Social information sampling and decision-making: An evolutionary and ontogenetic perspective

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

„Doktor rerum naturalium“

der Georg-August-Universität Göttingen

im Promotionsstudiengang Behavior and Cognition (BeCog)

der Georg-August University School of Science (GAUSS)

vorgelegt von

Rowan Elizabeth Titchener

aus Frankfurt am Main

Göttingen, 2022

Betreuungsausschuss

Prof. Dr. Julia Fischer, Abt. Kognitive Ethologie, Deutsches Primatenzentrum Göttingen

Prof. Dr. Hannes Rakoczy, Abt. Kognitive Entwicklungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Dr. Igor Kagan, Abt. Kognitive Neurowissenschaften, Decision and Awareness Group, Deutsches Primatenzentrum Göttingen

Mitglieder der Prüfungskommission

Referentin: **Prof. Dr. Julia Fischer**, Abt. Kognitive Ethologie, Deutsches Primatenzentrum Göttingen

Korreferent: **Prof. Dr. Hannes Rakoczy**, Abt. Kognitive Entwicklungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskomission:

Dr. Stefanie Keupp, Abt. Kognitive Ethologie, Deutsches Primatenzentrum Göttingen

Dr. Igor Kagan, Abt. Kognitive Neurowissenschaften, Decision and Awareness Group, Deutsches Primatenzentrum Göttingen

Dr. Tanya Behne, Abt. Kognitive Entwicklungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Julia Ostner, Abt. Verhaltensökologie, Johann-Friedrich-Blumenbach Institut für Zoologie und Anthropologie, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 5. Dezember 2022

Contents

List of Figures

List of Tables

Summary

An advantage of being a group-living-, as opposed to a solitary-species, is having the opportunity to engage with conspecifics in mutually beneficial cooperative activities. Choice of cooperation partner often decides whether a cooperative venture succeeds or fails. Cooperation therefore, calls for the careful evaluation of conspecifics. The ideal partner treats others well (shows commitment and values fairness) and has the requisite competence to be able to contribute to the task at hand. This dissertation, comprised of two studies (three experiments), aimed to further what we know about the social evaluation skills of humans and non-human primates. Specifically, I investigated how individuals evaluate others with respect to fairness and competence.

In Study 1, I investigated whether a catarrhine species - the long-tailed macaque (*Macaca fascicularis*) - is sensitive to fairness. Experiments have shown that non-human primates refuse food if they are poorly paid for work compared to a conspecific. I investigated whether refusal behaviour is due to social comparison with the well-paid conspecific, or disappointment in the distributor who rewards poorly. I manipulated two factors: (1) partner presence and (2) type of distributor; half of the subjects experienced a human distributor, the other half a machine distributor. In inequality test conditions, subjects received low-value food while the partner (when present) received high-value food. In an equality control condition, subject and partner received the same low-value food. I measured food refusal behaviour. Study 1 revealed (1) an effect of distributor under conditions of inequality; monkeys who were rewarded by the human, refused food more often compared to those rewarded by the machine (2) no effect of distributor under conditions of equality; irrespective of distributor type, food refusal rates were low when both animals received the same value food and (3) monkeys worked faster when the partner was present, compared to when they were alone.

In Study 2, I investigated how four- to seven-year-old children evaluate and apply competence information in different co-action settings. I designed two online experiments to investigate whether the inference processes thought to be at play in cooperative contexts (trait-reasoning and simple heuristics), are evident in the competitive context - a second setting that calls for a similar strategic use of competence information. In Experiment 1, participants learned about the strength of two models - one model was strong, the other weak. In test trials, participants had to recruit a co-action partner for competitive and co-

operative strength games and a co-action partner for either a competitive or a cooperative knowledge game. Experiment 1 revealed that the older children recruited rationally in both contexts for the strength games, and that children of all ages generalized in both contexts, transferring strength competence information to the knowledge domain. In Experiment 2, participants learned about the strength and object-labelling competence of two models (one model was physically strong but inaccurate at object-labelling, while the other model was physically weak but accurate at object-labelling). In test trials, participants had to recruit a co-action partner for competitive and cooperative strength and knowledge games. Experiment 2 was inconclusive. The children showed little rational recruitment behaviour in either co-action context.

It is possible to draw two conclusions from these studies. Study 1 shows that multiple factors mediate subjects' food refusal behaviour in an impunity experiment; social disappointment in the human distributor, social comparison with the conspecific and food competition all had an effect on refusal behaviour in this experiment. A study limitation (the test conditions were run before the control conditions) however, means that it would be premature to conclude that the social comparison effect we observe, is evidence that long-tailed macaques perceive inequality (are sensitive to fairness). It is possible that by the end of the experiment, subjects were simply tired of refusing food. Future work that counterbalances condition presentation order, could clarify whether we are looking at a three-factor explanation in this species, or whether a two-factor explanation would suffice. From Study 2, I tentatively conclude that the inferential processes (trait-reasoning and a simple decision heuristic) thought to explain children's partner recruitment decisions in cooperative contexts, are also present in the competitive context. Experiment 2 would have been necessary to confirm that trait-reasoning was behind the rational recruitment choices in Experiment 1. These two child experiments were run online. It would be valuable to obtain in-presence data to establish whether demands of the online setting masked children's trait-reasoning abilities in Experiment 2.

Zusammenfassung

Ein Vorteil des Gruppenlebens im Gegensatz zu einer solitär lebenden Art ist die Möglichkeit, sich mit Artgenossen zu kooperativen Aktivitäten zusammenzuschließen, die für beide Seiten von Vorteil sind. Die Wahl des Kooperationspartners entscheidet oft über Erfolg oder Misserfolg eines kooperativen Vorhabens. Erfolgreiche Zusammenarbeit erfordert daher eine sorgfältige Bewertung der Artgenossen. Der ideale Partner behandelt andere gut (zeigt Engagement, legt Wert auf Fairness) und verfügt über die erforderlichen Kompetenzen, um zur gestellten Aufgabe beitragen zu können. Mit dieser Dissertation, die aus zwei Studien besteht, soll das Wissen über die sozialen Bewertungsfähigkeiten von Menschen und nichtmenschlichen Primaten erweitert werden. Genauer gesagt untersuchte ich, wie sich Individuen gegenseitig in Bezug auf Fairness und Kompetenz bewerten.

In Studie 1 habe ich untersucht, ob Javaneraffen (*Macaca fascicularis*) Fairness wahrnehmen. Experimente haben gezeigt, dass nicht-menschliche Primaten Nahrung verweigern, wenn sie für ihre Arbeit im Vergleich zu einem Artgenossen schlecht bezahlt werden. Ich habe untersucht, ob das Verweigerungsverhalten auf den sozialen Vergleich mit dem gut bezahlten Artgenossen oder auf die Enttäuschung über den Verteiler der Nahrung zurückzuführen ist, der die Arbeit schlecht belohnt. Ich habe zwei Faktoren manipuliert: (1) Anwesenheit des Partners und (2) Art des Verteilers; die Hälfte der Versuchsindividuen erlebte einen menschlichen, die andere Hälfte einen maschinellen Verteiler. In der Variante mit ungleicher Belohnung erhielten die Versuchstiere Nahrungsmittel mit geringem Wert, während der Partner (sofern anwesend) Nahrungsmittel mit hohem Wert erhielt. In der Kontrolle mit ausgeglichener Belohnung erhielten die Versuchstiere und ihre Partner dasselbe geringwertige Lebensmittel. Ich habe das Nahrungsverweigerungsverhalten gemessen. Studie 1 ergab (1) eine Auswirkung des Verteilers unter ungleichen Bedingungen; Affen, die vom Menschen belohnt wurden, verweigerten häufiger das Futter im Vergleich zu denen, die von der Maschine belohnt wurden, (2) keine Auswirkung des Verteilers unter gleichen Bedingungen; unabhängig vom Verteilertyp waren die Futterverweigerungsraten niedrig, wenn beide Tiere das gleichwertige Futter erhielten, und (3) die Affen bedienten das Versuchsgerät schneller, wenn der Partner anwesend war, als wenn sie allein waren.

In Studie 2 habe ich untersucht, wie vier- bis siebenjährige Kinder Kompetenzinformationen in unterschiedlichen Handlungskontexten bewerten und anwenden. Ich habe zwei Online-Experimente konzipiert, um zu untersuchen, ob die Inferenzprozesse, von denen

man annimmt, dass sie in kooperativen Kontexten eine Rolle spielen (“trait-reasoning und einfache Heuristiken), auch im kompetitiven Kontext zu beobachten sind – einer zweiten Situation, die eine ähnliche strategische Nutzung von Kompetenzinformationen erfordert. In Experiment 1 erfuhren die Teilnehmer das Maß der körperlichen Stärke von zwei Personen - eine Person war stark, die andere schwach. In den Testdurchläufen mussten die Testpersonen einen Mitspieler für kompetitive und kooperative, körperliche Stärke erfordernde Spiele und einen Mitspieler für ein kompetitives oder kooperatives, Wissen erforderndes Spiel auswählen. Experiment 1 zeigte, dass die älteren Kinder in beiden Kontexten für die Stärke erfordernden Spiele eine rationale Wahl trafen und dass Kinder aller Altersgruppen in beiden Kontexten generalisierten und die Informationen über die körperliche Stärke auf das vermutete Maß an Wissen übertrugen. In Experiment 2 erfuhren die Testpersonen wie stark zwei Personen jeweils sind und wie gut sie Objekte benennen können (eine Person war stark, aber ungenau im Benennen von Objekten, die zweite Person das Gegenteil). In Testdurchläufen mussten die Testpersonen einen Mitspieler für kompetitive und kooperative Stärke oder Wissen erfordernde Spiele auswählen. Die Ergebnisse von Experiment 2 waren nicht aussagekräftig, da die Kinder unabhängig vom Kontext wenig rationales Auswahlverhalten zeigten.

Aus diesen Studien lassen sich zwei Schlussfolgerungen ziehen. Aus Studie 1 schließe ich, dass mehrere Faktoren das Verweigerungsverhalten der Versuchstiere in unserem Testparadigma beeinflussten: soziale Enttäuschung über den menschlichen Verteiler, sozialer Vergleich mit dem Artgenossen sowie Nahrungskonkurrenz wirken sich auf das Verweigerungsverhalten aus. Durch eine Einschränkung der Studie (die Testbedingungen wurden vor den Kontrollbedingungen durchgeführt) wäre es verfrüht zu schlussfolgern, dass der von uns beobachtete Effekt des sozialen Vergleichs ein Beweis dafür ist, dass Javaneraffen Ungleichheit wahrnehmen (für Fairness empfänglich sind). Es ist möglich, dass die Versuchsindividuen am Ende des Experiments der Futterverweigerung überdrüssig waren. Künftige Arbeiten, bei denen die Reihenfolge der Präsentation der Bedingungen ausgeglichen wird, könnten klären, ob wir es mit einer Drei-Faktoren-Erklärung zu tun haben, oder ob eine Zwei-Faktoren-Erklärung ausreichen würde. Aus Studie 2 schließe ich vorläufig, dass die Schlussfolgerungsprozesse (“trait-reasoning” und eine einfache Entscheidungsheuristik), von denen man annimmt, dass sie die Entscheidungen zur Partnerwahl der Kinder in kooperativen Kontexten erklären, auch im Wettbewerbskontext vorhanden sind. Experiment 2 wäre notwendig gewesen, um zu bestätigen, dass die rationalen Entscheidungen bei der Partnerwahl in Experiment 1 auf der Grundlage von Merkmalen getroffen wurden. Diese beiden Experimente wurden online durchgeführt. Es wäre wertvoll, Präsenzdaten zu erhalten, um festzustellen, ob die speziellen Bedingungen des Online-Settings die Fähigkeiten der Kinder in Experiment 2 verdeckt haben.

Chapter 1

General Introduction

> *"In choosing a partner for a collaborative effort, early human individuals wanted to choose someone who would live up to role-specific ideals and who would divide the spoils fairly."* [Tomasello, 2019]

Humans and the majority of non-human primates are cooperative group-living species. Cooperative species band together not only at the group-level but also at the subgroup level where non-kin group members frequently form small transient units for the purposes of cooperative mutualism (Mesterton-Gibbons and Dugatkin, 1992). Cooperation is no simple endeavour. Every individual that attempts to initiate some form of cooperative activity faces a dilemma, namely, which partner(s) to choose. Group-living animals have the choice of any number of conspecifics and group members compete with one another for the opportunity to be selected for cooperative activities (Barclay and Willer, 2007; Herrmann et al., 2019; McNamara et al., 2008; Noë and Hammerstein, 1994; Noë and Hammerstein, 1995). Individuals compete with one another by displaying behaviour that is desirable or sought after in a cooperative partner. Humans for example, often go out of their way to engage in costly prosocial acts in the knowledge that being seen as generous increases one's social standing in the community (Bliege Bird and Power, 2015; Hardy and Van Vugt, 2006). Non-human primates that are high in rank often engage in costly displays of dominance (Milich and Maestripieri, 2016). High-ranking individuals are often sought after for coalitions and alliances (e.g., Perry et al., 2004; Silk, 1999), and (while serving a variety of functions) such dominance displays can be interpreted as an individual signalling that they are partner material.

Advertising oneself – signalling one's worth – only works as a strategy, if there is a perceptive audience. Fortunately anthropoids, with their "social brain", have evolved the skills that are needed to track and monitor the behaviour of group members (Humphrey, 1976; Jolly, 1966). Individuals are always "online" when it comes to gathering social information; individuals monitor one another during the dyadic interactions that they are involved in, and they monitor one another by "eavesdropping" on third-party social interactions (e.g., Anderson et al., 2013). Information that an individual acquires about another agent, is a form of social capital that is constantly revised and updated, and then applied to partner recruitment decisions. Accurate appraisal of another individual will often result in success in the cooperative task. Accurate evaluation is thought to have the added benefit of doing away with the need to constantly monitor partner performance during an interaction (Schino and Aureli, 2017; Sterelny, 2007). Inaccurate positive appraisal of an individual on the other hand, may result in failure in the task at hand, or if the task is a success, may result in an unacceptable (unfair) division of resources.

In this dissertation, I focus on the social evaluation abilities of human and non-human primates. Specifically, I consider how individuals perceive others with respect to fairness and with respect to competence - two qualities that are desirable in a cooperative partner. In the section that follows (section 1.1), I summarize what we know about how human and non-human primates evaluate actions with regard to fairness. Following this (in section 1.2), I summarize what we know about how human and non-human primates evaluate actions that convey competence. To conclude this chapter, I will outline the concrete aims of the two projects that comprise the main body of this dissertation (section 1.3).

1.1 Social evaluation and fairness

Cooperative activities that are undertaken to acquire resources can be contentious. If the cooperation is successful, then before collaborators can part company, the communally-acquired rewards must be amicably divided. If collaborators do not derive (what they perceive to be) an equitable return on their effort, then there is little incentive to re-engage with the same partner(s) in the future. To ensure that they come away with their fair share, individuals engage in social comparison – they monitor their own payoffs relative to those of their collaborators – and they challenge inequity. Truly fair individuals will challenge underpayment, a form of protest that is referred to as "disadvantageous inequity aversion", and they will challenge overpayment – "advantageous inequity aversion" (Fehr and Schmidt, 1999). In instances where protest does not deter an individual from taking more than his share, punishment follows protest (Fehr and Fischbacher, 2004; Henrich et al., 2006). Punishment often takes the form of ostracism (Cinyabuguma et al., 2005). Individuals who short-change others gain in the short-term, but social exclusion ensures that they lose out in the long-run, as they are left out of potentially lucrative cooperative endeavours (Panchanathan and Boyd, 2004).

1.1.1 Fairness in humans

In an effort to understand the rational limits of human decision-making, behavioural economists have investigated how humans perceive and respond to fair and unfair distributive acts in different social scenarios (e.g., Fehr and Gächter, 2000). Economists have made extensive use of three economic games in their work – the Dictator-, Ultimatum-, and Impunity-game (Bolton and Zwick, 1995; Güth et al., 1982; Kahneman et al., 1986).

In the Dictator game, Actor A receives an endowment and the instruction to split the reward between himself and a second person, Actor B. Actor B has no say and must accept whatever decision Actor A reaches. The rational decision here, is for Actor A to hold on to the endowment. In humans however, this seldom happens - Actor A will generally share resources with Actor B (e.g., Bolton et al., 1998). In the Ultimatum game, Actor A again receives an endowment and the instruction to split the reward with a second person, Actor B. This time however, Actor B has been granted some limited agency – he can choose to accept or reject Actor A's offer. If he accepts the offer, the split plays out but if Actor B rejects the offer, then both players lose – neither player receives a reward. If Actor B responds in a rational manner, then he should accept anything that he is offered. The Ultimatum game however reveals that humans frequently reject low offers to punish unfair split proposals (Henrich et al., 2001; Roth et al., 1991). The Impunity game is very similar to the Ultimatum game; Actor A is granted an endowment that he must divide between himself and Actor B. Actor A makes an offer and if Actor B is satisfied, both of the players receive the rewards. If Actor B is dissatisfied however, and rejects the offer, then Actor B leaves empty-handed, but Actor A still receives his share. Rationality dictates that Actor B accept any offer, no matter how low. Again though, this is not the case - humans are quick to reject offers that are not fair, even though this serves to widen rather than close the inequity gap (Yamagishi et al., 2009). Economic game theory provides reliable evidence not only that humans are sensitive to reward distribution, but that humans value fair reward division to such an extent that they are willing to punish others and forfeit rewards.

Sensitivity to fairness is evident early in ontogeny. Studies demonstrate that well before their second year, infants discriminate between fair and unfair distributive acts in third-party interactions (e.g., Schmidt and Sommerville, 2011; Sloane et al., 2012). For example, 15-month-old infants look significantly longer at an image that shows that an agent distributed rewards unfairly as opposed to fairly (Schmidt and Sommerville, 2011). In addition to evidence from such looking time studies that show that infants form expectations around fairness, there is also evidence that infants act upon fairness information (e.g., Burns and Sommerville, 2014; Lucca et al., 2018). For instance, 13- and 17-month-old infants who have observed an individual distribute rewards fairly, and an individual distribute rewards unfairly, prefer to interact with – accept a toy from – the individual who behaved fairly (Lucca et al., 2018).

Infant research has focused on how children respond when they witness third-party distributive acts. As children mature, it becomes possible to examine how children behave when they themselves play a role in the social interaction, i.e. when they are on the receiving end of a fair or an unfair action, or when it is their own actions that create the fair or unfair situation. Research suggests that from three to four years of age, children show disadvantageous inequity aversion – they protest in the face of underpayment (e.g., Blake and McAuliffe, 2011; Fehr et al., 2008; Ulber et al., 2017). Studies initially suggested that it is only at a much later stage, sometime between six and nine years of age, that children begin to shy away from overpayment – to exhibit advantageous inequity aversion (e.g., Blake and McAuliffe, 2011; Blake et al., 2015; Shaw and Olson, 2012). A recent study however, indicates that advantageous inequity aversion may emerge earlier than previously thought (Ulber et al., 2017). Ulber and colleagues demonstrated that three-year-old children who take part in a collaborative task and are overpaid for their efforts relative to another agent, often elect to share or forego rewards rather than allow the inequitable distribution (Study 1., 2017).

Although questions about the developmental timeline remain, the fact that humans show these two forms of inequity aversion is not up for debate. Given that sensitivity to fairness is evident so early in childhood (before the age of two), it is likely that socialization and cultural-learning cannot fully explain this phenomenon; there is likely an evolutionary component to this behaviour (Schmidt and Sommerville, 2011). Recognizing this, in the last two decades, ethologists have begun to investigate whether non-human animals also show evidence of disadvantageous and advantageous inequity aversion.

1.1.2 Fairness in non-human primates?

The first study to investigate sensitivity to fairness in a non-human primate species used a version of the impunity game (Brosnan and de Waal, 2003). Brosnan and de Waal tested brown capuchin monkeys (*Cebus apella*) in a simple token exchange paradigm in which they manipulated the value of the food that a subject received, relative to a conspecific. The subject and a "partner" monkey were placed in neighbouring test compartments that were separated by mesh, such that the subject could fully observe the partner. A human distributor took it in turns to exchange a token with each monkey under two different conditions. In an inequality condition, the subject received low-value food for her effort, while the partner received high-value food. In an equality condition, the subject and partner both received the same low-value food. Brosnan and de Waal reported that female brown capuchin monkeys show frustration behaviour and become demotivated (more often refuse to token-exchange) in the inequality as compared to the equality condition. The authors concluded that capuchin monkeys engage in social comparison with conspecifics – that monkeys show disadvantageous inequity aversion.

Since Brosnan and de Waal's publication almost two decades ago, primatologists have investigated an impressive array of species but with mixed results (see Table 1, Oberliessen and Kalenscher, 2019). Some primatologists corroborate the findings of Brosnan and de Waal. For example, it has been claimed that the following primate species show disadvantageous inequity averse protest behaviour: capuchin monkeys (van Wolkenten et al., 2007; Fletcher, 2008), rhesus macaques (*Macaca mulatta*; Hopper et al., 2013), long-tailed macaques (*Macaca fascicularis*; Massen et al., 2012), chimpanzees (*Pan troglodytes*; Brosnan et al., 2005; Brosnan et al., 2010) and tamarin monkeys (*Saguinus ursulas*; Neiworth et al., 2009). Other scientists report null findings. The following primate species have also been tested, and these species do not appear to protest underpayment relative to a conspecific: capuchin monkeys (Dubreuil et al., 2006; Fontenot et al., 2007; Roma et al., 2006; Silberberg et al., 2009), chimpanzees (Bräuer et al., 2006; Bräuer et al., 2009; Jensen et al., 2007; Kaiser et al., 2012), bonobos (*Pan paniscus*; Bräuer et al., 2006; Bräuer et al., 2009), orangutans (*Pongo pygmaeus*; Bräuer et al., 2006; Bräuer et al., 2009; Brosnan et al., 2011), gorillas (*Gorilla gorilla*; Bräuer et al., 2006; Bräuer et al., 2009), marmosets (*Callithrix jacchus*), owl monkeys (*Aotus spp.*), and squirrel monkeys (*Saimiri boliviensis*; Freeman et al., 2013). Notably, there is overlap between these two lists. Tests on different populations of the same species have returned divergent findings.

There is scepticism about the existence of inequity aversion in non-human primates. Since Brosnan and de Waal's report, a number of alternative explanations have been put forward to explain why subjects might show elevated food refusal under conditions of social inequality (Engelmann et al., 2017; Roma et al., 2006; Wynne, 2004). Two of these explanations – the *food expectation* hypothesis and the *frustration* hypothesis – centre around food. The food expectation hypothesis contends that the prominent display of high-value food in the inequality condition (a food display that is absent in the equality condition) could explain why subjects refuse food in the inequality, but not the equality condition (Wynne, 2004). The frustration hypothesis contends that past receipt of high-value rewards may explain the refusal patterns (Roma et al., 2006). An animal that was previously accustomed to a certain level of food quality is thought to react negatively to a drop in food quality. A third, more recent (and arguably more complex) hypothesis – the *social disappointment* hypothesis – has also been put forward. The social disappointment hypothesis contends that food refusals are not about the partner monkey at all, they are rather an expression of disappointment in the human experimenter who consistently hands out low-value food when they *could* hand out high-value food (Engelmann et al., 2017). Engelmann and colleagues tested the social disappointment hypothesis in chimpanzees with some success. As is evident from these many hypotheses, whether non-human primates are sensitive to fairness, that is, whether non-human primates show inequity aversion, remains a topic of much debate.

1.2 Social evaluation and competence

Cooperation is undertaken with a goal in mind. A cooperative unit that is strategically formed, i.e. comprised of competent individuals, stands a better chance of success – is more likely to achieve their goal, compared to a unit that has formed out of convenience. Within a population, there is natural interindividual variation in terms of the competence people have or acquire in different domains; some individuals are born with superior physical coordination skills (e.g., Beunen et al., 2014), others perform better in cognitive tasks (Vernon, 1985). Competence is time-consuming to develop and maintain, and an individual therefore tends to develop a niche skill set over their lifetime (Becker and Murphy, 1992). In light of the fact that cooperative ventures among competent individuals are more likely to succeed, it pays for individuals to advertise competence, and it pays for individuals to be able to recognize competence in others.

1.2.1 Competence evaluation by humans

Adults recognize competence in others (e.g., Baumann and Bonner, 2013; Bonner, 2004), and prefer to work alongside competent rather than incompetent individuals (Bor, 2017). Developmental psychologists have sought to understand the ontogeny of such selectivity and thus far, their endeavours have returned mixed findings.

A large body of literature indicates that preschool children (three- to four-year-olds) are not epistemically vigilant. Rather than showing an understanding that there are interindividual differences in competence, preschool children often naïvely act as though *all* adults are competent (e.g., Jaswal et al., 2010; Jaswal et al., 2014). This phenomenon can be explained by the fact that, although children possess some foundational ("core") knowledge (Kinzler and Spelke, 2007), in their early years they are heavily reliant on adults to supply them with information that is not possible to learn first-hand (Csibra and Gergely, 2011; Harris, 2012). This frequent experience – that adults have the answer – is thought to result in young children perceiving adults as all-knowing, all-perceiving entities (Mossler et al., 1976); adults are viewed as a knowledgeable class of people – the competence assessment is at the group-, rather than at the individual level.

A similarly large body of literature indicates that preschool children (three- to four-year-olds) *are* epistemically vigilant (see Mills, 2013 and Robinson and Einav, 2014 for a review of the selective trust literature). Preschoolers do seem to be aware that different individuals have different knowledge bases (e.g., Kushnir et al., 2013; Lutz and Keil, 2002; VanderBorght and Jaswal, 2009). For example, preschoolers consider peers a better source for information on toys, and adults a better source for information on food (VanderBorght and Jaswal, 2009). Children also prefer to imitate and learn from individuals who have demonstrated competence, as opposed to incompetence, in the past (e.g., Koenig et al., 2004; Wilks et al., 2015). Koenig and colleagues for example have demonstrated that three- and four-year-old children prefer to learn new vocabulary from a previously accurate, as opposed to inaccurate, object-labeller (2004).

In recent years, developmental psychologists have shifted their focus, from investigating the social and epistemic cues that are thought to influence children's trust, to theorizing about the inferential processes likely to underlie these divergent patterns of trust behaviour (Fusaro et al., 2011; Hermes et al., 2018; Sobel and Kushnir, 2013). The puzzle that developmental psychologists are grappling with, is the fact that in some situations three- to four-year-old children seem to have an unsophisticated grasp on the limits to people's knowledge, while in other situations children of this age seem to monitor competence and apply the information in a rational manner.

1.2.2 Competence evaluation by non-human primates

The majority of the social evaluation studies primatologists have carried out to date, have investigated how different non-human primate species evaluate third-party interactions between humans (see Chijiiwa, 2021 for a review). There is evidence to suggest that non-human primates discriminate between people on the basis of generous and non-generous behaviour (Russell et al., 2008; Subiaul et al., 2008), whether individuals reciprocate (Kawai et al., 2014; Kawai et al., 2019), and whether individuals help or hinder others (e.g., Anderson et al., 2013; Krupenye and Hare, 2018). Chimpanzees prefer to spend time near a human who has shared food in the past, as opposed to a human who has not (Russell et al., 2008). Common marmosets prefer to accept rewards from an individual who reciprocates item-gifting, as opposed to an individual who collects rewards that are handed to him and gifts no item back; this preference disappears when both of the people engage in reciprocal exchange (Kawai et al., 2014). Capuchin monkeys prefer to accept food from a person who has helped another person in the past, as opposed to someone who has repeatedly ignored solicitations for help (Anderson et al., 2013). Such studies are informative – they provide evidence that non-human primates evaluate social agents and act upon their social evaluations. One limitation of these studies however, is that they look at how non-human primates evaluate another species; the subjects in these studies evaluate humans, the focus is not on conspecific evaluation.

One key experiment has looked at evaluation of competence within a non-human primate species in the context of partner choice (Melis et al., 2006; Experiment 2). Melis and colleagues investigated whether chimpanzees discriminate between competent and incompetent conspecifics when recruiting a partner for a cooperative task (2006). Each subject took part in two sessions - an introductory session and a test session. The procedure in these two sessions was identical, but the introductory session took place one day prior to the test session. In each session, the subject was placed in a test compartment, within reach of a cooperative rope-pulling apparatus – a baited plank that was out of reach, but that could be hauled to within reach by two individuals simultaneously taking either end of a rope. Two conspecifics were locked into test rooms adjacent to the subject's compartment and the subject used a key to release the conspecific they wished to interact with. In the introductory session, the subject had six trials to learn that one of the conspecifics was competent at the task and that the other conspecific was incompetent. Melis and colleagues reported a significant interaction between partner and session. Subjects were more likely to recruit the effective partner, but only in the test session i.e., only after the subject had sufficient opportunity to contrast and evaluate the performance of each of the individuals. This study nicely demonstrates that non-human primates, like humans, evaluate others with regard to competence, and prefer to work together on a cooperative task with a more skilled individual.

1.3 Dissertation aims and objectives

The aim of this dissertation was to augment what we know about the social evaluation abilities of human and non-human primates. Specifically, I aimed to fill two knowledge gaps. The first knowledge gap concerns how non-human primates evaluate actions with regard to fairness. The second knowledge gap concerns the inferential processes thought to underlie how children evaluate competence.

As outlined in section 1.1, there is consensus that humans show inequity aversion. In contrast, primatologists have been unable (as yet) to reach consensus on whether non-human primates show inequity aversion - this very much remains an open question. In the first project of this dissertation (detailed in chapter 2), I investigated inequity aversion in a non-human primate species - the long-tailed macaque. Specifically, I set out to contrast two social hypotheses. I ran an experiment in which a subject worked for low-value food in two types of conditions: (1) alone and (2) alongside a more well rewarded conspecific. Half of the subjects were offered rewards by a human, the other half by a machine. The behavioural measure was food refusal. If food refusal is an expression of disappointment in a human who rewards poorly when they could reward well, then subjects who experience a human distributor should refuse food more often compared to subjects who experience a machine distributor. If on the other hand, subjects' food refusals are due to social comparison processes, then irrespective of distributor type, food refusals should be elevated when subjects are working in the presence of a more well rewarded conspecific compared to subjects who work alone.

As outlined in section 1.2, there is ample evidence to demonstrate that young children are discerning in terms of whom they will trust. Simultaneously (and puzzlingly) however, there is also evidence that indicates the opposite, that young children often show a naive, blind trust in their dealings with others. A dual-process account has been proposed to explain this behavioural incongruence. Formal testing has begun (e.g., Hermes et al., 2020), but this account has yet to be fully explored. In the second project of this dissertation (detailed in chapter 3), I ran two online experiments in which I contrasted partner recruitment behaviour in the cooperative context, with partner recruitment behaviour in a novel co-action setting – the competitive context. Successful navigation of the competitive context, like the cooperative context, calls for the strategic use of social information. If there is evidence that children use sophisticated trait-based reasoning when they have sufficient social information, but fall back to using simple heuristics when they have insufficient social information, this would indicate that a dual-process account (explained in chapter 3) has a wider explanatory reach than previously thought – that it's reach extends further than the cooperative context.

Chapter 2

Social disappointment and partner presence affect long-tailed macaque behaviour in an 'inequity aversion' experiment

Rowan Titchener[1,3,4], Constance Thiriau[5], Timo Hüser[2], Hansjörg Scherberger[2,4,6], Julia Fischer[1,4,7], Stefanie Keupp[1,4,7]

[1] *Cognitive Ethology Laboratory, and* [2] *Neurobiology Laboratory, Deutsches Primatenzentrum GmbH, Kellnerweg 4, 37073 Göttingen, Germany*
[3] *Institute of Psychology, University of Göttingen, Waldweg 26, 37073 Göttingen, Germany*
[4] *Leibniz ScienceCampus Primate Cognition, 37077 Göttingen, Germany*
[5] *Université Paris Nord, 99 Avenue Jean Baptiste Clément, 93430 Villetaneuse, France*
[6] *Faculty of Biology and Psychology, and* [7] *Department for Primate Cognition, University of Göttingen, 37077 Göttingen, Germany*

Royal Society Open Science (2023) 10: 221225.
DOI: 10.1098/rsos.221225

Chapter 3

Social evaluation strategies in four- to seven-year-old children: Age-related changes in partner choice in competitive and cooperative settings

Rowan Titchener[1,2,3], Jonas Hermes, Julia Fischer[1,3,4], Hannes Rakoczy[2], Stefanie Keupp[1,3,4]

[1] *Cognitive Ethology Laboratory, Deutsches Primatenzentrum GmbH, Kellnerweg 4, 37073 Göttingen, Germany*

[2] *Institute of Psychology, University of Göttingen, Waldweg 26, 37073 Göttingen, Germany*

[3] *Leibniz ScienceCampus Primate Cognition, 37077 Göttingen, Germany*

[4] *Department of Primate Cognition, University of Göttingen, 37077 Göttingen, Germany*

Abstract

Given sufficient information, children make rational choices about whom to copy or approach for help in cooperative experimental settings. Trait-based reasoning is thought to explain this selectivity. Given insufficient information, children fall back on simple heuristics and generalize evidence of competence in one domain, to unrelated domains. Motivated to establish whether social selectivity and generalization behaviour is context-sensitive, we devised two online experiments for 4- to 7-year-old children. In Experiment 1, we manipulated the degree of competence of two co-action partners. Children were introduced to a weak and to a strong model and had the opportunity to apply this information strategically in competitive and cooperative domain-relevant games (requiring strength) and in a domain-irrelevant game (requiring object-labelling knowledge). In Experiment 2, children were introduced to a weak but smart model and a strong but ignorant model and had the opportunity to apply this information in games of different domain-context combinations (e.g. a cooperative knowledge game or a competitive strength game). In Experiment 1, we were able to demonstrate that from approximately five years of age, children recruit the rational partner for the competitive context (the weak model) and the rational partner for the cooperative context (the strong model). Experiment 1 also revealed that given insufficient information, children generalize partner expertise and pick the strong partner for a cooperative knowledge task and the weak partner for a competitive knowledge task. In Experiment 2, all children performed surprisingly poorly irrespective of context, and we discuss aspects of the online test environment that may account for this outcome. The present study highlights the importance of accounting for experimenter identities during statistical analyses. As is not uncommon in this field, multiple experimenters were involved in data collection. Despite standardized training protocols and balanced assignment of participant age and experimental condition to each experimenter, we found that experimenter ID significantly influenced the results of both experiments.

Keywords: Social evaluation – Partner choice – Strategy – Cooperation – Competition

3.1 Introduction

Children gain valuable knowledge from observing and communicating with the people in their surrounds (Csibra and Gergely, 2009; Harris, 2012; Harris et al., 2018). Blind trust in all social agents would, however, be maladaptive as, whether by design or accident, people spread misinformation and can differ markedly from one another in terms of expertise and competence. Preschoolers are aware that such inter-individual differences exist (e.g., Kushnir et al., 2013; Lutz and Keil, 2002) and from approximately three to four years of age, preschoolers are highly selective about whom they attend to, imitate and trust when making use of socially-disseminated information (for reviews on selective trust see Mills, 2013; Robinson and Einav, 2014).

The bulk of the evidence on children's selective trust has been derived from laboratory studies implementing the so-called "two-informant" paradigm (developed by Koenig et al., 2004). There are two phases to the procedure. In a familiarization phase, children are shown two models who provide contrasting information (e.g., labels for familiar objects); one model provides accurate labels and the other model, inaccurate labels. In the test phase, children watch the models provide different labels for an unfamiliar object and are asked to endorse one of the two models. From studies that have employed the two-informant paradigm, we know that preschoolers tune into epistemic cues (e.g., Birch et al., 2008; Jaswal and Neely, 2006; Koenig et al., 2004; Koenig and Harris, 2005; Pasquini et al., 2007). For example, children prefer to learn new information from a previously accurate as opposed to an inaccurate model (Koenig et al., 2004). Via the two-informant paradigm, we know that, in addition to epistemic cues, preschooler selectivity is mediated by social cues (Jaswal and Kondrad, 2016). For example, children prefer to learn from similar (Elashi and Mills, 2014; MacDonald et al., 2013), familiar (Corriveau and Harris, 2009), attractive (Bascandziev and Harris, 2014) and dominant, authoritative individuals (Bernard et al., 2016).

Recently, the focus has shifted from identifying cues that mediate selectivity, to theorizing about the inferential processes likely to underlie selectivity (Fusaro et al., 2011; Hermes et al., 2015; Sobel and Kushnir, 2013). So far, three candidate processes have been put forward (outlined in Fusaro et al., 2011): (1) behaviour-matching, (2) global impression formation and (3) trait-reasoning. Behaviour-matching and global impression formation count as low-level inference processes. A child that engages in behaviour-matching infers that a model shows behavioural consistency over time, i.e., someone who has demonstrated accuracy in object-labelling will provide accurate labels in the future. An inference based on behaviour-matching is narrow as the child does not go so far as to assign domain-wide competence to the model. A child that forms a global impression of a model (often referred to as a "halo/pitchfork" effect) draws a wider conclusion than is warranted based on the model's behaviour. For example, a child might infer that someone who is strong is also an accurate object-labeller. Relative to behaviour-matching and global impression formation, trait-reasoning is a more complex inference process. A child who engages in trait reasoning succeeds in a two-component process (Liu et al., 2007). The first component involves mapping observed behaviour to a trait (e.g., Model A was accurate in object-labelling; therefore, Model A is an intelligent person). The second component involves mapping the assigned trait to future behaviour (Model A is an intelligent person; therefore, Model A will be good at many knowledge-tasks, for instance object-labelling tasks).

There is evidence to support the idea that children engage in global impression formation in certain situations and trait-reasoning in others (Brosseau-Liard and Birch, 2010; Fusaro et al., 2011). For example, we know that preschoolers are predisposed to learn from people who behave prosocially (Brosseau-Liard and Birch, 2010) – this is an example of global impression formation given that prosocial behaviour is not linked to whether a person is knowledgeable. There is also, however, ample evidence that young children have a grasp on the fact that people have different domains of expertise – children appear to understand that it is more appropriate to consult certain people in some situations as opposed to others (e.g., Kushnir et al., 2013; Lane and Harris, 2015; Lutz and Keil, 2002; VanderBorght and Jaswal, 2009). For example, given the choice between adults and peers, children consult peers for toy-related information and adults on the topic of food (VanderBorght and Jaswal, 2009).

In an effort to reconcile these two apparently divergent findings – the fact that young children show both sophisticated and unsophisticated forms of trust – Hermes and colleagues recently suggested a dual-process account (2018). They theorize that very early in development, children's trust decisions may be mediated by simple inference strategies, simple decision heuristics termed "Type-I" processes. At some later stage, as children acquire a more sound concept of knowledge, a better understanding of how different domains relate to one another, children's trust decisions come to be mediated by a second type of process – a "Type-II" process (sophisticated trait-reasoning). Type-I processes are intuitive – they are automated and implicit. Type-II processes involve computation and reasoning – they are slow, effortful and explicit. The crux of Hermes and colleagues theory is that when children show unsophisticated trust behaviour (e.g., generalize about competence from one domain to another), they are using simple Type-I processes (e.g., a "trust the better" heuristic). And when children exhibit sophisticated trust behaviour (e.g., selecting a competent above an incompetent person for help in a domain-relevant task), they are using a Type-II process – sophisticated trait-reasoning. The dual-process account contends that neither of these inference processes are mutually exclusive. Type-I and Type-II processes co-exist and which of these two inference strategies is tapped depends upon the particulars of the situation (e.g., the amount of information the child has on hand).

A recent study indicates that the dual-process account may gain traction (Hermes et al., 2020). Hermes and colleagues demonstrated via a two-informant paradigm that inference strategy can be manipulated as a function of cognitive load. Under normal conditions (older) children and adults were shown to engage in effortful trait-reasoning but following an increase in cognitive load (participants were asked to do an *n-back* task on top of a selective trust task), participants fell back to generalization behaviour. One way to establish further support for the dual-process account, is to demonstrate that the Type-I and Type-II strategies used in the cooperative context, are used in a second context that calls for similar problem-solving. Children grow up surrounded by cooperative and well-meaning adults, and it is natural that to date studies on social selectivity have focussed on the cooperative context (e.g., Hermes, Behne, Studte, et al., 2016). Competition is a second context however, in which children can profit from applying trait-information. Competition is by no means foreign to preschool and school-age children; children often compete with siblings and peers for resources and attention, and they take part in simple competitive games, for instance team sports (Sheridan and Williams, 2006). From approximately four years of age children begin to show verbal and physical competitive behaviour in experimental settings (e.g., Leuba, 1933; for a review see Tsiakara and Digelidis, 2021). We think it logical that trait-reasoning would be at play in the competitive context and that in this setting, given insufficient information, children may also fall back to generalizing about competence. Evidence of partner selectivity and generalization behaviour in the competitive setting could provide further support for the dual-process account.

We designed two experiments to investigate how four- to seven-year-old children recruit co-action partners in the competitive as compared to the cooperative context, when they have (a) degree-of-competence information, (b) insufficient and (c) domain-of-competence information about a pair of models. In Experiment 1, we examined situations (a) and (b). We employed a two-informant "degree- of-competence" paradigm. Children were given information about two models who differed from one another in terms of physical strength. Children then played a series of competitive and cooperative strength games, for which they had the necessary information to select well. We were interested to see whether children would systematically choose the weak model as an opponent in the competitive strength games, and the strong model as a teammate in the cooperative strength games. In addition to these strength games, each child played a knowledge game for which they had insufficient information to recruit well, and where they might therefore be expected to revert to generalization behaviour (e.g., recruiting the physically weak model as an opponent in the competitive knowledge game).

We had two predictions for Experiment 1. We predicted that age and context would interact to affect rationality of recruitment in the strength games. We expected that the older children would be more successful at rational partner recruitment than the younger children, and that older children would show equally high levels of selectivity in the two contexts. In comparison, we expected younger children to be more selective in the cooperative as opposed to the competitive context. We thought the younger children would struggle in the competitive context as this context poses more complex perspective-taking demands compared to the cooperative context (Perner et al., 2005; Priewasser et al., 2013). In the cooperative context, the perspective of the child and the recruited team member align, but in the competitive context the child has to keep in mind that their goal, and the goal of the recruited character, are in conflict. The second prediction for Experiment 1 concerns the generalization test trials. We predicted that both age and context would interact to affect tendency to generalize in the knowledge games. We predicted that older children would generalize at low (chance) levels compared to the younger children. Older children have more life experience, which means they have a better grasp on how different domains potentially relate to one another (Lutz and Keil, 2002). We predicted that the younger children would be more likely to generalize about partner competence across domains. For the younger children, we predicted a context-mediated effect, namely that they would generalize more often in the cooperative as compared to the competitive context. This prediction is again based on the fact that young children seem to struggle to understand competitive as compared to cooperative games (Priewasser et al., 2013).

In Experiment 2, we examined situation (c). We employed a two-informant "domain-of-competence" paradigm. Children were given complete information about two models who differed from one another in terms of both physical strength and knowledge (accuracy in labelling objects). Experiment 2 was designed to provide information about the inference process underlying rational recruitment choices in the competitive context – to clarify whether trait-reasoning could explain selectivity in this setting. In Experiment 2, we provided full model information to see whether children would recruit partners flexibly, depending on the domain-context combination they were presented with. Flexible recruitment behaviour would be evidence that children are reasoning rationally about the trait information they have been given.

In Experiment 2, we predicted that age and context would interact to affect rationality of recruitment choices in the domain-of-competence test trials. We predicted that older children would be equally competent in the cooperative and the competitive context, while younger children would struggle more often, in particular in the competitive context. According to the dual-process theory, it takes some time before the Type-II process, sophisticated trait-reasoning, is fully-fledged (Hermes et al., 2018). We think it is likely, therefore, that young children will use a mix of trait-reasoning and simple strategies when playing the recruitment games; this would result in the younger children more frequently making irrational recruitment calls compared to the older children.

3.2 Experiment 1

In Experiment 1 participants were given "degree-of-competence" information. We introduced participants to two characters that differed from one another in terms of strength; one character was strong, the other was weak. We then monitored the partner recruitment choices of each participant in a series of competitive and cooperative "intra-domain" strength games, and either a competitive or a cooperative "inter-domain" knowledge game.

3.2.1 Material and Methods (preregistered)

Participants

One hundred and twenty-three 4- to 7-year-old children were recruited from a departmental database. The data of 107 children were analysed (mean age = 72.64 months, 51 girls). Sixteen test sessions were not analysed for the following reasons: (i.) experimenter error (2 sessions), (ii.) the session had to be aborted (2 sessions), (iii.) the child failed a control question (5 sessions), (iv.) the child was a sibling of a previous participant i.e., the child was invited in error (1 session), (v.) technical issues (2 sessions), (vi.) the child was ill and could not complete the session (1 session) and (vii.) the child did not indicate during the concluding questions that they identified with their assigned onscreen character (3 sessions); we elaborate on the reasoning behind this final exclusion criterion in the Electronic Supplementary Material (ESM – stored in Appendix B). A majority of the children were residing in or near Göttingen, a medium-sized town in Lower Saxony, Germany. Each child participated voluntarily and was thanked with a certificate of participation.

So as to have a comparison standard, we ran a validation study with adult participants. Forty-two adult participants were recruited via a local university website and through word-of-mouth. We analysed the data of 41 adults (mean age = 25.46 years, 21 women). One test session was not analysed as the demographic information provided by the participant was not valid. All participants took part voluntarily; undergraduate students could receive compensation in the form of university credits.

Materials

All video stimuli were created using the *Vyond* animation software. An overview of the stimuli can be found in the ESM (subsection B.1.2).

Design

We designed an online experiment in *Labvanced* (Finger et al., 2017). The sessions were moderated, i.e., the child met with an experimenter via virtual conferencing software (usually *BigBlueButton* but occasionally *Zoom*). The experimenter kept her audio on during the session, but her camera was switched off most of the time so as to provide the child with a full-screen view of the stimuli. Each session was recorded either using screen-capture software (*Open Broadcaster Software*) or an external camcorder. Three female experimenters collected the data. To minimize experimenter bias, the first author thoroughly trained all experimenters on how to present the material and interact with parents and children. Participant age, experimental condition, and presentation order was balanced between experimenters.

As we used a repeated-measure, within-subject design for the strength games (the "degree-of-competence" test trials), we counterbalanced the context of the first test trial (53 of the 107 participants experienced a competition strength game first). The knowledge game (generalization test trial) was implemented as a between-subject factor. Children who had received a competition strength game first, experienced a competition knowledge game and vice versa for those who had received a cooperation strength game first.

We elected not to counterbalance a number of elements of the experimental design. We did not counterbalance character-strength assignment, the order in which the characters appeared on the screen and the location of the characters on screen. Mr. Blue was always the strong character, Mr. Blue always appeared first (e.g., his block of three strength demonstration videos were shown first) and Mr. Blue always stood on the left side of the screen when children were asked to choose between the characters. The reason why we chose not to counterbalance these elements, is that this experiment investigates whether children rationally choose the character most suited to the context. Children who are influenced by colour, or who show primary or recency effects, or a side-bias do not show rational behaviour.

The adult version of the experiment was also run in *Labvanced*. The stimuli were identical to those used in the child study, although bridging text was added so that the participants could navigate the experiment alone (the adult experiment was not moderated). Some sections of the experiment were slightly modified (e.g., we prefaced the experiment with an adult-appropriate formalities section) or removed (e.g., the children were shown a video to celebrate the prizes they had accumulated in the experiment; we did not include this video in the adult experiment). Given that the child and adult versions of the experiment were near-identical, in the section that follows, we restrict our description to the child experiment.

Procedure

There were seven stages to Experiment 1 (summarized in Table 3.1). For detailed information on each stage, see subsection B.1.1 of the ESM.

Coding

The keyboard keys the experimenter used to navigate the experiment automatically logged data. One person (RT) transcribed the answers children gave to the two identification questions and the two justification questions. RT categorized the answers according to several detailed coding schemes (summarized in Tables B.10 and B.11 of the ESM).

A person blind to the hypothesis of the study coded 20% of the child videos. This person transcribed and categorized the identification and justification answers. Overall inter-coder agreement was excellent. Coder 1 and 2 matched perfectly in their assessment of the answers given to the two identification questions (Cohen's Kappa coefficient (k) = 1). Inter-coder agreement on the category assigned to the degree-of-competence justification answers, and the category assigned to the generalization justification answers was excellent (k = 1 and 0.934 respectively).

The same person coded 20 rows of the adult data, categorizing the written identification and justification answers the adults had provided. Inter-coder agreement was fine. Coder 1 and 2 matched perfectly in their assessment of the answers given to the two strength-of-identification questions (Maxwell's coefficient (RE) = 1). Inter-coder agreement on the degree-of-competence justification answers was adequate (k = 0.429). Inter-coder agreement on the category assigned to the generalization justification answers was good (k = 0.797).

Analyses

All statistical analyses were carried out in R (R-Core-Team, 2021, version 4.1.1). We fitted a total of six models. Three Generalized Linear Mixed Models (GLMMs; Baayen, 2008) were used to analyse the degree-of-competence test trials (partner recruitment choices in strength games). Two Generalized Linear Models (GLMs; Baayen, 2008) and one GLMM were used to analyse the generalization test trials (partner recruitment choices in knowledge games). Four of these analyses (Models 1 through 4) were preregistered on OSF (https://archive.org/details/osf-registrations-q8myj-v1). Two non-preregistered analyses (Models 1B and 3B) were run to check for a potential effect of experimenter ID (given that three experimenters collected the child data).

The models were fitted using functions from the *lme4* package (Bates et al., 2015; version $1.1 - 27.1$). GLMMs were fitted using the *glmer* function and GLMs were fitted using the *glm* function. In the case of the GLMMs, the *bobyqa* optimizer was used to aid convergence. Confidence intervals (95% CI) were generated using the *bootMer* function of the *lme4* package (in the case of GLMMs), or the *confint* function of the *stats* package (for GLMs).

In each case the response variable was binary, and all models were therefore fitted with a binomial error structure and logit link function (McCullagh and Nelder, 1989). In the subsections that follow, we concisely describe the structure of each model but for detailed information concerning model structure, sample size, variable standardization, model diagnostics (stability, collinearity), and to view the output tables see subsection B.1.3 of the ESM. Reliable *p*-values were obtained by dropping the fixed effects from the full model one at a time and comparing the full model with each respective reduced model (achieved via the *drop1* R function).

For each model constructed, we used a likelihood ratio test to ascertain the effect of our variable(s) of interest (Dobson, 2002). The likelihood ratio test compares the fit of the full model with that of a null model. In each case the structure of the null model was identical to that of the full model, but it lacked the variable of interest.

The qualitative data generated by the open-ended "identification" and "justification" questions were used in several exploratory analyses. Qualitative data were categorized according to several coding schemes (detailed in subsection B.1.4 of the ESM).

Table 3.1: Experiment 1 procedure (degree-of-competence paradigm)

Stage	Description
Demonstration	The child met two models who differed from one another in degree-of-competence in the strength domain. The child learned via six videos that Mr. Blue was strong and Mr. Green was weak. *[We asked two control questions to check that the child had formed the correct character-strength associations.]*
Character assignment	The child was assigned an onscreen representative. *[We asked a control question to check that the child understood that the character onscreen was "them"]*
Game familiarization	The child learned the rules and possible outcomes of the games that would be used in the degree-of-competence test trials. The experimenter used four videos to explain how a competitive and a cooperative version of a strength game worked.
Degree-of-competence test trials	The child had the chance to apply the information they had learned about the two models. Each child played six strength games (three per context). The goal was to win as many virtual prizes as possible.
Generalization	The child was asked to recruit a partner (either Mr. Blue or Mr. Green) for either a competitive or a cooperative knowledge game.
Identification questions	The child was asked two questions to gauge whether they had identified with their assigned character. Q1: The child was asked to label a series of pictures of characters/objects. The child's character appeared in the picture sequence. Q2: The child was asked to describe a still-image from a competition strength game. The child's character appeared in the scene. *[Children who used a first-person, first-person possessive or a third-person reference in their answer "identified" with their character]*
Justification questions	The child was asked for the reasons behind their partner recruitment choices in the degree-of-competence test trials and in the generalization trial.

Analysis of degree-of-competence data

We constructed a GLMM, Model 1, to test whether either an interaction between context and age, or a main effect of context or age, had an effect on children's partner recruitment choices in the degree-of-competence test trials. The response variable was rationality of partner recruitment choice (the weak character was the rational choice in the competition strength game, vice versa for the cooperation strength game). Model 1 comprised a two-way interaction between context and age, in addition to their main effects and the fixed effects of gender, condition presentation order and trial number. Participant ID was included as a random effect and the random slopes of context and trial number were included in the model.

In addition to fitting Model 1, we fitted a second GLMM – Model 1B. The motivation in fitting Model 1B was to control for a potential effect of experimenter ID. The two models were nearly identical, differing only in terms of their random effects components (and therefore also with regard to their random slope structures). Model 1B included the random effect of experimenter ID, in addition to the random effect of participant ID. This meant Model 1B included the random slopes of context and trial number within participant ID and the random slopes of the context × age interaction term, gender, order and trial number within experimenter ID.

To establish whether context had an effect on adult partner recruitment choices in the degree-of-competence test trials, we fitted Model 2 (a GLMM). The response variable was rationality of partner choice. Model 2 comprised a main effect of context, in addition to the fixed effects of gender, order and trial number. Participant ID was included as a random effect, and the random slopes of context and trial number were included in the model.

Analysis of generalization data

We constructed a GLM, Model 3, to test whether either an interaction between context and age, or a main effect of context or age, had an effect on child generalization behaviour. The response variable was whether recruitment choice constituted a generalization; to generalize in the competition knowledge game was to choose the physically weak partner and to generalize in the cooperation knowledge game was to choose the strong partner. Model 3 comprised a two-way interaction between context and age, in addition to their main effects and the fixed effects of gender and the intercept-BLUPs (Best Linear Unbiased Predictors) generated in Model 1. These BLUPs were a proxy measure for how rationally each child had performed in the degree-of-competence test trials.

To estimate the effect of context and age on the likelihood that children generalize about competency while controlling for a potential effect of experimenter ID, we constructed another GLMM – Model 3B. Although the two models were similar in structure, Model 3B included the random effect of experimenter ID, and the BLUPs were taken from Model 1B rather than Model 1. Model 3B included the random slopes of the context × age interaction term, gender and BLUPs within experimenter ID.

We constructed a GLM, Model 4, to test whether context had an effect on adult generalization behaviour. The response variable was whether the recruitment choice constituted a generalization. Model 4 comprised the main effect of context, in addition to the fixed effects of gender and the BLUPs generated in Model 2.

3.2.2 Results

Partner recruitment in strength games

The younger participants were overall less rational in their partner recruitment choices compared to the older participants. As can be seen in Figure 3.2, the youngest children recruited poorly irrespective of condition while the oldest children performed well in both contexts. The developmental trajectory seems to comply with our predictions, namely that rational choice is apparent earlier in the cooperative, as compared to the competitive context (see also Figure 3.1, where we plotted the data according to age bins).

In Model 1 we tested whether context, age, or an interaction between context and age had an effect on child recruitment choices. The full-null model comparison was significant (likelihood ratio test: $\chi^2_3 = 37.125$, $p = <0.001$). The full model revealed that the context × age interaction was not significant ($p = 0.376$) and we therefore fitted a reduced model. The reduced model revealed a significant effect of context ($p = 0.035$). The children were significantly more likely to recruit rationally in the cooperative as compared to the competitive context. The reduced model also revealed a significant effect of age ($p = <0.001$). The older children were significantly more likely to recruit rationally compared to the younger children. In a side-finding, we report a significant effect of gender ($p = 0.019$); boys were significantly more likely to recruit rationally than girls.

Model 1B differed from Model 1 in that we controlled for the fact that multiple experimenters had collected the data. The full model differed significantly from the null model (likelihood ratio test: $\chi^2_3 = 11.643$, $p = 0.009$). As the full model revealed that the context × age interaction was not significant ($p = 0.370$), we fitted a reduced model. The reduced model revealed no significant effect of context ($p = 0.153$), but a significant effect of age ($p = 0.003$). Like Model 1, reduced Model 1B revealed a significant effect of gender ($p = 0.027$). It seems the identity of the experimenter played a role in children's response behaviour and this effect cannot be ignored. For this reason, we discuss the results of the more complex model – Model 1B – in subsequent sections of the paper.

In Model 2 we tested whether context had an effect on adult recruitment choices. The full-null model comparison revealed no significant difference between the two models (likelihood ratio test: $\chi^2_3 = 0.562$, $p = 0.454$). Context had no effect on adult recruitment behaviour. The adults recruited rationally at high levels in both contexts.

At the conclusion of the experiment we asked each participant whether they could explain their recruitment choices in the strength games. This question revealed that as the age of the participants increased, the frequency of answers that explicitly referenced strength competence also increased (see Figure 3.3). Compared to the other age groups, very few four-year-old children were either willing or able to provide a justification for their recruitment behaviour in the strength games. Of all the age groups, the four-year-olds most often offered a fairness-based justification (stating e.g., “I wanted to alternate [between the characters]”).

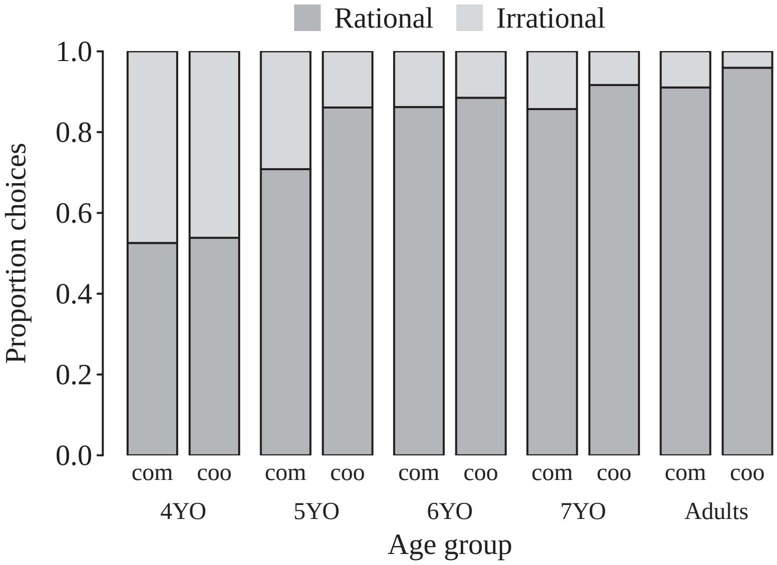

Figure 3.1: The older participants made more rational partner recruitment choices compared to the younger participants. The five-year-olds were the only age group to behave differently across the two contexts; the five-year-olds were more rational in the cooperative (*coo*) as compared to the competitive context (*com*).

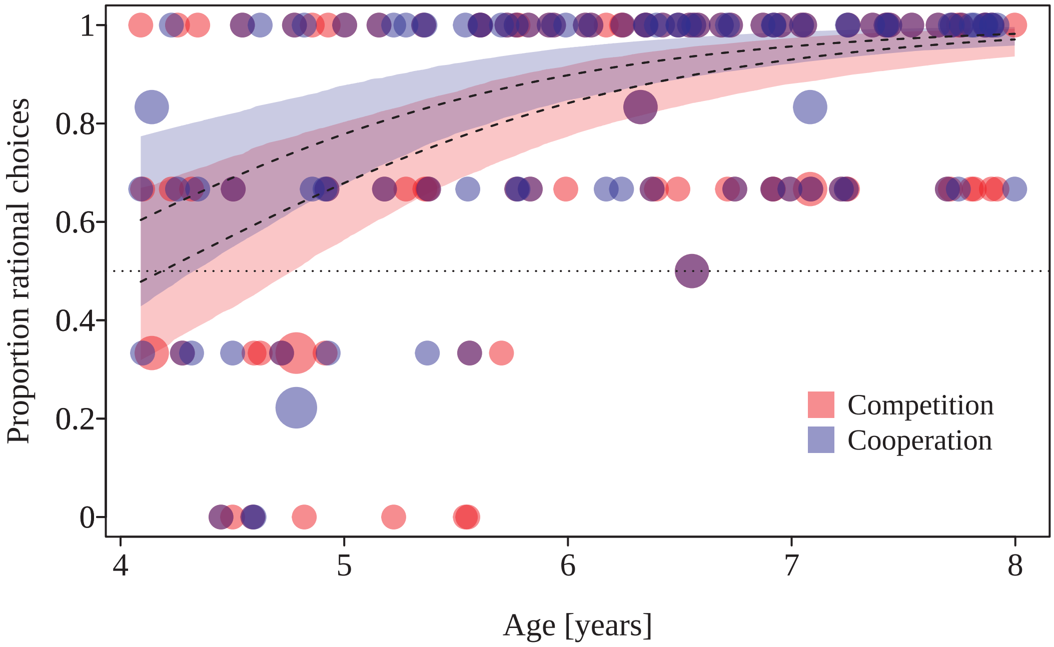

Figure 3.2: Data, model estimates (dotted lines) and 95% confidence intervals (shaded areas) for the degree-of-competence data. Children who always selected the strong model for the cooperative strength game scored a "1"; likewise for children who always selected the weak character for the competitive strength game. Several data points cover a larger area as some participants were the exact same age. Model estimates are from reduced Model 1B. The faint horizontal line that intersects the y-axis at 0.5 indicates chance level decision-making; once the estimate and boundary of the lower confidence interval sit above this threshold, participants of this age are no longer selecting at chance levels.

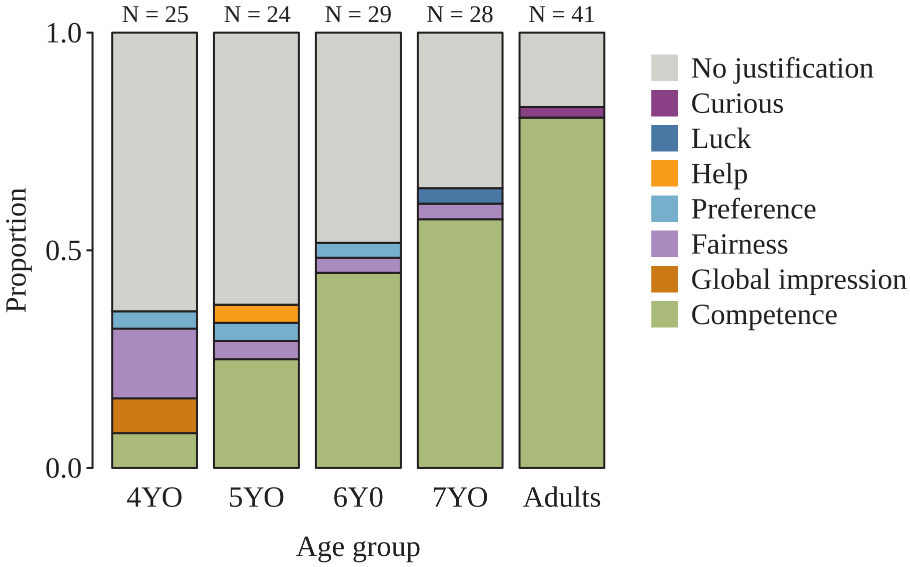

Figure 3.3: Reasons offered by participants when asked to explain their recruitment choices in the degree-of-competence test trials (the strength games). The participants' answers were categorized according to a standardized coding scheme (summarized in Table B.10 of the ESM).

Partner recruitment in knowledge game

Generalization was frequent in both contexts; overall 83% of the children generalized, picking the weak character for the competitive knowledge game, or the strong character for the cooperative knowledge game (see Figures 3.4 and 3.5). The younger children (the four- and five-year-olds) showed a similar distinct pattern of generalization behaviour; children who played the competition knowledge game generalized less often compared to the children who played the cooperation knowledge game. Contrary to our predictions, the six- and seven-year-old children generalized at high levels and they did so in both contexts. In contrast to the six- and seven-year-old children, the adults generalized at low levels.

In Model 3 we tested whether context, age, or an interaction between context and age, had an effect on tendency to generalize about competence. A full-null model comparison revealed there was a trend towards a difference between the fit of the full model and the fit of the null model (likelihood ratio test: $\chi^2_3 = -6.923$, $p = 0.074$). The full model revealed that the interaction between context and age was not significant ($p = 0.138$) and we therefore fitted a reduced model. The reduced model revealed neither a significant effect of context ($p = 0.109$), nor a significant effect of age ($p = 0.153$). In a side-finding, we report a significant effect of our rationality proxy measure – the intercept-BLUPs ($p = 0.011$). The more rational a child had recruited during the strength games, the more likely that child was to generalize when recruiting a partner for the knowledge game.

Model 3B differed from Model 3 in that we controlled for the fact that multiple experimenters had collected the data. In this case, the full and the null models were not significantly different from one another(likelihood ratio test: $\chi^2_3 = 4.415$, $p = 0.22$). Neither the context × age interaction term, nor the main effects of context or age affected the likelihood that a child would generalize about competency beyond the strength domain.

In Model 4 we tested whether context would affect the likelihood of adults to generalize about competence. The full and the null model were not significantly different from one another (likelihood ratio test: $\chi^2_3 = -1.127$, $p = 0.289$). Context had no effect on adult generalization behaviour.

We asked each participant whether they could explain their recruitment choice in the knowledge game (see Figure 3.6). The proportion of answers that fit the category "halo/ pitchfork effect" was relatively large relative to the other categories. Approximately 24% of the four-year-olds provided answers indicating that global impression formation was at play. In the case of the five-, six- and seven-year-olds, between 37% and 48% of the participants indicated that their recruitment choice was the result of global impression formation. Thirty-seven percent of the adults indicated that their choice of co-action partner in the knowledge game was influenced by valence of strength competence.

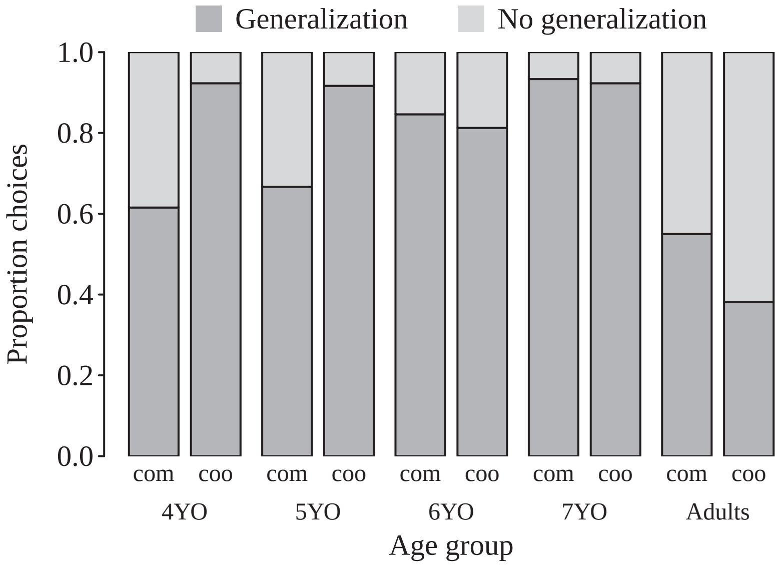

Figure 3.4: The children generalized at higher levels compared to the adults. The younger children recruited differently in the competitive as compared to the cooperative context. The four- and five-year-old children who played the cooperative (*coo*) knowledge game generalized more often compared to the four- and five-year-old children who played the competitive (*com*) knowledge game.

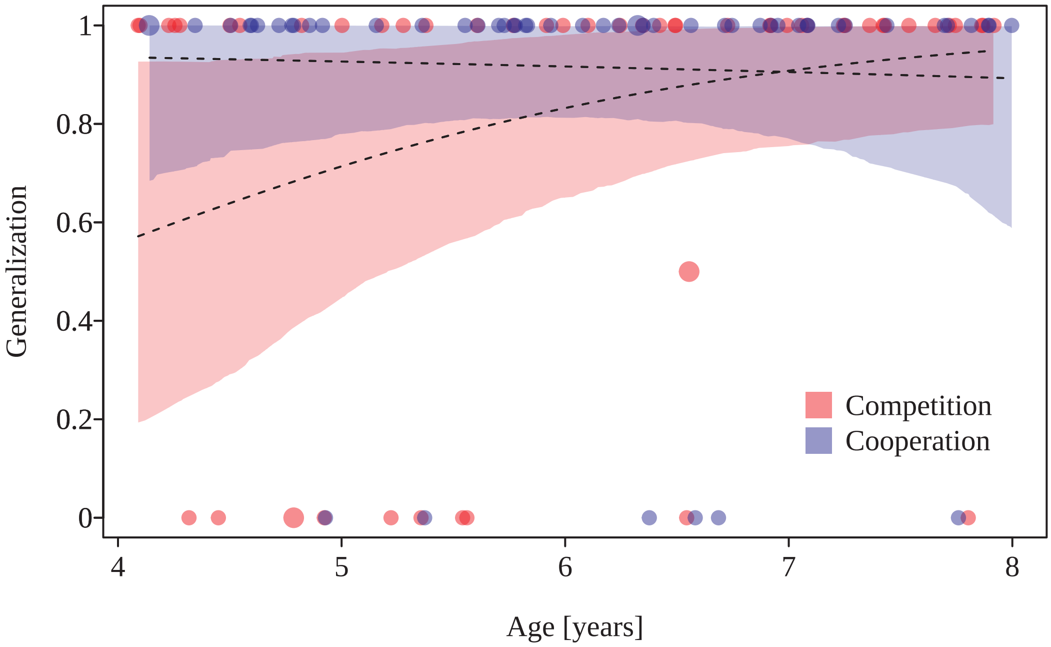

Figure 3.5: Data, model estimates (dotted lines) and 95% confidence intervals (shaded areas) for the generalization data. Children who selected the strong model for the cooperative knowledge game or the weak model for the competitive knowledge game scored a "1"; these children transferred information about competence in the strength domain, to the knowledge domain. Several data points cover a larger area as some participants were the exact same age.

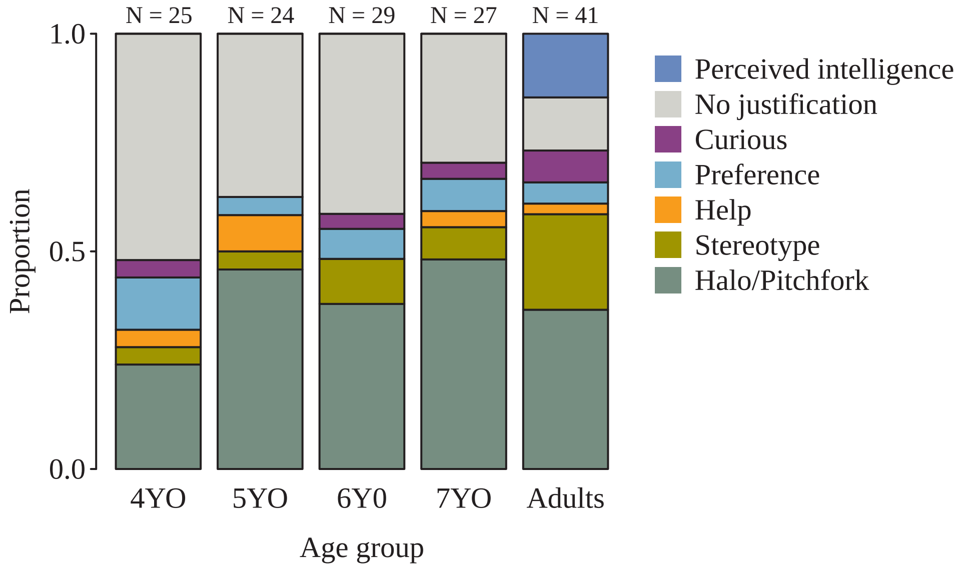

Figure 3.6: Reasons offered by participants when asked to explain their recruitment choice in the generalization test trial (the knowledge game). The participants' answers were categorized according to a standardized coding scheme (summarized in Table B.11 of the ESM).

Identification with assigned virtual character

Identification questions 1 and 2 confirmed that the majority of the participants "identified" with the virtual character that represented them onscreen during the test trials (see Figure 3.7). In Identification question 1 participants were asked to provide a label for a picture of their character. A participant who used a first person (e.g., "that's me"), a first-person possessive (e.g., "my character") or a third-person reference (i.e., the child used their name) in their answer counted as having identified with the character. All five-, six-, and seven-year-olds identified with the character. A small number (2 from 26 four-year-olds, and 1 from 41 adults) did not identify with their character. In Identification question 2, participants were asked to describe a scene in which their character appeared. Many of the participants (but especially the four-year-old and the adult participants) avoided mentioning their character in their answer. In comparison to Identification question 1, question 2 shows that some participants were hesitant to "claim" their character.

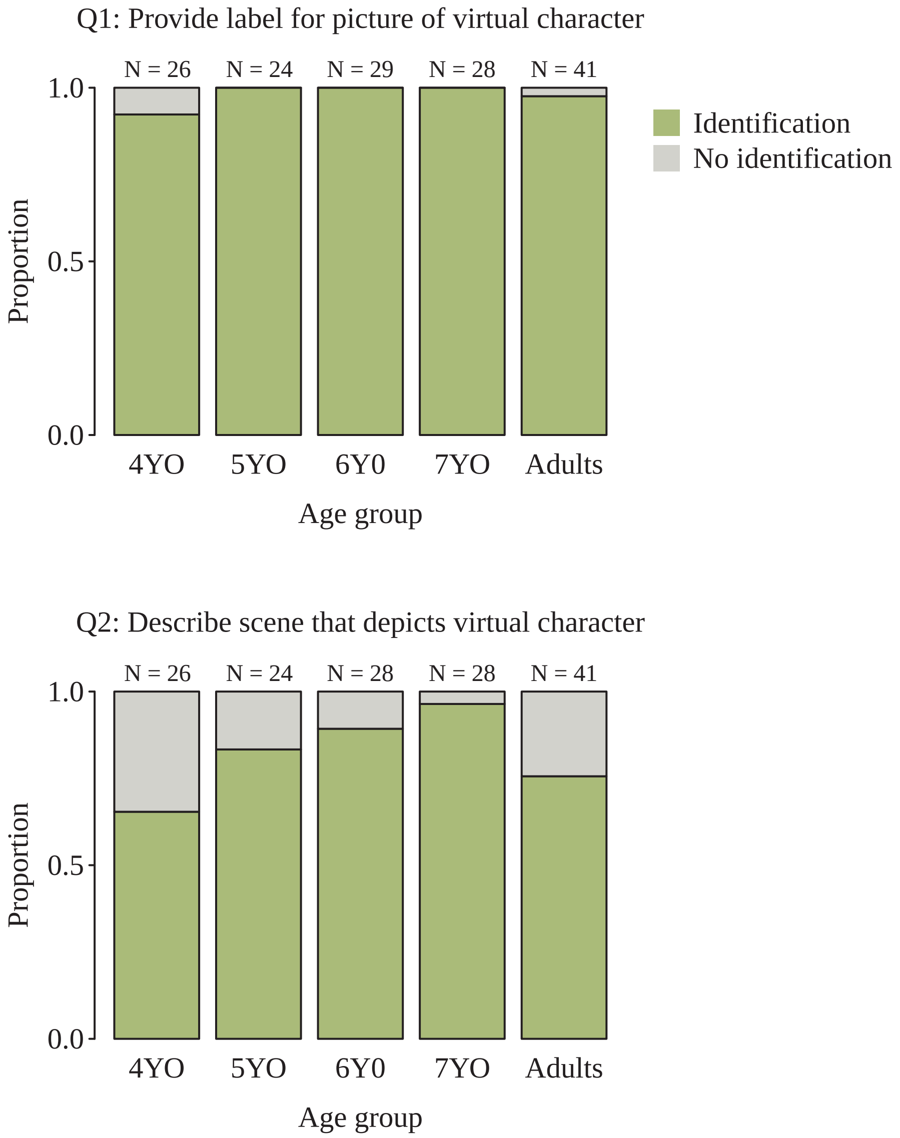

Figure 3.7: Results of the two "identification" questions; a participant who used a first-person, first-person possessive, or a third-person reference in their answer was seen as having identified with their virtual character. The overall level of identification was high, although question 1 resulted in a higher number of participants being classed as having identified with their virtual character compared to question 2. The two questions revealed that of all the participants, the four-year-old and the adult participants identified least often with their character.

3.2.3 Discussion

In Experiment 1, we investigated whether partner selectivity and generalization behaviour are context-sensitive. With regard to partner selectivity, we found a significant effect of age (p = 0.003). The older children recruited more rationally in the strength games compared to the younger children. In addition to age, context initially appeared to have a significant effect on recruitment behaviour ($p = 0.035$); the five-year-old children recruited more rationally in the cooperative as compared to the competitive context. This effect of context disappeared however, once experimenter ID was factored into the analysis (p = 0.153). The four-year-old children performed surprisingly poorly in the cooperative context, recruiting rationally at chance levels in both contexts. Overall, we can conclude that from approximately five years of age, children show selectivity in the competitive as well as in the cooperative context – they systematically select the physically weak model for the role of opponent and the strong model for the role of teammate; partner selectivity therefore, is not restricted to the cooperative context.

At first glance it appeared that context might mediate generalization behaviour (see Figure 3.4) as, in line with our predictions, the younger children (the four- and five-year-olds) generalized at lower levels in the competitive as compared to the cooperative knowledge game (see Figure 3.4). Via GLMs and a GLMM however, we established that there was no significant effect of context ($p = 0.109$). In addition, we found no significant effect of age on generalization behaviour ($p = 0.153$). We can conclude therefore, that similar to partner selectivity, generalization behaviour is not restricted to the cooperative context.

Numerous studies have demonstrated that four-year-old children are selective when it comes to choosing a partner or informant in cooperative settings (e.g., Hermes, Behne, Studte, et al., 2016; Koenig and Harris, 2005). In light of this fact, the four-year-olds in our study performed surprisingly poorly in the cooperation strength games. One possible explanation for their poor performance, is that this was an online experiment. Although many of the experiments in which four-year-old children succeed in demonstrating social selectivity have made use of video stimuli (e.g., Koenig and Harris, 2005), these experiments all took place in-person, either in laboratory or kindergarten settings. In the present experiment, the tester assigned the child an onscreen character and the child navigated the experiment in the virtual first-person. This aspect of the study design (use of the virtual first-person) adds a layer of complexity and detachment to the standard two-informant design because the child's experience becomes passive – the consequences of the child's recruitment choices are only indirectly experienced. Rather than the children themselves feeling true emotion after winning or losing a game, they watch their character display the appropriate response.

At the conclusion of the experiment, for exploratory purposes, we asked several identification questions to gauge whether participants had identified with their virtual character. These questions indicated that of all the participants, the four-year-old children were the group least likely to identify with their character. By the age of three, children have a sound concept of self (Marsh et al., 2002; Stipek et al., 1990) and it is possible that the four-year-old children were less inclined to entertain the idea that the character on the screen was their representative compared to the older child participants. This reasoning however, does little explain why some adults showed some resistance to claiming their character in Identification question 2. It could be that these adults were less willing to "enter" the onscreen world because this involved some form of pretence. Young children are generally willing to immerse themselves in pretend play, but this willingness tails off in mid-childhood (Smith and Lillard, 2012). In the general discussion we examine in some detail the potential consequences of implementing a paradigm that relies on the virtual first-person. In addition, we discuss more generally some aspects the online environment that may have influenced recruitment behaviour.

Success in the competitive setting calls for the strategic application of social information. Overall, our finding that partner selectivity and generalization behaviour are not context-sensitive is a first indication that the dual-process theory outlined by Hermes and colleagues (2018) may also account for recruitment patterns in the competitive setting. In situations where the children had the requisite information on hand, older children recruited partners rationally in the competitive setting. In situations where the children did not have the necessary information on hand, they used a simple heuristic – they based their decisions on the valence of competence a model had demonstrated in an unrelated domain. The degree-of-competence paradigm we used in Experiment 1 allows us to identify patterns of partner selectivity, but it does not allow us to discern underlying inference strategy. At this point, it is an assumption that the older children who showed selectivity in the competitive strength games were engaging in trait reasoning - a Type-II process. Theoretically these children could be using simple heuristics, for example an "oppose-the-worse [model]" heuristic – would result in a similar pattern of selectivity. Motivated to determine the inference strategy at play, and establish whether trait-reasoning or rather a simpler inference strategy could explain selectivity in the competitive setting, we ran Experiment 2.

3.3 Experiment 2

In Experiment 2 participants were given "domain-of-competence" information. We introduced participants to two characters that differed from one another not only in terms of strength, but also in terms of knowledge; one character was strong but ignorant, the other weak but knowledgeable. Having provided this information, we monitored the partner recruitment choices of each participant in a series of competitive and cooperative strength and knowledge games.

3.3.1 Material and Methods (preregistered)

Participants

One hundred and twenty-three 4- to 7-year-old children were recruited from a departmental database. Children who had taken part in Experiment 1 were not called for Experiment 2. The data of 98 children were analysed (mean age = 5.5 months, 48 girls). Twenty-five test sessions were not analysed for the following reasons: (i.) experimenter error (1 session), (ii.) the child failed a control question (strength control question = 2 sessions; knowledge control question = 3 sessions, and character control question = 8 sessions), (iii.) there was interference from a third-party (sibling interference = 1 session and parental interference = 1 session), (iv.) technical issues (7 sessions) and (v.) the participant did not supply information about their gender (information that was needed for the model; 2 sessions). A majority of the children were residing in or near Göttingen. Each child participated voluntarily and was thanked with a certificate of participation.

So as to have a comparison standard, we ran a validation study with adult participants. Fifty adult participants were recruited via a local university website and through word-of-mouth. We analysed the data of 48 adults (mean age = 23.83 years, 24 females). The data from two participants was not analysed; the data from one participant was incomplete (this person did not answer our open-ended questions), and another participant provided an invalid date of birth. All adults took part voluntarily. Undergraduate students were eligible to claim payment in the form of university credits.

Materials

As we recruited a new pool of subjects, we were able to upcycle many of the Experiment 1 video stimuli and re-use them in Experiment 2. Some changes were made to the Experiment 1 videos (e.g., sound effects were added) and additional stimuli were created using *Vyond* animation software. For an overview of the stimuli see Figures B.1, B.5 and B.6 in the ESM.

Design

We designed an online experiment in *Labvanced* (Finger et al., 2017). The sessions were moderated; the child met with an experimenter (we again had three – one experimenter was retained from Experiment 1) using a virtual conference software (*BigBlueButton*). The experimenter was present via audio for the whole session however, her camera was switched off most of the time so that the child had a full-screen view of the stimuli. Each session was recorded using screen-capture software (*Open Broadcaster Software*).

As we used a repeated-measure, within-subject design for our test trials, we counterbalanced the test trial that appeared first (51 of the 98 participants experienced the competition context first and of these, 25 participants started with a strength game; 47 participants experienced the cooperation context first and of these, 24 participants started with a strength game). At the close of the experiment we asked the children several questions to probe the rationale behind their partner recruitment choices. Children who had received a competition test trial first, were asked about the competition setting first in these questions and vice versa for the children who had received a cooperation test trial first.

In Experiment 2 (as in Experiment 1), we elected not to counterbalance a number of elements of experimental design. We did not counterbalance character-strength assignment, the order in which the characters appeared on the screen and the location of the characters on screen. Mr. Blue was always the strong character, Mr. Blue always appeared first (e.g., his blocks of demonstration videos were shown first), and Mr. Blue always stood on the left side of the screen when children were asked to choose between the characters. Again, we chose not to counterbalance these elements as this experiment investigates rational behaviour – whether children choose the character most suitable for the context. Children who are influenced by colour, or who show primary or recency effects, or a side-bias do not show rational behaviour.

The adult version of the experiment was run in *Labvanced.* The stimuli were identical to those used in the child study, although bridging text was added so the participants could navigate the experiment alone. Some sections of the experiment were slightly modified to make the experiment more age-appropriate (e.g., we adapted the formalities section), and other sections were removed (e.g., the children were shown a video to celebrate the prizes they had accumulated in the experiment; we did not include this celebration video in the adult experiment). In the section that follows, we restrict our description to the child version of the experiment.

Procedure

There were five stages to Experiment 2 (summarized in Table 3.2). For detailed information on each stage see subsection B.2.1 of the ESM.

Coding

The keys the experimenter used to navigate the experiment automatically logged data. One person (RT) transcribed the answers children gave to the two advice questions. RT categorized the answers according to a detailed coding scheme (summarized in Table B.16 of the ESM).

A person blind to the hypothesis of the study coded 20 of the child videos. This person transcribed and categorized the answers provided in response to the two advice questions. Inter-coder agreement was excellent. Coder 1 and 2 matched in their categorization of the answers given to the competition and the cooperation advice questions (k = 0.933 and 1 respectively). The same person coded data from 20 adults, categorizing the written answers adult participants gave in response to the two advice questions. Again here the inter-coder agreement was excellent (k = 1 for each advice question).

Table 3.2: Experiment 2 procedure (domain-of-competence paradigm)

Stage	Description
Demonstration	The child met two models who differed from one another in terms of their domains of expertise. The child watched twelve videos and learned that Mr. Blue was strong and ignorant, while Mr. Green was weak but knowledgeable. *[We asked two control questions per domain to check that the child had formed the correct character-strength/-knowledge associations].*
Character assignment	The child was assigned an onscreen representative. *[We asked a control question to check that the child understood that the character onscreen was "them"].*
Game familiarization	The child learned the rules and possible outcomes of the games that would be used in the domain-of-competence test trials. The experimenter used eight videos to explain how a competitive and a cooperative version of a strength and a knowledge game worked.
Domain of competence trials	The child had the chance to apply the information they had learned about the two models. Each child played four games (one per context and domain). The goal was to win as many virtual prizes as possible.
Advice questions	To see whether the child could reason about recruitment, the child was shown a still-image from a competitive and cooperative strength game, and asked to offer advice to a playmate about to play the game.

Analyses

Statistical analyses were carried out in R (R-Core-Team, 2021, version 4.1.1). We analysed the domain-of-competence data via three GLMMs. Two of these GLMMs (Models 5 and 6) were preregistered on OSF (https://archive.org/details/osf-registrations-q8myj-v1). One non-preregistered GLMM – Model 5B, was run to check for a potential effect of experimenter ID.

The models were fitted using the *glmer* function from the *lme4* package (Bates et al., 2015; version 1.1 - 27.1); the *bobyqa* optimizer was used to aid convergence. Confidence intervals (95% CI) were generated using the *bootMer* function of the *lme4* package.

In each case the response variable was binary, and all models were fitted with a binomial error structure and logit link function (McCullagh and Nelder, 1989). In the subsections that follow, we concisely describe the structure of each model (Models 5, 5B and 6). For detailed information concerning model structure, sample size, variable standardization, model diagnostics (stability, collinearity), and to view the output tables, see subsection B.2.3 in the ESM. *P*-values were obtained via the *drop1* R function. For each model constructed, we used a likelihood ratio test to ascertain the effect of our variable(s) of interest (Dobson, 2002).

The qualitative data generated by the open-ended "advice" questions were used in several exploratory analyses. These data were categorized according to a standardized coding scheme (detailed in subsection B.2.4 of the ESM).

Analysis of domain-of-competence data

We constructed a GLMM, Model 5, to test whether an interaction between context and age, or a main effect of context or age, had an effect on children's partner recruitment choices in the domain-of-competence test trials. The response variable was rationality of partner recruitment choice. Model 5 comprised a two-way interaction between context and age, in addition to their main effects and the fixed effects of gender, order (whether the child received the competition-knowledge, competition-strength, cooperation-knowledge, or cooperation-strength test trial first) and trial number. Participant ID was included as a random effect, and the random slopes of context and trial number within participant ID were included in the model.

We constructed another GLMM, Model 5B, to test whether an interaction between context and age, or a main effect of context or age, had an effect on children's partner recruitment choices. The motivation in running Model 5B was to control for a potential effect of experimenter ID. The structure of Model 5B was similar to that of Model 5, but Model 5B included the random effect of experimenter ID in addition to the random effect of participant ID. Model 5B included the random slopes of context and trial within participant ID, and the random slopes of the context × age interaction term, gender and trial number within experimenter ID.

We constructed a GLMM, Model 6, to test whether context had an effect on adult partner recruitment choices in the domain-of-competence test trials. The response variable was rationality of partner recruitment choice. Model 6 comprised the main effect of context in addition to the fixed effects of gender, order and trial number. Participant ID was included as a random effect, and the random slopes of context and trial number within participant ID were included in the model.

3.3.2 Results

Partner recruitment in strength and knowledge games

Overall, the children performed poorly in the domain-of-competence test trials (see Figures 3.8 and 3.9); the children recruited rationally in just 57% of the test trials. Interestingly, the adults' recruitment choices were less rational in the competitive as compared to the cooperative context.

In Model 5 we tested whether, given domain-of-competence information, context and age interact to affect children's recruitment choices. Comparison of the full with the null model revealed there was a significant difference between the two models (likelihood ratio test: $\chi^2_3 = 7.974$, $p = <0.047$). The full model revealed that the interaction between context and age was not significant ($p = 0.261$), and we therefore fitted a reduced model. The reduced model revealed no significant effect of context ($p = 0.289$), but a significant effect of age ($p = 0.019$). The older children were significantly more likely to recruit rationally compared to the younger children.

In Model 5B we tested whether, given domain-of-competence information, context and age interact to affect children's recruitment choices. Model 5B controlled for a potential effect of experimenter ID. A comparison revealed that the full and null models were not significantly different from one another (likelihood ratio test: $\chi^2_3 = 5.474$, $p = 0.14$). Neither the context × age interaction term, nor the main effects of context or age had an effect on how rationally children recruited partners for the co-action games.

In Model 6 we tested whether context affected the rationality of adult partner recruitment choices. A full-null model comparison revealed that the two models differed from one another (likelihood ratio test: $\chi^2_3 = 16.071$, $p = <0.001$). Context had a significant effect on the likelihood that an adult would recruit rationally in a domain-of-competence test trials ($p = <0.001$) with adults less likely to recruit rationally in a competitive, as opposed to a cooperative test trial.

We asked each participant to provide advice to a playmate about to encounter a competition and a cooperation strength game. We were interested to see whether participants would provide rational tips that might reveal how they themselves had approached their recruitment decisions in the domain-of-competence test trials. The advice questions revealed that with increasing age, children provided increasingly rational advice (see Figure 3.10). For example, the seven-year-olds gave proportionally more rational advice (made more references to character competence) compared to the four-, five- and six-year-olds, and this pattern was apparent for both the competition and the cooperation advice question. Interestingly, the "rationale-context mismatch" category featured more often in answers given to the competition, as compared to the cooperation, advice question. An answer that was classified as a rationale-context mismatch, was one in which a participant correctly identified the valence of a character's competence, but their use of this information did not suit the context at hand. The fact that this answer category occurred more frequently in response to the competition advice question indicates that, of the two co-action contexts, children may have found the competitive setting more challenging to navigate.

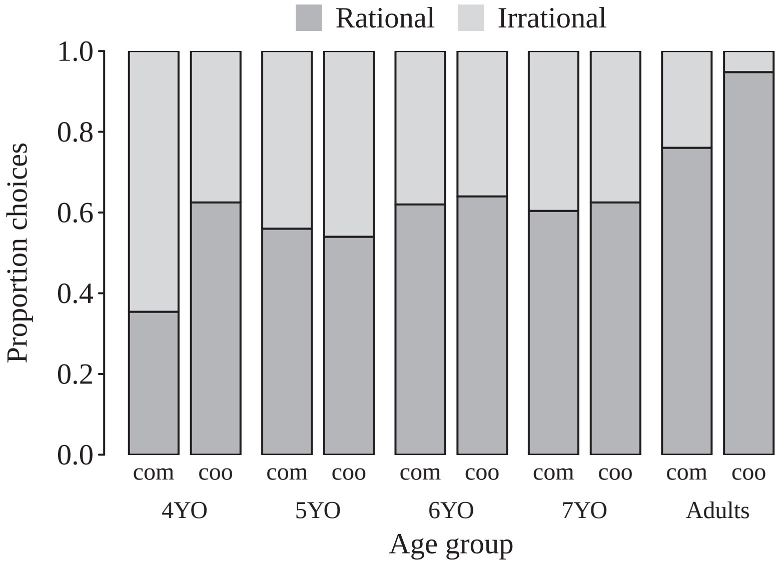

Figure 3.8: Children performed poorly in the domain-of-competence test trials. A context difference is only apparent in the four-year-olds and in the adults; these participants recruited more poorly in the competitive as opposed to the cooperative games.

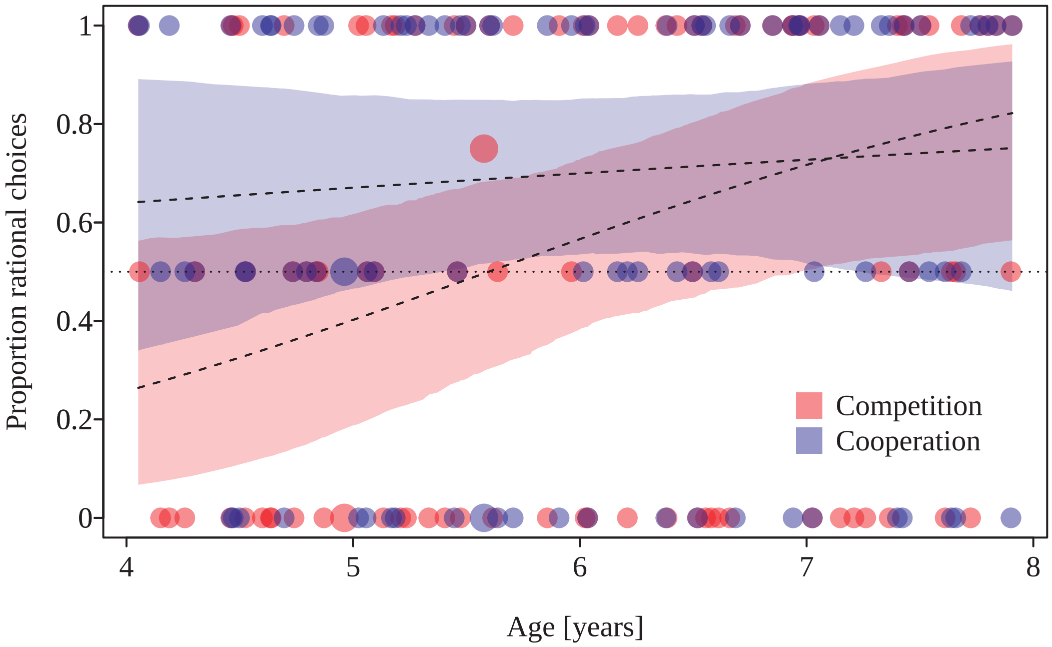

Figure 3.9: Data, model estimates (dotted lines) and 95% confidence intervals (shaded areas) for the domain-of-competence data. Children who selected the strong model for the cooperative strength game and the knowledgeable model for the cooperative knowledge game scored a "1" overall for the cooperative context. Children who selected the weak model for the competitive strength game and the ignorant model for the competitive knowledge game, scored a "1" overall for the competitive context. The faint horizontal line that intersects the y-axis at 0.5 indicates chance level decision-making; once the estimate and boundary of the lower confidence interval sit above this threshold, participants of this age are no longer selecting at chance levels.

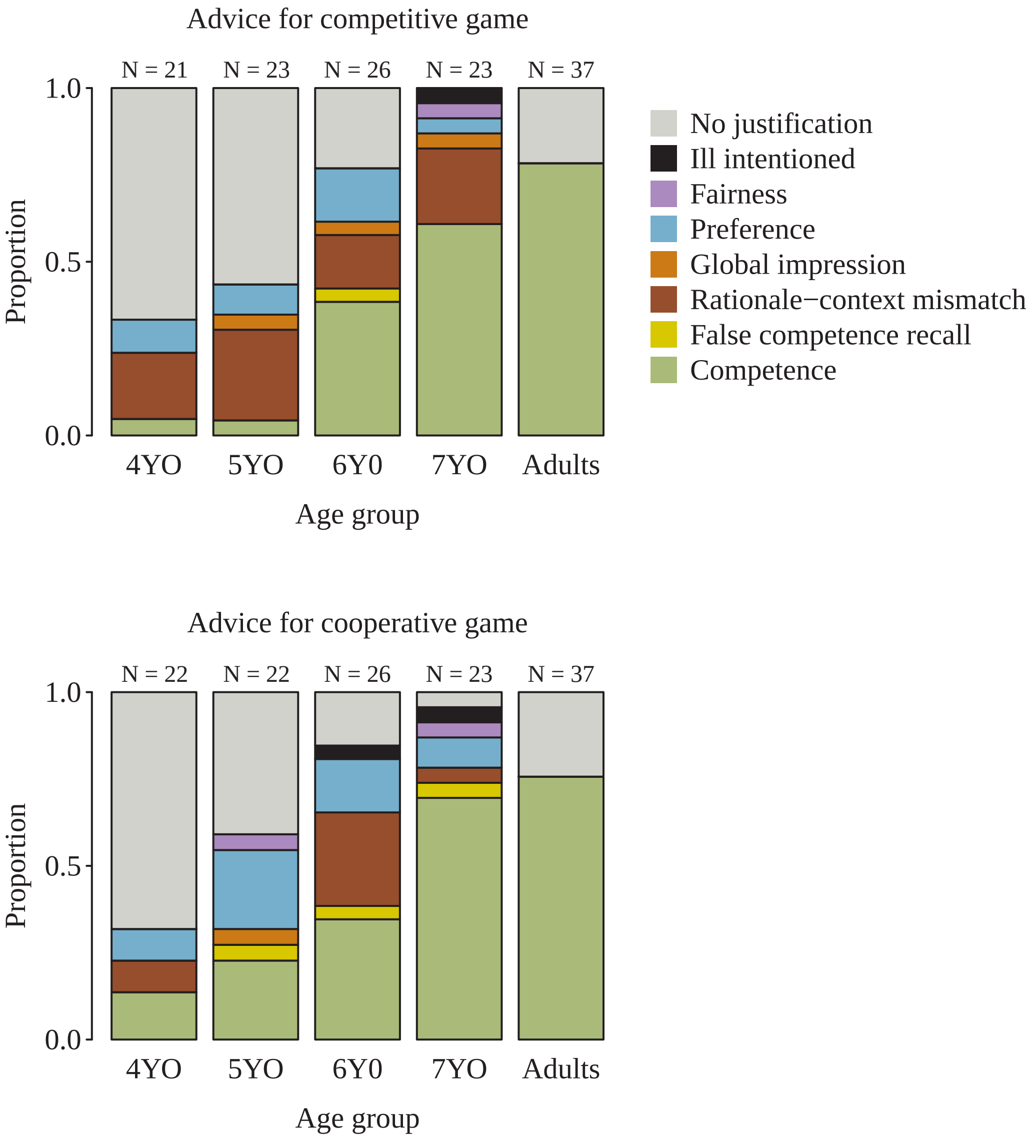

Figure 3.10: Two advice questions were used to probe the rationale behind participants' recruitment decisions in the domain-of-competence test trials (refer to the coding scheme in Table B.16 for category descriptions).

3.3.3 Discussion

In Experiment 2 we investigated whether trait-reasoning is evident in the competitive context, in addition to the cooperative context. We provided participants with contrasting information about two models in two domains – each model had a strength and a weakness (i.e., Model 1 was physically strong but inaccurate at object-labelling, vice-versa for Model 2). Experiment 2 proved inconclusive; the children did not select rationally at high levels in either of the contexts. The four-year-old participants, performed particularly poorly in the competitive context, and while the adult participants performed rationally in the cooperative context, we also observed a drop in their performance in the competitive context. Via GLMMs we established that neither context, nor age, nor an interaction between context and age had an effect on child recruitment behaviour, at least when Experimenter ID was included as a random effect in the analysis (likelihood ratio test: $p = 0.14$). When Experimenter ID was not included as a random effect in the model, we reported an effect of age ($p = 0.019$) but no effect of context ($p = 0.289$); older children were more likely to recruit rationally than younger children and recruitment behaviour was similar in the competitive and the cooperative contexts. Via a GLM, we established that context indeed had a significant effect on the behaviour of the adults ($p = <0.001$). The adults were significantly more likely to recruit rationally in the cooperative as opposed to the competitive context.

In light of the fact that the older children who took part in Experiment 1 performed well, the poor performance of the older children who took part in Experiment 2 raises questions. Ample studies have demonstrated that from the age of four years, children are capable of flexibly switching between models that show expertise in different domains in the cooperative context (e.g., Hermes, Behne, Studte, et al., 2016). Yet in the present experiment, not even the oldest children (seven-year-olds) performed well in the cooperative context. One possible explanation is that the memory demands of Experiment 2 affected performance. The children who took part in Experiment 2 had to memorize double the amount of character competence information compared to the children who took part in Experiment 1. Other studies that have tested domain-of-competence have implemented a between-subject design (Fusaro et al., 2011; Hermes, Behne, Studte, et al., 2016). Rather than all participants receiving the complete set of model information, half of the participants in these studies were familiarized with information about accuracy competence, the other half with information about strength competence. While we cannot rule out that memory played a role in the poor performance of the children in Experiment 2, the answers that the seven-year-old children provided in answer to the concluding advice questions (which pertained to the strength domain) imply that the children did not have difficulty in distinguishing between the models (at least based on their strength competence). The children viewed the strength demonstration videos early in the experiment and answered two advice questions at the very end of the experiment. Many of the seven-year-old children correctly referenced character competence in their answers to these advice questions. This implies that children of this age group did not perform poorly in the test trials due to memory overload. We turn then to an alternative explanation.

A second potential explanation for the poor performance of the children in Experiment 2 concerns the fact that children navigated the experiment in the virtual first-person and the fact that prior to their first test trial, participants had no information about how the knowledge or strength competence of their virtual character. Left in ignorance, it is possible that some children sacrificed success in a trial in order to find out more about their own character's competence, this applies especially to the competitive trials (since the teammate largely masks the performance of the virtual character in the cooperative context). From the social comparison literature we know that people tend to use agents who appear slightly more competent as comparison standards (Festinger, 1954). In our experiment, racing the character you perceive as strong is informative if you wish to learn more about your character's strength. Each child experienced four domain-of-competence test trials, one trial per domain-context combination, so in total two trials per context. With just these two test trials per context, a participant shows up as recruiting rationally at chance levels as soon as they make one suboptimal (investigative) recruitment decision. Curiosity and social comparison effects could theoretically explain why the four-year-olds and the adults performed poorly in the competitive test trials compared to the cooperative test trials.

The remainder of the participants, the five-, six- and seven-year-olds, performed equally poorly in the two contexts. Their performance therefore cannot be explained by social comparison effects and curiosity about the competence of their virtual character. Owing to the fact that the results of Experiment 1 (partially) and Experiment 2 (fully) deviate from the fairly robust selectivity patterns reported in studies conducted in-person, in the general discussion that follows, we discuss aspects of the online environment that may have influenced the children's behaviour.

3.4 General Discussion

The present study consisted of two experiments. In Experiment 1 we used a two-informant degree-of-competence paradigm to investigate whether rational partner selectivity and generalization behaviour are context-sensitive (whether these behaviours are restricted to cooperative contexts, or are also apparent in competitive contexts). We found that age but not context had a significant effect on the rationality of a child's recruitment choices. After observing two models that differed in terms of physical strength, the older children were more likely to make rational partner recruitment choices in the competitive and the cooperative strength games compared to the younger children. Neither age, nor context, had an effect on generalization behaviour. Children of all ages generalized often, and they did so in both contexts. From Experiment 1, we conclude therefore, that partner selectivity and generalization behaviour are not restricted to the cooperative context. In Experiment 2 we used a two-informant domain-of-competence paradigm to investigate whether trait-reasoning can explain partner selectivity in the competitive context. Experiment 2 was inconclusive. Neither age nor context (nor an interaction between age and context) had an effect on the likelihood that a child would flexibly switch between co-action partners depending on the domain and the context at hand. From four- through to seven-years of age, the children were at, or near, chance-level performance.

Depending upon the inference strategy children were using in the Experiment 1 competition strength games, we present tentative evidence that the dual-process account holds for the competitive context. The older children were selective in the competitive strength games and these same children generalized about competence in a competitive knowledge game. These behaviour patterns suggest that a Type-I process (global impression formation) and a Type-II process (sophisticated trait-reasoning) may be at play in the competitive context. Here we emphasize that we have tentative evidence as it is not possible to confirm via a degree-of-competence paradigm, whether the selectivity we observe is due to sophisticated trait-reasoning; attention to valence – an "oppose-the-worse" heuristic – would also explain the recruitment patterns we see.

Further research is needed before we can identify which inferential process underpins the selectivity we observed in the older children in Experiment 1. Whether this future research should take the form of an online experiment is unclear. We do not underestimate the contribution that online research has made, and will undoubtedly continue to make, to the developmental field. With the arrival of SARS-CoV-2, we have seen rapid advancement in terms of the methods and platforms that are available to run online test sessions (for a review see Tsuji et al., 2022 and Zaadnoordijk and Cusack, 2022). A growing number of studies document that the results achieved in an online test setting are comparable to those achieved in-presence in the laboratory (e.g., Nussenbaum et al., 2020; Schidelko et al., 2021). That said, a number of experiments document discrepancies between data collected on- versus offline (e.g., Bochynska and Dillon, 2021; Lapidow et al., 2021; Scott et al., 2017). For example, in a study that pre-dates COVID, Scott and colleagues adapted an in-presence study (Pasquini et al., 2007) and tested three- and four-year-olds in an online two-informant selective trust paradigm (Scott et al., 2017; Experiment 3). While the overall pattern of behaviour Scott and colleagues report was the same (three and four-year-olds preferred a more accurate above an inaccurate informant), the children tested online performed moderately worse compared to the children who took part in the original (offline) study. We highlighted the fact that in Experiment 1, in contrast to what we know from offline studies, the four-year-old children performed surprisingly poorly in the cooperative setting. Similarly, in Experiment 2, the five-, six- and seven-year-olds performed poorly in both co-action contexts. As we have alluded to, several aspects of the online test environment that we created in our studies may have masked the children's abilities to strategically recruit between the two models. In the paragraphs that follow we discuss several aspects of our experimental design that may have been problematic.

Firstly, as was briefly touched upon earlier, participants navigated the experiment in the virtual first-person. In a deviation from the standard two-informant study design, we provided feedback after each test trial; children watched their recruited co-action partner either race or collaborate with their own virtual character. It is possible that being assigned this virtual avatar, led to a certain level of detachment among participants. Children who were not emotionally invested in their character, might have been less motivated to recruit rationally. Prior to the test trials, we introduced each child to their character (to simplify the statistical analysis, each participant was assigned the same gender-neutral character). After the child had met their character, we asked two control questions to check that the children understood that they had an online representative. Children had two chances to pass the character control question. Five children in Experiment 1 and eight children in Experiment 2 were excluded for failing the character control questions. This fact, interpreted alongside the fact that some responses children gave to the two "identification" questions asked at the conclusion of Experiment 1 indicated that some children were reluctant to take ownership of their character (Figure 3.7, Q2), indicates that the virtual character was problematic. Fifteen children in Experiment 1, and 12 children in Experiment 2 required the second character control question. It cannot be ruled out that these children passed the character control question on the second attempt simply because they realized that a certain answer was expected of them by the experimenter.

For the children who were reluctant to embrace their character, the fact that the character was gender-neutral may have been an issue. There is evidence to suggest that the appearance and characteristics of an online avatar are important. Not only does the opportunity to personally customize an avatar positively influence levels of identification (Zarei et al., 2020), but scientists have reported that the outward appearance of avatars can mediate behavioural change in participants – this is referred to as the Proteus Effect (Yee and Bailenson, 2007). How the children viewed their avatar then, may have influenced their recruitment choices; if a child viewed their character negatively, they might not have felt that their character deserved to win. Future research that makes use of the virtual first-person, should consider allowing each participant to build their own virtual representative or, if this is not feasible, one could offer a selection of characters for participants to choose between.

An effect of experimenter ID was apparent in both Experiment 1 and Experiment 2. Three experimenters collected the data in each experiment (there were five different experimenter identities in total). In Experiment 1 we reported a significant effect of age, irrespective of whether Experimenter ID was included as a random effect in the GLMM analysis. We initially reported a significant effect of context (Model 1, p = 0.035) however, this significant finding disappeared once Experimenter ID was factored into the analysis (Model 1B, p = 0.153). When Experimenter ID was left out of the random effects structure of the model, context had a significant effect on the rationality of recruitment behaviour of the participants. Similarly, in Experiment 2, our findings also changed depending upon whether the random effect of experimenter ID was factored into the statistical analysis (likelihood ratio test statistic: p (Model 5) = 0.047; p (Model 5B) = 0.14). We made the decision to include experimenter ID in the random effects structure of each GLMM in light of a recent publication that indicated it is possible to include a factor with just three levels, as a random effect (Gomes, 2022). We were surprised to find that even in 'remote' research where one might think the experimenter could have little influence (in the present study the experimenters were only audible, not visible), there is still the danger of unaccounted for experimenter bias. For logistic reasons the use of multiple experimenter identities is common practice in the field of developmental psychology and despite awareness that experimenters can bias the results of studies (Gallo and Dale, 1968), it is not common to see the number of experimenters reported in papers or data divided up by experimenter ID in supplementary materials. We are aware that the use of a single experimenter comes with its own problems, as it might contribute to the generalizability crisis (Yarkoni, 2022). Yet, information about number of experimenter identities and how this detail has been controlled for in statistical analyses would help readers to assess the robustness of findings.

Via two online experiments we investigated whether children show partner selectivity and generalization behaviour in a competitive as well as in a cooperative context. We report that from approximately five years of age, children exhibit partner selectivity in both contexts. We additionally observed frequent generalization behaviour across the age range we tested (four- to seven-year-olds), again in both co-action contexts. Experimental refinement is needed before we can conclude that the recruitment selectivity we report in the competitive context is due to sophisticated trait-reasoning. Due to the SARS-CoV-2 pandemic, we ran these two experiments online, but it would be valuable to have an in-presence, laboratory-run comparison-standard for each of these experiments.

Data accessibility

Supporting material (the Electronic Supplementary Material (Appendix B), datasets and R code) have been made available on OSF at the following link:
https://osf.io/65fv3/?view_only=cb1dc0b626084a1780faf8e4db1c3d6d

Conflict of interest declaration

The authors have no conflict of interest to declare.

Authors' contributions

R.T.: conceptualization, formal analysis, investigation, methodology, resources, visualization, writing – original draft and writing – review and editing; J.H.: resources; J.F.: conceptualization, funding acquisition, methodology, investigation, supervision and writing – review and editing; H.R.: conceptualization, funding acquisition, methodology, investigation, supervision and writing – review and editing; S.K.: conceptualization, formal analysis, investigation, methodology, resources, visualization, supervision and writing – review and editing.

Funding

This work was carried out with funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; project number 254142454/GRK 2070).

Acknowledgements

We thank Imke Koch, Christiane Decker, Lia Künnemann, Josephine Quirbach and Mellory Kripzak for help with data collection and Eva-Maria Wellmer for help with reliability coding. Our thanks to Marlen Kaufmann for help with participant recruitment and Roger Mundry, for valuable statistical guidance. Thank you to the children, parents and students who volunteered to take part in these experiments.

Chapter 4

General Discussion

The aim of this dissertation was to further what we know about the social evaluation abilities of human and non-human primates. Study 1 (chapter 2) investigated how long-tailed macaques evaluate actions with respect to fairness. Study 2 (chapter 3) considered how young children evaluate and apply competence information in different co-action contexts. I begin this final chapter by briefly summarizing the main findings of these two studies (section 4.1). Following this, in sections 4.2 and 4.3, I situate my findings within the wider literature before reflecting on the directions that future studies could take in order to move research on social evaluation forward.

4.1 Summary of results

Study 1 investigated whether long-tailed macaques evaluate actions with respect to fairness. Specifically, Study 1 contrasted two competing social hypotheses that have been put forward to explain the food refusal patterns non-human primates show when tested in an "inequity aversion" paradigm. In this paradigm, a subject works to obtain rewards (in our case, the subject pulled a lever) and they do so alone or alongside a well-rewarded conspecific. The inequity aversion hypothesis contends that social comparison processes drive food refusal behaviour (Brosnan and de Waal, 2003). The social disappointment hypothesis contends that food refusals reflect discontent in the human distributor who distributes low- as opposed to high-value food (Engelmann et al., 2017). In four test conditions, we systematically manipulated partner presence (whether a well-rewarded conspecific was present or absent) and distributor type (whether a subject was rewarded by a human experimenter or a machine). Data from the test conditions revealed support for the social disappointment hypothesis. We reported a trend-level interaction between distributor and partner presence ($p = 0.075$) and a reduced model revealed a significant effect of distributor ($p = 0.003$). Subjects who were rewarded by the human refused food more often, in comparison to the subjects whose food was dispensed by a machine. At the conclusion of the experiment, each subject experienced one control condition in which both partner and subject received low-value food. The data from these control conditions indicated that social comparison processes may still play a role in mediating subjects' refusal behaviour. When the partner was also poorly paid for his efforts, subjects (in both distributor conditions) seldom refused the low-value food offered by the human or the machine. Finally, in an exploratory analysis we investigated pull-latency behaviour. Subjects had two minutes to pull the lever or they forfeit their low-value reward. Pull-latencies in the test conditions were shorter when a partner was present as opposed to when the subject was alone. This finding indicates that a third process – food competition – was also mediating our subjects' behaviour in this paradigm.

Study 2 investigated whether the inference processes thought to underlie children's selective trust (sophisticated trait-reasoning and simple heuristics), are discernible in the competitive as well as in the cooperative context. The cooperative context calls for the selective use of social information; an individual will succeed in a cooperative task if they can identify and recruit a competent teammate. Success in the competitive context depends upon a similar logic – an individual will succeed in a competitive task if they can identify and "recruit" an incompetent opponent. Study 2 consisted of two online experiments. Each experiment was run with four- to seven-year-old children. These experiments returned mixed results. Experiment 1 revealed that, given degree-of-competence information, the older children recruited rationally in the competitive as well as the cooperative context. Figure 3.2 showed that from the age of five onwards, the participants were selecting rationally significantly more often than would be predicted by chance. Experiment 1 additionally revealed that, given degree-of-competence information in a domain unrelated to the activity at hand (i.e., given strength information and then asked to engage in a knowledge activity), all age groups generalized about competence and did so at high levels in both co-action contexts. Experiment 2 revealed that given domain-of-competence information (i.e., given both strength and knowledge competence information about two models), and asked to play domain-relevant games, four- to seven-year-old children did not recruit rationally from between the two models in either the competitive or the cooperative context.

4.2 Fairness in non-human primates

Engelmann and colleagues (2017) were the first to systematically investigate the role of the human distributor in Brosnan and Waal's classic inequity aversion paradigm (2003). Engelmann *et al.* reported a significant interaction between distributor and partner presence, and post-hoc pairwise comparisons revealed that contrary to the predictions of the inequity aversion hypothesis, chimpanzee subjects refused low-value food more often in the absence rather than the presence of a conspecific (this behaviour was only evident in subjects who experienced the human distributor condition). Our test condition data neatly matches this chimpanzee data. We reported a trend-level interaction between distributor and partner presence, and a reduced model revealed a significant effect of distributor. Inspection of the raw data revealed that our long-tailed macaque subjects, like the chimpanzees, also refused food more often in the absence rather than in the presence of a conspecific. In our study, this food refusal pattern was apparent in subjects of both distributor groups (not just in subjects who experienced the human distributor condition, as was the case for Engelmann and colleagues). The study 1 test condition data provide clear support for the social disappointment hypothesis.

Massen and colleagues (2012) were the first, and to date are the only other group, to have investigated inequity aversion in the long-tailed macaque. Massen *et al.* ran two experiments in which they investigated how effort and relationship quality mediate food refusal behaviour in this species. They employed the classic "inequity aversion" paradigm – a subject and a partner monkey worked (reeled a baited platform within reach) to obtain rewards of different values; the subject always received low-value food, and the partner always received high-value food. In the first experiment, Massen *et al.* systematically manipulated the effort the monkeys had to expend in order to attain their rewards. There was a *no-effort* condition (the human provisioned the subject and partner by simply pushing the baited platforms within reach), a *small effort* condition (a 0.5 kg counterweight was attached to the platforms), a *large effort* condition (a 2.3 kg counterweight was attached to the platforms), and a differential effort condition (a 2.3 kg weight was attached to the platform of the subject, while the partner was provisioned by the experimenter). In each dyad that was tested, the subject was higher-ranking than the partner. This experiment revealed that subjects refused food more often in an inequitable *small effort* condition compared to an equitable *small effort* condition. A second experiment examined whether relationship quality mediates food refusal behaviour. It is hypothesized that dyads with a strong affiliative bond will refuse food less often under conditions of inequity, compared to dyads with a weak affiliative bond. In this second experiment, Massen and colleagues were able to replicate their findings from Experiment 1. When subjects had to expend moderate effort, they refused low-value food more often under inequitable as compared to equitable conditions (relationship quality had no influence on refusal behaviour). From these two experiments, Massen and colleagues concluded that dominant long-tailed macaques show disadvantageous inequity aversion albeit, only under certain conditions.

Our conclusion that social disappointment and food competition, rather than inequity aversion drive food refusal behaviour in our test conditions is at odds with Massen and colleagues' conclusion that long-tailed macaques show inequity aversion. If long-tailed macaques showed inequity aversion in our experiment then the subjects should have refused food in all inequitable conditions where the partner was present (i.e., in the *human distributor/partner present* condition and in the *machine distributor/partner present condition*) significantly more often than in the partner absent conditions. Both our raw food refusal data (monkeys refused food less often in the presence of a well-rewarded conspecific) and the pull-latency data (monkeys worked faster in the presence of a conspecific) show that if anything, our subjects were more, not less willing to accept low-value food in the presence (as opposed to the absence) of a partner. One key difference between our experiment and the experiment of Massen and colleagues, is that Massen *et al.* carefully selected their dyads such the partner monkey always ranked lower than the subject. The long-tailed macaque is a hierarchical species with a relatively limited capacity for social tolerance (Thierry, 2000). By ensuring that the partner was lower ranking than the subject, the experimenters created somewhat of an “unnatural” context that was more likely to result in protest behaviour. It has been suggested that in hierarchical species such as the long-tailed macaque, inequity averse protest behaviour may only be apparent in such a situation – where a dominant individual is poorly treated relative to a lower-ranked individual (Brosnan, 2006). Inequity averse protest behaviour by a low-ranking long-tailed macaque would be more likely to invite aggression than lead to any form of food-sharing.

We did not systematically manipulate the rank relationship between the subject and partner and it is possible that this detail accounts for the fact that our findings diverge from those of Massen and colleagues. We did not have the rank data for the social group at the time the experiment was run, but anecdotally we know that some subjects ranked higher, others lower, than the partner monkeys. The two monkeys, Mars and Lukas, that refused food most often in the test conditions were both high-ranking individuals (Mars was the alpha male) who both happened to be assigned to the human distributor group. It is possible therefore, that our support for the social disappointment hypothesis hinges on these two individuals, and is an artificial effect of rank. It would have been interesting to observe how these high-ranking monkeys would have responded in the machine distributor condition. A future experiment might manipulate distributor as a within-subject factor.

Our control condition data align with the findings of Massen and colleagues. We reported that under conditions of equality (i.e., once the partner received the same low-value food as the subject), refusal rates dropped. In light of the fact that we ran our equality control conditions at the end of the experiment however, one needs to interpret the control condition data with caution. Engelmann and colleagues did not run any equality control conditions, and we were motivated to replicate their experimental procedure as closely as possible. Our decision to run the control conditions at the end of the experiment was a compromise – we gathered equality condition data, but not at the expense of being able to directly compare our results with those of Engelmann and colleagues (2017). This compromise means we cannot cleanly interpret the food refusal behaviour in our equality conditions; we cannot rule out that the drop in refusal rate that we see in the *human distributor/partner present* control condition, as compared to the *human distributor/partner present* test condition, might be due to the subjects having tired of refusing food or having grown accustomed to receiving low pay. A future experiment might correct this condition presentation order limitation.

In light of our mixed findings – support for the social disappointment account, *tentative* support for the inequity aversion account and an effect of food competition – it seems logical to conclude that multiple factors mediate long-tailed macaque food refusal behaviour in the standard “inequity aversion” paradigm. A multifactorial finding such as this is hardly problematic. The inequity aversion and the social disappointment hypotheses are not mutually exclusive and indeed, there seems to be some evidence in the human literature to suggest that both accounts (inequity aversion and social disappointment) do co-exist (Sanfey, 2003 but see McAuliffe et al., 2013). Sanfey and colleagues found that participants who took part in an ultimatum game were more likely to refuse unequal offers that originated from a human, as opposed to a computer (2003). Sanfey *et al.*'s finding that the animacy of the distributor affected refusal rates does not cancel out the fact that humans engage in social comparison processes and protest unfair distributive acts, it just means that protest behaviour is amplified under certain social conditions.

The human experimenter has now been shown to affect food refusal behaviour in both a great ape species (chimpanzees) and in a catarrhine species (the long-tailed macaque). Future research that is interested in establishing whether a species shows inequity averse protest behaviour should omit the human factor from the experiment. One option would be to test animals in a non-social version of the classic "inequity aversion" paradigm (i.e., only use machine test conditions). A second option would be to employ a different type of task, one that is less reliant on the actions of a third agent. The *choice task*, an adapted version of the Dictator game, would be one option here (e.g., Fletcher, 2008; see Oberliessen and Kalenscher, 2019 for a list of the non-human species that have been tested in choice tasks thus far). In the choice task paradigm, the subject is faced with two options; the subject can choose between an equal reward outlay (e.g., subject and partner each receive 1 reward), and an unequal reward outlay that favours the partner (e.g., subject receives 1 reward and the partner receives 3 rewards). The subject makes this decision in both a social and a non-social condition (i.e., in the presence and in the absence of a conspecific). Species that are sensitive to reward division and that compare their own payoffs with those of conspecifics should prefer the equal reward scheme (1:1), but this preference should be weaker when the conspecific is absent. This choice task has the advantage that it removes the human agency from the picture; any inequality that is created originates from the actions of the subject. The ideal task with which to investigate inequity aversion in animals, is one that omits the possibility for food expectation effects, frustration effects and social disappointment in a human agent. Eliminating the potential for these effects to influence subjects will help scientists to more easily identify inequity-averse protest behaviour should it exist.

4.3 Competence evaluation by children

Study 2 is but one of a number of studies to have investigated the inference processes theorized to underlie children's trust decisions (e.g., Hermes et al., 2015; Hermes, Behne, Bich, et al., 2016; Hermes et al., 2018; Sobel and Kushnir, 2013). The recent proposal of Hermes *et al.*, that dual-processes may mediate children's selective trust behaviour, has the potential to bring together two strands of findings in the selective trust literature (2018). The first strand of research indicates that children's selective social learning – children's selective trust – is based on simple decision heuristics (e.g., Brosseau-Liard and Birch, 2010; Fusaro et al., 2011). The second strand of research indicates that children's selective social learning is rather based on a more sophisticated process, namely trait-reasoning (e.g., Hermes et al., 2015). Hermes *et al.* have presented some early evidence that their dual-process account may hold (Hermes et al., 2020).

Study 2 was an attempt to ascertain whether the dual-process account might extend beyond the cooperative context. Unfortunately, this pivotal experiment in our set of experiments (Experiment 2, the domain-of-competence paradigm), was inconclusive. Future work is needed before we can clarify whether trait-reasoning is evident in the competitive context. At the close of chapter 3, we discussed at some length several possible reasons for the null findings in Experiment 2, and the particularly poor performance of the 4- year-old participants in Experiment 1. Some points that were touched upon were: (1) the fact that these were online experiments, (2) the fact that we used the virtual first-person and (3) the fact that the virtual character the children were assigned was gender-neutral. These are all details that future online studies that examine selective recruitment may want to take into consideration. In addition to these three points, a further aspect of study design that may be important for future work concerns incentivization.

We motivated our participants by offering a virtual prize, a trophy, for each rational recruitment decision. At the conclusion of the experiment, the experimenter showed each child the trophies they had collected over the course of the experiment. It is well-documented that two types of motivation govern behaviour (summarized in Ryan and Deci, 2000). There is intrinsic motivation, which is a self-propelled drive to perform. In the case of intrinsic motivation a person's interest in the task is high and a sense of enjoyment or satisfaction is derived from completing the task. The second type of motivation is extrinsic motivation. In the case of extrinsic motivation someone else, an external influence, provides the impetus to perform and the person's interest level in the task is much lower. We created an extrinsic motivational climate in our two experiments – the experimenter told the child that their goal was to collect as many trophies as possible. A number of developmental studies indicate that offering external rewards can alter children's behaviour (e.g., Ulber et al., 2016; Warneken and Tomasello, 2008). Future studies that look at children's recruitment behaviour, need to carefully consider how they implement incentives – ideally it will be possible to design a study that taps intrinsic motivation, as under "natural" conditions, in everyday competitive and cooperative interactions, it is intrinsic motivation that is mediating people's social selectivity.

Many of the online studies that were run during the pandemic already have in-presence equivalents (e.g., Schidelko et al., 2021). In our case there is, as yet, no comparable in-presence data. The online experience we created was relatively passive, and we relied on a verbal response measure. An interactive, in-presence experiment would likely be more engaging for the children, and would have the added benefit that we could use more implicit behavioural response measures. Future research that translates this study paradigm to the laboratory setting would provide a valuable comparison standard – not just for developmental psychologists, but also for scientists working on comparative cognition for whom a task with a verbal response is no option.

4.4 Conclusion

The two studies in this dissertation have sought to move our understanding of the social evaluation abilities of human and non-human primates forward. Study 1 contributes to the now expansive body of research that has investigated whether nonhuman primates are sensitive to fairness. Study 1 confirmed that the presence and actions of the human experimenter in the classic “inequity aversion” paradigm influences the food refusal behaviour of a second non-human primate species – the long-tailed macaque. Study 2 was a first step toward understanding how children apply competence information in the competitive context, and although the results were inconclusive, this study serves as a prototype that future studies can build upon. An in-presence version of Study 2 would create a valuable reference dataset for the psychologists and primatologists who seek to better understand primate social cognition.

Appendix A

Supplementary material for: “Social disappointment and partner presence affect long-tailed macaque refusal behaviour in an ‘inequity aversion’ experiment”

Article:
Titchener R, Thiriau C, Hüser T, Scherberger H, Fischer J, Keupp S. 2023 Social disappointment and partner presence affect long-tailed macaque refusal behaviour in an ‘inequity aversion’ experiment. *R. Soc. Open Sci.* 10: 221225. https://doi.org/10.1098/rsos.221225

Supplementary material has been made available online:
https://royalsocietypublishing.org/doi/suppl/10.1098/rsos.221225

Appendix B

Supplementary material for: “Social evaluation strategies in four- to seven-year-old children: Age-related changes in partner choice in competitive and cooperative settings”

Contents

B.1 Experiment 1

B.1.1 Procedure

There were seven stages to Experiment 1: (1) a demonstration phase where the child was introduced to two animated characters, (2) a character assignment phase where the child was introduced to the character who would represent them onscreen, (3) a familiarization phase where the child learned the rules and possible outcomes of the games they would encounter in the test trials, (4) six "degree-of-competence" test trials, (5) a "generalization" phase in which the child was introduced to a novel activity in the knowledge domain (object-labelling) and the child experienced one test trial in this novel domain, before moving on to (6) a series of open-ended questions designed to probe the extent to which the child had identified with their assigned character and (7) a series of open-ended questions to see whether the child could outline the reason behind their co-action partner recruitment choices in the strength games and in the knowledge game. We elaborate on each of these seven stages below.

1. Demonstration phase

In a series of short videos (see Figure B.1), we introduced the child to two animated characters – Mr. Blue and Mr. Green. Mr. Blue was portrayed as strong, Mr. Green as weak. The child watched three video clips per character. Mr. Blue succeeded in each strength task, while Mr. Green struggled and always had to settle for a work-around. After watching the videos, the child was asked two control questions to check that they had correctly linked strength competence and character. We asked the child to name the character who had performed the tasks in the videos well and to name the character who had struggled. Children who were unable to answer both control questions correctly were shown two extra demonstration videos and the questions were repeated. Three children needed the two extra videos but ultimately all children were able to correctly discern the weak from the strong character.

2. Character assignment phase

We assigned the child an animated character who would represent them onscreen in the test trials (the same character was used for all participants). The child watched their character engage in a couple of different activities (dancing and playing soccer) and was then asked a control question to make sure they understood that the character on the screen was "them". In the control question, the child saw their character jumping skip-rope alongside a novel character playing the drums. The experimenter asked the child to state the activity associated with their character. If the child answered incorrectly, the experimenter corrected them and the child was given another chance – this time they saw their character doing jumping-jacks alongside the "novel" character, who was eating ice-cream. If the child was still unable to correctly state the activity associated with their character, their data were excluded from analysis. Fifteen children required the second character identification question. Five children were excluded for failing the character control questions.

3. Game familiarization phase

With the aid of four short video clips containing black-and-white figures, the experimenter explained how a competitive and a cooperative version of a strength game worked. In the competitive version of the strength game, two characters raced one another to shift a heavy object from left to right across the screen; the character that was first to reach the right-hand side won a virtual trophy. In the cooperative version of the strength game, two characters worked together to shift a heavy object. To win, the characters needed to succeed at the task before a timer ran out. We counterbalanced the context that was introduced first; 53 of the 107 children learned about the competitive game first. After these two versions of the strength game had been explained and the child had seen the two possible outcomes of each game (a win and a loss), the child was ready to move on to the degree-of-competence test trials.

4. "Degree-of-competence" test trials

There were a total of six degree-of-competence test trials - three competitive and three cooperative strength games (see Figure B.2). Prior to the first test trial, the experimenter told the child that their goal was to collect as many virtual prizes as possible. The two contexts were presented alternately, i.e., the presentation order was competition-cooperation-competition-cooperation-competition-cooperation or vice versa. Prior to each game, the experimenter told the child which context they were about to encounter. In addition, we provided a visual cue; a pink background cued the competitive context, a yellow background the cooperative context. Also prior to each game the experimenter showed the child the heavy object that they would have to shift. We used three different objects (listed in order of presentation): a chest of drawers, a stack of books and a treasure chest. In each test trial, the child was asked to select a co-action partner from between Mr. Blue and Mr. Green. The child then watched the consequence of their choice – whether their character won or lost the game.

5. "Generalization" phase

At this stage we switched from the strength domain to a knowledge domain (object labelling). To introduce the new domain, we showed the child several video clips in which their character correctly labelled some common objects (a ball, a car and a flower) and then a video in which their character was unable to name a tricky, fictitious object.

The child then played one final game, a knowledge game – this was the generalization test trial (see Figure B.3). The experimenter told the child the context that they were about to encounter (competition or cooperation). Here we retained the visual cue we had used in the degree-of-competence games – pink cued competition and yellow cooperation. Prior to the game, the experimenter showed the child the object their character would have to label. In the competitive version of the game, two characters strove to be the first to correctly label the object. In the cooperative version of the game, the child selected a teammate to help them come up with the object's name before a timer ran out. The child was asked to select a co-action partner from between Mr. Blue and Mr. Green. Importantly, the child had no information about the competence of Mr. Blue and Mr. Green in the knowledge domain. A virtual reward was again at stake; however, to end the experiment on a positive note, the game was rigged in favour of the child, i.e., the recruited character supplied the child with an accurate label in the cooperation setting, and was too slow to call out the label in the competitive setting.

6. "Identification" questions

We asked the child two questions to gauge whether they had identified with their assigned character. Firstly, we asked the child to call out the name of some common objects and some people (Mr. Blue, Mr. Green, and their own assigned character) that appeared on the screen one after the other. Here we were interested in the answer the child would supply when their assigned character appeared on the screen. Secondly, we showed the child a still-image taken from one of the competition strength games and asked them to describe what they saw. The child's character was depicted in the still-image. We were interested in whether the child would mention their character, and if so, how they would refer to it, i.e., whether they would use a first-person reference (e.g., "that's me") or a vague third-person reference (e.g., "a person").

7. "Justification" questions

We concluded the experiment by asking the child two questions to see whether the children could offer a reason for their recruitment choices in the seven test trials. In the first justification question, we reminded the child of the strength games (the degree-of-competence test trials) and asked if they tell us why they had recruited Mr. Blue and Mr. Green there (in case the child had only ever recruited one character, the question was adjusted to reflect this). In the second justification question we reminded the child of the object-labelling game (the generalization test trial), and asked if they could tell us why they had elected to play with [Mr. Blue/Mr. Green].

B.1.2 Stimuli

All stimuli were created using the *Vyond* animation software. The scenarios used in the strength demonstration videos are depicted in Figure B.1. The degree-of-competence test trials are shown in Figure B.2 and the generalization test trials in Figure B.3. The pink images in Figures B.2 and B.3 show the competitive context; yellow images show the cooperative context.

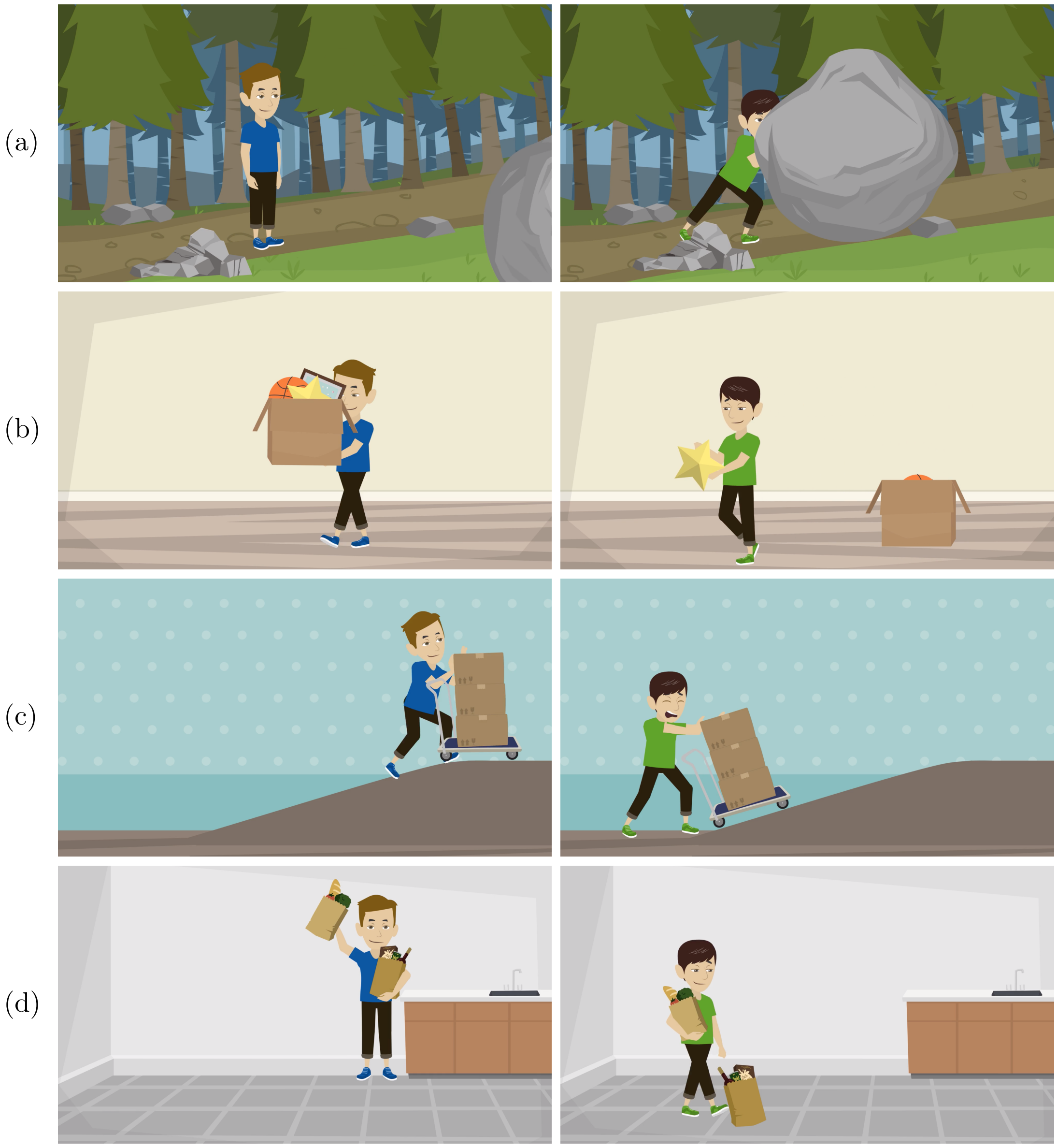

Figure B.1: Still-images from the strength demonstration scenarios. All children saw scenarios (a) through (c). Only children who failed a set of control questions were shown scenario (d).

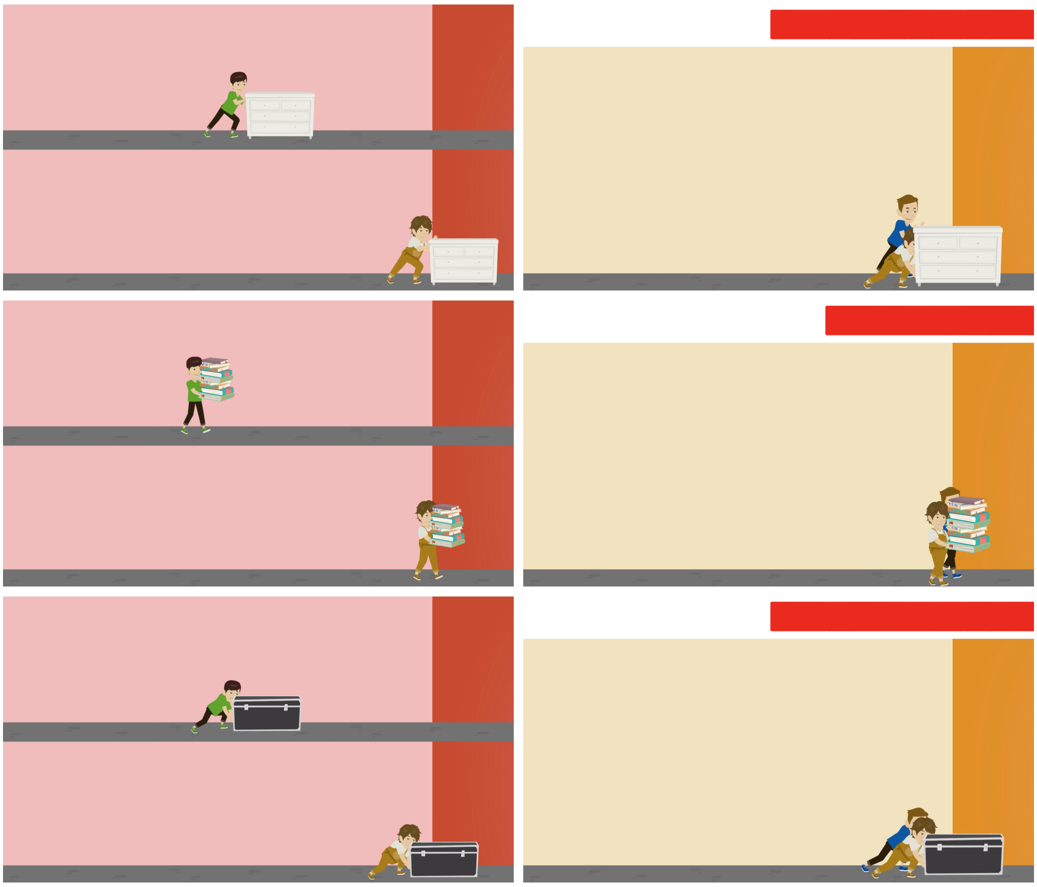

Figure B.2: The six degree-of-competence test trials. In these strength games, the participant had the opportunity to recruit rationally from between Mr. Blue and Mr. Green in the competitive and the cooperative context. These six still-images depict rational recruitment behaviour; if a participant is motivated to win virtual prizes, then they should recruit the weak character (Mr. Green) for the competitive context and the strong character (Mr. Blue) for the cooperative context.

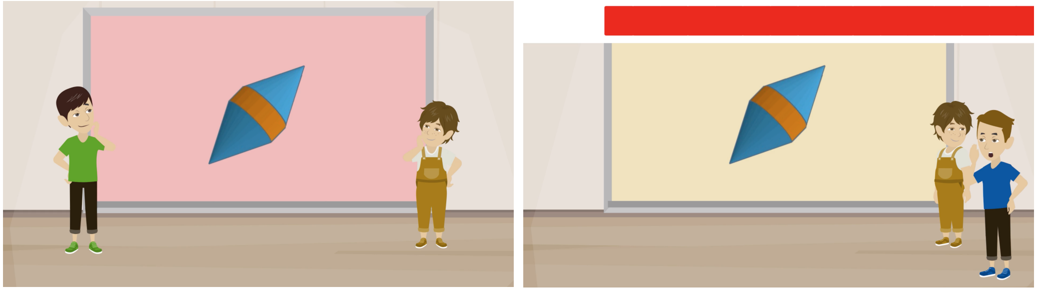

Figure B.3: The two generalization test trials. Each participant experienced either the competition or the cooperation knowledge game. These two still-images depict generalization behaviour; if a participant generalized information about strength competence to the unrelated knowledge domain, they would recruit the weak character (Mr. Green) for the competitive knowledge game or the strong character (Mr. Blue) for the cooperative knowledge game.

B.1.3 Confirmatory analyses

All analyses were conducted in R (R-Core-Team, 2021, version 4.1.1). We fitted a total of six models. Three Generalized Linear Mixed Models (GLMMs; Baayen, 2008) were used to analyse the degree-of-competence data. Two Generalized Linear Models (GLMs; Baayen, 2008) and one GLMM were used to analyse the generalization data. The majority of these analyses were preregistered on OSF (https://archive.org/details/osf-registrations-8r2hs-v1). We highlight the instances where it was necessary to deviate from our preregistration.

All GLMMs were fitted using the *glmer* function of the *lme4* package (Bates et al., 2015; version $1.1 - 27.1$), and to aid convergence we used the *bobyqa* optimizer. In the case of the GLMMs, 95% confidence intervals were derived using the *bootMer* function of the *lme4* package. The GLMs were fitted using the *glm* function of the *lme4* package and 95% confidence intervals were derived using the *confint* function of the *stats* package.

The response variable in each model was binary. All models were therefore fitted with a binomial error structure and logit link function (McCullagh and Nelder, 1989). Covariates included in the models were z-transformed to have a mean of zero and a standard deviation of one. In order to keep type I error rates low (at the nominal level of 0.05), and to avoid an 'over-confident' model, where possible we included random slopes in the models (Barr et al., 2013; Schielzeth and Forstmeier, 2009). Factors that appeared in the random slopes structure of a model were dummy-coded and centred prior to inclusion.

As per our preregistration, we originally intended to include a "strength-of-ID" index as a fixed effect in the models used to analyse the child data. However, non-identification with the virtual assigned character proved rare (only three from 110 children did not identify with the character) and including this variable led to issues with complete separation (Field, 2005). Our solution was to exclude the data of these three children from analysis, rendering inclusion of the "strength-of-ID" variable in the child models redundant.

For each model that was fitted, we used a likelihood ratio test (Dobson, 2002) to establish the overall effect(s) of our factor(s) of interest. Use of the likelihood ratio test meant that in cases where there were multiple factors of interest, we avoided running into issues with 'cryptic multiple testing' (Forstmeier and Schielzeth, 2011).

Prior to inference we assessed model stability and checked for absence of collinearity. In the case of the GLMMs, we checked model stability by dropping the levels of the random effects (one at a time) and comparing the estimates derived from each reduced dataset with those obtained from the full dataset. In the case of the GLMs, we used DFBeta values to evaluate the range of estimated coefficients that were obtained following case-wise data deletions. In each case model stability was fine, and we provide information on stability in each table that documents a full model (this information is stored in table columns headed *min.* and *max.*).

Prior to inference we also checked for absence of collinearity. We inspected Variance Inflation Factors for each model (VIFs; Field, 2005). These VIFs were obtained using the *vif* function of the *car* package (Fox and Weisberg, 2011); in each case this function was applied to a linear model lacking random effects. The largest VIF we encountered was 1.016; that is to say, collinearity was not an issue in any of the models we fitted.

P-values are often retrieved using the *summary* function in R. More reliable *p*-values are derived by dropping the fixed effects from the model one at a time and comparing the full model with each respective reduced model (achieved using the *drop1* function in R). All *p*-values we report were obtained via this second approach.

Three experimenters collected the child data, and we therefore needed to be able to rule out an effect of experimenter ID. In our preregistration, we had proposed that we would split the child data according to experimenter ID and run our analyses with each respective experimenter data subset. The aim was to qualitatively compare the estimated coefficients generated by these respective models. Unfortunately in running the analysis with these smaller subsets, we encountered issues with complete separation. In light of a recent publication (Gomes, 2022) that indicated it is possible to include a three-level factor as random effect, we ran each child analysis ("Model 1" and "Model 3") as originally preregistered (as detailed in subsections below), but we also ran supplementary analyses in which we controlled for a potential effect of experimenter ID.

All Experiment 1 confirmatory analyses are outlined in detail in the subsections below. The data and R-scripts used in the analyses have been made available on OSF (https://osf.io/65fv3/?view_only=cb1dc0b626084a1780faf8e4db1c3d6d). A number of R functions (written by Roger Mundry) were sourced at various stages of these analyses. These functions are available on Zenodo (https://doi.org/10.5281/zenodo.7670524).

Models 1/1B: Analysis of child degree-of-competence data

To estimate the effect of context and age on children's partner recruitment behaviour, we constructed Model 1 (a GLMM). The response variable was partner recruitment choice. The model comprised a two-way interaction between context and age, in addition to their fixed main effects and the fixed effects of gender, order (whether the competition or the cooperation test trial appeared first) and trial number (1 – 6; note that inclusion of trial number as a fixed effect was overlooked in the preregistration). Participant ID was included as a random effect given that repeated measures were taken from the same individuals. The random slopes of context and trial number within participant ID were included in the model. Originally, we included estimates of the correlations between the random intercept and slopes in the model, however as a correlation was high, we went on to remove these from the model. This led to a slight decrease in model fit (log-likelihoods: full model with correlations: -288.378; full model lacking the correlations: -288.721). A "singular fit" message indicated that some random effects terms were unidentifiable, however, as we were interested in the fixed effects this message was not a concern.

We used a likelihood ratio test to establish the overall effect of context, age and the interaction between context and age. Here we compared the full model with a null model lacking the interaction term and their main effects. The likelihood ratio test revealed the two models differed significantly from one another ($\chi^2_3 = 37.125$, $p = <0.001$). The output of the full model (Table B.1) revealed the interaction term (context × age) was not significant ($p = 0.376$). We therefore fitted a reduced model (output in Table B.2). The reduced model revealed a significant effect of context ($p = 0.035$) and of age ($p = <0.001$).

To estimate the effect of context and age on children's partner recruitment behaviour, while controlling for a potential effect of experimenter ID, we constructed Model 1B (a GLMM). Model 1B was near-identical to Model 1, the two models differed only in terms of their random effects and random slopes structures. In Model 1B, in addition to participant ID, experimenter ID was included as a random effect. As a consequence, not only the random slopes of context and trial number within participant ID but also the context × age interaction term, gender, order, and trial number within experimenter ID were included in the model. We did not include estimates of the correlations between the random intercept and slopes in the model. A "singular fit" message indicated that some random effects terms were unidentifiable, however, as we were interested in the fixed effects this message was not a concern.

We used a likelihood ratio test to establish the overall effect of context, age and the interaction between context and age. Here we compared the full model with a null model lacking the interaction term and their main effects. The likelihood ratio test revealed the two models differed significantly from one another ($\chi^2_3 = 11.643$, $p = 0.009$). The output of the full model (Table B.3) revealed the interaction term (context × age) was not significant ($p = 0.370$). We therefore fitted a reduced model (output in Table B.4). The reduced model revealed no significant effect of context ($p = 0.153$) but a significant effect of age ($p = 0.003$). Given that the output of reduced Model 1B differed from that of reduced Model 1, we conclude there is an effect of experimenter ID, and for this reason we discuss the results of reduced Model 1B as our main finding in the article.

Table B.1: Model 1 (full model)
Preregistered analysis of child degree-of-competence data (642 trials from 107 children): Testing the effect of the interaction between context and age and their main effects, as well as the fixed effects of gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	1.293	0.341	0.668	2.197			(1)	1.207	1.391
Context (coop)	0.577	0.258	0.115	1.166			(1)	0.509	0.649
Age(2)	0.960	0.220	0.567	1.585			(1)	0.900	1.025
Gender (male)	0.841	0.363	0.127	1.659	5.536	1	0.019	0.736	0.928
Order (coop)	-0.033	0.355	-0.732	0.730	0.009	1	0.926	-0.140	0.056
Trial number(3)	0.247	0.160	-0.076	0.586	2.396	1	0.122	0.209	0.284
Context:Age	0.222	0.252	-0.283	0.781	0.783	1	0.376	0.167	0.316

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2225.178 and 420.584 respectively.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 3.5 and 1.709 respectively.

Table B.2: Model 1 (reduced model)
Preregistered analysis of child degree-of-competence data (642 trials from 107 children): Testing the effects of context and age, as well as the fixed effects of gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals and test results).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value
Intercept	1.313	0.341	0.704	2.214			(1)
Context (coop)	0.507	0.243	0.039	1.022	4.444	1	0.035
Age (2)	1.056	0.194	0.708	1.566	31.808	1	<0.001
Gender (male)	0.840	0.362	0.138	1.663	5.537	1	0.019
Order (coop)	-0.029	0.355	-0.741	0.730	0.007	1	0.936
Trial number(3)	0.246	0.159	-0.078	0.573	2.389	1	0.122

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2225.178 and 420.584 respectively.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 3.5 and 1.709 respectively.

Table B.3: Model 1B (full)
Non-preregistered analysis of child degree-of-competence data in which we account for the fact that three experimenters collected the data: Testing the effect of the interaction between context and age and their main effects, as well as the fixed effects of gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	1.316	0.371	0.670	2.255			(1)	0.788	1.700
Context (coop)	0.580	0.313	-0.014	1.249			(1)	0.246	0.809
Age(2)	0.961	0.220	0.575	1.536			(1)	0.759	1.230
Gender (male)	0.837	0.362	0.170	1.688	4.859	1	0.027	0.691	0.986
Order (coop)	-0.044	0.419	-0.943	0.815	0.011	1	0.916	-0.575	0.355
Trial number(3)	0.249	0.160	-0.057	0.615	2.413	1	0.120	0.211	0.286
Context:Age	0.226	0.253	-0.311	0.769	0.803	1	0.370	0.164	0.328

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2225.178 and 420.584 respectively.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 3.5 and 1.709 respectively.

Table B.4: Model 1B (reduced model)
Non-preregistered analysis of child degree-of-competence data in which we account for the fact that three experimenters collected the data: Testing the effects of context and age, as well as the fixed effects of gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals and test results).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value
Intercept	1.334	0.371	0.678	2.253			(1)
Context (coop)	0.508	0.299	-0.047	1.233	2.038	1	0.153
Age(2)	1.058	0.194	0.697	1.568	8.771	1	0.003
Gender (male)	0.837	0.361	0.164	1.567	4.859	1	0.027
Order (coop)	-0.039	0.417	-0.856	0.811	0.009	1	0.925
Trial number(3)	0.248	0.160	-0.085	0.583	2.405	1	0.121

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable were 2225.178 and 420.584 respectively.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable were 3.5 and 1.709 respectively.

Model 2: Analysis of adult degree-of-competence data

To estimate the effect of context on adults' partner recruitment behaviour, we constructed Model 2 (a GLMM). The response variable was partner recruitment choice. The model comprised the main effect of context as well as the fixed effects of gender, order (whether the competition or the cooperation test trial appeared first) and trial number (1 – 6). Participant ID was included as a random effect. The random slopes of context and trial number within participant ID were included in the model. Originally we included estimates of the correlations between the random intercept and slopes in the model, however, to match our approach in Model 1 we removed these from the model. This led to a slight decrease in model fit (log-likelihoods: full model with correlations: -50.798; full model lacking the correlations: -51.505).

We used a likelihood ratio test to establish the effect of context. Comparison of the full model with a null model lacking our factor of interest (i.e., lacking context) revealed the two models were not significantly different from one another ($\chi^2_3 = 0.562$, $p = 0.454$). The output of the full model is provided in Table B.5.

Table B.5: Model 2 (full model)
Analysis of adult degree-of-competence data (246 trials from 41 participants): Testing the effect of context, and the fixed effects of gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	4.728	2.826	7.686	23.869			[1]	3.268	8.762
Context (coop)	0.778	0.899	-3.322	8.369	0.562	1	0.454	0.457	1.123
Gender (male)	-1.134	0.915	-4.018	1.207	1.074	1	0.300	-1.448	-0.552
Order (coop)	0.880	0.891	-2.382	4.319	0.717	1	0.397	0.361	1.154
Trial number[2]	0.661	0.451	-0.332	7.613	1.535	1	0.215	0.563	0.956

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 3.5 and 1.711 respectively.

Models 3/3B: Analysis of child generalization data

To estimate the effect of context and age on the likelihood that children generalize and ascribe competence or incompetence beyond the strength domain, we constructed Model 3 (a GLM; in the preregistration we erroneously stated this would be a GLMM). The response – "generalization" – was a binary variable. Selection of the physically weak model for the competition knowledge game or the strong model for the cooperation knowledge test trial constituted generalization behaviour. The model comprised a two-way interaction between context and age, in addition to their fixed main effects and the fixed effects of gender and the Best Linear Unbiased Predictors (BLUPs) extracted from Model 1. These BLUPs were a proxy measure for how rationally the child had chosen partners in the six degree-of-competence test trials.

We used a likelihood ratio test to establish the overall effect of context, age and the interaction between context and age. Here we compared the full model with a null model lacking the interaction term and their main effects. The likelihood ratio test revealed a borderline significant difference between the two models ($\chi^2_3 = -6.923$, $p = 0.074$). The output of the full model (Table B.6) revealed the interaction term (context × age) was not significant ($p = 0.138$). As the full-null model comparison revealed a borderline statistic, we proceeded with a reduced model (output in Table B.7). The reduced model revealed no significant effect of context ($p = 0.109$) and no significant effect of age ($p = 0.153$).

To estimate the effect of context and age on the likelihood that children generalize and ascribe competence or incompetence beyond the strength domain, while controlling for a potential effect of experimenter ID, we constructed another GLMM, Model 3B. Although the structure of Model 3B closely approximated that of Model 3, the models differed in two regards. Firstly, Model 3B included the random effect of experimenter ID – this random effect was absent from Model 3. Secondly, the BLUPs used in Model 3B were extracted from Model 1B (for Model 1 the BLUPs were taken from Model 1). The random slopes of the context × age interaction term, gender and BLUPs within experimenter ID were included in the model. We did not include estimates of the correlations between the random intercept and slopes in the model. A "singular fit" message indicated that some random effects terms were unidentifiable, however, as we were interested in the fixed effects this message was not a concern.

We used a likelihood ratio test to establish the overall effect of context, age and the interaction between context and age. We compared the full model with a null model lacking the interaction term and their main effects. The likelihood ratio test revealed the two models did not differ significantly from one another ($\chi^2_3 = 4.415$, $p = 0.22$). The output of the full model is provided in Table B.8. We discuss the results of Model 3B in the main article.

Table B.6: Model 3 (full model)
Analysis of child generalization data (107 trials from 107 participants): Testing the effect of the interaction between context and age and their main effects, as well as the fixed effects of gender and BLUPs on the probability to generalize competence beyond the strength domain (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	1.428	0.485	0.542	2.470			(1)	1.249	1.522
Context (coop)	0.797	0.589	-0.346	2.010			(1)	0.643	0.963
Age(2)	0.695	0.359	0.024	1.455			(1)	0.526	0.802
Gender (male)	0.070	0.562	-1.050	1.181	0.016	1	0.901	-0.049	0.202
BLUPs(3)	0.675	0.274	0.156	1.244	6.544	1	0.011	0.569	0.801
Context:Age	-0.849	0.580	-2.032	0.272	2.200	1	0.138	-1.028	-0.690

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2225.178 and 422.234 respectively.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was -0.157 and 0.806 respectively.

Table B.7: Model 3 (reduced model)
Analysis of child generalization data (107 trials from 107 participants): Testing the effect of context and age, and the fixed effects of gender and BLUPs on the probability to generalize competence beyond the strength domain (table shows estimates, standard errors, confidence intervals and test results).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value
Intercept	1.302	0.450	0.468	2.252			(1)
Context (coop)	0.890	0.568	-0.194	2.068	2.574	1	0.109
Age(2)	0.383	0.272	-0.141	0.936	2.044	1	0.153
Gender (male)	0.141	0.551	-0.952	1.236	0.065	1	0.798
BLUPs(3)	0.656	0.267	0.150	1.210	6.499	1	0.011

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2225.178 and 422.234 respectively.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was -0.157 and 0.806 respectively.

Table B.8: Model 3B (full model)
Analysis of child generalization data (107 trials from 107 participants) in which we account for the fact that three experimenters collected the data: Testing the effect of the interaction between context and age, and their main effects, as well as the fixed effects of gender and BLUPs on the probability to generalize competence beyond the strength domain (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	1.642	0.673	0.596	4.135			(1)	1.017	2.117
Context (coop)	0.720	0.931	-1.297	5.113			(1)	-0.293	2.651
Age(2)	0.794	0.477	-0.036	2.352			(1)	0.504	1.041
Gender (male)	0.042	0.583	-1.484	1.430	0.005	1	0.943	-0.493	0.881
BLUPs(3)	0.620	0.288	0.074	1.559	4.231	1	0.040	0.534	0.893
Context:Age	-0.955	0.704	-3.380	0.757	1.514	1	0.219	-1.492	0.470

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was -0.261 and 0.628 respectively.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was -0.157 and 0.806 respectively.

Model 4: Adult generalization data

To estimate the effect of context and age on the likelihood that adults generalize and ascribe competence or incompetence beyond the strength domain, we constructed Model 4 (a GLM). The response variable was binary (generalization = 1). The model comprised the fixed effect of context, as well as the fixed effects of gender and the BLUPs extracted from Model 2.

We used a likelihood ratio test to establish the overall effect of our factor of interest (context). Comparison of the full model with a null model that lacked our factor of interest (i.e., that lacked "context") revealed that the two models were not significantly different from one another ($\chi^2_3 = -1.127$, $p = 0.289$). The output of the full model is provided in Table B.9.

Table B.9: Model 4 (full model)
Analysis of adult generalization data (41 trials from 41 participants): Testing the effect of context, in addition to the fixed effects of gender and BLUPs on the probability to generalize competence beyond the strength domain (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	0.271	0.553	-0.814	1.394			(1)	0.008	0.451
Context (coop)	-0.673	0.639	-1.957	0.568	1.127	1	0.289	-0.824	-0.508
Gender (male)	-0.158	0.642	-1.438	1.103	0.061	1	0.805	-0.304	0.048
BLUPs(2)	-0.153	0.326	-0.827	0.488	0.224	1	0.636	-0.355	0.023

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was -0.261 and 0.628 respectively.

B.1.4 Exploratory analyses

We concluded the experiment by asking each participant four questions (the children responded verbally, the adults typed their answers). There were two "identification" questions and two "justification" questions. Participants' answers were explored in two ways: (1) we categorized answer content in order to check whether content varied systematically with age, and (2) we checked whether justification content correlated with partner recruitment decisions in the degree-of-competence test trials.

1. Identification with onscreen character

We asked the participants two questions to gauge whether they had identified with their assigned virtual character. In Identification question 1, the participant was shown a sequence of pictures of objects and characters, and asked to label each picture ("What do you see here?"). The participant's character appeared as part of the sequence. We were interested to see what answer the participant would supply for this picture. Participants who used a first-person reference (e.g., "me"), a first-person possessive (e.g., "my character") or a third-person reference (i.e., their name), were categorized as having identified with the character. Participants who used a non-specific, "other-reference" (e.g., said "a person") were categorized as not having identified with their character. Figure 3.7 (in main article) shows the proportion of participants per age group who identified with their onscreen character according to this question measure.

In Identification question 2, the participant was asked to describe a still-image taken from the starting scene of a competition degree-of-competence test trial. The participant's character appeared in the scene, and we were again interested in the words the participant would use to refer to their character. Participants who used a first-person reference (e.g., "me"), a first-person possessive (e.g., "my character") or a third-person reference (i.e., their name) were categorized as having identified with their character. Participants who used a non-specific "other-reference" (e.g., "a person") were classified as not having identified with the character, as were participants who described all elements aside from their character (e.g., "there are two boxes") or who refused to describe the scene at all (e.g., saying "I don't know"). One six-year-old child's answer was inaudible (child 35) and was coded as an "NA". The experimenter sometimes used follow-up questions to prompt children to speak. To be systematic, we only coded children's pre-prompt utterances. Figure 3.7 (in chapter 3) shows the proportion of participants of each age group who identified with their character according to this question measure.

2. Justification content as a function age

We asked two justification questions. In Justification question 1, the experimenter asked the participant about their choices in the strength games:

> E: "In the first game – remember, where you and Mr. Blue and Mr. Green were

> pushing those heavy things – sometimes you chose Mr. Blue and sometimes Mr. Green - can you tell me why you chose them?"

In case the child had only ever selected one character for the strength games (e.g., always Mr. Blue), the experimenter took this into account when framing the question. For younger participants, the experimenters often simplified the question and asked about each model separately. We classified the answers according to the coding scheme described in Table B.10.

As it was possible for an answer to span multiple categories (the answers provided by two children and one adult spanned two categories), categorization became a two-step process. In step one we listed all categories. In step two we assigned a rank to the categories; the "competence" category trumped all other categories, and all other categories trumped the "no justification" category (i.e., *competence* > remaining categories > *no justification*). In case an answer fit two categories, but the content fit neither the "competence" nor the "no justification" category, then the answer was assigned to the category that appeared first in the answer – in this way we could systematically compress multicategory answers such that each participant was represented just once in the justification plots (Figure 3.3). We took a similar approach for the binary classification – whether a response was counted as rational/irrational (column three in Table B.10). In case an answer was a mix of rational and irrational categories, the rational category trumped the irrational and the child's answer was classified as rational overall. One child's answer (child 40) was inaudible; this child was excluded from analysis.

In Justification question 2, the experimenter asked the participant about their choice in the knowledge game:

> E: "In the second game, remember, where you and Mr. Blue or Mr. Green had to label a picture, you chose Mr. [Blue/Green]. Can you tell me why you chose Mr. [Blue/Green]?"

We classified the answers according to the coding scheme described in Table B.11. While it was also theoretically possible for a participant's answer to span multiple categories, such an answer proved rare. No children gave multicategory answers and only one of the adults provided an answer that spanned two categories (*curious*, *stereotype*). In this case, as per the compression rules outlined above, since the "curious" content appeared first in the answer, this participant was represented by the "curious" category in the data. Due to experimenter error (the experimenter accidentally referred to the wrong character while framing the question), one child's answer was excluded from analysis (child 72).

3. Does recruitment behaviour correlate with justification provided?

We explored whether there was any correlation between recruitment behaviour in the degree-of-competence test trials and whether a participants' justification answer contained rational or irrational content. Figure B.4 shows the BLUPs (Best Linear Unbiased Predic-

Table B.10: Coding scheme used to categorize participants' answers to the justification question concerning the degree-of-competence test trials (Justification question 1).

Category	Description	Rational
Competence	Participant commented on the strength/speed of a character, referenced performance during the strength demonstration videos *or* commented on the outcome (win/loss) generally associated with a character	Y
Global impression	Participant commented that a character was good/bad/better/worse in general	Y
Help	Participant commented that the character was a help	N
Preference	Participant stated they liked one character/colour more	N
Fairness	Participant stated they had been alternating in their character selection	N
Curious	Participant indicated that they were interested to see what would happen	N
Luck	Participant commented on the good/bad fortune of a character	N
No justification	Participant was unable to provide a reason, gave an ambiguous reason, answered with a clarification question *or* supplied an answer lacking discernible logic	N

Table B.11: Coding scheme used to categorize participants' answers to the justification question concerning the generalization test trial (Justification question 2).

Category	Description	Generalization
Halo/Pitchfork	Participant either commented that one of the characters was better/worse than the other *or* commented on the outcome (win/loss) generally associated with the character	Y
Perceived intelligence	Participant stated that Mr. Green's superior problem-solving abilities in the strength demonstration videos was evidence that Mr. Green was clever	N
Stereotype	Participant stated the character was less clever because he was strong or vice versa	N
Help	Participant commented that the character had been a help previously	N
Preference	Participant stated they liked one character/colour more	N
Fairness	Participant stated they had been alternating in their character selection	N
Curious	Participant indicated they were interested to see what would happen	N
No justification	Participant was unable to provide a reason, gave an ambiguous reason, answered with a clarification question *or* supplied an answer lacking discernible logic	N

tors) for the participants who provided a rational explanation for their recruitment choices and for those who did not. BLUPs are a proxy measure for how rationally an individual performed in the degree-of-competence test trials – the higher the BLUP value, the more rationally that individual recruited across the six test trials. The child BLUPs were extracted from Model 1B, the adult BLUPs from Model 2.

We used a t-test to explore whether there was a correlation between behaviour and justification. On average the children who provided a rational justification did indeed, at the group-level, behave more rationally during the degree-of-competence test trials (mean BLUP = 0.037) compared to the children who provided an irrational justification (mean BLUP = -0.252). This difference between the two groups was statistically significant $t(100.085) = -2.012$, $p = 0.047$.

We ran the same analysis with the adult data and found that on average, the adults who provided us with a rational justification behaved more rationally during the degree-of-competence test trials (mean BLUP = -0.163) compared to the adults who provided an irrational justification (mean BLUP = -0.670). This difference between the two groups was, however, not statistically significant $t(8.133) = -1.532$, $p = 0.164$.

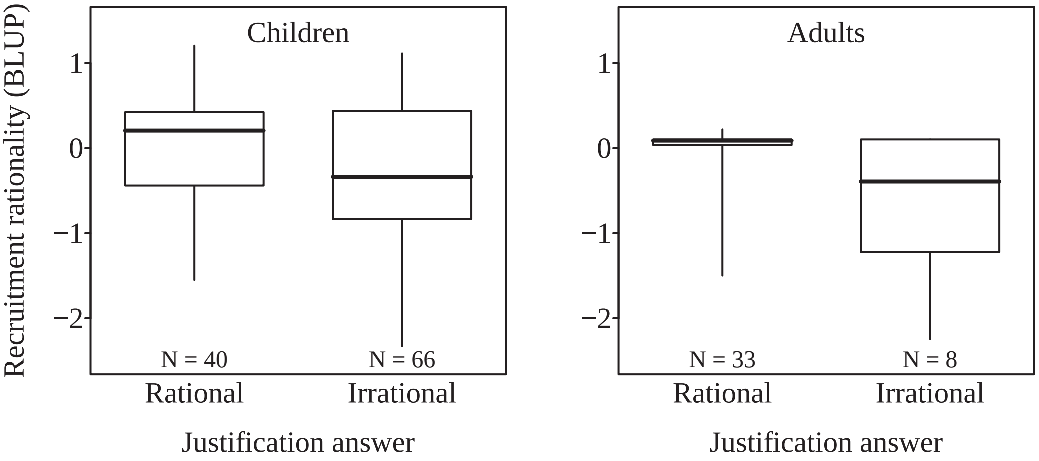

Figure B.4: Correlation between action (decision-making) in the degree-of-competence test trials and justification provided at the conclusion of the experiment. The thick horizontal line in each box shows the group median; the vertical line shows the range.

B.2 Experiment 2

B.2.1 Procedure

There were five stages to Experiment 2: (1) a demonstration phase where the child was introduced to two animated characters, (2) a character assignment phase where the child was introduced to the character who would represent them onscreen, (3) a familiarization phase where the child learned the rules and possible outcomes of the games they would encounter in the test trials, (4) four "domain-of-competence" test trials, and finally (5) the experimenter asked a series of open-ended advice questions designed to probe whether the child would offer rational advice to a peer who was about to play the strength games. We elaborate on each of these five stages below.

1. Demonstration phase

In a series of short videos, we introduced the child to two animated characters. One character ("Mr. Blue") was portrayed as strong but ignorant (he provided incorrect object labels), while the other character ("Mr. Green") was portrayed as weak but knowledgeable (he provided correct object labels).

The demonstration phase began with the strength domain. The child watched three video clips per character (refer back to Figure B.1). Mr. Blue succeeded in each strength task while Mr. Green struggled and had to settle for a work-around. After watching the videos, the child was asked two control questions to check that they had correctly linked strength competence and character; we asked the child to name the character who had performed the tasks in the videos well, and the character who had struggled. Children who were unable to answer both control questions correctly were shown two extra demonstration videos and the questions were repeated. Of the 98 children, three children required the extra strength demonstration videos in order to correctly differentiate the strong from the weak character. Two children were excluded for failing the strength demonstration control questions.

Next, we introduced the knowledge domain. We began by asking the child to label three familiar objects (a cup, a dog and a sock); in case a child's label was wrong or too generic, the experimenter gently corrected the child. Next the child watched three video clips per character in which the character offered a label for the same three objects (see Figure B.5). Mr. Blue offered incorrect labels (plate, cat, shoe) while Mr. Green labelled the objects correctly. After watching the videos, the child was asked two control questions to check that they had correctly linked knowledge competence and character; we asked the child to name the character who had performed the tasks in the videos well and the character who had struggled. Children who were unable to answer both control questions correctly were shown two extra demonstration videos and the questions were repeated. Three children required the extra demonstration videos and all three of these children were excluded for failing the knowledge demonstration control questions.

2. Character assignment phase

The character assignment phase and character control questions in Experiment 2 were identical to the procedure of Experiment 1. Twelve children required the extra character control question to correctly identify their character. Eight children were excluded for failing the character control questions.

3. Game familiarization phase

With the aid of eight short video clips containing black-and-white figures, the experimenter explained how a competitive and a cooperative version of a strength game and a competitive and cooperative version of a knowledge game worked. In the competitive strength game, two characters raced one another to shift a heavy object from left to right across the screen; the character that was first to reach the right-hand side won a virtual trophy. In the cooperative strength game, two characters worked together to shift a heavy object. To win, the characters needed to succeed at the task before a timer ran out. In the competitive knowledge game, two characters raced to be the first to label a novel object; the character that was first to call out the correct label won a virtual trophy. In the cooperative knowledge game, two characters worked together to label a novel object; one character suggested a label and the other character endorsed it. If the label suggestion was correct, the characters won a virtual trophy. Although the experimenter always introduced the strength games first, we counterbalanced the context that was introduced first; 51 of the 98 children learned about the competitive strength game first. After all four games had been explained and the child had seen the two possible outcomes associated with each game, the child was ready to move on to the domain-of-competence test trials.

4. "Domain-of-competence" test trials

In total, there were four domain-of-competence test trials – two competitive and two co-operative games, i.e., one test trial per context per domain (see Figure B.6). Prior to the first test trial, the experimenter told the child that their goal was to collect as many virtual prizes as possible. Context was presented as a block, i.e., the presentation order was either competition-competition-cooperation-cooperation or cooperation-cooperation-competition-competition. Prior to each game the experimenter told the child which context they were about to encounter, and we additionally provided a visual context cue; a pink background cued the competitive context, and a yellow background cued the cooperative context. Also prior to each game, the experimenter showed the child the heavy object that they would have to shift (a chest of drawers or a stack of books), or the novel object they would have to label (a fictitious geometric shape). In each test trial the child was asked to select a co-action partner from between Mr. Blue and Mr. Green and they watched the consequence of their choice – whether their character won or lost the game.

5. "Advice" questions

We asked two questions to gauge how children were reasoning in their partner recruitment choices. For each question, the experimenter showed the child a still-image taken from a strength test trial (one image per context). No characters appeared in the still-image but a picture of Mr. Blue and Mr. Green was overlaid over the lower right-hand corner of the image. The child was asked what advice they would give to a highly motivated playmate about to play the game. We were interested to see whether the child would advise their playmate to choose the rational partner for the context and whether the child could explain why this character was a good choice. Children who had received a competition domain-of-competence test trial first, were asked to give advice on the competition context first, and vice versa for the children who had received a cooperation domain-of-competence test trial first.

B.2.2 Stimuli

All stimuli were created in *Vyond*. The strength demonstration videos from Experiment 1 (refer back to Figure B.1) were re-used in Experiment 2. Still-images from the knowledge demonstration videos are shown in Figure B.5 and the domain-of-competence test trials are shown in Figure B.6.

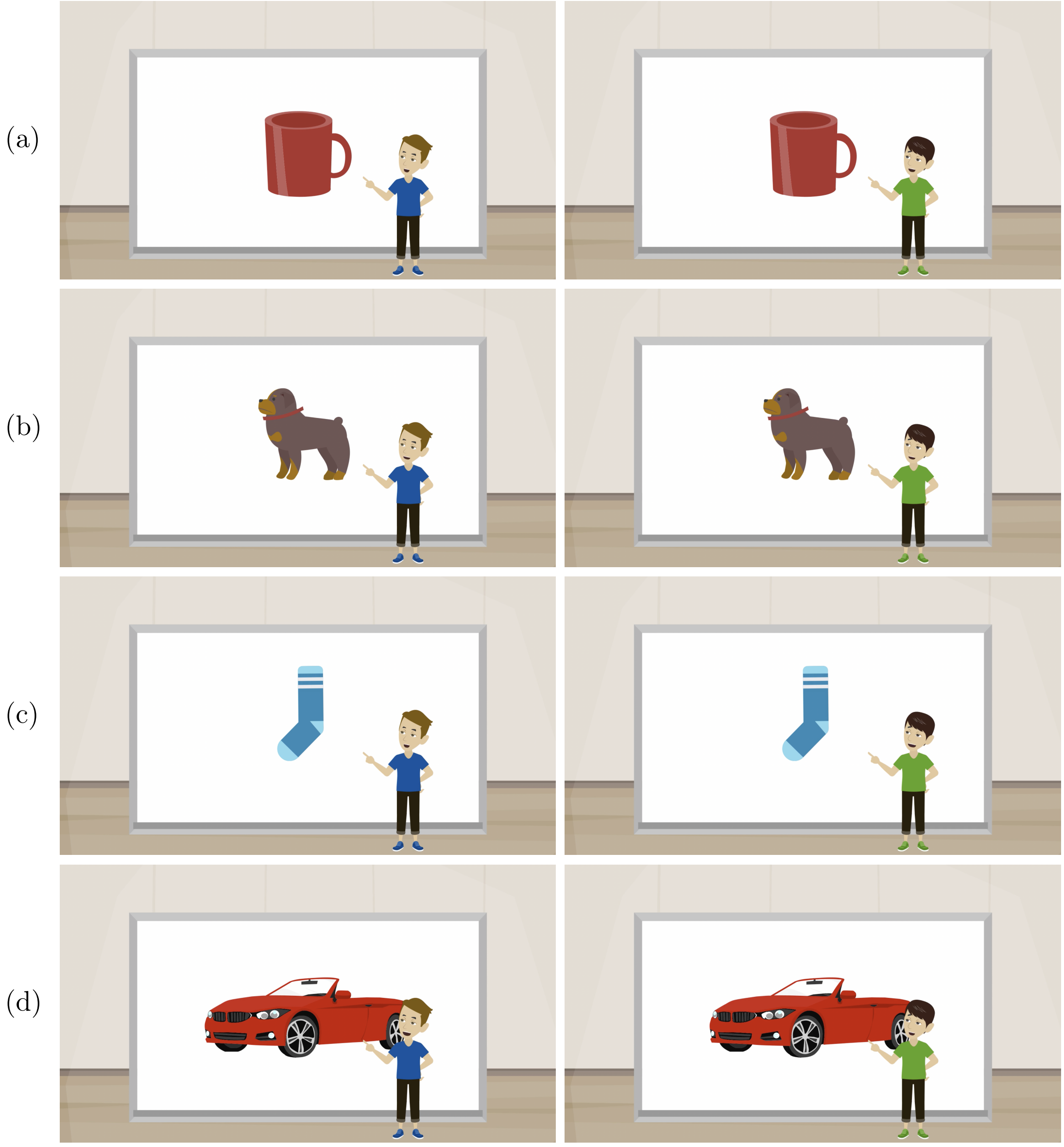

Figure B.5: Still-images from the knowledge demonstration videos. All children saw scenarios (a) through (c). Only children who failed a set of control questions were shown scenario (d). Mr. Blue provided incorrect object labels (plate, cat, shoe, bus) in contrast to Mr. Green (cup, dog, sock, car).

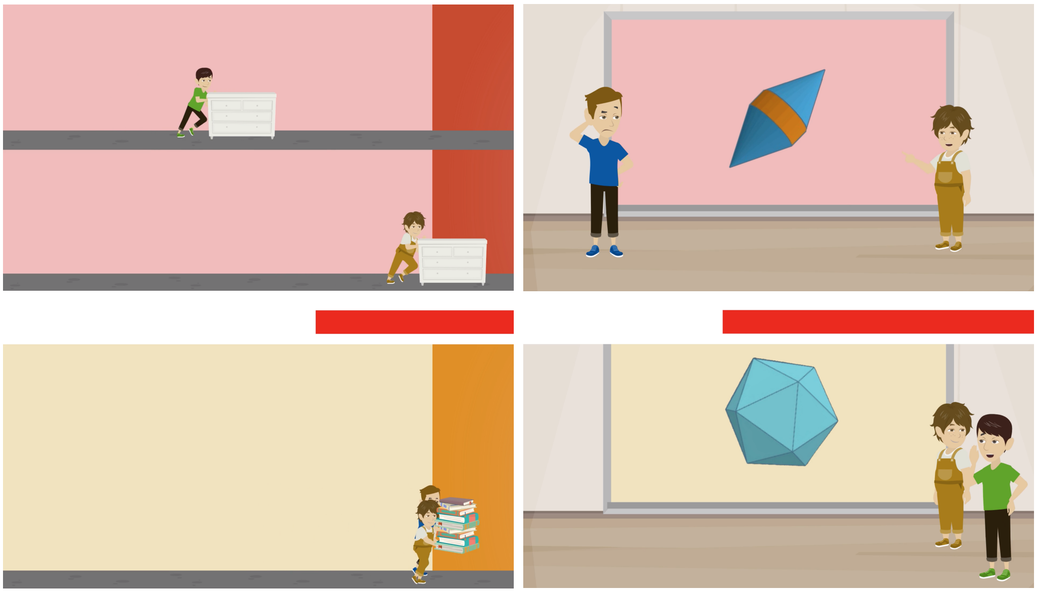

Figure B.6: The four domain-of-competence test trials. In these strength and knowledge games, the participant had the opportunity to recruit from between Mr. Blue and Mr. Green. These four still-images depict rational recruitment behaviour, as per our definition of "rational" in this study. A participant who is motivated to win virtual prizes should recruit the weak character (Mr. Green) for the competitive strength game (top-left), the ignorant character (Mr. Blue) for the competitive knowledge game (top-right), the strong character (Mr. Blue) for the cooperative strength game (bottom-left) and the knowledgeable character (Mr. Green) for the cooperative knowledge game (bottom-right).

B.2.3 Confirmatory analyses

All analyses were conducted in R (R-Core-Team, 2021, version 4.1.1). We fitted a total of three GLMMs. Two of the three models were preregistered on OSF (https://archive.org/details/osf-registrations-q8myj-v1). We take care to highlight where the analyses deviate from the preregistration.

All models were fitted using the *glmer* function of the *lme4* package (Bates et al., 2015; version $1.1-27.1$). To aid convergence, we used the *bobyqa* optimizer. Confidence intervals were derived using the *bootMer* function of the *lme4* package.

The response variable in each model was binary. All models were therefore fitted with a binomial error structure and logit link function (McCullagh and Nelder, 1989). Covariates included in the models were z-transformed to have a mean of zero and a standard deviation of one. In order to keep type I error rates low (at the nominal level of 0.05), and to avoid an 'over-confident' model, we included random slopes in the models (Barr et al., 2013; Schielzeth and Forstmeier, 2009). Factors that appeared in the random slopes structure of a model were dummy-coded and centred prior to inclusion.

For each model we fitted, we used a likelihood ratio test (Dobson, 2002) to establish the overall effect(s) of our factor(s) of interest. Use of the likelihood ratio test meant that in cases where there were multiple factors of interest, we were able to avoid running into ‘cryptic multiple testing’ issues (Forstmeier and Schielzeth, 2011).

Prior to inference, we assessed model stability and checked for absence of collinearity. We checked model stability by dropping the levels of the random effects (one at a time) and comparing the estimates derived from each reduced dataset with those obtained from the full dataset. In each case model stability appeared to be fine (see columns *min.* and *max.* provided in each table associated with a full model). To check whether we needed to be concerned about collinearity, we inspected Variance Inflation Factors for each model (VIFs; Field, 2005). To obtain the VIFs we used the *vif* function of the *car* package (Fox and Weisberg, 2011); this function was applied to a linear model lacking the random effects. The largest VIF encountered was 1.654, i.e., collinearity was not an issue.

P-values were derived by dropping the fixed effects from the model one at a time and comparing the full model with each respective reduced model (achieved using the *drop1* function in R).

As was the case in Experiment 1, in Experiment 2 we had three experimenters collect the child data. It was therefore necessary to rule out an effect of experimenter ID. As in Experiment 1, our original plan involved subsetting the data according to experimenter ID and running each subset through the preregistered model (Model 5). Issues with complete separation arose, however, and we instead elected to run a supplementary (i.e., non-preregistered) model (Model 5B) in which we controlled for an effect of experimenter ID (experimenter ID was included in Model 5B as a random effect).

All Experiment 2 confirmatory analyses are outlined in detail in the subsections below. The data and R-scripts used in the analyses have been made available on OSF (https://osf.io/65fv3/?view_only=cb1dc0b626084a1780faf8e4db1c3d6d). Numerous R functions (written by Roger Mundry) were sourced at various stages of the analyses. These functions have been made available on Zenodo (https://doi.org/10.5281/zenodo.7670524).

Model 5/5B: Analysis of child domain-of-competence data

To estimate the effect of context and age on children’s partner recruitment behaviour given domain-of-competence information, we constructed Model 5 (a GLMM). The response variable was partner recruitment choice. The model comprised a two-way interaction between context and age in addition to their fixed main effects and the fixed effects of domain, gender, order (whether the child received the competition-knowledge, competition-strength, cooperation-knowledge or cooperation-strength test trial first) and trial number (1 – 4). Participant ID was included as a random effect given that repeated measures were taken from the same individuals. The random slopes of context and trial number within participant ID were included in the model. As outlined in our preregistration, we did not include parameters for the correlations among random intercepts and slopes. With only four data points from each individual, it was likely that including these correlations would lead to an overparameterized random effects structure with unidentifiable random effects components. A "singular fit" message indicated that some random effects terms were unidentifiable, however, as we were interested in the fixed effects, this message was not a concern.

We used a likelihood ratio test to establish the overall effect of context, age and the interaction between context and age. Here we compared the full model with a null model lacking the interaction term and their main effects. The likelihood ratio test revealed the two models differed significantly from one another ($\chi^2_3 = 7.974$, $p = 0.047$). The output of the full model (Table B.12) revealed the interaction term (context × age) was not significant ($p = 0.261$). We therefore fitted a reduced model (output in Table B.13). The reduced model revealed no significant effect of context ($p = 0.289$) but a significant effect of age ($p = 0.019$).

To estimate the effect of context and age on children's partner recruitment behaviour, while controlling for a potential effect of experimenter ID, we constructed another GLMM – Model 5B. The structure of Model 5B was near-identical to that of Model 5, but there was one key difference. Namely, Model 5B included the random effect of experimenter ID in addition to the random effect of participant ID. The random slopes included in Model 5B were, therefore, context and trial number within participant ID, and the context × age interaction term, gender, and trial number within experimenter ID. A "singular fit" message indicated that some random effects terms were unidentifiable.

We used a likelihood ratio test to establish the overall effect of context, age and the interaction between context and age. We compared the full model with a null model lacking the interaction term and their main effects. The likelihood ratio test revealed the two models did not differ significantly from one another ($\chi^2_3 = 5.474$, $p = 0.14$). We report and discuss this null finding in the main article. The output of the full model is provided in Table B.14.

Table B.12: Model 5 (full model)
Analysis of child domain-of-competence data (392 trials from 98 participants): Testing the effect of the interaction between context and age, and their main effects, as well as the fixed effects of domain, gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	-0.056	0.534	-1.077	0.933			(1)	-0.233	0.160
Context (coop)	0.588	0.550	-0.485	1.759			(1)	0.483	0.685
Age(2)	0.769	0.352	0.137	1.544			(1)	0.701	0.860
Domain (strength)	0.248	0.289	-0.261	0.762	0.743	1	0.389	0.171	0.346
Gender (male)	-0.210	0.399	-0.999	0.525	0.278	1	0.598	-0.320	-0.079
Order (comp strength)(3)	0.760	0.568	-0.260	1.866	2.775	3	0.428	0.562	0.960
Order (coop knowledge)(3)	-0.004	0.563	-1.087	1.068				-0.191	0.227
Order (coop strength)(3)	0.526	0.556	-0.499	1.554				0.332	0.755
Trial number(4)	0.135	0.224	-0.280	0.574	0.368	1	0.544	0.092	0.197
Context:Age	-0.610	0.548	-1.739	0.445	1.261	1	0.261	-0.753	-0.490

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2187.449 and 413.884 respectively.

[3] Order was dummy-coded with *competition knowledge* as the reference level; the indicated test is for the overall effect of order.

[4] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2.5 and 1.119 respectively.

Table B.13: Model 5 (reduced model)
Analysis of child domain-of-competence data (392 trials from 98 participants): Testing the effect of context and age and the fixed effects of domain, gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals and test results).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value
Intercept	-0.053	0.535	-1.003	1.027			(1)
Context (coop)	0.581	0.553	-0.454	1.639	1.124	1	0.289
Age(2)	0.464	0.207	0.063	0.889	5.526	1	0.019
Domain (strength)	0.249	0.289	-0.344	0.783	0.746	1	0.388
Gender (male)	-0.211	0.398	-1.086	0.482	0.281	1	0.596
Order (comp strength)(3)	0.761	0.568	-0.238	1.900	2.807	3	0.422
Order (coop knowledge)(3)	-0.006	0.563	-1.065	1.075			
Order (coop strength)(3)	0.528	0.556	-0.427	1.701			
Trial number(4)	0.136	0.225	-0.322	0.638	0.371	1	0.542

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2187.449 and 413.884 respectively.

[3] Order was dummy-coded with *competition knowledge* as the reference level; the indicated test is for the overall effect of order.

[4] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2.5 and 1.119 respectively.

Table B.14: Model 5B (full model)
Analysis of child domain-of-competence data (392 trials from 98 participants) in which we account for the fact that three experimenters collected the data: Testing the effect of the interaction between context and age, and their main effects as well as the fixed effects of domain, gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	-0.104	0.568	-1.250	0.921			[1]	-0.646	0.441
Context (coop)	0.585	0.544	-0.456	1.693			[1]	0.125	1.228
Age[2]	0.751	0.346	0.090	1.516			[1]	0.582	1.052
Domain (strength)	0.247	0.288	-0.258	0.786	0.738	1	0.390	-0.214	0.473
Gender (male)	-0.208	0.442	-1.068	0.643	0.216	1	0.642	-0.694	0.276
Order (comp strength)[3]	0.789	0.549	-0.190	1.873	2.711	3	0.438	0.525	1.091
Order (coop knowledge)[3]	0.055	0.594	-1.020	1.274				-0.560	0.850
Order (coop strength)[3]	0.536	0.564	-0.483	1.673				0.307	1.050
Trial number[4]	0.132	0.223	-0.309	0.550	0.356	1	0.551	0.009	0.430
Context:Age	-0.598	0.542	-1.731	0.413	1.234	1	0.267	-0.780	-0.401

[1] Not indicated as of limited interpretability.

[2] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2187.449 and 413.884 respectively.

[3] Order was dummy-coded with *competition knowledge* as the reference level; the indicated test is for the overall effect of order.

[4] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2.5 and 1.119 respectively.

Model 6: Adult domain-of-competence data

To estimate the effect of context on adults' partner recruitment behaviour given domain-of-competence information, we constructed Model 6 (a GLMM). The response variable was partner recruitment choice. The model comprised a two-way interaction between context and age in addition to their fixed main effects and the fixed effects of domain, gender, order (whether the participant received the competition-knowledge, competition-strength, cooperation-knowledge or cooperation-strength test trial first) and trial number (1 – 4). Participant ID was included as a random effect. The random slopes of context and trial number within participant ID were included in the model. We did not include parameters for the correlations among random intercepts and slopes. A "singular fit" message indicated that some random effects terms were unidentifiable.

We used a likelihood ratio test to compare the full model with a null model lacking the fixed effect of context. This test revealed that the two models differed significantly from one another ($\chi^2_3 = 16.071$, $p = <0.001$). The output of the full model (Table B.15) showed that context (our variable of interest) was significant ($p = <0.001$).

Table B.15: Model 6 (full model)
Analysis of adult domain-of-competence data (192 trials from 48 participants): Testing the effect of context, in addition to the fixed effects of domain, gender, order and trial number on the probability to select the rational co-action partner (table shows estimates, standard errors, confidence intervals, test results and minimum and maximum of model estimates obtained after dropping levels of random effects one at a time).

Term	Estimate	SE	CI_{Lower}	CI_{Upper}	LRT	Df	p-value	min.	max.
Intercept	0.735	0.548	-0.344	6.220			(1)	0.410	1.201
Context (coop)	1.961	0.589	1.114	12.534	16.071	1	<0.001	1.832	2.481
Domain (strength)	0.264	0.451	-1.015	1.575	0.344	1	0.557	0.039	0.464
Gender (male)	-0.043	0.439	-1.073	1.007	0.009	1	0.922	-0.166	0.209
Order (comp strength)(2)	1.275	0.755	-0.079	14.801	5.588	3	0.133	1.022	1.701
Order (coop knowledge)(2)	-0.181	0.615	-3.124	1.182				-0.762	0.176
Order (coop strength)(2)	0.681	0.756	-1.709	3.677				0.092	0.954
Trial number(3)	0.308	0.282	-0.308	2.176	1.281	1	0.258	0.217	0.540

[1] Not indicated as of limited interpretability.

[2] Order was dummy-coded with *competition knowledge* as the reference level; the indicated test is for the overall effect of order.

[3] Z-transformed to a mean of 0 and a standard deviation of 1; mean and SD of the original variable was 2.5 and 1.121 respectively.

B.2.4 Exploratory analyses

To conclude the experiment, we asked each participant two advice questions. The purpose of these questions was to explore how children were reasoning about partner recruitment. Similar to Experiment 1, the answers the participants provided were used in two ways: (1) we categorized the advice offered by the participants, to see whether advice content differed systematically with age, and (2) we checked whether rational advice content was positively correlated with rational performance in the test trials.

1. Advice content as a function of age

We asked the participants two advice questions – a question about the competitive context and a question about the cooperative context (context presented first was counterbalanced between participants). The child was asked for the name of a playmate [X]. The experimenter then showed the image of one of the strength games (Mr. Blue and Mr. Green were depicted in the lower right-hand corner as a visual aid) and asked the following question:

> "If [X] was about to play this game, and they really wanted to win a prize – what tips would you give him/her?"

We categorized the participants' answers according to the coding scheme described in Table B.16. Multicategory answers were "compressed" according to a set of rules outlined in section B.1.4. The complete advice data of three children (child 93, 107 and 109) was excluded from analysis due to parental interference.

In response to being asked about the competition context, two children provided advice that spanned two categories (child 4 and child 50). The adults provided advice that was able to be neatly assigned to one category. In addition to the three children whose data could not be analysed due to parental interference, the competition advice of two children (child 57 and 67) was excluded; child 57 was unable to grasp the question and the sibling of child 67 interfered. Figure 3.10 (chapter 3) shows the advice content of the answers given in response to the competition advice question.

In response to being asked about the cooperation context, two children provided advice that spanned two categories (child 50 and child 78). None of the adults provided a multicategory answer. In addition to the three children whose data could not be analysed due to parental interference, the cooperation answers of two children (child 57 and 112) were inaudible. Figure 3.10 (chapter 3) shows the advice content of the answers given to the cooperation advice question.

Table B.16: Coding scheme used to categorize participants' advice answers.

Category	Description	Rational
Competence	Participant commented on the strength/speed of a character, referenced a character's performance during the strength demonstration videos *or* commented on whether a win or loss was generally associated with a character	Y
False competence recall	Participant incorrectly stated a model's competence (e.g., claimed Mr. Blue was weak)	N
Rationale-context mismatch	Participant demonstrated knowledge of a model's competence but applied the knowledge incorrectly for the context at hand	N
Global impression	Participant commented that a character was good/bad/better/worse in general	N
Preference	Participant stated that they have a bias for a character/colour or that their playmate has such a bias	N
Fairness	Participant indicated that they were alternating in their partner selection so as to give each character equal playing time	N
Ill-intentioned	Participant selected a character to prevent their friend from winning a prize	N
No justification	Participant was unable to provide a reason, gave an ambiguous reason, made an incorrect statement *or* supplied an answer lacking discernible logic	N

2. Does recruitment behaviour correlate with advice content?

We explored whether there was any correlation between participant recruitment behaviour and whether a participants' advice answer contained rational or irrational content. Figure B.7 shows the BLUPs of the participants who provided a playmate with rational advice as compared to the BLUPs of participants who provided irrational or no advice. The child BLUPs were extracted from Model 5B and the adult BLUPs from Model 6.

We used t-tests to explore whether rational recruitment behaviour was positively correlated with rational advice. We ran one t-test per context. With regard to the competition context, there was little to suggest that the children who provided a playmate with rational advice behaved more rationally during the domain-of-competence test trials (mean BLUP $= -0.012$) compared to the children who provided irrational advice (mean BLUP $= -0.008$), and the t-test confirmed there was no significant difference between these two groups $t(100.085) = 0.04$, $p = 0.968$. With regard to the cooperation context, on average children who offered to their playmate rational advice performed similarly in terms of rationality of recruitment (mean BLUP $= -0.079$) compared to the children who offered irrational advice (mean BLUP $= 0.036$). Again, the t-test confirmed that there was no significant difference between the two groups $t(56.196) = 1.204$, $p = 0.234$.

We ran the same analysis with the adult data. With regard to the competition context, there was little to suggest that the adults who provided a friend with rational advice behaved more rationally during the domain-of-competence test trials (mean BLUP $= -0.001$) compared to the adults who gave irrational or no advice (mean BLUP $= -0.007$) and indeed, here there was also no difference between the two groups $t(8.111) = -1.1$, $p = 0.303$. We report a similar finding with regard to the cooperation context t(-1.625), $p = 0.137$.

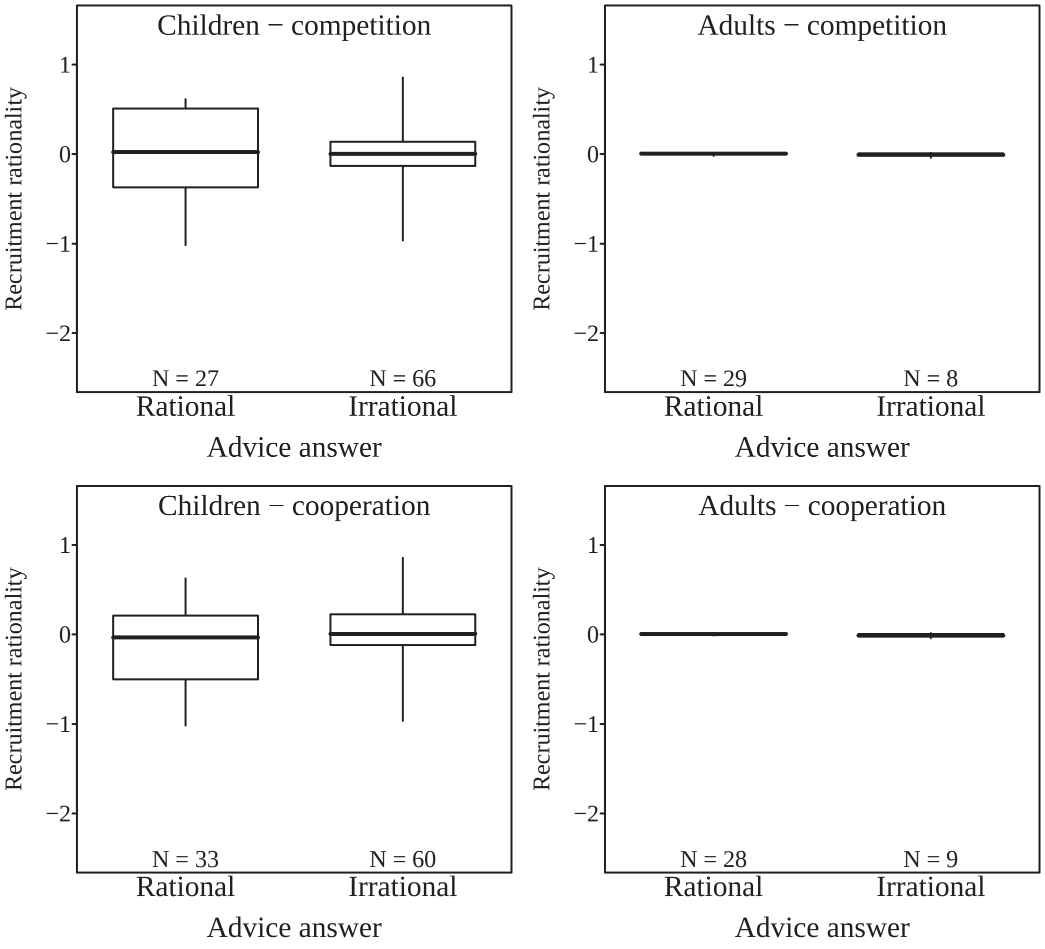

Figure B.7: Exploring the correlation between behaviour and advice content. The thick black line in each box shows the group median; the vertical line indicates the range.

B.3 Effect of Experimenter ID

Three experimenters collected the child data in each experiment and in total there were five experimenter identities (one tester took part in both experiments). In the three plots that follow, we show the data split by experimenter ID. Figure B.8 shows the degree-of-competence data, Figure B.9 the generalization data and Figure B.10 the domain-of-competence data. As is apparent from inspecting the plots, it is difficult to discern a distinct pattern in the raw data – there is considerable noise. Here it is important to keep in mind that with the data split by experimenter ID and context within age group, the sample size in each column is relatively small; each column shows the data of four to six children.

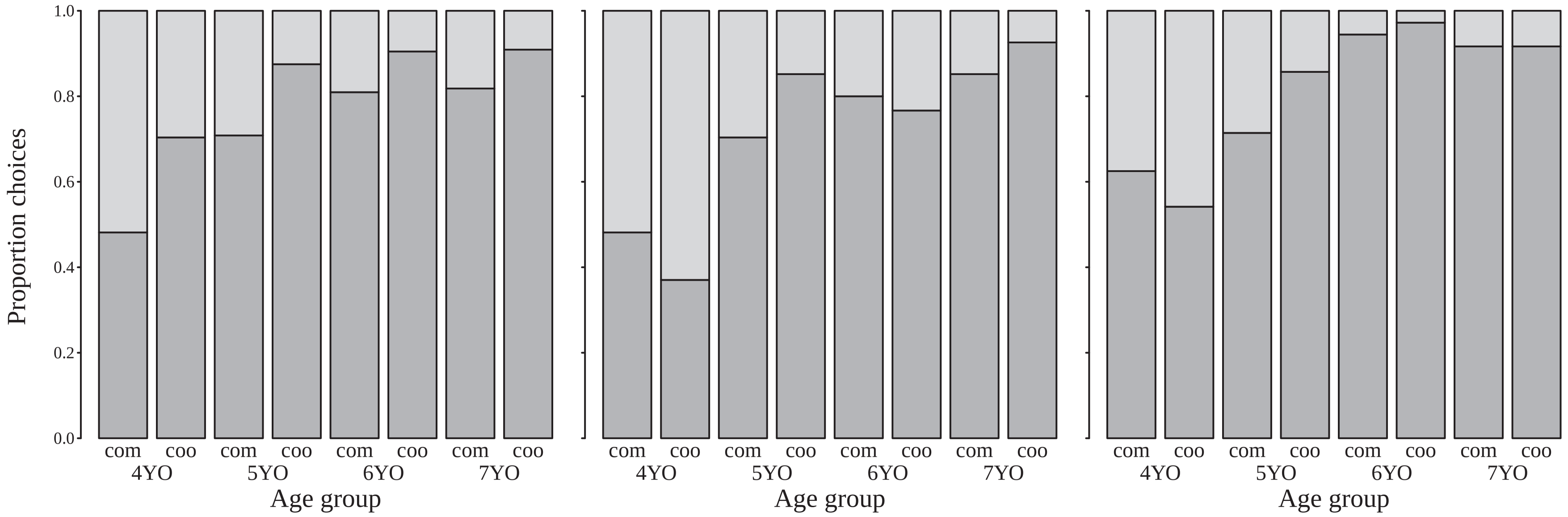

Figure B.8: The degree-of-competence test trials split by experimenter ID.

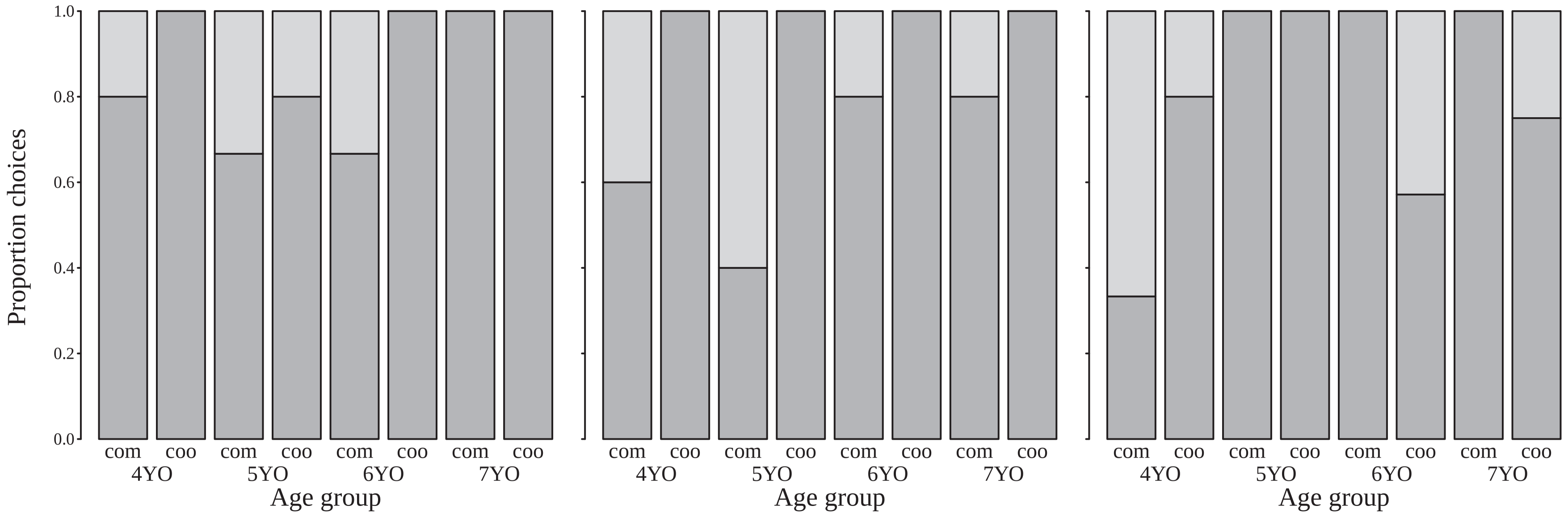

Figure B.9: The generalization test trials split by experimenter ID

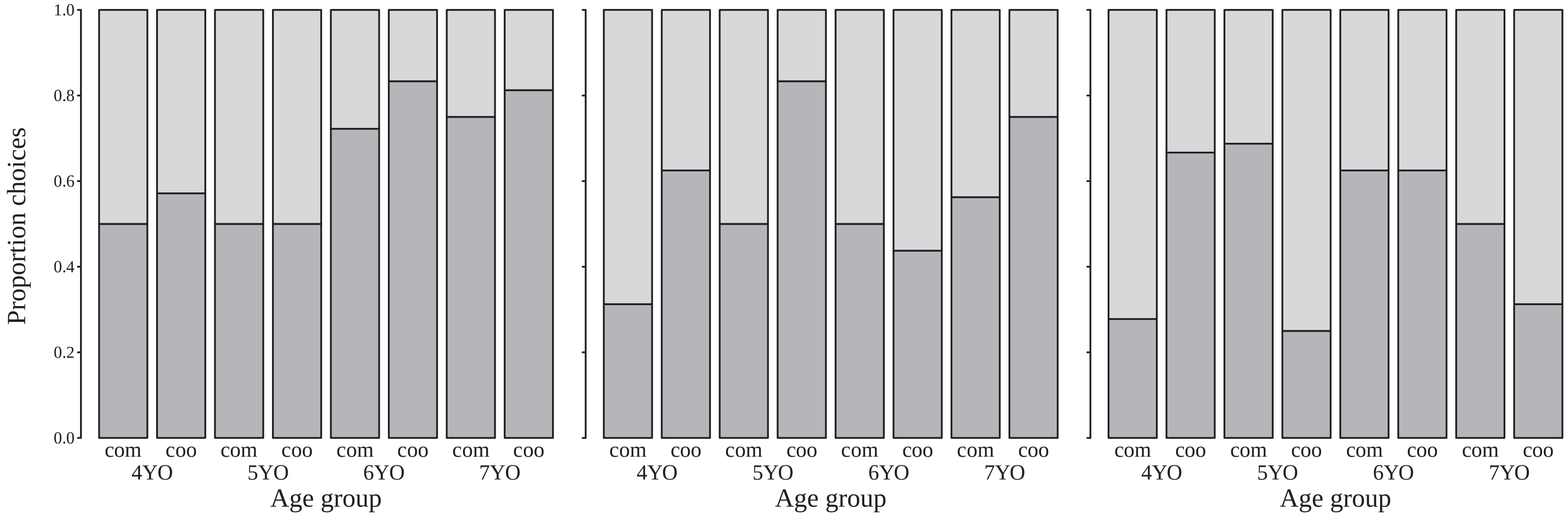

Figure B.10: The domain-of-competence test trials split by experimenter ID

References

Anderson, J. R., Kuroshima, H., Takimoto, A., & Fujita, K. (2013). Third-party social evaluation of humans by monkeys. *Nature Communications*, *4*(1), 1561. https://doi.org/10.1038/ncomms2495

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749–753. https://doi.org/10.1098/rspb.2006.0209

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. https://doi.org/10.1016/j.jml.2012.11.001

Bascandziev, I., & Harris, P. L. (2014). In beauty we trust: Children prefer information from more attractive informants. *British Journal of Developmental Psychology*, *32*(1), 94–99. https://doi.org/10.1111/bjdp.12022

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Lme4: Fitting linear mixed-effects models using lme4. *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumann, M. R., & Bonner, B. L. (2013). Member awareness of expertise, information sharing, information weighting, and group decision making. *Small Group Research*, *44*(5), 532–562. https://doi.org/10.1177/1046496413494415

Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, *107*(4), 1137–1160. https://doi.org/10.2307/2118383

Bernard, S., Castelain, T., Mercier, H., Kaufmann, L., Van der Henst, J.-B., & Clément, F. (2016). The boss is always right: Preschoolers endorse the testimony of a dominant over that of a subordinate. *Journal of Experimental Child Psychology*, *152*, 307–317. https://doi.org/10.1016/j.jecp.2016.08.007

Beunen, G. P., Thomis, M., & Peeters, M. W. (2014). Genetic variation in physical performance. *The Open Sports Sciences Journal*, *3*(1), 77–80. https://doi.org/10.2174/1875399X010030100077

Birch, S. A., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018–1034. https://doi.org/10.1016/j.cognition.2007.12.008

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., Kleutsch, L., Kramer, K. L., Ross, E., Vongsachang, H., Wrangham, R., & Warneken, F. (2015). The ontogeny of fairness in seven societies. *Nature*, *528*(7581), 258–261. https://doi.org/10.1038/nature15703

Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair": Eight-year-olds reject two forms of inequity. *Cognition*, *120*(2), 215–224. https://doi.org/10.1016/j.cognition.2011.04.006

Bliege Bird, R., & Power, E. A. (2015). Prosocial signaling and cooperation among Martu hunters. *Evolution and Human Behavior*, *36*(5), 389–397. https://doi.org/10.1016/j.evolhumbehav.2015.02.003

Bochynska, A., & Dillon, M. R. (2021). Bringing home baby Euclid: Testing infants' basic shape discrimination online. *Frontiers in Psychology*, *12*, 734592. https://doi.org/10.3389/fpsyg.2021.734592

Bolton, G. E., Katok, E., & Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory*, *27*(2), 269–299. https://doi.org/10.1007/s001820050072

Bolton, G. E., & Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic Behavior*, *10*(1), 95–121. https://doi.org/10.1006/game.1995.1026

Bonner, B. L. (2004). Expertise in group problem solving: Recognition, social combination, and performance. *Group Dynamics: Theory, Research, and Practice*, *8*(4), 277–290. https://doi.org/10.1037/1089-2699.8.4.277

Bor, A. (2017). Spontaneous categorization along competence in partner and leader evaluations. *Evolution and Human Behavior*, *38*(4), 468–473. https://doi.org/10.1016/j.evolhumbehav.2017.03.006

Bräuer, J., Call, J., & Tomasello, M. (2006). Are apes really inequity averse? *Proceedings of the Royal Society B: Biological Sciences*, *273*(1605), 3123–3128. https://doi.org/10.1098/rspb.2006.3693

Bräuer, J., Call, J., & Tomasello, M. (2009). Are apes inequity averse? New data on the token-exchange paradigm. *American Journal of Primatology*, *71*(2), 175–181. https://doi.org/10.1002/ajp.20639

Brosnan, S. F. (2006). Nonhuman species' reactions to inequity and their implications for fairness. *Social Justice Research*, *19*(2), 153–185. https://doi.org/10.1007/PL00022136

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature*, *425*(6955), 297–299. https://doi.org/10.1038/nature01963

Brosnan, S. F., Flemming, T., Talbot, C. F., Mayo, L., & Stoinski, T. (2011). Orangutans (*Pongo pygmaeus*) do not form expectations based on their partner's outcomes. *Folia Primatologica*, *82*(1), 56–70. https://doi.org/10.1159/000328142

Brosnan, S. F., Schiff, H. C., & de Waal, F. B. M. (2005). Tolerance for inequity may increase with social closeness in chimpanzees. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1560), 253–258. https://doi.org/10.1098/rspb.2004.2947

Brosnan, S. F., Talbot, C., Ahlgren, M., Lambeth, S. P., & Schapiro, S. J. (2010). Mechanisms underlying responses to inequitable outcomes in chimpanzees, *Pan troglodytes*. *Animal Behaviour*, *79*(6), 1229–1237. https://doi.org/10.1016/j.anbehav.2010.02.019

Brosseau-Liard, P. E., & Birch, S. A. (2010). 'I bet you know more and are nicer too!': What children infer from others' accuracy. *Developmental Science*, *13*(5), 772–778. https://doi.org/10.1111/j.1467-7687.2009.00932.x

Burns, M. P., & Sommerville, J. A. (2014). "I pick you": The impact of fairness and race on infants' selection of social partners. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00093

Chijiiwa, H. (2021). Social evaluation in non-human animals. In *Comparative cognition* (pp. 221–232). Springer Singapore. https://doi.org/10.1007/978-981-16-2028-7_13

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, *89*(8), 1421–1435. https://doi.org/10.1016/j.jpubeco.2004.05.011

Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, *12*(3), 426–437. https://doi.org/10.1111/j.1467-7687.2008.00792.x

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1149–1157. https://doi.org/10.1098/rstb.2010.0319

Dobson, A. J. (2002). *An introduction to generalized linear models* (2nd ed). Chapman & Hall/CRC.

Dubreuil, D., Gentile, M. S., & Visalberghi, E. (2006). Are capuchin monkeys (*Cebus apella*) inequity averse? *Proceedings of the Royal Society B: Biological Sciences*, *273*(1591), 1223–1228. https://doi.org/10.1098/rspb.2005.3433

Elashi, F. B., & Mills, C. M. (2014). Do children trust based on group membership or prior accuracy? The role of novel group membership in children's trust decisions. *Journal of Experimental Child Psychology*, *128*, 88–104. https://doi.org/10.1016/j.jecp.2014.07.003

Engelmann, J. M., Clift, J. B., Herrmann, E., & Tomasello, M. (2017). Social disappointment explains chimpanzees' behaviour in the inequity aversion task. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1861), 20171502. https://doi.org/10.1098/rspb.2017.1502

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*(7208), 1079–1083. https://doi.org/10.1038/nature07155

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Field, A. (2005). *Discovering statistics using SPSS*. Sage Publications.

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies.

Fletcher, G. E. (2008). Attending to the outcome of others: disadvantageous inequity aversion in male capuchin monkeys (*Cebus apella*). *American Journal of Primatology*, *70*(9), 901–905. https://doi.org/10.1002/ajp.20576

Fontenot, M., Watson, S., Roberts, K., & Miller, R. (2007). Effects of food preferences on token exchange and behavioural responses to inequality in tufted capuchin monkeys, *Cebus apella*. *Animal Behaviour*, *74*(3), 487–496. https://doi.org/10.1016/j.anbehav.2007.01.015

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology*, *65*(1), 47–55. https://doi.org/10.1007/s00265-010-1038-5

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed). SAGE Publications.

Freeman, H. D., Sullivan, J., Hopper, L. M., Talbot, C. F., Holmes, A. N., Schultz-Darken, N., Williams, L. E., & Brosnan, S. F. (2013). Different responses to reward comparisons by three primate species. *PLoS ONE*, *8*(10), e76297. https://doi.org/10.1371/journal.pone.0076297

Fusaro, M., Corriveau, K. H., & Harris, P. L. (2011). The good, the strong, and the accurate: Preschoolers' evaluations of informant attributes. *Journal of Experimental Child Psychology*, *110*(4), 561–574. https://doi.org/10.1016/j.jecp.2011.06.008

Gallo, P. S., & Dale, I. A. (1968). Experimenter bias in the prisoner's dilemma game. *Psychonomic Science*, *13*(6), 340–340. https://doi.org/10.3758/BF03342616

Gomes, D. G. (2022). Should I use fixed effects or random effects when I have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ*, *10*, e12794. https://doi.org/10.7717/peerj.12794

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Hardy, C., & Van Vugt, M. (2006). Giving for glory in social dilemmas: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, *32*, 1402–1413.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Belknap Press of Harvard University Press.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, *69*(1), 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of *Homo economicus*: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78. https://doi.org/10.1257/aer.91.2.73

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Hermes, J., Behne, T., Bich, A. E., Thielert, C., & Rakoczy, H. (2016). Children's selective trust decisions: Rational competence and limiting performance factors. *Developmental Science*, *21*(2), e12527. https://doi.org/10.1111/desc.12527

Hermes, J., Behne, T., & Rakoczy, H. (2015). The role of trait reasoning in young children's selective trust. *Developmental Psychology*, *51*(11), 1574–1587. https://doi.org/10.1037/dev0000042

Hermes, J., Behne, T., & Rakoczy, H. (2018). The development of selective trust: Prospects for a dual-process account. *Child Development Perspectives*, *12*(2), 134–138. https://doi.org/10.1111/cdep.12274

Hermes, J., Behne, T., Studte, K., Zeyen, A.-M., Gräfenhain, M., & Rakoczy, H. (2016). Selective cooperation in early childhood – How to choose models and partners. *PLOS ONE*, *11*(8), e0160881. https://doi.org/10.1371/journal.pone.0160881

Hermes, J., Brugger, F., Illner, T., Plate, A., Rakoczy, H., & Behne, T. (2020). *Selective trust in young children and distracted adults: Halo-effects outweigh rational choices* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/fr84t

Herrmann, E., Engelmann, J. M., & Tomasello, M. (2019). Children engage in competitive altruism. *Journal of Experimental Child Psychology*, *179*, 176–189. https://doi.org/10.1016/j.jecp.2018.11.008

Hopper, L. M., Lambeth, S. P., Schapiro, S. J., Bernacky, B. J., & Brosnan, S. F. (2013). The ontogeny of social comparisons in rhesus macaques (*Macaca mulatta*). *Journal of Primatology*, *2*(1). https://doi.org/10.4172/2167-6801.1000109

Humphrey, N. K. (1976). The social function of intellect. In *Growing points in ethology* (pp. 303–317). CUP Archive.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, *21*(10), 1541–1547. https://doi.org/10.1177/0956797610383438

Jaswal, V. K., & Kondrad, R. L. (2016). Why children are not always epistemically vigilant: Cognitive limits and social considerations. *Child Development Perspectives*, *10*(4), 240–244. https://doi.org/10.1111/cdep.12187

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science*, *17*(9), 757–758. https://doi.org/10.1111/j.1467-9280.2006.01778.x

Jaswal, V. K., Pérez-Edgar, K., Kondrad, R. L., Palmquist, C. M., Cole, C. A., & Cole, C. E. (2014). Can't stop believing: Inhibitory control and resistance to misleading testimony. *Developmental Science*, *17*(6), 965–976. https://doi.org/10.1111/desc.12187

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are rational maximizers in an ultimatum game. *Science*, *318*(5847), 107–109. https://doi.org/10.1126/science.1145850

Jolly, A. (1966). Lemur social behavior and primate intelligence: The step from prosimian to monkey intelligence probably took place in a social context. *Science*, *153*(3735), 501–506. https://doi.org/10.1126/science.153.3735.501

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *The Journal of Business*, *59*(4), S285–S300.

Kaiser, I., Jensen, K., Call, J., & Tomasello, M. (2012). Theft in an ultimatum game: Chimpanzees and bonobos are insensitive to unfairness. *Biology Letters*, *8*(6), 942–945. https://doi.org/10.1098/rsbl.2012.0519

Kawai, N., Nakagami, A., Yasue, M., Koda, H., & Ichinohe, N. (2019). Common marmosets (*Callithrix jacchus*) evaluate third-party social interactions of human actors but Japanese monkeys (*Macaca fuscata*) do not. *Journal of Comparative Psychology*, *133*(4), 488–495. https://doi.org/10.1037/com0000182

Kawai, N., Yasue, M., Banno, T., & Ichinohe, N. (2014). Marmoset monkeys evaluate third-party reciprocity. *Biology Letters*, *10*(5), 20140058. https://doi.org/10.1098/rsbl.2014.0058

Kinzler, K. D., & Spelke, E. S. (2007). Core systems in human cognition. In *Progress in brain research* (pp. 257–264). Elsevier. https://doi.org/10.1016/S0079-6123(07)64014-X

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694–698. https://doi.org/10.1111/j.0956-7976.2004.00742.x

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Krupenye, C., & Hare, B. (2018). Bonobos prefer individuals that hinder others over those that help. *Current Biology*, *28*(2), 280–286.e5. https://doi.org/10.1016/j.cub.2017.11.061

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). “Who can help me fix this toy?” The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology*, *49*(3), 446–453. https://doi.org/10.1037/a0031649

Lane, J. D., & Harris, P. L. (2015). The roles of intuition and informants' expertise in children's epistemic trust. *Child Development*, *86*(3), 919–926. https://doi.org/10.1111/cdev.12324

Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A tale of three platforms: Investigating preschoolers' second-order inferences using in-person, Zoom, and Lookit methodologies. *Frontiers in Psychology*, *12*, 731404. https://doi.org/10.3389/fpsyg.2021.731404

Leuba, C. (1933). An experimental study of rivalry in young children. *Journal of Comparative Psychology*, *16*(3), 367–378. https://doi.org/10.1037/h0070972

Liu, D., Gelman, S. A., & Wellman, H. M. (2007). Components of young children's trait understanding: Behavior-to-Trait inferences and trait-to-behavior predictions. *Child Development*, *78*(5), 1543–1558. https://doi.org/10.1111/j.1467-8624.2007.01082.x

Lucca, K., Pospisil, J., & Sommerville, J. A. (2018). Fairness informs social decision making in infancy. *PLOS ONE*, *13*(2), e0192848. https://doi.org/10.1371/journal.pone.0192848

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*(4), 1073–1084. https://doi.org/10.1111/1467-8624.00458

MacDonald, K., Schug, M., Chase, E., & Barth, H. (2013). My people, right or wrong? Minimal group membership disrupts preschoolers' selective trust. *Cognitive Development*, *28*(3), 247–259. https://doi.org/10.1016/j.cogdev.2012.11.001

Marsh, H. W., Ellis, L. A., & Craven, R. G. (2002). How do preschool children feel about themselves? Unraveling measurement and multidimensional self-concept structure. *Developmental Psychology*, *38*(3), 376–393. https://doi.org/10.1037/0012-1649.38.3.376

Massen, J. J. M., Van Den Berg, L. M., Spruijt, B. M., & Sterck, E. H. M. (2012). Inequity aversion in relation to effort and relationship quality in long-tailed macaques (*Macaca fascicularis*). *American Journal of Primatology*, *74*(2), 145–156. https://doi.org/10.1002/ajp.21014

McAuliffe, K., Blake, P. R., Kim, G., Wrangham, R. W., & Warneken, F. (2013). Social influences on inequity aversion in children. *PLoS ONE*, *8*(12), e80966. https://doi.org/10.1371/journal.pone.0080966

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. Chapman; Hall.

McNamara, J. M., Barta, Z., Fromhage, L., & Houston, A. I. (2008). The coevolution of choosiness and cooperation. *Nature*, *451*(7175), 189–192. https://doi.org/10.1038/nature06455

Melis, A. P., Hare, B., & Tomasello, M. (2006). Chimpanzees recruit the best collaborators. *Science*, *311*(5765), 1297–1300. https://doi.org/10.1126/science.1123007

Mesterton-Gibbons, M., & Dugatkin, L. A. (1992). Cooperation among unrelated individuals: Evolutionary factors. *The Quarterly Review of Biology*, *67*(3), 267–281. https://doi.org/10.1086/417658

Milich, K. M., & Maestripieri, D. (2016). Sex or power? The function of male displays in rhesus macaques. *Behaviour*, *153*(3), 245–261. https://doi.org/10.1163/1568539X-00003340

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology*, *49*(3), 404–418. https://doi.org/10.1037/a0029500

Mossler, D. G., Marvin, R. S., & Greenberg, M. T. (1976). Conceptual perspective taking in 2- to 6-year-old children. *Developmental Psychology*, *12*(1), 85–86. https://doi.org/10.1037/0012-1649.12.1.85

Neiworth, J. J., Johnson, E. T., Whillock, K., Greenberg, J., & Brown, V. (2009). Is a sense of inequity an ancestral primate trait? Testing social inequity in cotton top tamarins (*Saguinus oedipus*). *Journal of Comparative Psychology*, *123*(1), 10–17. https://doi.org/10.1037/a0012662

Noë, R., & Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11. https://doi.org/10.1007/BF00167053

Noë, R., & Hammerstein, P. (1995). Biological markets. *Trends in Ecology & Evolution*, *10*(8), 336–339. https://doi.org/10.1016/S0169-5347(00)89123-5

Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., & Hartley, C. A. (2020). Moving developmental research online: Comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology*, *6*(1), 17213. https://doi.org/10.1525/collabra.17213

Oberliessen, L., & Kalenscher, T. (2019). Social and non-social mechanisms of inequity aversion in non-human animals. *Frontiers in Behavioral Neuroscience*, *13*, 133. https://doi.org/10.3389/fnbeh.2019.00133

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*(7016), 499–502. https://doi.org/10.1038/nature02978

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216–1226. https://doi.org/10.1037/0012-1649.43.5.1216

Perner, J., Zauner, P., & Sprung, M. (2005). What does 'that' have to do with point of view? Conflicting desires and 'want' in German. In *Why language matters for theory of mind*. Oxford University Press.

Perry, S., Barrett, H., & Manson, J. H. (2004). White-faced capuchin monkeys show triadic awareness in their choice of allies. *Animal Behaviour*, *67*(1), 165–170. https://doi.org/10.1016/j.anbehav.2003.04.005

Priewasser, B., Roessler, J., & Perner, J. (2013). Competition as rational action: Why young children cannot appreciate competitive games. *Journal of Experimental Child Psychology*, *116*(2), 545–559. https://doi.org/10.1016/j.jecp.2012.10.008

R-Core-Team. (2021). *R: a language and environment for statistical computing*. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/

Robinson, E. J., & Einav, S. (Eds.). (2014). *Trust and skepticism: Children's selective learning from testimony*. Psychology Press.

Roma, P. G., Silberberg, A., Ruggiero, A. M., & Suomi, S. J. (2006). Capuchin monkeys, inequity aversion, and the frustration effect. *Journal of Comparative Psychology*, *120*(1), 67–73. https://doi.org/10.1037/0735-7036.120.1.67

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, *81*(5), 1068–1095. http://www.jstor.org/stable/2006907

Russell, Y. I., Call, J., & Dunbar, R. I. (2008). Image scoring in great apes. *Behavioural Processes*, *78*(1), 108–111. https://doi.org/10.1016/j.beproc.2007.10.009

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Sanfey, A. G. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755–1758. https://doi.org/10.1126/science.1082976

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online testing yields the same results as lab testing: A validation study with the false belief task. *Frontiers in Psychology*, *12*, 703238. https://doi.org/10.3389/fpsyg.2021.703238

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. https://doi.org/10.1093/beheco/arn145

Schino, G., & Aureli, F. (2017). Reciprocity in group-living animals: partner control versus partner choice. *Biological Reviews*, *92*(2), 665–672. https://doi.org/10.1111/brv.12248

Schmidt, M. F. H., & Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. *PLoS ONE*, *6*(10), e23223. https://doi.org/10.1371/journal.pone.0023223

Scott, K., Chu, J., & Schulz, L. (2017). Lookit (part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind*, *1*(1), 15–29. https://doi.org/10.1162/OPMI_a_00001

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, *141*(2), 382–395. https://doi.org/10.1037/a0025907

Sheridan, S., & Williams, P. (2006). Constructive competition in preschool. *Journal of Early Childhood Research*, *4*(3), 291–310. https://doi.org/10.1177/1476718X06067581

Silberberg, A., Crescimbene, L., Addessi, E., Anderson, J. R., & Visalberghi, E. (2009). Does inequity aversion depend on a frustration effect? A test with capuchin monkeys (*Cebus apella*). *Animal Cognition*, *12*(3), 505–509. https://doi.org/10.1007/s10071-009-0211-6

Silk, J. B. (1999). Male bonnet macaques use information about third-party rank relationship to recruit allies. *Animal Behaviour*, *58*, 45–51.

Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science*, *23*(2), 196–204. https://doi.org/10.1177/0956797611422072

Smith, E. D., & Lillard, A. S. (2012). Play on: Retrospective reports of the persistence of pretend play into middle childhood. *Journal of Cognition and Development*, *13*(4), 524–549. https://doi.org/10.1080/15248372.2011.608199

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4), 779–797. https://doi.org/10.1037/a0034191

Sterelny, K. (2007). Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 719–730. https://doi.org/10.1098/rstb.2006.2006

Stipek, D. J., Gralinski, J. H., & Kopp, C. B. (1990). Self-concept development in the toddler years. *Developmental Psychology*, *26*(6), 972–977. https://doi.org/10.1037/0012-1649.26.6.972

Subiaul, F., Vonk, J., Okamoto-Barth, S., & Barth, J. (2008). Do chimpanzees learn reputation by observation? Evidence from direct and indirect experience with generous and selfish strangers. *Animal Cognition*, *11*(4), 611–623. https://doi.org/10.1007/s10071-008-0151-6

Thierry, B. (2000). Covariation of conflict management patterns across macaque species. In F. Aureli & F. B. M. de Waal (Eds.), *Natural conflict resolution* (p. 409). University of California Press.

Tsiakara, A. A., & Digelidis, N. M. (2021). The competition in preschool age: a short review. *European Journal of Education Studies*, *8*(8). https://doi.org/10.46827/ejes.v8i8.3859

Tsuji, S., Amso, D., Cusack, R., Kirkham, N., & Oakes, L. M. (2022). Empirical research at a distance: New methods for developmental science. *Frontiers in Psychology*, *13*, 938995. https://doi.org/10.3389/fpsyg.2022.938995

Ulber, J., Hamann, K., & Tomasello, M. (2016). Extrinsic rewards diminish costly sharing in 3-year-olds. *Child Development*, *87*(4), 1192–1203. https://doi.org/10.1111/cdev.12534

Ulber, J., Hamann, K., & Tomasello, M. (2017). Young children, but not chimpanzees, are averse to disadvantageous and advantageous inequities. *Journal of Experimental Child Psychology*, *155*, 48–66. https://doi.org/10.1016/j.jecp.2016.10.013

VanderBorght, M., & Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development*, *18*(1), 61–71. https://doi.org/10.1002/icd.591

van Wolkenten, M., Brosnan, S. F., & de Waal, F. B. M. (2007). Inequity responses of monkeys modified by effort. *Proceedings of the National Academy of Sciences*, *104*(47), 18854–18859. https://doi.org/10.1073/pnas.0707182104

Vernon, P. A. (1985). Individual differences in general cognitive ability. In *The neurophyschology of individual differences* (pp. 125–150). Springer Science+Business Media.

Warneken, F., & Tomasello, M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology*, *44*(6), 1785–1788. https://doi.org/10.1037/a0013860

Wilks, M., Collier-Baker, E., & Nielsen, M. (2015). Preschool children favor copying a successful individual over an unsuccessful group. *Developmental Science*, *18*(6), 1014–1024. https://doi.org/10.1111/desc.12274

Wynne, C. D. L. (2004). Fair refusal by capuchin monkeys. *Nature*, *428*(6979), 140–140. https://doi.org/10.1038/428140a

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences*, *106*(28), 11520–11523. https://doi.org/10.1073/pnas.0900636106

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Yee, N., & Bailenson, J. (2007). The proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research*, *33*(3), 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x

Zaadnoordijk, L., & Cusack, R. (2022). Online testing in developmental science: A guide to design and implementation. In *Advances in child development and behavior* (pp. 93–125). Elsevier. https://doi.org/10.1016/bs.acdb.2022.01.002

Zarei, N., Chu, S. L., Quek, F., Rao, N. J., & Brown, S. A. (2020). Investigating the effects of self-avatars and story-relevant avatars on children's creative storytelling. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3313831.3376331

Acknowledgements

First and foremost, I would like to acknowledge my three supervisors – Julia Fischer, Hannes Rakoczy and Stefanie Keupp. Julia, I feel truly privileged to have been able to count myself a Koggie. Thank you for the many helpful discussions, the honest feedback (the steady supply of commas), and the unwavering encouragement. Hannes, in my second year your department became my second home – thank you for the opportunity to work in your lab, for the guidance these past years and for so generously sharing your knowledge and time. Steffi, your base changed so many times these last years, yet you could not have been more of a constant in my PhD journey. Thank you for being there to share in and guide me through the many ups and downs of this journey.

I would like to thank Igor Kagan, Julia Ostner and Tanya Behne for having agreed to appear on my examination board. Additionally, I would like to thank Igor Kagan for encouragement and constructive feedback during thesis committee meetings.

I am grateful for the support I received as a member of the Research Training Group 2070 – thank you to my peers in the second cohort, the PIs who facilitated interdisciplinary exchange and to the coordination team – Rebecca Jürgens and Valerie Liebs.

A number of people were key to the success of the social disappointment project. Thank you Timo Hüser and Matthias Dörge for help in constructing the macaque apparatus (the “tables of doom”). I'm especially grateful to Timo who spent a lot of time maintaining the tables toward the end of testing. Thank you, Constance Thiriau, Lisa Ohrndorf and Carolin Kade for help with data collection. My thanks to Ludwig Ehrenreich for technical support.

I am grateful to the students who worked alongside me on the child project: Imke Koch, Christiane Decker, Lia Künnemann, Josephine Quirbach and Mellory Kripzak. Thank you also to Marlen Kaufmann for help in recruiting participants.

Although the project folded, it would be remiss of me not to mention the risk team. Thank you, Louis Frank, Judit Stolla and Carina Bruchmann for the countless hours you put into the risk(y) project that *almost* made it off the ground.

My thanks to Roger Mundry for always having an open door (or zoom room), and the patience to answer yet another plotting question.

I was fortunate to have two departments to call home, and I'd like to thank my colleagues and friends in the Cog Etho and the Kindsköpfe labs for the camaraderie and support these last years. Special thanks go to my finishing cohort – Anaïs Aviles de Diego, Dominique Treschnak and Lukas Schad.

Last but not least, I gratefully acknowledge the love and support of my family – Alexander Pfaff and Mark, Maureen and Hugh Titchener.