspective on pragmatic processing. It allows empirical findings, such as variability across comprehenders or robustness under cognitive load, to be related to different parts of the derivation process, rather than treating inference derivation as a single outcome. More generally, it highlights the importance of studying pragmatic phenomena that are not anchored in lexical alternatives, and therefore require a broader view of how context, knowledge, and reasoning interact during comprehension.

11.2 Accommodation of atypicality inferences

According to the proposed derivation account, atypicality inferences do not end with lowering beliefs about an otherwise predictable event. Once the event atypicality has been inferred, it must be integrated into the discourse representation. This accommodation step is non-trivial, as the inferred atypicality conflicts with expectations derived from script knowledge. As a result, interpretation requires accommodating the inference in a way that renders the utterance meaningful in context.

Accommodation here goes beyond a simple adjustment of event typicality. It involves constructing an interpretation that reconciles the inferred atypicality with the activated script. This often takes some form of explanatory enrichment, where comprehenders introduce assumptions that, as shown in Chapter 7, vary widely across

items, often involving inferences about the character's habits, preferences, or circumstances. Importantly, this process is not limited to a single outcome. Multiple explanations may be compatible with the same utterance, and different comprehenders may arrive at different interpretations.

The accommodations of atypicality observed in Chapter 7 suggest that participants do not converge on a single interpretation, but instead construct a range of context-dependent explanations. The analysis of participants' explanations showed that, beyond adjusting their typicality judgments, participants actively engaged in interpreting the utterances by generating explanations that integrate the inferred atypicality into the discourse context. These explanations revealed substantial variability in reasoning. In some cases, participants accepted the atypicality inference and elaborated on it. In others, they considered the inference but rejected it or reinterpreted the utterance in a way that avoided the inference altogether. This indicates that approaching informational redundancy involves different strategies.

These findings also highlight a limitation of rating-based measures used by Kravtchenko (2022). Typicality ratings capture the final outcome of interpretation, but they do not reveal how that outcome was reached. The same rating may result from different reasoning processes. For example, a high typicality judgment may reflect a failure to derive the inference, but it may also reflect a case where the inference was considered and subsequently rejected. Conversely, lower typicality ratings may arise from different reasoning paths too. In some cases, participants may derive the atypicality inference and accommodate it by constructing an explanation that directly supports the conclusion that the event is not typical for the individual (e.g., by assuming habits or circumstances beyond the given context). In other cases, lower ratings may reflect more tentative interpretations, where the inference is entertained but not fully integrated, or where participants express uncertainty about how to reconcile the utterance with the script. As a result, similar rating outcomes may correspond to qualitatively different processes.

The use of explanation elicitation allowed me to access these differences. Participants were asked to justify their judgments, and these justifications formed the basis of the annotation scheme used in the analysis. This approach allowed for the identification of reasoning strategies that would not have been observable from ratings alone. In this respect, the method complements the standard measures employed by Kravtchenko (2022) and offers further details on how participants interpret informationally redundant utterances.

The present findings also relate to recent work that employs strategy elicitation in the study of pragmatic reference games (Mayn & Demberg, 2023, 2026). These studies show that collecting explanations alongside performance measures can reveal differences in how participants approach a task which are not fully captured by accuracy or participants' responses alone. Explanations have been shown to distinguish between participants who arrive at the same responses but rely on different reasoning strategies. The results of the present dissertation extend this observation to atypi-

cality inferences, showing that explanation elicitation can uncover variability in how inferences are derived, evaluated, and integrated into discourse representations.

At the same time, I share the view of Mayn & Demberg (2023) that explanation-based methods come with limitations. As they discuss, individuals may lack direct access to their cognitive processes and instead provide post-hoc rationalizations, with explanations reflecting constructed justifications rather than the actual mechanisms that produced a response. Moreover, known biases in reasoning are not always reported in such explanations, suggesting that participants are not fully aware of the factors influencing their behavior (Nisbett & Wilson, 1977; Cushman, 2020; Evans, 2019). Despite these limitations, the explanations collected in this dissertation nevertheless revealed a wide range of ways in which participants accommodated atypicality inferences and allowed a glimpse into the reasoning underlying their typicality judgments.

11.3 Consistency and variability across comprehenders and time

Previous studies on pragmatic phenomena such as numerals or scalar implicatures have demonstrated within-subject consistency and between-subject variability in a single experimental session (Singh et al., 2016; Marty et al., 2013; Tavano & Kaiser, 2010; Panizza et al., 2009). However, no such evidence exists for atypicality inferences. In addition, the question of whether response preferences remain stable over time has, in general, received little attention in the literature, particularly in adult populations (cf. Taguchi, 2012, but in the context of L2 populations).

Given that previous studies on atypicality inferences have used a one-shot approach (Kravtchenko, 2022; Kravtchenko & Demberg, 2022a), it remained unclear to what extent the derivation of the inference was consistent within individual participants. Thus, in Chapter 7 I tested whether participants consistently derive atypicality inferences, providing the basis for exploring their relationship to individual traits. The results showed that participants differed in their tendency to derive atypicality inferences. Some participants consistently interpreted redundant utterances as signaling atypicality inferences, while others were less likely to do so or adopted alternative interpretations. This was reflected in both the rating data and the explanations provided for the typicality judgments. Additionally, I provide evidence that responses to atypicality inferences are stable over time. The longitudinal data from Chapter 10 indicated that participants' responses did not change across sessions, and no overall shift in responses was observed.

One way to interpret the observed tendencies to derive or not derive atypicality inferences is to relate them to the derivation process discussed in Section 11.1. Variability may arise at different stages of this process. Some comprehenders may be more likely to identify redundancy relative to script knowledge, others may differ in

whether they treat such redundancy as pragmatically marked, and others still may vary in how they accommodate the resulting inference. As discussed in Section 11.2, these differences are also reflected in the range of explanations that participants construct.

More generally, the empirical results presented in Chapters 7 and 10 extend the evidence on within-comprehender consistency and between-comprehender variability beyond pragmatic phenomena that rely on lexical scales. The results suggest that atypicality inferences exhibit variability in responses that is consistent within individuals and stable over time. This supports the view that their derivation reflects comprehender-specific tendencies rather than random variation.

11.4 Atypicality inferences under cognitive load

The main goal of this dissertation was to identify the factors that influence derivation of atypicality inferences. To address this question, I combined two complementary approaches. One approach examined natural variability across comprehenders by relating inference derivation to their cognitive and personality-related traits (discussed in the next section). The other approach, discussed here, investigated whether inference derivation was sensitive to externally imposed constraints on cognitive resources.

In Chapter 4, I first reviewed major accounts of pragmatic processing and processing cost, including the Default account (Levinson, 2000), the Relevance theory (Sperber & Wilson, 1996), and the Constraint-Based account of Degen & Tanenhaus (2019), to derive their predictions for atypicality inferences. While these frameworks differ in their treatment of scalar implicatures, they converge on atypicality inferences. Across accounts, I argued that atypicality inferences should be cognitively costly, because they arise in neutral contexts that do not explicitly support a pragmatic interpretation and because their alternatives are not lexically constrained but must be constructed in a highly context-dependent way.

Across three well-powered dual-task experiments, I tested whether atypicality inferences are cognitively costly by manipulating cognitive load, following the paradigm previously used for scalar implicatures (De Neys & Schaeken, 2007; Fairchild & Papafragou, 2021; Marty et al., 2013; Cho, 2020). I presented the materials in two modalities and employed two types of secondary tasks. In Section 9.2, participants listened to stories containing atypicality inferences while simultaneously performing a visuo-motor tracking task that taxed attentional resources. A visuo-motor tracking task was a novel task that has not been used before in the study of pragmatic processing cost. In Chapters 9.3 and 9.4, participants read the stories and, between reading and providing their typicality judgments, performed a reading span task in which they memorized words and recalled them after giving their judgments. This task, similar to that used by Cho (2020), taxed verbal working memory capacity. Contrary to the predictions outlined above, I found no evidence that cognitive load reduces the

strength of atypicality inferences. Participants derived these inferences equally under high and low/no load conditions across all three experiments, despite the two types of load manipulation. These null effects were further supported by Bayesian analysis showing evidence in favor of the null hypothesis.

The only exception was a small trade-off effect observed in Section 9.2, where increased load led to greater tracking deviation in the condition in which atypicality inferences were expected to be derived. However, this effect is likely attributable to exclamatory prosody rather than to pragmatic processing itself, as it did not replicate in the experiments in which atypicality inference materials were presented textually rather than auditorily. Importantly, the load manipulations were successful, as comprehension accuracy decreased under higher load across all experiments. Overall, the results indicate that atypicality inferences were not sensitive to working memory limitations in the way predicted by existing accounts.

These findings challenge the common assumption that particularized implicatures are necessarily effortful, at least for atypicality inferences as the existing accounts describe. One possible explanation is that these inferences are based on script knowledge, which is strongly embedded in long-term memory and can be rapidly accessed and integrated with context (Hagoort et al., 2004; Chwilla & Kolk, 2005). This may allow listeners to quickly detect informational redundancy without substantial demands on working memory.

More broadly, the results align with findings suggesting that pragmatic processing does not always depend on executive resources such as working memory. For instance, Foppolo et al. (2021) show that inference-relevant information can, under some conditions, be obtained without incurring additional cost. In parallel, Fairchild & Papafragou (2021) report no unique contribution of working memory to implicature derivation and likewise fail to find effects of cognitive load, instead arguing that executive resources may only play an indirect role.

Taken together, atypicality inferences may not be modulated by working memory resource constraints, but they are nonetheless likely sensitive to other cognitive factors. In particular, while some aspects of these inferences, such as redundancy detection, may be relatively automatic and grounded in readily accessible knowledge, other aspects, such as generating and evaluating explanatory alternatives, may draw on different cognitive resources. This perspective motivated my second line of research that focused on individual differences in the cognitive and personality-related traits of comprehenders. I discuss the findings in the next section.

11.5 Effects of natural variability in cognitive and personality-related traits

As a second perspective, I adopted an individual-differences approach to pragmatic processing, investigating how variability in cognitive and personality-related traits

of comprehenders modulates the derivation of atypicality inferences. In contrast to Chapter 9, where external constraints were imposed on attentional and verbal working memory resources uniformly across participants, in Chapters 8 and 10 I directly measured participants' cognitive traits. Drawing on prior literature on individual differences in pragmatic processing, I identified a subset of factors that may influence the derivation of atypicality inferences. Below, I summarize the main findings and discuss their implications.

11.5.1 Effects of executive functions

Across experiments presented in Chapters 8 and 10, I tested several components of executive functions, namely working memory capacity, inhibition, and memory updating (Miyake et al., 2000). Across the two studies, I did not find an effect of any of them, which aligns with the results of the dual-tasking experiments discussed in the previous section. Neither verbal working memory capacity (tested in Chapters 8 and 10), nor inhibition or memory updating (both tested only in Chapter 10) emerged as significant predictors of the derivation of atypicality inferences.

A large body of work in experimental pragmatics has reported that working memory capacity plays an important role in pragmatic processing, particularly for implicatures based on lexical scales. For example, lower working memory capacity has been shown to predict a tendency to give logical (i.e., non-pragmatic) responses (Feeney et al., 2004), while individuals with higher working memory capacity have been argued to better suppress default responses under cognitive load (Dieussaert et al., 2011). In addition, taxing working memory through a secondary task has been shown to reduce the rate of pragmatic responses (e.g., De Neys & Schaeken, 2007; Marty et al., 2013; Cho, 2020). Taken together, it has been widely assumed that working memory supports pragmatic enrichment, such that individuals with higher working memory capacity should be more likely to derive pragmatic implicatures.

The present findings do not support this assumption for atypicality inferences. Across both individual-differences analyses and dual-task experiments, no evidence was found that working memory capacity modulated their derivation. This convergence across two methodologies suggests that working memory is not a central limiting factor for atypicality inferences.

Similarly, I found no evidence for the role of inhibition or memory updating. Although these components of executive function have been less frequently investigated in experimental pragmatics, they are theoretically related to working memory and broader cognitive processes (Miyake et al., 2000), which motivated testing of whether they contribute to variability in the derivation of atypicality inferences.

In the existing literature, some evidence links inhibition to aspects of pragmatic and figurative language processing. For example, in the context of social communication in children with ADHD, Rints et al. (2015) showed that individual differences in inhibition predicted certain aspects of pragmatic language development. However,

inhibition in that study was measured using the Statue task, which targets motor inhibition. This differs from the Stroop task used in Chapter 10, which captures cognitive inhibition. In a different domain, Chiappe & Chiappe (2007) found that Stroop performance and inhibition errors were related to both the speed and the quality of metaphor interpretation. For memory updating, Schuster et al. (2023) reported that updating capacity correlates with the magnitude of semantic-pragmatic adaptation in the use of uncertainty expressions. Taken together, these findings suggest that inhibition and memory updating may be involved in certain types of pragmatic processing. However, the present results do not provide evidence that these mechanisms play a substantial role in the derivation of atypicality inferences.

As discussed in the previous section, one possible interpretation is that atypicality inferences rely on script knowledge, which can be rapidly accessed and integrated with linguistic input (Hagoort et al., 2004; Chwilla & Kolk, 2005). Detecting informational redundancy relative to such knowledge may therefore place relatively low demands on executive resources. At the same time, these findings are compatible with previous work suggesting that executive functions may play a more indirect or supportive role in pragmatic processing (Fairchild & Papafragou, 2021).

Taken together, my results suggest that working memory, inhibition, and memory updating are not strong predictors of variability in atypicality inference derivation. Rather than being primarily driven by differences in executive functions, the derivation of atypicality inferences may depend on other cognitive mechanisms.

11.5.2 Effects of higher-level reasoning abilities

In Chapter 8, I found an effect of reasoning ability (operationalized as fluid intelligence and cognitive reflection ability) on the derivation of atypicality inferences. Comprehenders with higher reasoning ability were more likely to derive atypicality inferences, both in typicality judgments and in their explanations, where they more frequently produced responses signifying an atypicality inference.

Fluid intelligence was measured using Raven's Progressive Matrices (Raven, 1936), while cognitive reflection ability was assessed using a Cognitive Reflection Test (CRT; Frederick, 2005). These two measures were shown to exhibit substantial shared variance (Welsh, 2022). In line with this, I combined them using a principal component analysis and referred to the resulting composite as *reasoning ability* in Chapter 8.

Empirical work in pragmatic processing provides evidence that reasoning-related abilities can modulate pragmatic behavior. For example, a positive effect of Raven's matrices and CRT scores on pragmatic responding has been observed in reference game paradigms (Mayn & Demberg, 2022, 2026, 2023; Duff et al., 2025). Similarly, Heyman & Schaeken (2015) report that participants with higher CRT scores show more consistent responses to underinformative utterances. The present findings extend this line of work by demonstrating that reasoning ability is relevant not only for

non-social problem-solving tasks, but also for the derivation of atypicality inferences, in narrative contexts.

A plausible interpretation is that reasoning ability primarily affects the *repair* process once informational redundancy has been detected, rather than the detection itself. In the present materials, the critical utterance describes an event that is already highly expected given script knowledge. To derive an atypicality inference, comprehenders must go beyond utterance's literal meaning and search for an alternative context in which the utterance becomes informative. This involves, at a minimum, detecting a mismatch between what is said and what would normally need to be said, and then constructing a non-obvious explanation that reconciles the utterance with the discourse context. Individuals with higher reasoning ability may be more effective in sustaining this interpretive process, insofar as they can maintain, compare, and selectively elaborate multiple candidate explanations during comprehension.

At the same time, although Raven's matrices and the CRT have been shown to exhibit substantial shared variance, suggesting partial overlap in the cognitive processes they engage (Welsh, 2022), they are often argued to place somewhat different functional demands on higher-level cognition (Stanovich, 2012). Raven's matrices are typically taken to index abstract pattern detection, efficient allocation of attention within a problem space, and persistence in exploring the solution space before giving up (Stocco et al., 2021; Liu & Demberg, 2026). In contrast, CRT – beyond capturing the ability to resist an immediately available response (Frederick, 2005) – has also been argued to reflect aspects of thinking disposition, such as the tendency to engage in reflective, open-minded, and effortful thought (Toplak et al., 2011, 2014). From this perspective, it is plausible that CRT-like ability is more closely related to the **willingness** to engage in the generation of accommodation candidates – in other words, the willingness to perform the pragmatic reasoning itself. In contrast, Raven-like ability supports sustained exploration of the hypothesis space and facilitates disengagement from unproductive interpretations. This distinction, however, should not be interpreted as suggesting that the two measures rely on separate mechanisms; instead, they most likely capture complementary aspects of the same underlying reasoning process.

A point of concern is that the effect observed in Chapter 8 was not replicated in Chapter 10, where atypicality inferences were tested alongside other types of implicatures (e.g., those based on lexical scales such as *some* or *or*). One possible explanation is that differences in task design altered the processing demands of the task. In particular, presenting multiple types of implicatures within the same experiment may have encouraged a more generalized or heuristic processing strategy, reducing the extent to which participants engaged in the deeper reasoning processes required for deriving atypicality inferences. Another possibility is that the presence of multiple inference types introduced additional variability or interference, thereby attenuating the effect of reasoning ability. Finally, it must also be acknowledged that the effect observed in Chapter 8 may not be fully robust.

At the same time, recent ongoing work by Bila (2026), which closely replicates the design of Chapter 8, reports a similar effect of reasoning ability on both typicality judgments and annotated explanations. This finding supports the interpretation that the effect of reasoning is present when the conditions closely resemble those of Chapter 8.

11.5.3 Effects of socio-pragmatic abilities

Theory of mind (ToM; the ability to attribute mental states such as beliefs, intentions, and knowledge to others, Goldman, 2012; Carlson et al., 2013) and autistic traits (Baron-Cohen et al., 2001b) have both been proposed to influence how comprehenders go beyond literal meaning. The general idea is that pragmatic interpretation may, at least in some cases, involve sensitivity to the speaker's perspective and communicative intentions. Empirically, some studies report that ToM abilities and autistic traits are associated with differences in implicature derivation (Nieuwland et al., 2010; Yang et al., 2018; Mazzaggio & Surian, 2018; Fairchild & Papafragou, 2021), although findings are not fully consistent (Antoniou et al., 2016; Heyman & Schaeken, 2015). This makes both ToM and autistic traits good candidates for explaining variability in the derivation of atypicality inferences. In particular, I hypothesized that these factors may play a role in recognizing the pragmatic markedness of informationally redundant utterances or in supporting the accommodation of atypical behavior.

More specifically, I tested the role of theory of mind, measured via the Reading the Mind in the Eyes Test (RMET; Baron-Cohen et al., 2001a) in Chapter 10, and autistic traits, as measured by the Autism Spectrum Quotient (AQ; Baron-Cohen et al., 2001b), in Chapters 8 and 10. Below, I discuss the findings.

Theory of Mind. No effect of theory of mind (ToM) was observed. One possible explanation is that atypicality inferences may differ from other implicatures for which ToM effects have been reported, in that they can be derived without detailed mentalizing, at least at an initial stage (Brown & Dell, 1987; Grigoroglou & Papafragou, 2019; Lockridge & Brennan, 2002). In particular, listeners, acting as overhearers, may be able to recognize informational redundancy and infer its pragmatic markedness without explicitly reasoning about the speaker's communicative intentions or knowledge states. At the same time, it remains an open question whether later stages of the derivation process, such as constructing and evaluating possible explanations for the inferred atypicality, can proceed without engaging such mentalizing processes.

Another possible explanation concerns the measure I used to assess ToM, namely the Reading the Mind in the Eyes Test (RMET; Baron-Cohen et al., 2001a). There is an ongoing debate on what different ToM tasks capture and, in particular, whether RMET primarily measures emotion recognition rather than the ability to infer beliefs and intentions (see, e.g., Quesque & Rossetti, 2020, for discussion). This raises the possibility that RMET may not capture the aspect of ToM that is most relevant for

atypicality inferences derivation. In line with this, ToM has been argued to comprise at least two subcomponents: cognitive ToM (inferring others' beliefs and intentions) and affective ToM (reasoning about others' emotional states); see, e.g., Shamay-Tsoory et al. (2006); Tager-Flusberg & Sullivan (2000). The RMET is typically taken to index the latter. It is therefore possible that the absence of an RMET effect reflects a mismatch between the construct measured and the processes involved in deriving atypicality inferences, rather than the absence of a role for ToM more generally.

This interpretation is supported by recent (ongoing) work (Bila, 2026), which used a similar paradigm as in Chapter 8 and included both RMET (affective ToM) and a measure of cognitive ToM (e.g., the Faux Pas task; Stone et al., 1998). The results showed no effect of RMET, consistent with my findings, but did reveal an effect of cognitive ToM on both typicality ratings and given explanations. Participants with higher cognitive ToM were more likely to derive atypicality inferences and to provide atypical explanations. These results suggest that mentalizing abilities related to reasoning about beliefs and intentions, rather than emotion recognition, may be relevant for atypicality inference derivation, potentially by supporting the recognition of pragmatic markedness or the construction of fitting explanations.

Autistic traits. Given that autistic traits have often been associated with difficulties in mentalizing abilities (see e.g., Baron-Cohen et al., 1985; Bosco et al., 2018, for discussion), one might expect that higher levels of such traits would be associated with reduced sensitivity to pragmatic cues and, consequently, with weaker derivation of atypicality inferences. In Chapter 8, I found an effect of participants' Autism-Spectrum Quotient (AQ) scores on the derivation of atypicality inferences, such that participants with higher levels of autistic traits were more likely to derive atypicality inferences. This finding was against the predictions.

One possible explanation concerns the multifaceted nature of the AQ measure (Kloosterman et al., 2011). The Autism-Spectrum Quotient was designed to capture a range of traits (Baron-Cohen et al., 2001b), including differences in social communication as well as attention to detail. In the context of atypicality inferences, while the former may be associated with difficulties in reasoning about speaker intentions, the latter may facilitate sensitivity to subtle mismatches between an informationally redundant utterance and its context. From this perspective, future work is needed to determine which specific subcomponents or items of the AQ drive the observed effect.

At the same time, this interpretation remains tentative. The AQ effect was not replicated in Chapter 10, which may be due to differences in task design. Alternatively, the lack of replication raises the possibility that the observed effect is not robust. Further research is therefore required to establish whether and how autistic traits, as measured by the AQ test, are related to atypicality inference derivation.

11.5.4 Effects of linguistic experience

I tested the effect of exposure to print in Chapters 8 and 10. I used the Author Recognition Test (ART) to measure print exposure (Stanovich & West, 1989). The test measures exposure to print indirectly and can be viewed as a proxy. As Scholman et al. (2020) note, higher reading and literacy skills can still be acquired through means other than extensive reading of fiction. Nevertheless, despite this limitation, the Author Recognition Test has been widely used in studies exploring diverse linguistic phenomena (Schuster et al., 2023; Johnson & Arnold, 2021; Freed et al., 2017). More generally, linguistic experience has been shown to support various aspects of language and cognition (Scribner & Cole, 1981; Stanovich & Cunningham, 1993; Stanovich et al., 1995; Cipielewski & Stanovich, 1992; Freed et al., 2017). In the field of pragmatic processing, language skills have been hypothesized to play a role in some types of pragmatic phenomena, particularly those that arise in linguistically rich contexts, where comprehenders must process more complex discourse and integrate larger amounts of textual input (Yang et al., 2018).

I found no effect of ART on the derivation of atypicality inferences. The only effect observed was a main effect of ART, such that participants tended to give higher typicality judgments overall, regardless of the presence of an informationally redundant utterance. However, this effect was not replicated in the analysis based on participants' explanations, which limits its interpretability. One possible explanation is that greater exposure to print is associated with increased tolerance for overinformativity, or with differences in how redundancy is evaluated at the rating level. At present, however, this interpretation remains speculative and requires further investigation.

These results go against the hypotheses outlined by Yang et al. (2018), at least insofar as atypicality inferences arise in linguistically rich contexts. They also appear to contrast with findings reported by Çiftlikli & Demirel (2022), who showed that performance on some conversational implicatures (e.g., indirect requests) is positively correlated with language skills in language learners. However, it is important to note that the measure of language skills used in that study differs substantially from the ART. In particular, Çiftlikli & Demirel (2022) employed the IELTS reading test, which more directly targets reading comprehension abilities. In contrast, the ART is best understood as a proxy for exposure to print and does not provide a fine-grained assessment of language proficiency.

Taken together, these results suggest that there may be no direct relationship between linguistic experience and the derivation of atypicality inferences. At the same time, future work is needed to disentangle different aspects of linguistic experience. In particular, future research should aim to separate reading experience from reading proficiency and to decompose these constructs into more specific functional components. This could be achieved by employing a broader range of measures of language skills and exposure to print.

11.6 Comparison with other implicature types

In the final research question, I examined how atypicality inferences relate to other types of pragmatic implicature. This question is motivated by ongoing debates about whether different implicature types rely on shared cognitive mechanisms or instead reflect distinct inferential processes. While scalar implicatures are often treated as generalized conversational implicatures, atypicality inferences are more closely aligned with context-dependent reasoning and have been argued to resemble particularized implicatures (Levinson, 2000). However, it remains unclear whether this theoretical distinction is reflected in comprehenders' behavior. Moreover, previous studies rarely compare multiple implicature types within the same participant sample, limiting direct empirical evidence. Examining how atypicality inferences pattern relative to other implicature types therefore provides a useful reference point for understanding the processes involved in their derivation.

To address this question, I designed an experiment in which atypicality inferences were tested alongside several types of scale-based implicatures, including scalar implicatures, numerals, disjunctions, and their variants embedded in non-monotonic contexts (Chapter 10). This design allowed me to directly compare response tendencies across implicature types within the same individuals.

The results showed that atypicality inferences exhibit only limited correlations with implicatures tied to lexical scales. While some implicature types cluster together – particularly those based on similar lexical scales – atypicality inferences display a distinct response pattern. This finding is consistent with Floyd et al. (2025), who show that pragmatic language use is not supported by a single unified ability but instead comprises several dissociable components (Floyd et al., 2025).

Chapter 12

Directions for future work

12.1 Variability of script knowledge

Script knowledge serves as an effective proxy for world knowledge, aligning comprehenders on what might be perceived as informationally redundant (Kravtchenko, 2022). However, it is important to acknowledge that scripts have limitations: they are not entirely universal and are often culturally specific. An activity scripted in one culture may not have a corresponding script in another, or the sequence of events associated with that activity may vary across cultures. Harris et al. (1988) showed that cultural differences affect memory for narratives, such that stories that do not match a participant's cultural scripts are often recalled in a way that makes them more similar to familiar scripts.

For the purposes of this dissertation, this suggests that comprehenders from different cultural backgrounds may engage in different processes when interpreting stories based on unfamiliar scripts. In such cases, atypicality inferences may fail to arise if the event described in the target utterance is not perceived as redundant relative to the comprehender's own script knowledge. Alternatively, if the inference is derived, it may involve additional effort, for instance due to competition between different script representations or the need to retrieve a less accessible script.

In addition to cross-cultural variation, scripts also change over time as a result of technological and social developments. This has consequences for the materials used in experimental studies. Changes in everyday practices can make certain alternatives more or less salient, thereby affecting the strength of the pragmatic effect. For example, in the *going shopping* story, I observed that explanations such as "John does not usually pay the cashier because he uses self-checkout machines" became more frequent over the years during which these materials were used. This alternative is also more plausible and socially acceptable than assuming that John regularly steals goods. Since the task concerns typical behavior, participants may therefore be more

inclined to accept the implicature when such alternatives are readily available. I observed an even more interesting effect in a story about a visit to a medical practice, where the target utterance was "He was examined by the doctor!". This item behaved differently across time: during the COVID-19 period, it no longer reliably triggered the expected pragmatic effect. One possible explanation is that expectations about medical care shifted, such that being examined by a doctor was no longer taken for granted under conditions of system overload. Alternatively, participants may still have detected the redundancy but accommodated it by appealing to the COVID-19 context, in which explicitly mentioning such an event becomes informative.

Together, these observations suggest that it should not be assumed that the results of this dissertation are generalizable across all cultures and contexts. Both cultural background and temporal changes in everyday practices may affect whether atypicality inferences are derived and how strong they are. Future research should therefore consider cross-cultural variation more explicitly and ensure that experimental materials are appropriately normed and adapted to the relevant population and time period.

12.2 Item-level variability in atypicality inferences

Across the experiments, I observed variability between items in the strength of atypicality inferences. Some stories showed a strong difference between conditions with and without informational redundancy, whereas others showed only weak or, sometimes, no effects. This variability was visible not only in the typicality ratings, but also in the explanations provided by participants, as participants differed in how they accommodated the inferred atypicality and generally how easy it was for them to arrive at a plausible explanation. The observed variability should not necessarily be viewed as a limitation. A diverse set of experimental items rather helped to highlight the richness of human reasoning, as different stories allowed for different types of interpretations and accommodations. At the same time, it points to a direction for future work. One possible approach would be to investigate item-level properties more directly, for example by eliciting possible explanations in a separate task (e.g., a cloze-like task or by directly asking participants what typically happens **instead**, similarly to how LLMs were prompted in Kurch et al., 2024, where the derivation scheme was tested step by step). These responses could then be classified in terms of their frequency and along dimensions such as plausibility or social acceptability. For instance, although stealing may be a salient candidate in the context of grocery shopping, it may be less acceptable as a long-term explanation of typical behavior, which in turn may affect how readily participants adopt the corresponding inference. Such analyses could help to better understand the interpretation of individual-difference effects.

12.3 Online measures of atypicality inference derivation

The experiments in this dissertation relied on offline measures, in particular typicality judgments and explanation elicitation. While these measures provide insight into participants' final interpretations, it remains unclear whether atypicality inferences are derived during the reading of the story or only at a later stage, for example when participants are asked to provide a judgment or an explanation. Future research could address these questions by employing online methods such as eye-tracking (Parola & Bosco, 2022). Online measures would make it possible to investigate the time course of atypicality inference derivation and to provide additional insight into the cognitive factors involved in this process.

12.4 Atypicality inferences in interactive and social contexts

The present experiments were conducted in a setting in which participants acted as overhearers, rather than as active participants in a communicative interaction. Therefore, a direction for future work would be to investigate atypicality inferences in more interactive settings, where comprehenders are more directly involved in conversational exchange (e.g., Bonnefon et al., 2009).

Such settings would also make it possible to examine more naturally how flexible comprehenders are in adapting to different speaker characteristics. For example, speakers may differ in age, linguistic competence, or communicative style (e.g., adult vs. child, native speaker vs. language learner, speakers with talkative vs. quiet personality; Fairchild et al., 2020; Reksnes et al., 2024; Rees et al., 2026). Investigating this would help clarify how expectations about speakers influence atypicality inferences and how such effects interact with comprehender-specific cognitive and socio-pragmatic traits.

Finally, it would be interesting to extend this line of work to developmental populations. Previous research has examined how children respond to informativity in communication (e.g., Rees & Rohde, 2024), but it remains an open question whether and how they derive atypicality inferences. Addressing this question would require either adapting the materials to age-appropriate scripts or, alternatively, using the current materials to trace the development of sensitivity to informational redundancy.

12.5 Role of additional personality-related traits

Beyond the factors considered here, other personality-related traits may also play a role in pragmatic processing. Previous work has explored traits such as those captured

by the Big Five, as well as self-perceived honesty, particularly in the context of scalar implicatures. However, the evidence for such effects is mixed. Some studies find little to no contribution of personality-based measures once cognitive factors are taken into account (e.g., Heyman & Schaeken, 2015), while others suggest that traits such as self-rated honesty may modulate interpretation under certain conditions, for instance via expectations about speaker's honesty (Feeney & Bonnefon, 2012).

One possible direction for future work is to examine whether such traits become more relevant in richer or more socially grounded contexts. For example, if listeners expect speakers to be generally honest and informative, this may increase the likelihood that a redundant utterance is interpreted as meaningful rather than as noise. At the same time, it remains unclear whether such expectations would affect only early stages of interpretation or also the subsequent derivation and accommodation of atypicality. More systematic investigation of these factors, ideally across different task types and discourse settings, would help to clarify their role in pragmatic reasoning.

12.6 Atypicality inferences and large language models

Another direction for future work concerns the comparison between human comprehenders and large language models (LLMs). In previous work, we tested whether LLMs derive atypicality inferences in a human-like manner (Kurch et al., 2024). Despite extensive prompting (e.g., one-shot and few-shot settings, as well as step-by-step prompting of the derivation process), the models did not reliably reproduce humans and appeared to perform pattern matching rather than pragmatic reasoning.

Given the rapid development of LLMs, it would be informative to revisit these findings with more recent models. This would make it possible to assess whether sensitivity to informational redundancy and the ability to derive atypicality inferences emerge with improved architectures, or whether these inferences remain difficult to capture for current models.

In addition, future work could explore how atypicality inferences are affected when the speaker is modeled as a human vs. an artificial agent. As LLM-based systems are increasingly used as conversational partners, comprehenders may or may not attribute different communicative intentions to them. This, in turn, may influence how informational redundancy is interpreted and whether atypicality inferences are derived.

List of Figures

List of Tables

Appendix A

Complete set of experimental materials for atypicality inferences

In this appendix, a complete set of experimental materials used to assess atypicality inferences is provided.

The materials comprise twenty-four different stories describing everyday situations, originally taken from Kravtchenko (2022) (Table A.1, Context version I).

The stories used in the experiments described in Chapters 7 and 8, Section 9.2, and Chapter 10 were identical to those in Kravtchenko (2022) (Table A.1, Context version I).

In the experiments reported in Sections 9.3 and 9.4, the stories were modified so that responses to the revised comprehension questions could be provided on a continuous scale (Table A.1, Context version II). All modifications to the original texts are indicated in bold.

The target and comprehension questions associated with each story are presented in Table A.2. Filler questions are listed in Table A.3. The target question assessing the target event typicality, as well as Filler 1, 2, and 3 questions, were taken from Kravtchenko (2022).

Table A.1: Materials

Story	Context version I	Context version II	IR-utterance
grocery	John often goes to the grocery store around the corner from his apartment. Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.	John often goes to the grocery store around the corner from his apartment. Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway **preparing for some yoga exercises, as she does all the time at home. John** started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.	Susan said to Peter: "John just came back from the grocery store. He paid the cashier!"
restaurant	Mary is a journalist who often goes to restaurants after her interviews. Yesterday, she went to a popular Chinese place. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.	Mary is a journalist who often goes to restaurants after her interviews. Yesterday, she went to a popular **Italian place, where they always have nice live music**. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.	David said to Sally: "I ran into Mary leaving that Chinese place. She ate there!"
feed dog	Jim lives in a shared apartment, where it's his job to feed the dog in the evenings The other day he was feeding the dog some canned food, as his roommate Lucy came into the kitchen, and made herself a snack while chatting with him. Later in the evening, she settled down to watch TV alone with their roommate Carl.	Jim lives in a shared apartment, where it's his job to **always** feed the dog in the evenings. The other day he was feeding the dog some canned food, as his roommate Lucy came into the kitchen, and made herself a snack while chatting with him. Later in the evening, she settled down to watch TV alone with their roommate Carl.	Lucy said to Carl: "Jim was feeding the dog earlier. He threw the can away!"

subway	Jane takes the subway all the time to get around the city. Today she was entering a subway station when she ran into her friend Don, and they took the train together as they were heading in the same direction. Later that day, Don ran into Beth, Jane's sister, on the street.	no changes	Don said to Beth: "I took a train with Jane today. She bought a subway ticket!"
swimming	Lisa likes to go swimming at a nearby pool after work. A couple days ago she was at the pool when she saw Harvey, another regular member, and they stopped to chat. After Harvey changed and went out into the pool area, he ran into Jen, another swimmer and a friend of Lisa's.	Lisa likes to go swimming at a nearby pool after work**, as they are always open when her working day is over**. A couple days ago she was at the pool when she saw Harvey, another regular member, and they stopped to chat. After Harvey changed and went out into the pool area, he ran into Jen, another swimmer and a friend of Lisa's.	Harvey said to Jen: "Lisa's here to swim, too. She brought her swimsuit!"
train	Brian takes the train most mornings, although the commute takes a long time. Last week when he was getting on the train, he ran into his old colleague Rachel, and they chatted until Brian got off. When Rachel got to work, she saw Oliver, who also used to work with Brian.	Brian takes the train most mornings, although the commute takes a long time. Last week**, as always, he bought a cup of coffee, and** when he was getting on the train, he ran into his old colleague Rachel. They chatted until Brian got off. When Rachel got to work, she saw Oliver, who also used to work with Brian.	Rachel said to Oliver: "I saw Brian on the train this morning. He got off at his stop!"
work	Laura works as a software engineer at a large company. A couple of days ago she was getting ready to leave for work together with her husband Dustin. After they both left the house, he ran to catch his bus, and met up with Courtney, an acquaintance who took the same bus with him every day.	Laura works as a software engineer at a large company. **She likes her job, because she never has any business trips.** A couple of days ago she was getting ready to leave for work together with her husband Dustin. After they both left the house, he ran to catch his bus, and met up with Courtney, an acquaintance who took the same bus with him every day.	Dustin said to Courtney: "Laura was just getting ready for work with me. She grabbed her house keys!"

doctor	Bruce goes to his local medical practice every few years. Yesterday after leaving the practice he ran into his friend Sarah on the street, and they stopped to catch up. After they parted, Sarah walked on and soon saw Bruce's brother Drake on the street. She stopped to say Hi.	Bruce goes to his local medical practice every few years. Yesterday after leaving the practice he **was going, as always, to take the bus home, and** ran into his friend Sarah walked on the street. They stopped to catch up. After they parted, Sarah walked on and she soon saw Bruce's brother Drake on the street. She stopped to say Hi.	Sarah said to Drake: "Bruce was just leaving the medical practice. He got examined by the doctor!"
shampoo	Olivia has beautiful hair, and pays a lot of attention to it. Today, when she was leaving the bathroom after showering, she ran into her roommate and best friend Thomas. She talked to him briefly about her hair, as she tends to do. Later that day, when their housemate Jill came home, she and Thomas started talking about Olivia.	Olivia has beautiful hair. **She pays a lot of attention to it, and never dyes her hair.** Today, when she was leaving the bathroom after showering, she ran into her roommate and best friend Thomas. She talked to him briefly about her hair, as she tends to do. Later that day, when their housemate Jill came home, she and Thomas started talking about Olivia.	Thomas said to Jill: "Olivia was talking to me about washing her hair. She used shampoo!"
skydiving	Jared takes skydiving courses at the local airfield, when he has free time. Last week he was at the skydiving center, with his friend Stella in the same group as him. They spent the day together, and when Stella went home in the evening, she texted Jared's brother Don, who was also a good friend of hers.	Jared takes skydiving courses at the local airfield, when he has free time. Last week he was at the skydiving center, with his friend Stella in the same group as him. **As always,** they spent the day together, and when Stella went home in the evening, she texted Jared's brother Don, who was also a good friend of hers.	Stella said to Don: "Jared was in the skydiving course today. He jumped out of the plane!"

letter	Amy enjoys writing letters to people she is close to, especially around holidays. About two days ago, she wrote a letter to her cousin Michelle, and today she talked about it with her brother Steve. In the evening, Steve got a call from Michelle, and they started talking about family.	Amy enjoys writing letters to people she is close to, especially around holidays. **She always uses fountain pens for writing letters, as it is something deep and old fashioned.** About two days ago, she wrote a letter to her cousin Michelle, and today she talked about it with her brother Steve. In the evening, Steve got a call from Michelle, and they started talking about family. Steve said to Michelle: Amy wrote you a letter. She mailed it!	Steve said to Michelle: "Amy wrote you a letter. She mailed it!"
bus	Adam usually takes the bus to work, as the stop is a few blocks from his house. Last week, after he got off the bus, he ran into Virginia, his ex-girlfriend. They stopped for a little while to catch up.	**Adam lives quite far from his office.** Last week**, as always,** he took the bus to work, as the stop is a few blocks from his house. After he got off the bus, he ran into Virginia, his ex-girlfriend. They stopped for a little while to catch up.	Adam said to Virginia: "I took the bus this morning. I walked to the bus stop!"
clothes	Esther often goes along with her friends when they go clothes shopping, as it's something she also enjoys. Today, when she was walking out of a mall after spending time with her friends, she ran into George, another old friend of hers. They decided to catch up while walking to the bus stop.	Esther often goes along with her friends when they go clothes shopping, as it's something she also enjoys. Today, when she was walking out of the mall after spending time with her friends, she **bought an ice-cream, as she always does, and** ran into George, another old friend of hers. They decided to catch up while walking to the bus stop.	Esther said to George: "I was out clothes shopping. I tried something on!"
airplane	Greg frequently travels by air, to see family and attend conferences. Last week he flew to a conference, and met up with Helen, an old colleague he occasionally traveled with. They went to breakfast together, and started talking about their travel.	Greg frequently travels by air, to see family and attend conferences**, although he never spends his miles, when booking a ticket**. Last week he flew to a conference, and met up with Helen, an old colleague he occasionally traveled with. They went to breakfast together, and started talking about their travel.	Greg said to Helen: "I flew here. I took my cell phone on board with me!"

hair	Sandy usually cuts her own hair, although she has no training. Two days ago, after she gave herself another haircut, she went for a walk along her street. She quickly ran into her ex, Patrick, and they stopped to catch up for a few minutes.	Sandy usually cuts her own hair, although she has no **formal** training **and never dyes her hair at home**. Two days ago, after she gave herself another haircut, she went for a walk along her street. She quickly ran into her ex, Patrick, and they stopped to catch up for a few minutes.	Sandy said to Patrick: "I just cut my hair. I used scissors!"
exhibit	Henry often goes to art exhibitions, as there's an art museum a short walk from his place. Last week, after going to a new photography exhibition, he encountered his friend Max on his way home. They paused on the street and chatted for a while.	Henry often goes to art exhibitions, as there's an art museum a short walk from his place. Last week, after going to a new photography exhibition, he encountered his friend Max on his way home. **They almost missed each other, since Henry always listens to music on headphones when walking on the streets and at first he did not hear Max calling him out. Finally** they paused on the street and chatted for a while.	Henry said to Max: I just went to the new photo exhibit. "I looked at the photographs!"
drive	Helen works hard at her job, and enjoys the challenges she's given at work. Today, after driving her car to work as usual, she ran into her office-mate Peter while walking into the building. They stopped briefly to say hello.	Helen works hard at her job. She enjoys the challenges she's given at work **and never leaves the office before her boss.** Today, after driving her car to work as usual, she ran into her office-mate Peter while walking into the building. They stopped briefly to say hello.	Helen said to Peter: "I just parked my car. I locked it!"
pizza	Gary often orders pizza at work, from a famous pizzeria nearby. A few days ago, after he placed an order, his colleague Stephanie walked over to his cubicle to chat.	Gary often orders pizza at work, from a famous pizzeria nearby**, as he always gets some discount there**. A few days ago, after he placed an order, his colleague Stephanie walked over to his cubicle to chat.	Gary said to Stephanie: "I just ordered pizza. I picked the toppings!"

dishes	Julia always tries to wash the dishes after eating, to avoid annoying her roommates. A few days ago, she was getting ready to go out after doing the dishes. She ran into her roommate Justin on her way out, and started talking to him.	Julia always tries to wash the dishes after eating, to avoid annoying her roommates**, as they always complain if there are some dishes left in the sink**. A few days ago, she was getting ready to go out after doing the dishes. She ran into her roommate Justin on her way out, and started talking to him.	Julia said to Justin: "I just did the dishes. I rinsed them!"
library	Emma often borrows books from the library, as she doesn't have much spare cash to spend. Last week, after going to the library, she was heading home with several books, and ran into her best friend Tim on the street. They stopped to quickly say hello.	no changes	Emma said to Tim: "I just got some books at the library. I checked them out!"
fueling	Rick prefers to go to the local gas station to fuel up, though they overcharge him. Yesterday he was fueling up there when he saw his friend Annie. They talked until he was finished, then said goodbye. After he was gone Annie went inside to buy some cigarettes, and ran into Sean, a mutual friend of theirs.	–	Annie said to Sean: "Rick just fueled up here. He closed the fuel cap!"
pasta	Nick enjoys making pasta dishes for his roommates, as it's an easy way to contribute to the household. Yesterday he was preparing pasta in the kitchen, to sit in the fridge until a party tomorrow. When he was done and cleaning up, his roommate Clara came into the kitchen, and they started talking about his dish.	–	Nick said to Clara: "I made some pasta for the meal. I boiled it in water!"

baking	Grace enjoys baking, as it's a great way to make new friends. A few days ago she was baking a cake in her kitchen. After she had put it in the oven, her roommate Kyle came into the kitchen to make a salad for himself. They started chatting about food.	–	Grace said to Kyle: "I'm baking a cake right now. I preheated the oven!"
laundry	Logan recently started doing his own laundry, after moving out of his parents' house. Yesterday, after doing a load, he went to the living room to watch some TV. Soon his roommate Sophia came home, and asked about his day while taking off her coat.	–	Logan said to Sophia: "I just did the laundry. I used detergent!"

Table A.2: Target and comprehension questions. **Target** = critical question. **Compr. 1** = comprehension questions used in Section 9.2. **Compr. 2** = comprehension questions used in Section 9.3 and Section 9.4.

Story	Target	Compr. 1	Compr. 2
grocery	How often do you think John usually pays the cashier, when going shopping?	Where did John meet his roommate Susan?	How often do you think John's flatmate Susan does yoga at home?
restaurant	How often do you think Mary usually eats, when going to a restaurant?	What is Mary's occupation?	How often do you think there is live music in the restaurant where Mary went yesterday?
feed dog	How often do you think Jim usually throws the can away, when feeding the dog?	What pet does Jim have to feed?	How often do you think Jim feeds the dog in the evenings?
subway	How often do you think Jane usually buys a ticket, when taking the subway?	Where did Jane meet her friend Don today?	How often do you think Jane takes the subway to get around the city?
swimming	How often do you think Lisa usually brings her swimsuit, when going swimming?	What does Lisa like to do after work?	How often do you think the swimming pool nearby Lisa's office is open when she finishes working?
train	How often do you think Brian usually gets off at his stop, when taking the train?	Where did Brian meet his colleague Rachel?	How often do you think Brian buys a cup of coffee in the mornings before getting on the train?
work	How often do you think Laura usually grabs her house keys, when getting ready for work in the morning?	What is Laura's occupation?	How often do you think Laura has business trips at her job?
doctor	How often do you think Bruce usually gets examined by the doctor, when going to the medical practice?	Where did Bruce go yesterday?	How often do you think Bruce takes the bus home after going to his local medical practice?
shampoo	How often do you think Olivia usually uses shampoo, when washing her hair?	What did Olivia talk about with her best friend Thomas?	How often do you think Olivia dyes her hair?
skydiving	How often do you think Jared usually jumps out of a plane, when going skydiving?	Where does Jared take skydiving courses?	How often do you think Jared and his friend Stella spend the day together when going to a skydiving center?
letter	How often do you think Amy usually mails a letter, after writing it?	What does Amy enjoy doing?	How often do you think Amy uses simple rollerball pens for writing letters to people she is close to?

bus	How often do you think Adam usually walks to the bus stop, when taking the bus in the morning?	How does Adam usually get to work?	How often do you think Adam takes the bus to work?
clothes	How often do you think Esther usually tries something on, when going clothes shopping?	Where did Esther meet her friend George?	How often do you think Esther buys an ice-cream after spending time with her friends in the mall?
airplane	How often do you think Greg usually carries his cell phone on board with him, when flying on a plane?	Where did Greg fly last week?	How often do you think Greg spends his miles when booking an airplane ticket?
hair	How often do you think Sandy usually uses scissors, when cutting her hair?	Who usually cuts Sandy's hair?	How often do you think Sandy dyes her hair at home?
exhibit	How often do you think Henry usually looks at photographs, when going to a photo exhibit?	What exhibition did Henry attend last week?	How often do you think Henry listens to music on headphones when walking on the streets?
drive	How often do you think Helen usually locks her car after parking it?	How does Helen usually get to work?	How often do you think Helen leaves the office before her boss?
pizza	How often do you think Gary usually picks the toppings when ordering pizza?	What does Gary often order at work?	How often do you think Gary gets a discount when ordering pizza from the famous pizzeria nearby?
dishes	How often do you think Julia usually rinses the dishes when doing them?	What does Julia always try to do after eating?	How often do you think Julia's flatmates complain if there are some dishes left in the sink?
library	How often do you think Emma usually checks out the books when getting some books at the library?	Where was Emma coming from when she met her best friend Tim?	How often do you think Emma buys books?
fueling	How often do you think Rick usually closes the fuel cap, after fueling up?	–	–
pasta	How often do you think Nick usually boils pasta in water, when making it?	–	–
baking	How often do you think Grace usually preheats the oven, when baking a cake?	–	–
laundry	How often do you think Logan usually uses detergent, when doing the laundry?	–	–

Table A.3: Filler questions. **Filler 1** = filler questions used in Chapter 9 and Chapter 10. **Filler 2** = filler questions used in Sections 9.3 and 9.4, and in Chapter 10. **Filler 3** = fillers questions used in Chapter 10.

Story	Filler 1	Filler 2	Filler 3
grocery	How often do you think John usually gets apples, when going shopping?	How often do you think Susan and Peter usually talk to each other?	How often do you think John usually goes to the grocery store?
restaurant	How often do you think Mary usually gets to see the kitchen, when going to a restaurant?	How often do you think Sally and David usually run into each other?	How often do you think Mary usually goes to restaurants?
feed dog	How often do you think Jim usually adds some medicine to the food, when feeding the dog?	How often do you think Carl and Lucy usually chat?	How often do you think Jim usually feeds the dog?
subway	How often do you think Jane usually comes close to falling off the platform, when taking the subway?	How often do you think Beth and Don usually meet?	How often do you think Jane usually takes the subway?
swimming	How often do you think Lisa usually brings her children, when going swimming?	How often do you think Jen and Harvey usually see each other at the pool?	How often do you think Lisa usually goes swimming at the pool?
train	How often do you think Brian usually gets to work late, when taking the train?	How often do you think Oliver and Rachel usually see each other?	How often do you think Brian usually takes the train?
work	How often do you think Laura usually puts on several layers of clothing, when getting ready for work in the morning?	How often do you think Dustin and Courtney usually talk to each other?	How often do you think Dustin usually sees Laura getting ready for work?
doctor	How often do you think Bruce usually gets fitted with a heart rate monitor, when going to the medical practice?	How often do you think Drake and Sarah usually run into each other?	How often do you think Bruce usually goes to the medical practice?
shampoo	How often do you think Olivia usually finds some split ends, when washing her hair?	How often do you think Thomas and Jill usually chat with each other?	How often do you think Olivia usually washes her hair?
skydiving	How often do you think Jared is usually the first to jump, when going skydiving?	How often do you think Stella and Don usually talk?	How often do you think Jared usually goes to the skydiving course?
letter	How often do you think Amy usually writes letters?	How often do you think Steve and Michelle usually talk to each other?	How often do you think Amy usually writes letters?

bus	How often do you think Adam usually barely has room to stand, when taking the bus in the morning?	How often do you think Adam and Virginia usually run into each other?	How often do you think Adam usually takes the bus to work?
clothes	How often do you think Esther usually comes across a big sale, when going clothes shopping?	How often do you think Esther and George usually see each other?	How often do you think Esther usually goes to clothing stores?
airplane	How often do you think Greg usually gets into business class, when flying on a plane?	How often do you think Greg and Helen usually meet up?	How often do you think Greg usually travels by airplane?
hair	How often do you think Sandy usually cuts her hair a bit shorter than intended, when cutting it?	How often do you think Sandy and Patrick usually see each other?	How often do you think Sandy usually cuts her own hair?
exhibit	How often do you think Henry usually decides to buy a photograph, when going to a photo exhibit?	How often do you think Henry and Max usually talk to each other?	How often do you think Henry usually goes to photo exhibitions?
drive	How often do you think Helen usually discovers that one of her tail lights has gone out, when parking her car?	How often do you think Helen and Peter usually run into each other?	How often do you think Helen usually drives to work?
pizza	How often do you think Gary usually orders pizza at work?	How often do you think Gary and Stephanie usually talk?	How often do you think Gary usually orders pizza at work?
dishes	How often do you think Julia usually polishes the dishes, when doing them?	How often do you think Julia and Justin usually see each other?	How often do you think Julia usually does the dishes?
library	How often do you think Emma usually looks at the library's exhibit, when getting some books at the library?	How often do you think Emma and Tim usually run into each other?	How often do you think Emma usually gets books at the library?
fueling	How often do you think Rick usually gets some discounted gas, when fueling up?	How often do you think Annie and Sean usually talk?	How often do you think Rick usually fuels up at the gas station?
pasta	How often do you think Nick usually adds some vegetables, when making pasta?	How often do you think Nick and Clara usually talk?	How often do you think Nick usually makes pasta for meals?

baking	How often do you think Grace usually adds chocolate chips to the recipe, when baking a cake?	How often do you think Grace and Kyle usually talk to each other?	How often do you think Grace and Kyle usually talk to each other?
laundry	How often do you think Logan usually adds some softener to the wash, when doing the laundry?	How often do you think Logan and Sophia usually talk to each other?	How often do you think Logan usually does the laundry?

Appendix B

Supplementary materials for Chapter 10

B.1 Examples of experimental items

B.1.1 Bare *some*

The item pool consisted of 40 sentences in total: 20 underinformative target sentences, 10 literal control sentences, and 10 filler sentences. These items were distributed across ten experimental lists.

Control and filler sentences used the same semantic categories as the target items (birds, trees, flowers, fish, and insects), but different lexical instances were used so that no sentence appeared both as a target and as a control or filler item.

Items were distributed across the lists such that, across the two experimental sessions, participants encountered approximately balanced numbers of items from the different semantic categories.

The example of experimental items in each condition is presented in Table B.1.

Table B.1: Bare *some*. Example of experimental items in the target, control, and filler conditions. Source: De Neys & Schaeken (2007).

Condition	**Example sentence**
Target (Underinformative some)	*Some oaks are trees.*
Control (True all)	*All cockroaches are insects.*
Control (True some)	*Some birds are woodpeckers.*
Filler (False all)	*All tuna are birds.*
Filler (False some)	*Some insects are pigeons.*

B.1.2 Bare numerals

Stimuli followed the generation procedure of Marty et al. (2013). Only pictures in which both colors were present were included. Trials were excluded if both the target color and the alternative color occurred exactly N times (e.g., a cube with four red and four green dots paired with the sentence "4 dots are red"), regardless of the configuration of the remaining dots.

For stimulus generation, the starting number was either 3 or 4, as in Marty et al. (2013). The numeral used in the sentence was always one higher than the starting number. For example, in Table B.2, the picture in the target condition has a starting number of 4.

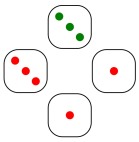
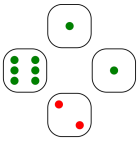
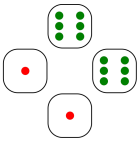
Targets. In the target condition, the number of dots of the target color in the picture was four higher than the starting number.

Controls. In the control condition, the number of dots of the target color matched the numeral in the sentence.

Fillers. In filler trials, the number of dots of the target color was lower than the starting number (for starting number = 4: -1 or -2; for starting number = 3: -1).

Example items are shown in Table B.2. Approximately 200 pictures were randomly generated per condition.

Table B.2: Bare numerals. Example of experimental items in the target and control conditions. Source: Marty et al. (2013).

Condition	Illustration	Expected answer
Control	 5 dots are red.	High rating (the sentence is true with respect to the picture).
Target	 5 dots are green.	**Pragmatic response:** low rating (the sentence is false with respect to the picture). **Literal response:** high rating (the sentence is true with respect to the picture).
Filler	 5 dots are red.	**Pragmatic response:** low rating (the sentence is false with respect to the picture). **Literal response:** high rating (the sentence is true with respect to the picture).

B.1.3 Bare disjunctions

The visual materials were adapted from Frank et al. (2016) and generated following the paradigm of Singh et al. (2016). The stimuli consisted of pictures of either a man

or a snowman wearing different combinations of clothing items.

In total, nine versions of a snowman were used (varying in the direction of the hands and the direction of the carrot nose) and three versions of a man (varying in face color). A man could wear three items (a bowtie, a hat, and glasses), while a snowman could wear four items (a hat, a scarf, mittens, and a belt).

Prior to the experiment, visual scenes were generated by randomly permuting these features across target, control, and filler conditions. Rather than generating the full space of all possible stimulus combinations, around 200 trials were generated by uniformly assigning the domain (man or snowman) across all combinations of experimental conditions. In the experiment, individual trials were therefore randomly sampled from the joint set of man and snowman scenes for each condition. As a result, the number of man and snowman stimuli was approximately balanced across conditions. Filler items were distributed across the two experimental sessions as follows. Participants in the first half of the experimental lists saw a no-connector filler (man) in Session 1 and a conjunction filler (snowman) in Session 2, whereas the order was reversed for the second half of the lists. The same counterbalancing was applied to the target and control items, such that the number of man and snowman stimuli was balanced across the two experimental sessions. Examples of experimental items in all conditions are presented in Table B.3.

Table B.3: Bare disjunctions. Example of experimental items in the target, control, and filler conditions. Source: Singh et al. (2016), visuals taken from Frank et al. (2016).

Condition	Illustration	Expected answer
Control	 The man is wearing glasses or a bowtie	High rating (the sentence is true with respect to the picture).
Target	 The man is wearing a hat or a bowtie	**Pragmatic response:** low rating (the sentence is false with respect to the picture); **Literal response:** high rating (the sentence is true with respect to the picture).
Filler no connector	 The man is wearing glasses.	Low rating (the sentence is false with respect to the picture).
Filler conjunction	 The snowman is wearing a hat and a scarf.	Low rating (the sentence is false with respect to the picture).

B.1.4 Embedded disjunctions

Examples of experimental items in all conditions are presented in Table B.4. Scene generation followed the same permutation procedure as for the bare disjunction materials. In addition, scenes with men and snowmen were distributed across the experimental lists so that the number of man and snowman stimuli was balanced across bare and embedded disjunctions, as well as across conditions, and in total for two sessions.

Table B.4: Embedded disjunctions. Example of experimental items in the target, control, and filler conditions. Source: Singh et al. (2016), visuals taken from Frank et al. (2016).

Condition	Illustration	Description
Control	 Every snowman is wearing mittens or a scarf	High rating (the sentence is true with respect to the picture).
Target	 Every snowman is wearing a belt or a scarf	**Pragmatic response:** low rating (the sentence is false with respect to the picture); **Literal response:** high rating (the sentence is true with respect to the picture).
Filler no connector	 Every man is wearing glasses.	Low rating (the sentence is false with respect to the picture).
Filler conjunction	 Every snowman is wearing a belt and a scarf.	Low rating (the sentence is false with respect to the picture).

B.1.5 Embedded *some*

To generate the scenes for embedded *some*, I first selected five colors of sports uniforms (pink, brown, seafoam, royal blue, and yellow) from the materials of Potts et al. (2016). Using these features, I generated the full set of scene permutations. In total, 180 items were generated for the NNA condition and 360 items for the NSA condition, covering all combinations of ball orders and player uniform colors. Within each scene, the players always wore different uniform colors. Examples of experimental items in the NNA and NSA conditions are presented in Table B.5. Filler items are presented in Table B.6. No additional control items were included for this implicature type.

Table B.5: Embedded *some*. Example of experimental items in the NNA and NSA conditions. Source: Potts et al. (2016).

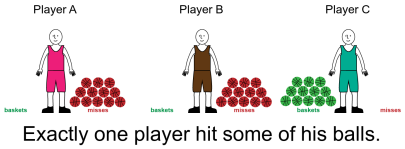
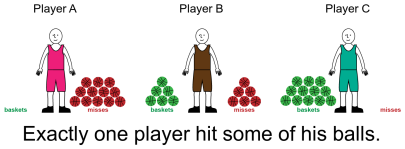
Condition	Illustration	Description
NNA		**Pragmatic response:** low rating (the sentence is false with respect to the picture); **Literal response:** high rating (the sentence is true with respect to the picture).
NSA		**Pragmatic response (local enrichment):** high rating (the sentence is true with respect to the picture); **Literal response:** low rating (the sentence is false with respect to the picture).

Table B.6: Embedded *some*. Filler items. **Note:** N = none, A = all, S = some, M = most, F = few.

Sentence	World	Answer
List 1-5		
Player A shot perfectly.	ANN	TRUE
Most of the players made their shots.	NAM	TRUE
Someone did better than Player B.	FMN	FALSE
Player B tied with Player C.	MNF	FALSE
List 6-10		
We have a clear winner.	NAF	TRUE
All players are tied.	SSS	TRUE
Player C made no successful shots.	FAF	FALSE
Player A placed second.	NFA	FALSE

B.1.6 Politeness implicatures

Following Bonnefon et al. (2009), two story topics were included in the analysis: one about organizing a trip and another about giving a speech – see Table B.7. Each participant saw each story topic and condition only once in a given session. In the second session, the same topics were presented in the opposite condition compared

to Session 1. The first half of the experimental lists contained the speech story in the target condition and the trip story in the control condition. The second half of the lists contained the speech story in the control condition and the trip story in the target condition. There were no filler items.

Table B.7: Politeness implicatures. Example of experimental items in the target and control conditions. Source: Bonnefon et al. (2009).

Condition	Story	Question	Expected response
Control	Imagine you organized a group trip. You are discussing the trip with Alice, who was in the group. There were 6 other people who went on this trip and you know that Alice spoke with all of them about it later. You tell Alice that you would like to know group's feedback. Hearing this, Alice tells you that "Some people hated the way the trip was organized."	Given what Alice told you, do you think that it is possible that everybody hated your trip?	Higher rating compared to the target condition
Target	Imagine you gave a speech at a small political meeting. You are discussing your speech with Denise, who was also there. There were 6 other people in the audience that day and you know that Denise spoke with all of them about it later. You tell Denise that you would like to know the audience's reaction. Hearing this, Denise tells you that "Some people loved your speech."	Given what Denise told you, do you think that it is possible that everybody loved your speech?	Lower rating compared to the control condition

B.2 Exploratory and confirmatory factor analyses for Section 10.6.2

Exploratory factor analysis

To examine whether different implicature tasks reflect common underlying dimensions, an exploratory factor analysis (EFA) was conducted on the participant-level mean ratings for seven implicature types: atypicality inferences, bare *some*, bare numerals, bare disjunctions, embedded disjunctions, and the two embedded *some* conditions (NNA and NSA). Prior to analysis, all variables were centered and scaled.

The number of factors was assessed using parallel analysis. Although the parallel analysis occasionally suggested either a two- or a three-factor solution across runs, a two-factor solution was retained, since the third factor was unstable and did not yield a clearly interpretable structure, whereas the two-factor solution was more parsimonious and theoretically more coherent. Accordingly, an exploratory factor analysis was conducted using the minimum residual (*minres*) extraction method with *oblimin* rotation.

The resulting factor loadings are shown in Table B.8 and visualized in Figure B.1. The first factor (MR1) received strong loadings from bare disjunctions (1.00) and embedded disjunctions (0.86), as well as a moderate loading from bare some (0.52). The second factor (MR2) was defined by the two embedded some conditions, with loadings of 0.99 for the NNA condition and 0.72 for the NSA condition.

Table B.8: Factor loadings from the exploratory factor analysis of the seven implicature types. Only loadings $\geq .20$ are shown. Factors were extracted using the minimum residual (*minres*) method with *oblimin* rotation. h^2 indicates the communality estimate (the proportion of variance in each measure explained by the extracted factors).

Implicature	Factor 1 (MR1)	Factor 2 (MR2)	h^2
Atypicality inferences			0.002
Bare some	0.52		0.33
Bare numerals			0.02
Bare disjunctions	1.00		0.96
Embedded disjunctions	0.86		0.77
Embedded some (NNA)		0.99	1.00
Embedded some (NSA)		0.72	0.48

In contrast, atypicality inferences and bare numerals showed negligible loadings on either factor, indicating that these measures do not share substantial variance with

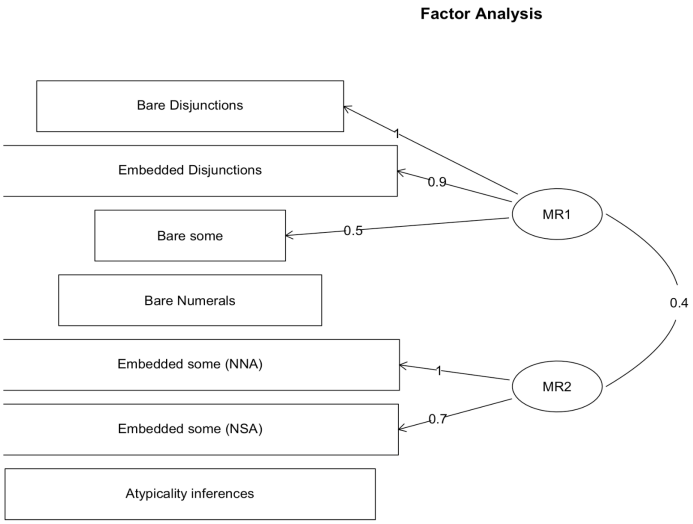


Figure B.1: Diagram of the exploratory factor analysis. Factors were extracted using the minimum residual (*minres*) method with *oblimin* rotation. Numbers on the arrows indicate standardized relationships between the latent factors and the observed measures; the curved arrow indicates the correlation between the two factors.

the other implicature tasks.

Overall, the two factors accounted for 51% of the total variance (Factor 1: 29%, Factor 2: 22%). The factors were moderately correlated ($r = 0.44$). Model fit indices indicated a good fit of the two-factor solution ($RMSR = 0.03$, $RMSEA = 0.054$, $TLI = 0.975$).

The EFA should nevertheless be interpreted with caution. First, the instability of the parallel analysis at the boundary between two and three factors suggests that the latent structure is not entirely robust. Second, the analysis is based on only seven implicature types, which limits the number of indicators available per factor. The EFA is therefore best understood as exploratory evidence concerning the clustering of the implicature tasks, rather than as strong evidence for a fixed latent structure.

Confirmatory factor analysis

A confirmatory factor analysis (CFA) was conducted to test the factor structure suggested by the exploratory analysis. The model specified two latent factors: a disjunction-related factor, defined primarily by bare disjunctions and embedded disjunctions, with a weaker contribution from bare *some*, and an embedded *some* factor, defined by the NNA and NSA conditions.

The model was estimated using the robust maximum likelihood estimator (MLR). The model showed a good fit to the data ($\chi^2(4) = 5.59$, $p = .23$, $CFI = 0.99$, $TLI = 0.99$, $RMSEA = 0.05$, $SRMR = 0.03$). Factor loadings were strong for all

indicators. Within the disjunction factor, loadings were highest for bare disjunctions and embedded disjunctions, with a somewhat smaller loading for bare *some*. The embedded *some* factor was defined by the two experimental conditions (NNA and NSA). The two latent factors were moderately correlated ($r = 0.25$). The CFA structure is illustrated in Figure B.2.

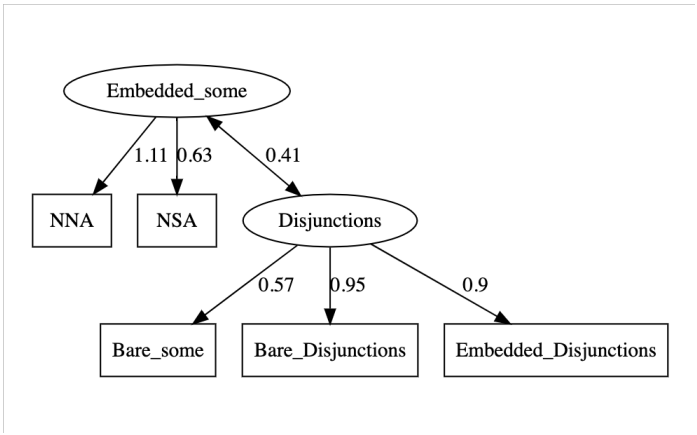


Figure B.2: Diagram of the confirmatory factor analysis. The model specifies two correlated latent factors: one factor defined by bare *some*, bare disjunctions, and embedded disjunctions, and one factor defined by the two embedded *some* conditions (NNA and NSA). Numbers on the arrows indicate standardized factor loadings; the curved arrow indicates the correlation between the two latent factors.

The CFA results should also be interpreted with caution. The model estimation produced a warning indicating a negative residual variance for one observed variable, namely the NNA condition. This constitutes a Heywood case and most likely reflects the combination of a very strong loading and the small number of indicators defining the *Embedded some* factor (Kline, 2011, p. 158). In addition, one factor in the CFA is represented by only two indicators, which is a relatively weak basis for confirmatory modeling. The CFA therefore does not establish the factor structure conclusively, but provides supportive evidence that the two-factor structure suggested by the EFA is compatible with the observed covariance pattern.

Taken together, the exploratory and confirmatory factor analyses suggest that the implicature tasks do not form a single set. Instead, the clearest common structure involves a disjunction-related dimension, defined primarily by bare and embedded disjunctions, and a separate embedded *some* dimension, defined by the NNA and NSA conditions. Bare *some* shows a more moderate association with the disjunction-related factor. By contrast, atypicality inferences and bare numerals were not meaningfully captured by either extracted factor. With respect to the main research question, these results are consistent with the correlational analysis in suggesting that participants' responses in the atypicality inference task do not pattern together with responses in other implicature tasks. At the same time, given the limited number

of implicature measures included in the factor analysis, this conclusion should be interpreted as exploratory support rather than as definitive evidence about the latent organization of pragmatic inferences tested in the present study.

B.3 PCA for the replication analysis presented in Chapter 8

The experiment reported in Chapter 8 provides evidence that the derivation of atypicality inferences may be modulated by a latent Reasoning factor, which combines non-verbal intelligence and cognitive reflection (i.e., the ability to suppress intuitive but incorrect responses).

The data collected in the experiment reported in Chapter 10 provide an opportunity to test the robustness of this finding. In particular, this experiment includes all individual difference measures used in Chapter 8, as well as multiple responses per participant in the atypicality inference task.

To make replication as comparable as possible to the earlier analysis, I conducted a principal component analysis (PCA) using the same subset of measures as in Chapter 8. PCA was performed in R using the `psych` package (Revelle, 2022, version 2.3.6;), following the procedure described by Tanner (2019).

The following measures were included in the PCA: AQ (Autism-Spectrum Quotient; autistic traits), RSpan (Reading Span Task; verbal working memory span), CRT (Cognitive Reflection Test; ability to suppress intuitive but incorrect responses), ART (Author Recognition Test; exposure to print), and Raven's IQ (Raven's Progressive Matrices; non-verbal intelligence). This experiment excluded the measures collected but not included in the previous study, specifically KTT (Keep Track Task; memory updating), RMET (Reading the Mind in the Eyes; theory of mind) and Stroop (Stroop Task; inhibition).

Following the PCA procedure described in Chapter 8, I first fitted an exploratory PCA extracting five components to inspect the distribution of explained variance. Inspection of the cumulative variance plot indicated that four components accounted for 87.6% of the variance in the data (see Figure B.3).

Based on this inspection, a PCA with four components was fitted. The resulting components were rotated using the Varimax method to maximize the variance of squared loadings and obtain orthogonal components that are easier to interpret (Kaiser, 1958).

The results of the four-component PCA are shown in Table B.9. Overall, the structure of the components is comparable to the PCA reported in Chapter 8. In particular, CRT and Raven's IQ load on the same component (Component 1), while the remaining measures each form distinct components. As discussed in Chapter 8, the grouping of CRT and Raven's IQ is consistent with theoretical accounts linking cognitive reflection and fluid intelligence (e.g., Welsh, 2022).

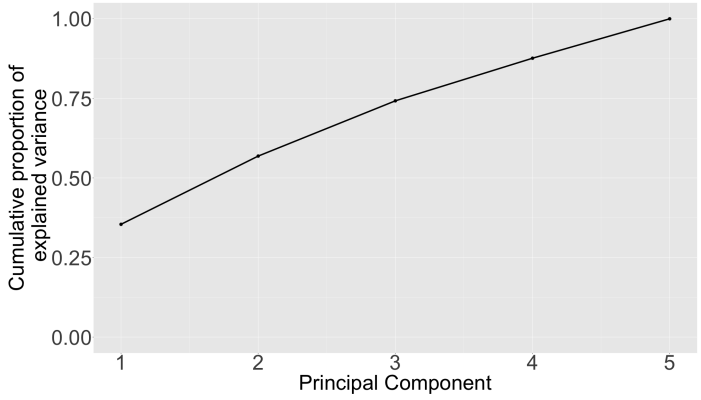


Figure B.3: A scree plot of the cumulative proportion of explained variance in the exploratory PCA.

The resulting components were used in the analysis reported in Section 10.6.5 to test whether the reasoning-related component identified in Chapter 8 replicates in the present dataset.

B.4 Individual differences models for other implicature types

This section contains the model descriptions for bare *some* (Table B.10), bare numerals (Table B.11), disjunctions (Table B.12), and embedded *some* (Table B.13), which are discussed in Section 10.6.5.

Figures B.4 and B.5 also show the observed trial order effect in the models for bare *some* and bare numerals, respectively.

Table B.9: Principal Components Analysis (varimax rotation) of cognitive and personality measures ($N = 156$) for the measures collected in Chapter 10 but for a subset of measures collected in Chapter 8. Standardized component loadings are shown.

Measure	Component 1	Component 2	Component 3	Component 4
AQ (overall)	0.10	**0.99**	-0.03	0.01
RSpan	0.15	0.00	0.01	**0.96**
ART	0.08	-0.04	**0.97**	0.03
CRT	**0.57**	0.11	0.31	0.39
Raven's IQ	**0.94**	0.07	-0.01	0.07

Note: Primary loadings (absolute value $\geq$ .50) are shown in **bold**. **AQ (overall)** = Autism-Spectrum Quotient (overall score; autistic traits); **RSpan** = Reading Span Task (verbal working memory span); **ART** = Author Recognition Test (exposure to print); **CRT** = Cognitive Reflection Test (ability to suppress intuitive but incorrect responses); **Raven's IQ** = Raven's Progressive Matrices (non-verbal intelligence).

Table B.10: Bare *some* model. Effect sizes (b), standard errors (SE), z-values, and p-values for the mixed-effects beta regression model of participants' ratings in the target condition (ratings transformed to fit a beta distribution). The dispersion parameter is 0.92.

	b	SE	z	p
Intercept	0.52	0.09	5.80	<.001
Trial order	-0.12	0.03	-3.36	**<.001**
AQ	-0.08	0.09	-0.83	.41
RMET	0.06	0.09	0.67	.50
Stroop	0.07	0.09	0.76	.45
KTT	-0.01	0.10	-0.12	.91
RSpan	-0.05	0.10	-0.51	.61
ART	0.01	0.10	0.10	.92
CRT	0.17	0.10	1.65	.10
Raven's IQ	-0.05	0.11	-0.43	.67
Random effects	Variance			
Subject	1.05			
Item	0.00			

Table B.11: Bare numerals model. Effect sizes (b), standard errors (SE), z-values, and p-values for the mixed-effects beta regression model of participants' ratings in the target condition (ratings transformed to fit a beta distribution). The dispersion parameter is 0.86.

	b	SE	z	p
Intercept	-0.24	0.09	-2.65	.008
Trial order	0.17	0.04	4.91	**<.001**
AQ	-0.06	0.10	-0.62	.54
RMET	-0.07	0.10	-0.75	.46
Stroop	0.05	0.10	0.52	.61
KTT	0.05	0.11	0.51	.61
RSpan	-0.09	0.10	-0.92	.36
ART	0.06	0.10	0.65	.51
CRT	0.03	0.10	0.31	.76
Raven's IQ	0.19	0.11	1.70	.09
Random effects	Variance			
Subject	1.13			

Table B.12: Disjunctions model. Effect sizes (b), standard errors (SE), z-values, and p-values for the mixed-effects beta regression model of participants' ratings in the target condition (ratings transformed to fit a beta distribution). The dispersion parameter is 1.12.

	b	SE	z	p
Intercept	1.19	0.08	14.09	$<.001$
Inference type (embedded disjunctions)	0.27	0.07	4.17	$<$**.001**
Trial order	-0.06	0.03	-1.73	.08
AQ	-0.07	0.08	-0.87	.38
RMET	-0.11	0.09	-1.30	.19
Stroop	0.01	0.08	0.17	.86
KTT	-0.07	0.09	-0.73	.46
RSpan	-0.06	0.09	-0.71	.48
ART	0.04	0.09	0.41	.68
CRT	0.02	0.09	0.26	.80
Raven's IQ	0.03	0.10	0.30	.76
Inference type (embedded disjunctions) : Trial order	-0.05	0.07	-0.80	.42
Inference type (embedded disjunctions) : AQ	0.02	0.07	0.24	.81
Inference type (embedded disjunctions) : RMET	0.06	0.07	0.80	.42
Inference type (embedded disjunctions) : Stroop	0.03	0.07	0.50	.61
Inference type (embedded disjunctions) : KTT	-0.08	0.08	-1.01	.31
Inference type (embedded disjunctions) : RSpan	0.12	0.07	1.69	.09
Inference type (embedded disjunctions) : ART	0.00	0.07	0.05	.96
Inference type (embedded disjunctions) : CRT	-0.05	0.07	-0.68	.50
Inference type (embedded disjunctions) : Raven's IQ	0.00	0.08	0.04	.97
Random effects	Variance			
Subject	0.83			

Table B.13: Embedded *some* model. Effect sizes (b), standard errors (SE), z-values, and p-values for the mixed-effects beta regression model of participants' ratings (ratings transformed to fit a beta distribution).The dispersion parameter is 0.605.

	b	SE	z	p
Intercept	0.50	0.07	7.48	<.001
Condition (NNA)	0.20	0.07	2.82	**.005**
Trial order	-0.06	0.04	-1.59	.113
AQ	-0.04	0.07	-0.51	.609
RMET	-0.11	0.07	-1.51	.130
Stroop	-0.04	0.07	-0.63	.527
KTT	-0.13	0.08	-1.69	.090
RSpan	0.00	0.07	0.04	.970
ART	0.28	0.07	3.99	**<.001**
CRT	0.01	0.07	0.18	.856
Raven's IQ	-0.01	0.08	-0.12	.906
Condition (NNA) : Trial order	-0.11	0.07	-1.54	.123
Condition (NNA) : AQ	0.05	0.08	0.64	.525
Condition (NNA) : RMET	-0.06	0.08	-0.78	.438
Condition (NNA) : Stroop	-0.04	0.08	-0.55	.585
Condition (NNA) : KTT	0.01	0.08	0.06	.951
Condition (NNA) : RSpan	-0.05	0.08	-0.58	.562
Condition (NNA) : ART	-0.07	0.08	-0.94	.349
Condition (NNA) : CRT	0.06	0.08	0.70	.487
Condition (NNA) : Raven's IQ	-0.01	0.09	-0.08	.940
Random effects	Variance			
Subject	0.47			

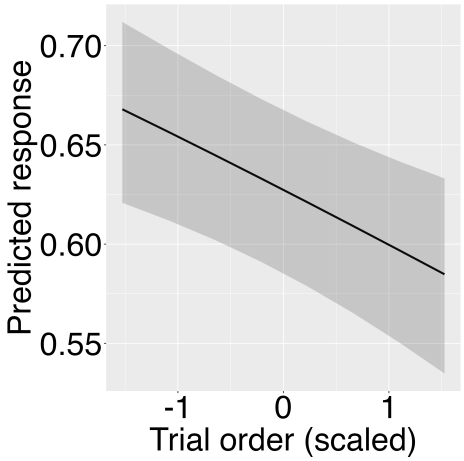


Figure B.4: Bare *some*. Predicted response as a function of trial order (scaled) based on the fitted model. The solid line represents the model-predicted values, and the shaded area indicates the 95% confidence interval.

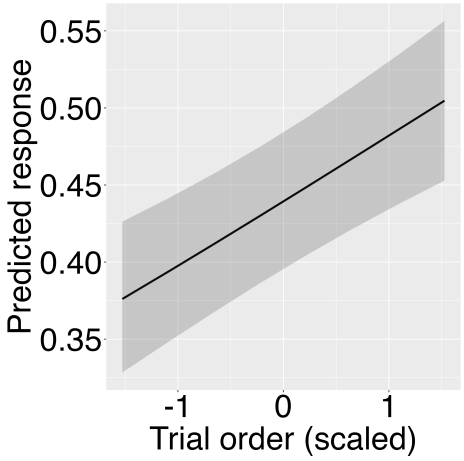


Figure B.5: Bare numerals. Predicted response as a function of trial order (scaled) based on the fitted model. The solid line represents the model-predicted values, and the shaded area indicates the 95% confidence interval.

Bibliography

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist*, *36*(7), 715–729. doi:10.1037/0003-066X.36.7.715.

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(1), 278–289. doi:10.3758/BRM.40.1.278.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV*. (4th ed.). Washington, DC: American Psychiatric Association. Available via PsychiatryOnline (APA Publishing): `https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890420614.dsm-iv`.

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, *99*, 78–95. doi:10.1016/j.pragma.2016.05.001.

Arnold, J. E., Strangmann, I. M., Hwang, H., Zerkle, S., & Nappa, R. (2018). Linguistic experience affects pronoun interpretation. *Journal of Memory and Language*, *102*, 41–54. doi:10.1016/j.jml.2018.05.002.

Austin, E. J. (2005). Personality correlates of the broader autism phenotype as assessed by the Autism Spectrum Quotient (AQ). *Personality and Individual Differences*, *38*(2), 451–460. doi:10.1016/j.paid.2004.04.022.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. (1st ed.). Psychology Press. doi:10.4324/9781315749860.

Baddeley, A. D., Bressi, S., Della Sala, S., Logie, R., & Spinnler, H. (1991). The decline of working memory in Alzheimer's disease: A longitudinal study. *Brain*, *114*(6), 2521–2542. doi:10.1093/brain/114.6.2521.

Baker, R., Gill, A., & Cassell, J. (2008). Reactive redundancy and listener comprehension in direction-giving. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 37–45). doi:10.3115/1622064.1622071.

Bambini, V., Arcara, G., Bechi, M., Buonocore, M., Cavallaro, R., & Bosia, M. (2016). The communicative impairment as a core feature of schizophrenia: Frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Comprehensive Psychiatry*, *71*, 106–120. doi:10.1016/j.comppsych.2016.08.012.

Bambini, V., & Domaneschi, F. (2024). Twenty years of experimental pragmatics. New advances in scalar implicature and metaphor processing. *Cognition*, *244*, 105708. doi:10.1016/j.cognition.2023.105708.

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (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. doi:10.1016/j.jarmac.2014.09.003.

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

Baron-Cohen, S. (1997). Are children with autism superior at folk physics? *New Directions for Child and Adolescent Development*, *75*, 45–54. doi:10.1002/cd.23219977504.

Baron-Cohen, S., Bowen, D. C., Holt, R. J., Allison, C., Auyeung, B., Lombardo, M. V., Smith, P., & Lai, M.-C. (2015). The "Reading the mind in the eyes" test: complete absence of typical sex difference in˜ 400 men and women with autism. *PloS one*, *10*(8), e0136521. doi:10.1371/journal.pone.0136521.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "Theory of Mind"? *Cognition*, *21*(1), 37–46. doi:10.1016/0010-0277(85)90022-8.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001a). 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. doi:10.1017/S0021963001006643.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001b). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. doi:10.1023/A:1005653411471.

Barr, A. (1981). Frames and scripts. In A. Barr, & E. A. Feigenbaum (Eds.), *The Handbook of Artificial Intelligence (Vol. 1)* (pp. 216–222). William Kaufmann, Inc. doi:10.1016/B978-0-86576-089-9.50008-9.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001.

Bartoń, K. (2024). MuMIn: Multi-Model Inference. R package version 1.48.4, https://CRAN.R-project.org/package=MuMIn.

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

Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 12–23. doi:10.1037/0096-1523.14.1.12.

Bila, N. (2026). *Theory of Mind Subcomponents and their Correlation with Pragmatic Tasks*. Master's thesis Saarland University, Saarbrücken, Germany.

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. doi:10.1177/1073191112446655.

Bjorklund, D. F., & Harnishfeger, K. K. (1995). The evolution of inhibition mechanisms and their role in human cognition and behavior. In F. N. Dempster, C. J. Brainerd, & C. J. Brainerd (Eds.), *Interference and Inhibition in Cognition* (pp. 141–173). San Diego: Academic Press. doi:10.1016/B978-012208930-5/50006-4.

Bonnefon, J.-F., Feeney, A., & De Neys, W. (2011). The risk of polite misunderstandings. *Current Directions in Psychological Science*, *20*(5), 321–324. doi:10.1177/0963721411418472.

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When *some* is actually *all*: Scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258. doi:10.1016/j.cognition.2009.05.005.

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. doi:10.1111/j.1467-9280.2006.01776.x.

Bosco, F. M., Bono, A., & Bara, B. G. (2012). Recognition and repair of communicative failures: The interaction between theory of mind and cognitive complexity in schizophrenic patients. *Journal of Communication Disorders*, *45*(3), 181–197. doi:10.1016/j.jcomdis.2012.01.005.

Bosco, F. M., Tirassa, M., & Gabbatore, I. (2018). Why pragmatics and theory of mind do not (completely) overlap. *Frontiers in Psychology*, *9*, 1453. doi:10.3389/fpsyg.2018.01453.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123–142. doi:10.1016/j.jml.2011.09.005.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457. doi:10.1016/j.jml.2004.05.006.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*(2), 177–220. doi:10.1016/0010-0285(79)90009-4.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378–400. doi:10.32614/RJ-2017-066.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*(4), 441–472. doi:10.1016/0010-0285(87)90015-6.

Burgoyne, A. P., Frank, D. J., & Macnamara, B. N. (2025). Which "working memory" are we talking about? Complex span tasks versus N-back. *Psychonomic Bulletin & Review*, *32*(3), 1337–1351. doi:10.3758/s13423-024-02622-0.

Butterfuss, R., & Kendeou, P. (2018). The role of executive functions in reading comprehension. *Educational Psychology Review*, *30*(3), 801–826. doi:10.1007/s10648-017-9422-6.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. doi:10.1037/0022-3514.42.1.116.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making*, *5*(3), 182–191. doi:10.1017/S1930297500001066.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(1), 77–94. doi:10.1017/S0140525X99001788.

Carlson, S. M., Koenig, M. A., & Harms, M. B. (2013). Theory of mind. *WIREs Cognitive Science*, *4*(4), 391–402. doi:10.1002/wcs.1232.

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension

difficulties: A meta-analysis. *Learning and Individual Differences*, *19*(2), 246–251. doi:10.1016/j.lindif.2008.10.002.

Carriedo, N., Corral, A., Montoro, P. R., Herrero, L., Ballestrino, P., & Sebastián, I. (2016). The development of metaphor comprehension and its relationship with relational verbal reasoning and executive function. *PLoS One*, *11*(3), e0150289. doi:10.1371/journal.pone.0150289.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. doi:10.1017/CBO9780511571312.

Castano, E., Paladino, M. P., Cadwell, O. G., Cuccio, V., & Perconti, P. (2021). Exposure to literary fiction is associated with lower psychological essentialism. *Frontiers in Psychology*, *12*, 662940. doi:10.3389/fpsyg.2021.662940.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. doi:10.1037/h0046743.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30(3)*, 687–696. doi:10.1037/0278-7393.30.3.687.

Champagne-Lavau, M., & Stip, E. (2010). Pragmatic and executive dysfunction in schizophrenia. *Journal of Neurolinguistics*, *23*(3), 285–296. doi:10.1016/j.jneuroling.2009.08.009.

Chiappe, D. L., & Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*, *56*(2), 172–188. doi:10.1016/j.jml.2006.11.006.

Chierchia, G. (2004). Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface. In A. Belletti (Ed.), *Structures and Beyond: The Cartography of Syntactic Structures, Volume 3* (pp. 39–103). New York, NY: Oxford University Press. doi:10.1093/oso/9780195171976.003.0003.

Cho, J. (2020). Memory Load Effect in the Real-Time Processing of Scalar Implicatures. *Journal of Psycholinguistic Research*, *49*, 865–884. doi:10.1007/s10936-020-09726-3.

Chwilla, D. J., & Kolk, H. H. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research*, *25*(3), 589–606. doi:10.1016/j.cogbrainres.2005.08.011.

Çiftlikli, S., & Demirel, Ö. (2022). The relationships between students' comprehension of conversational implicatures and their achievement in reading comprehension. *Frontiers in Psychology*, *13*, 977129. doi:10.3389/fpsyg.2022.977129.

Cipielewski, J., & Stanovich, K. E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology*, *54*(1), 74–89. doi:10.1016/0022-0965(92)90018-2.

Clark, H. H. (1990). [quantifying probabilistic expressions]: Comment. *Statistical Science*, *5*(1), 12–16. doi:10.1214/ss/1177012242.

Conway, A. R., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, *123*(4), 354–373. doi:10.1037/0096-3445.123.4.354.

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. doi:10.3758/BF03196772.

Cooper, R. P. (2010). Cognitive Control: Componential or Emergent? *Topics in Cognitive Science*, *2*(4), 598–613. doi:10.1111/j.1756-8765.2010.01110.x.

Cornoldi, C., Pra Baldi, A., & Rizzo, M. (1991). Prove Avanzate MT di Comprensione della Lettura [Advanced MT reading comprehension tests]. Florence, Italy: Organizzazioni Speciali. URL: `https://phaidra.cab.unipd.it/o:428441`.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, *26*, 197–223. doi:10.1007/s10648-013-9246-y.

Cowley, S. H., Pearson, L., & Barner, D. (2025). The relationships between students' comprehension of conversational implicatures and their achievement in reading comprehension. *Journal of Experimental Psychology: Learning, Memory, and cognition*, *51*(11), 1837–1850. doi:10.1037/xlm0001479.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, e28. doi:10.1017/S0140525X19001730.

Davies, C., & Katsos, N. (2010). Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua*, *120*(8), 1956–1972. doi:10.1016/j.lingua.2010.02.005.

Davies, C., & Katsos, N. (2013). Are speakers and listeners 'only moderately gricean'? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics*, *49*(1), 78–106. doi:10.1016/j.pragma.2013.01.004.

De Neys, W. (2006). Dual Processing in Reasoning: Two Systems but One Reasoner. *Psychological Science*, *17*(5), 428–433. doi:10.1111/j.1467-9280.2006.01723.x.

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. doi:10.1027/1618-3169.54.2.128.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710. doi:10.1111/cogs.12171.

Degen, J., & Tanenhaus, M. K. (2019). Constraint-based pragmatic processing. In C. Cummins, & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics*. Oxford University Press. doi:10.1093/oxfordhb/9780198791768.013.8.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*, 548–553. URL: https://escholarship.org/uc/item/9wn4w9zk.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367. doi:10.1080/17470218.2011.588799.

Dimitrov, M., Nakic, M., Elpern-Waxman, J., Granetz, J., O'Grady, J., Phipps, M., Milne, E., Logan, G. D., Hasher, L., & Grafman, J. (2003). Inhibitory attentional control in patients with frontal lobe damage. *Brain and Cognition*, *52*(2), 258–270. doi:10.1016/S0278-2626(03)00080-0.

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. URL: https://escholarship.org/uc/item/6mw6t8rc.

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), 170–189. doi:10.1037/a0017891.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of quantity? *Journal of Memory and Language*, *54*(4), 554–573. doi:10.1016/j.jml.2005.12.009.

Engle, R. W., Conway, A. R., Tuholski, S. W., & Shisler, R. J. (1995). A resource Account of Inhibition. *Psychological Science*, *6*(2), 122–125. doi:10.1111/j.1467-9280.1995.tb00318.x.

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), 309–331. doi:10.1037/0096-3445.128.3.309.

Engonopulos, N., Sayeed, A., & Demberg, V. (2013). Language and cognitive load in a dual task environment. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 2148–2153. URL: https://escholarship.org/uc/item/8p586904.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149. doi:10.3758/BF03203267.

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. doi:10.1177/1745691612460685.

Evans, J. S. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. doi:10.1080/13546783.2019.1623071.

Fairchild, S., Mathis, A., & Papafragou, A. (2020). Pragmatics and social meaning: Understanding under-informativeness in native and non-native speakers. *Cognition*, *200*, 104171. doi:10.1016/j.cognition.2019.104171.

Fairchild, S., & Papafragou, A. (2017). Flexible expectations of speaker informativeness shape pragmatic inference. *University of Pennsylvania Working Papers in Linguistics*, *23*(1), 7.

Fairchild, S., & Papafragou, A. (2021). The Role of Executive Function and Theory of Mind in Pragmatic Computations. *Cognitive Science*, *45*(2), e12938. doi:10.1111/cogs.12938.

Farmer, T. A., Fine, A. B., Misyak, J. B., & Christiansen, M. H. (2017). Reading Span Task Performance, Linguistic Experience, and the Processing of Unexpected Syntactic Events. *The Quarterly Journal of Experimental Psychology*, *70*(3), 413–433. doi:10.1080/17470218.2015.1131310.

Feeney, A., & Bonnefon, J.-F. (2012). Politeness and honesty contribute additively to the interpretation of scalar expressions. *Journal of Language and Social Psychology*, *32*(2), 181–190. doi:10.1177/0261927X12456840.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of *some*: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(2), 121–132. doi:10.1037/h0085792.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. doi:10.1080/0266476042000214501.

Fleva, E., Tsimpli, I. M., Fotiadou, G., & Katsiperi, M. (2017). The effect of print exposure upon performance on the raven progressive matrices test. *Selected papers on theoretical and applied linguistics*, *22*, 133–145.

Floyd, S., Jouravlev, O., Poliak, M., Mineroff, Z., Gibson, E., & Fedorenko, E. (2025). Three distinct components of pragmatic language use: Social conventions, intonation, and world knowledge-based causal reasoning. *Proceedings of the National Academy of Sciences*, *122*(50), e2424400122. doi:10.1073/pnas.2424400122.

Foppolo, F. (2007). Between "Cost" and "Default": A new approach to Scalar Implicature. In R. Artstein, & L. Vieu (Eds.), *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 137–150). Rovereto, Italy. URL: http://semdial.org/anthology/semdial2007_decalog_front_matter.pdf.

Foppolo, F., Mazzaggio, G., Panzeri, F., & Surian, L. (2021). Scalar and ad-hoc pragmatic inferences in children: guess which one is easier. *Journal of Child Language*, *48*(2), 350–372. doi:10.1017/S030500092000032X.

Fox, J. (2022). *RcmdrMisc: R Commander Miscellaneous Functions*. URL: https://CRAN.R-project.org/package=RcmdrMisc R package version 2.7-2.

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., & Potts, C. (2016). Rational speech act models of pragmatic reasoning in reference games. *PsyArXiv*. doi:10.31234/osf.io/f9y6b.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. doi:10.1257/089533005775196732.

Freed, E. M., Hamilton, S. T., & Long, D. L. (2017). Comprehension in proficient readers: The nature of individual variation. *Journal of Memory and Language*, *97*, 135–153. doi:10.1016/j.jml.2017.07.008.

Friedman, N. P., & Miyake, A. (2004). The Relations Among Inhibition and Interference Control Functions: A Latent-Variable Analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135. doi:10.1037/0096-3445.133.1.101.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590. doi:10.3758/BF03192728.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*(2), 172–179. doi:10.1111/j.1467-9280.2006.01681.x.

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—-225. doi:10.1037/0096-3445.137.2.201.

Frith, C. D., & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine*, *26*(3), 521–530. doi:10.1017/s0033291700035601.

Garmendia, J. (2023). Lies we don't say: Figurative language, commitment, and deniability. *Journal of Pragmatics*, *218*, 183–194. doi:10.1016/j.pragma.2023.11.003.

Gelman, A. (2018, March 15). You need 16 times the sample size to estimate an interaction than to estimate a main effect [blog post]. URL: https://statmodeling.stat.columbia.edu/2018/03/15/need16/.

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press. doi:10.1017/CBO9780511975158.

Goldman, A. I. (2012). Theory of Mind. In E. Margolis, R. Samuels, & S. P. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 402–424). Oxford University Press. doi:10.1093/oxfordhb/9780195309799.013.0017.

Graesser, A. C., Gordon, S. E., & Sawyer, J. D. (1979). Recognition memory for typical and atypical actions in scripted activities: Tests of a script pointer + Tag hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *18*(3), 319–332. doi:10.1016/S0022-5371(79)90182-8.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*, 163–189. doi:10.1146/annurev.psych.48.1.163.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. doi:10.1111/2041-210X.12504.

Grice, H. P. (1975). Logic and Conversation. In P. Cole, & J. L. Morgan (Eds.), *Speech Acts, Volume 3 of Syntax and Semantics* (pp. 41–58). Brill. doi:10.1163/9789004368811_003.

Grigoroglou, M., & Papafragou, A. (2016). Are children flexible speakers? Effects of typicality and listener needs in children's event descriptions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *38*, 782–787. URL: https://escholarship.org/uc/item/3s13v7sg.

Grigoroglou, M., & Papafragou, A. (2019). Children's (and Adults') Production Adjustments to Generic and Particular Listener Needs. *Cognitive Science*, *43*(10), e12790. doi:10.1111/cogs.12790.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55. doi:10.1016/j.cognition.2010.03.014.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science (New York, N.Y.)*, *304(5669)*, 438–441. doi:10.1126/science.1095455.

Hanaki, N., Jacquemet, N., Luchini, S., & Zylbersztejn, A. (2016). Fluid intelligence and cognitive reflection in a strategic environment: Evidence from dominance-solvable games. *Frontiers in Psychology*, *7*, 1188. doi:10.3389/fpsyg.2016.01188.

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. doi:10.1007/BF02172093.

Harnishfeger, K. K. (1995). The development of cognitive inhibition: Theories, definitions, and research evidence. In C. J. B. Frank N. Dempster, & C. J. Brainerd (Eds.), *Interference and Inhibition in Cognition* (pp. 175–204). Academic Press. doi:10.1016/B978-012208930-5/50007-6.

Harris, R. J., Lee, D. J., Hensley, D. L., & Schoen, L. M. (1988). The effect of cultural script knowledge on memory for stories over time. *Discourse Processes*, *11*(4), 413–431. doi:10.1080/01638538809544711.

Hasher, L., Lustig, C., & Zacks, R. (2007). Inhibitory mechanisms and the control of attention. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in Working Memory* (pp. 227–249). Oxford Academic. doi:10.1093/acprof:oso/9780195168648.003.0009.

Heyman, T., & Schaeken, W. (2015). Some differences in *some*: Examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, *55*(1), 1–18. doi:10.5334/pb.bc.

Holding, D. H. (1989). Counting backward during chess move choice. *Bulletin of the Psychonomic Society*, *27*(5), 421–424. doi:10.3758/bf03334644.

Holford, D. L., Juanchich, M., & Sirota, M. (2022). Characteristics of quantifiers moderate the framing effect. *Journal of Behavioral Decision Making*, *35*(1), e2251. doi:10.1002/bdm.2251.

Holtgraves, T., & Kraus, B. (2018). Processing scalar implicatures in conversational contexts: An ERP study. *Journal of Neurolinguistics*, *46*, 93–108. doi:10.1016/j.jneuroling.2017.12.008.

Hong, X., Ryzhova, M., Biondi, D., & Demberg, V. (2024). Do large language models and humans have similar behaviours in causal inference with script knowledge? In D. Bollegala, & V. Shwartz (Eds.), *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)* (pp. 421–437). Mexico City,

Mexico: Association for Computational Linguistics. doi:10.18653/v1/2024.starsem-1.34.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*, 107–129. doi:10.1016/0001-6918(67)90011-X.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Doctoral dissertation, University of California, Los Angeles, US. URL: https://linguistics.ucla.edu/ph-d-recipients/ Retrieved from linguistics.ucla.edu.

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 (GURT'84)* (pp. 11–42). Washington, DC.: Georgetown University Press. URL: https://repository.digital.georgetown.edu/handle/10822/555477 Retrieved from repository.digital.georgetown.edu.

Horn, L. R. (1991). Given as new: When redundant affirmation isn't. *Journal of Pragmatics*, *15*(4), 313–336. doi:10.1016/0378-2166(91)90034-U.

Horn, L. R. (1992). The said and the unsaid. In C. Baker, & D. Dowty (Eds.), *Proceedings from the Second Conference on Semantics and Linguistic Theory (SALT II)* (pp. 163–192). Ohio State University. doi:10.3765/salt.v2i0.3039.

Horn, L. R. (1993). Economy and redundancy in a dualistic model of natural language. *Finnish Journal of Linguistics*, (6), 33–72. URL: https://journal.fi/finjol/article/view/152740. Retrieved from journal.fi.

Horn, L. R. (2014). Information Structure and the Landscape of (Non-) at-issue Meaning. In C. Féry, & S. Ishihara (Eds.), *The Oxford handbook of information structure* (pp. 108–127). Oxford University Press. doi:10.1093/oxfordhb/9780199642670.013.009.

Huang, Y. (2012). *The Oxford dictionary of pragmatics*. Oxford, UK: Oxford University Press.

Huang, Y. (2014). *Pragmatics*. Oxford University Press, USA.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415. doi:10.1016/j.cogpsych.2008.09.001.

Hurst, R. M., Mitchell, J. T., Kimbrel, N. A., Kwapil, T. R., & Nelson-Gray, R. O. (2007). Examination of the reliability and factor structure of the Autism Spectrum Quotient (AQ) in a non-clinical sample. *Personality and Individual Differences*, *43*(7), 1938–1949. doi:10.1016/j.paid.2007.06.012.

Inohara, K., Ueda, A., Shioya, K., & Osanai, H. (2017). The effects of reading amount on letter reading skill: A longitudinal survey of japanese elementary schoolchildren. *Psychologia*, *60*(2), 85–96. doi:10.2117/psysoc.2017.85.

Johnson, E., & Arnold, J. E. (2021). Individual differences in print exposure predict use of implicit causality in pronoun comprehension and referential prediction. *Frontiers in Psychology*, *12:672109*. doi:10.3389/fpsyg.2021.672109.

Just, M. A., Carpenter, P. A., & Keller, T. A. (1996). The Capacity Theory of Comprehension: New Frontiers of Evidence and Arguments. *Psychological Review*, *103*(4), 773–780. doi:10.1037/0033-295X.103.4.773.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (p. 49–81). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511808098.004.

Kaiser, H. F. (1958). The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200. doi:10.1007/BF02289233.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*(1), 47–70. doi:10.1037/0096-3445.132.1.47.

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), 189–217. doi:10.1037/0096-3445.133.2.189.

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

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and Coreference Revisited. *Journal of Semantics*, *25*(1), 1–44. doi:10.1093/jos/ffm018.

Khorsheed, A., Price, J., & van Tiel, B. (2022). Sources of cognitive cost in scalar implicature processing: A review. *Frontiers in Communication*, *7*, 1–14. doi:10.3389/fcomm.2022.990044.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*(6156), 377–380. doi:10.1126/science.1239918.

Kim, S., Kim, J. S., Jin, M. J., Im, C.-H., & Lee, S.-H. (2018). Dysfunctional frontal lobe activity during inhibitory tasks in individuals with childhood

trauma: An event-related potential study. *NeuroImage: Clinical*, *17*, 935–942. doi:10.1016/j.nicl.2017.12.034.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. (3rd ed.). The Guilford Press. ISBN 978-1-60623-877-6.

Kloosterman, P. H., Keefer, K. V., Kelley, E. A., Summerfeldt, L. J., & Parker, J. D. (2011). Evaluation of the factor structure of the Autism-Spectrum Quotient. *Personality and Individual Differences*, *50*(2), 310–314. doi:10.1016/j.paid.2010.10.015.

Kravtchenko, E. (2022). *Integrating pragmatic reasoning in an efficiency-based theory of utterance choice*. Doctoral dissertation, Saarland University, Saarbrücken, Germany. doi:10.22028/D291-35858.

Kravtchenko, E., & Demberg, V. (2015). Semantically underinformative utterances trigger pragmatic inferences. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*, 1207–1212. URL: https://escholarship.org/uc/item/3f48n0tr.

Kravtchenko, E., & Demberg, V. (2022a). Informationally redundant utterances elicit pragmatic inferences. *Cognition*, *225*, 105159. doi:10.1016/j.cognition.2022.105159.

Kravtchenko, E., & Demberg, V. (2022b). Modeling atypicality inferences in pragmatic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 1918–1924. URL: https://escholarship.org/uc/item/7630p08b.

Kurch, C., Ryzhova, M., & Demberg, V. (2024). Large language models fail to derive atypicality inferences in a human-like manner. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, & Y. Oseki (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 86–100). Bangkok, Thailand: Association for Computational Linguistics. doi:10.18653/v1/2024.cmcl-1.8.

Kursat, L., & Degen, J. (2020). Probability and processing speed of scalar inferences is context-dependent. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*, 1236–1242. URL: https://escholarship.org/uc/item/3rt8763c.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–23. doi:10.18637/jss.v082.i13.

Kyllonen, P., & Kell, H. (2017). What is fluid intelligence? Can it be improved? In M. Rosén, K. Yang Hansen, & U. Wolff (Eds.), *Cognitive Abilities and Educational Outcomes: A Festschrift in Honour of Jan-Eric Gustafsson* (pp. 15–37). Cham: Springer International Publishing. doi:10.1007/978-3-319-43473-5$_2$.

Levinson, S. C. (1987). Pragmatics and the grammar of anaphora: a partial pragmatic reduction of Binding and Control phenomena1. *Journal of Linguistics*, *23*(2), 379–434. doi:10.1017/S0022226700011324.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT press. doi:10.7551/mitpress/5526.001.0001.

Li, M., Venhuizen, N. J., Jachmann, T. K., Drenhaus, H., & Crocker, M. W. (2023). Does informativity modulate linearization preferences in reference production? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*. URL: https://escholarship.org/uc/item/95v6j0sx.

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. doi:10.1016/j.jesp.2009.12.019.

Linares, R., & Pelegrina, S. (2023). The relationship between working memory updating components and reading comprehension. *Cognitive Processing*, *24*, 253–265. doi:10.1007/s10339-023-01127-3.

Liu, G., & Demberg, V. (2026). Modeling individual and item differences in atypicality inferences with ACT-R. Manuscript submitted for publication.

Liu, X., Stine-Morrow, E. A. L., & McCall, G. S. (2019). The role of print exposure in supporting cognitive ability among older adults. *Innovation in Aging*, *3*(Supplement 1), S651. doi:10.1093/geroni/igz038.2415.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review*, *9*(3), 550–557. doi:10.3758/BF03196312.

Lodi-Smith, J., Rodgers, J. D., Marquez Luna, V., Khan, S., Long, C. J., Kozlowski, K. F., Donnelly, J. P., Lopata, C., & Thomeer, M. L. (2021). The Relationship of Age with the Autism-Spectrum Quotient Scale in a Large Sample of Adults. *Autism in Adulthood*, *3*(2), 147–156. doi:10.1089/aut.2020.0010.

MacLeod, C. M. (2007). The concept of inhibition in cognition. In D. S. Gorfein, & C. M. MacLeod (Eds.), *Inhibition in Cognition* (pp. 3–23). American Psychological Association. doi:10.1037/11587-001.

Mahr, A., Feld, M., Moniri, M. M., & Math, R. (2012). The ConTRe (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In A. L. Kun, L. N. Boyle, B. Reimer, & A. Riener (Eds.), *Adjunct Proceedings of the 4th International Conference on Automotive*

User Interfaces and Interactive Vehicular Applications (AutomotiveUI-12) (pp. 88–91). Portsmouth, New Hampshire, USA: ACM Digital Library. URL: https://www.auto-ui.org/12/proceedings.php.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. URL: https://joss.theoj.org/papers/10.21105/joss.01541. doi:10.21105/joss.01541.

Martin-Chang, S. L., & Gould, O. N. (2008). Revisiting print exposure: Exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*(3), 273–284. doi:10.1111/j.1467-9817.2008.00371.x.

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163. doi:10.1016/j.lingua.2013.03.006.

Marty, P. P., & Chemla, E. (2013). Scalar implicatures: working memory and a comparison with *only*. *Frontiers in Psychology*, *4*, Article 403. doi:10.3389/fpsyg.2013.00403.

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition*, *27*(5), 759–767. doi:10.3758/bf03198529.

Mayn, A., & Demberg, V. (2022). Individual differences in a pragmatic reference game. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 3016–3022. URL: https://escholarship.org/uc/item/0hb5h0hk.

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. doi:10.1162/opmi_a_00077.

Mayn, A., & Demberg, V. (2026). Sources of individual variability in a pragmatic reference game: Effects of logical reasoning and Theory of Mind. *PLoS ONE*, *21*(2), e0339899. doi:10.1371/journal.pone.0339899.

Mazzaggio, G., Reboul, A., Henst, J.-B. v. d., Cheylus, A., Lorusso, P., & Stateva, P. (2022). On the cost of scalar implicatures: an eye-tracking study. *AMLaP 2022 [also] Architectures and Mechanisms for Language Processing*, (pp. 1–2).

Mazzaggio, G., & Surian, L. (2018). A diminished propensity to compute scalar implicatures is linked to autistic traits. *Acta Linguistica Academica*, *65*(4), 651–668. doi:10.1556/2062.2018.65.4.4.

Mazzarella, D. (2014). Is inference necessary to pragmatics? *Belgian Journal of Linguistics*, *28*(1), 71–95. doi:10.1075/bjl.28.04maz.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of Computer Vision: Computer Science Series* (pp. 211–267). New York, USA: McGraw-Hill. ISBN: 0-07-071048-1.

Mitchell, M., Reiter, E., & Van Deemter, K. (2013). Typicality and Object Reference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 3062–3067. URL: https://escholarship.org/uc/item/9mt2r0rb.

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. doi:10.1006/cogp.1999.0734.

Mo, S., Su, Y., Chan, R. C., & Liu, J. (2008). Comprehension of metaphor and irony in schizophrenia during remission: The role of theory of mind and IQ. *Psychiatry Research*, *157*(1-3), 21–29. doi:10.1016/j.psychres.2006.04.002.

Modi, A. (2017). *Modeling Common Sense Knowledge via Scripts*. Doctoral dissertation, Saarland University, Saarbrücken, Germany. doi:10.22028/D291-26779.

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology*, *81*(2), 111–121. doi:10.1111/j.2044-8295.1990.tb02349.x.

Morrow, D. G., Greenspan, S. L., & Bower, G. H. (1987). Accessibility and situation models in narrative comprehension. *Journal of Memory and Language*, *26*(2), 165–187. doi:10.1016/0749-596X(87)90122-7.

Muijselaar, M. M., & de Jong, P. F. (2015). The effects of updating ability and knowledge of reading strategies on reading comprehension. *Learning and Individual Differences*, *43*, 111–117. doi:10.1016/j.lindif.2015.08.011.

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63*(3), 324–346. doi:10.1016/j.jml.2010.06.005.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi:10.1037/0033-295X.84.3.231.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. doi:10.1016/S0010-0277(00)00114-1.

Noveck, I. A., Bianco, M., & Castry, A. (2001). The costs and benefits of metaphor. *Metaphor and Symbol*, *16*(1-2), 109–121. doi:10.1080/10926488.2001.9678889.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, *85*(2), 203–210. doi:10.1016/S0093-934X(03)00053-1.

Ostermann, S. (2020). *Script Knowledge for Natural Language Understanding*. Doctoral dissertation, Saarland University, Saarbrücken, Germany. doi:10.22028/D291-31301.

Pagliarini, E., Bill, C., Romoli, J., Tieu, L., & Crain, S. (2018). On children's variable success with scalar inferences: Insights from disjunction in the scope of a universal quantifier. *Cognition*, *178*, 178–192. doi:10.1016/j.cognition.2018.04.020.

Palladino, P., Cornoldi, C., De Beni, R., & Pazzaglia, F. (2001). Working memory and updating processes in reading comprehension. *Memory & Cognition*, *29*, 344–354. doi:10.3758/BF03194929.

Panizza, D., Chierchia, G., & Clifton Jr, C. (2009). On the role of entailment patterns and scalar implicatures in the processing of numerals. *Journal of Memory and Language*, *61*(4), 503–518. doi:10.1016/j.jml.2009.07.005.

Parola, A., Berardinelli, L., & Bosco, F. M. (2018). Cognitive abilities and theory of mind in explaining communicative-pragmatic disorders in patients with schizophrenia. *Psychiatry research*, *260*, 144–151. doi:10.1016/j.psychres.2017.11.051.

Parola, A., & Bosco, F. M. (2022). An eye-tracking investigation of the cognitive processes involved in the comprehension of simple and complex communicative acts. *Quarterly Journal of Experimental Psychology*, *75*(10), 1976–1995. doi:10.1177/17470218221079629.

Payne, B. R., Gao, X., Noh, S. R., Anderson, C. J., & Stine-Morrow, E. A. (2012). The effects of print exposure on sentence processing and memory in older adults: Evidence for efficiency and reserve. *Aging, Neuropsychology, and Cognition*, *19*(1-2), 122–149. doi:10.1080/13825585.2011.628376.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110. doi:10.1515/ling.1989.27.1.89.

Peng, P., Barnes, M., Wang, C.-C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76. doi:10.1037/bul0000124.

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*, 341–348. doi:10.3758/s13428-015-0576-1.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions. *Journal of Research in Personality*, *8*(1), 95–101. doi:10.1016/0092-6566(74)90049-X.

Perret, E. (1974). The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour. *Neuropsychologia*, *12*(3), 323–330. doi:10.1016/0028-3932(74)90047-5.

Plous, S. (1993). *The Psychology of Judgment and Decision Making*. New York: McGraw-Hill Book Company. ISBN-10: 0070504776.

Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, *33*(4), 755–802. doi:10.1093/jos/ffv012.

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. doi:10.1002/bdm.1883.

Pype, A., Lin, J., Murray, S., & Boynton, G. (2010). Individual differences in the shape of visual attention during object tracking. *Journal of Vision*, *10*(7), 315–315. doi:10.1167/10.7.315.

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), 384–396. doi:10.1177/1745691619896607.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `https://www.R-project.org/`.

Radvansky, G. A., & Copeland, D. E. (2001). Working memory and situation model updating. *Memory & Cognition*, *29*(8), 1073–1080. doi:10.3758/BF03206375.

Radvansky, G. A., Tamplin, A. K., Armendarez, J., & Thompson, A. N. (2014). Different kinds of causality in event cognition. *Discourse Processes*, *51*(7), 601–618. doi:10.1080/0163853X.2014.903366.

Ramotowska, S. (2022). *Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences*. Doctoral dissertation, University of Amsterdam. URL: `https://hdl.handle.net/11245.1/724cad22-3437-4535-9d23-2905ef5defb1`.

Raven, J. C. (1936). *Mental Tests Used in Genetic Studies: The Performance of Related Individuals on Tests Mainly Educative and Mainly Reproductive*. MSc Thesis, University of London, UK.

Recanati, F. (2004). *Literal meaning*. Cambridge University Press. doi:10.1017/CBO9780511615382.

Rees, A., Reksnes, V., & Rohde, H. (2026). Why are you telling me this? The availability and timing of relevance inferences. *Journal of Memory and Language*, *148*, 104741. doi:10.1016/j.jml.2026.104741.

Rees, A., & Rohde, H. (2023). Availability and timing of informativity inferences. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 1198–1205. URL: https://escholarship.org/uc/item/07n4k7rx.

Rees, A., & Rohde, H. (2024). Do children derive informativity inferences? In *Proceedings of AMLaP 2024 (Architectures and Mechanisms for Language Processing)*. Edinburgh, United Kingdom. URL: https://virtual.oxfordabstracts.com/event/31397/submission/218 conference abstract.

Reksnes, V. R. S., Rees, A., Cummins, C., & Rohde, H. (2024). Anticipating informativity in child-directed vs. adult-directed utterances. In *Proceedings of AMLaP 2024 (Architectures and Mechanisms for Language Processing)*. Edinburgh, United Kingdom. URL: https://virtual.oxfordabstracts.com/event/31397/submission/250 conference abstract.

Rett, J. (2011). Exclamatives, degrees and speech acts. *Linguistics and Philosophy*, *34*, 411–442. doi:10.1007/s10988-011-9103-8.

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. URL: https://CRAN.R-project.org/package=psych R package version 2.2.5.

Rints, A., McAuley, T., & Nilsen, E. S. (2015). Social communication is predicted by inhibitory ability and ADHD traits in preschool-aged children: A mediation model. *Journal of Attention Disorders*, *19*(10), 901–911. doi:10.1177/1087054714558873.

Ritchey, K. A. (2011). How generalization inferences are constructed in expository text comprehension. *Contemporary Educational Psychology*, *36*(4), 280–288. doi:10.1016/j.cedpsych.2011.03.002.

Robert, C., Borella, E., Fagot, D., Lecerf, T., & De Ribaupierre, A. (2009). Working memory and inhibitory control across the life span: Intrusion errors in the reading span test. *Memory & Cognition*, *37*(3), 336–345. doi:10.3758/MC.37.3.336.

Roberts, R. J., Hager, L. D., & Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General*, *123*(4), 374–393. doi:10.1037/0096-3445.123.4.374.

Rogowsky, B. A., Calhoun, B. M., & Tallal, P. (2016). Does modality matter? The effects of reading, listening, and dual modality on comprehension. *SAGE Open*, *6*(3). doi:10.1177/2158244016669550.

Rohde, H., Futrell, R., & Lucas, C. G. (2021). What's new? A comprehension bias in favor of informativity. *Cognition*, *209*, 104491. doi:10.1016/j.cognition.2020.104491.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 107–116). Paris, France. URL: https://archive.illc.uva.nl/semdial/ Retrieved from archive.illc.uva.nl.

Röhrig, S. (2010). *The acquisition of scalar implicatures*. Series: Göttinger Schriften zur Englischen Philologie (Vol. 3). Göttingen, Germany: Universitätsverlag Göttingen. doi:10.17875/gup2010-437.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:10.3758/PBR.16.2.225.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7:153*. doi:10.3389/fpsyg.2016.00153.

Rutter, M. (1978). Diagnosis and definition. In M. Rutter, & E. Schopler (Eds.), *Autism: A reappraisal of concepts and treatment* (pp. 1–25). New York: Plenum Press. doi:10.1007/978-1-4684-0787-7.

Ryzhova, M., & Demberg, V. (2020). Processing particularized pragmatic inferences under load. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*, 2594–2600. URL: https://escholarship.org/uc/item/31g0093d.

Ryzhova, M., & Demberg, V. (2023). Processing cost effects of atypicality inferences in a dual-task setup. *Journal of Pragmatics*, *211*, 47–80. doi:10.1016/j.pragma.2023.04.005.

Ryzhova, M., Loy, J., & Demberg, V. (2022). Pragmatic comprehension of implicatures – consistency within individuals across types and time. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*. URL: https://escholarship.org/uc/item/9z71453x.

Ryzhova, M., Mayn, A., & Demberg, V. (2023). What inferences do people actually make on encountering a redundant utterance? An individual differences study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 2631–2638. URL: https://escholarship.org/uc/item/88g7g5z0.

Sana, F., Park, J., Gagné, C. L., & Spalding, T. L. (2021). The interplay between inhibitory control and metaphor conventionality. *Memory & Cognition*, *49*, 1267–1284. doi:10.3758/s13421-021-01152-7.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2023). Workflow techniques for the robust use of bayes factors. *Psychological Methods*, *28*(6), 1404–1426. doi:10.1037/met0000472.

Schank, R. C. (1980). Language and memory. *Cognitive Science*, *4*(3), 243–284. doi:10.1207/s15516709cog0403$_2$.

Schank, R. C., & Abelson, R. P. (1975). Scripts, plans, and knowledge. In *Proceedings of the Fourth International Joint Conference on Artificial Intelligence (IJCAI-75)* (pp. 151–157). Tbilisi, Georgia, USSR.

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. doi:10.1080/0163853X.2020.1813492.

Schulz, K., & van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, *29*(2), 205–250. doi:10.1007/s10988-005-3760-4.

Schuster, S. (2020). *Semantic-Pragmatic Adaptation*. Doctoral dissertation, Stanford University. Retrieved from: https://purl.stanford.edu/rq433dy6205.

Schuster, S., Mayn, A., & Demberg, V. (2023). Working memory updating modulates adaptation to speaker-specific use of uncertainty expressions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 1213–1220. URL: https://escholarship.org/uc/item/8jr5s5fb.

Schwarz, F., Bill, C., & Romoli, J. (2016). Reluctant acceptance of the literal truth: Eye tracking in the covered box paradigm. *Proceedings of Sinn und Bedeutung*, *20*, 61–78. Retrieved from https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/286.

Schwarz, F., Romoli, J., & Bill, C. (2015). Scalar implicatures processing: slowly accepting the truth (literally). *Proceedings of Sinn und Bedeutung*, *19*, 573—590. doi:10.18148/sub/2015.v19i0.250.

Scribner, S., & Cole, M. (1981). *The Psychology of Literacy*. Cambridge, MA and London, England: Harvard University Press. doi:10.4159/harvard.9780674433014.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. doi:10.1023/A:1021928914454.

Sedivy, J. C. (2007). Implicature during real time conversation: A view from language processing research. *Philosophy Compass*, *2*(3), 475–496. doi:10.1111/j.1747-9991.2007.00082.x.

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, *1*(3-4), 149–166. doi:10.1080/17470910600985589.

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. doi:10.1037/h0020586.

Singh, R., Wexler, K., Astle-Rahim, A., Kamawar, D., & Fox, D. (2016). Children interpret disjunction as conjunction: Consequences for theories of implicature and child development. *Natural Language Semantics*, *24*(4), 305–352. doi:10.1007/s11050-016-9126-3.

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. doi:10.3758/s13428-018-1029-4.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. doi:10.1037/0033-2909.119.1.3.

Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, *7*(5), 273–294. doi:10.1111/lnc3.12018.

Sperber, D., & Wilson, D. (1996). *Relevance: Communication and Cognition*. (2nd ed.). Blackwell Publishing.

Spychalska, M., Kontinen, J., & Werning, M. (2014). Electrophysiology of pragmatic processing: Exploring the processing cost of the scalar implicature in the truth-value judgment task. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 1497–1502). volume 36. URL: https://escholarship.org/uc/item/6566q5j6.

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology*, *59*(4), 745–759. doi:10.1080/17470210500162854.

Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, *105*, 108–118. doi:10.1016/j.jml.2018.12.003.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 343–365). Oxford University Press. doi:10.1093/oxfordhb/9780199734689.013.0022.

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, *20*(1), 51–68.

Stanovich, K. E., & Cunningham, A. E. (1993). Where does knowledge come from? Specific associations between print exposure and information acquisition. *Journal of Educational Psychology*, *85*(2), 211–229. doi:10.1037/0022-0663.85.2.211.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*(4), 402–433. doi:10.2307/747605.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–726. doi:10.1017/s0140525x00003435.

Stanovich, K. E., West, R. F., & Harrison, M. R. (1995). Knowledge growth and maintenance across the life span: The role of print exposure. *Developmental Psychology*, *31*(5), 811–826. doi:10.1037/0012-1649.31.5.811.

Stewart, M. E., & Austin, E. J. (2009). The structure of the Autism Spectrum Quotient (AQ): Evidence from a student sample in Scotland. *Personality and Individual Differences*, *47*(3), 224–228. doi:10.1016/j.paid.2009.03.004.

Stieger, S., & Reips, U.-D. (2016). A limitation of the Cognitive Reflection Test: familiarity. *PeerJ*, *4:e2395*. doi:10.7717/peerj.2395.

Stocco, A., Prat, C. S., & Graham, L. K. (2021). Individual differences in reward-based learning predict fluid reasoning abilities. *Cognitive Science*, *45*(2), e12941. doi:10.1111/cogs.12941.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to Theory of Mind. *Journal of Cognitive Neuroscience*, *10*(5), 640–656. doi:10.1162/089892998562942.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. doi:10.1037/h0054651.

Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of Theory of Mind: Evidence from Williams syndrome. *Cognition*, *76*(1), 59–90. doi:10.1016/S0010-0277(00)00069-X.

Taguchi, N. (2012). *Context, individual differences and pragmatic competence*. Multilingual Maters.

Tanner, D. (2019). Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. *Cortex*, *111*, 210–237. doi:10.1016/j.cortex.2018.10.011.

Tavano, E., & Kaiser, E. (2010). Processing scalar implicature: What can individual differences tell us? *University of Pennsylvania Working Papers in Linguistics, Proceedings of the 33rd Annual Penn Linguistics Colloquium*, *16*(1), Article 24. URL: https://repository.upenn.edu/handle/20.500.14332/44764.

Teresa Guasti, M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, *20*(5), 667–696. doi:10.1080/01690960444000250.

Thomas, S., Shipp, N. J., & Ryder, N. (2022). Inhibition in preschool children at risk of developmental language disorder. *Child Language Teaching and Therapy*, *38*(3), 241–253. doi:10.1177/02656590221111341.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. doi:10.1017/S1930297500007622.

van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175.

van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 93–107.

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. doi:10.1080/13546783.2013.844729.

Tourtouri, E. N., Delogu, F., Sikos, L., & Crocker, M. W. (2019). Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science*, *3*(2), 175–202.

van den Broek, P. (1990a). The causal inference maker: Towards a process model of inference generation in text comprehension. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension Processes in Reading* (pp. 423–446). Routledge.

van den Broek, P. (1990b). Causal inferences and the comprehension of narrative texts. In A. C. Graesser, & G. H. Bower (Eds.), *Inferences and Text Comprehension (Psychology of Learning and Motivation, Vol. 25)* (pp. 175–196). Academic Press.

Veenstra, A., & Katsos, N. (2018). Assessing the comprehension of pragmatic language: Sentence judgment tasks. In A. H. Jucker, K. P. Schneider, & W. Bublitz (Eds.), *Methods in Pragmatics* (pp. 257–280). Berlin: De Gruyter Mouton. doi:10.1515/9783110424928-010.

Voeten, C. C. (2023). buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R package version 2.8, https://CRAN.R-project.org/package=buildmer.

Vogels, J., Demberg, V., & Kray, J. (2018). The index of cognitive activity as a measure of cognitive processing load in dual task settings. *Frontiers in Psychology*, (p. 2276).

Vogels, J., Howcroft, D. M., Tourtouri, E., & Demberg, V. (2020). How speakers adapt object descriptions to listeners under load. *Language, Cognition and Neuroscience*, *35*(1), 78–92.

Walker, M. A. (1993). *Informational redundancy and resource bounds in dialogue*. Ph.D. thesis Graduate School of Arts and Sciences, University of Pennsylvania.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*(4), 348. doi:10.1037/0096-3445.115.4.348.

Wanzare, L. D. A. (2020). *Script acquisition: a crowdsourcing and text mining approach*. Doctoral dissertation, Saarland University, Saarbrücken, Germany. doi:10.22028/D291-30163.

Waters, G. S. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 51–79.

Waters, G. S., Caplan, D., & Hildebrandt, N. (1987). Working memory and written sentence comprehension. In *Attention and performance XII: the psychology of reading* (pp. 531–555). Lawrence Erlbaum Associates, Inc.

Welsh, M. B. (2022). What is the CRT? Intelligence, Personality, Decision Style or Attention? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 607–614. URL: https://escholarship.org/uc/item/9gp1r0wd.

Whitely, C., & Colozzo, P. (2013). Who's who? Memory updating and character reference in children's narratives. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1625–1636. doi:10.1044/1092-4388(2013/12-0176).

Wing, L., & Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders*, *9*(1), 11–29.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3(1)*, 1–191. doi:10.1016/0010-0285(72)90002-3.

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology*, *9:1720*. doi:10.3389/fpsyg.2018.01720.

Yntema, D. B. (1963). Keeping track of several things at once. *Human factors*, *5*(1), 7–17.

Zufferey, S., Moeschler, J., & Reboul, A. (2019). *Implicatures*. Cambridge University Press.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 386.