Cognitive Development 66 (2023) 101313

Contents lists available at ScienceDirect

Cognitive Development

journal homepage: www.elsevier.com/locate/cogdev

Tokens of virtue: Replicating incentivized measures of children's prosocial behavior with online methods and virtual resources

Richard E. Ahl [a,*], Kelsey Hannan [a], Dorsa Amir [a,b], Aaron Baker [a], Mark Sheskin [c], Katherine McAuliffe [a]

[a] *Boston College, USA*
[b] *University of California, Berkeley, USA*
[c] *Minerva University, USA*

ARTICLE INFO

Keywords:
Resources
Prosocial development
Incentivized measures
Online replications
Fairness
Trustworthiness

ABSTRACT

Desirable resources are crucial for incentivized tasks of prosocial behavior. Developmentalists have often used tangible items, such as candy or stickers, as the resources in such tasks. However, such resources are infeasible for online testing, which has become popular in recent years. We investigated whether online methods, using virtual tokens to be traded for prizes, are viable for incentivized tasks with children. We conducted a pre-registered online replication ($n = 87$) of two tasks (Trustworthiness and Inequity Game) for children ages 6 through 11. We compared the results to a sample of participants ($n = 60$) we had tested in-person using candy. We successfully replicated our Trustworthiness results, but our Inequity Game results differed based on testing format. Older children appeared reluctant to "waste" resources in the online sample, suggesting greater efficiency concerns with tokens than candy. Implications for online methods and the use of diverse resource types are discussed.

1. Introduction

It is easy to say that one will be good, but it is harder to actually be good when being good comes at a cost to the self. Indeed, knowing what one should do does not always result in the choice to do it. This is true of children as well as adults (Baumeister, Vohs, & Funder, 2007; Blake, 2018). Because people may say one thing and do another, many psychologists and behavioral economists use incentivized tasks to draw out actual decisions in domains such as fairness (Blake, McAuliffe, & Warneken, 2014) and generosity (Smith, Blake, & Harris, 2013).

1.1. Using incentives to study prosocial behavior with children

Incentivized tasks such as the Prisoner's Dilemma (Rand & Nowak, 2013), Ultimatum Game (Güth, Schmittberger, & Schwarze, 1982;), and Dictator Game (Engel, 2011) are well-suited for studying prosocial behavior in adults because they enable the measurement of costly actions themselves. In such tasks, participants make choices that they believe have consequences for themselves and others, and doing the right thing carries costs for oneself. Incentivized tasks can also illuminate children's behavior and have been

[*] Correspondence to: Boston College, Department of Psychology and Neuroscience, McGuinn Hall, 275 Beacon St., Chestnut Hill, MA 02467, USA.
E-mail address: richard.ahl@bc.edu (R.E. Ahl).

https://doi.org/10.1016/j.cogdev.2023.101313
Received 30 June 2022; Received in revised form 19 January 2023; Accepted 8 February 2023

modified successfully for use in developmental studies (Benenson, Pascoe, & Radmore, 2007; Blake, Rand, Tingley, & Warneken, 2015; Gonzalez, Ahl, Cordes, & McAuliffe, 2022). These measures work because, much like adults, children are acutely aware of costs and benefits (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Liu, Gonzalez, & Warneken, 2019) and are motivated to earn rewards. In order for a resource to be an effective reward, the resource must be viewed as valuable and desirable.

It stands to reason that the specific resource used in a study will influence how much participants value that resource, with important consequences for their allocation decisions. Highly valued resources are more difficult to part with, and impose greater costs on someone who shares them, than less valued resources. Highly valued resources also raise special efficiency concerns. This is particularly true for tasks in which certain decisions cause resources to be wasted (e.g., versions of the Ultimatum Game and tasks measuring inequity aversion; see Blake & McAuliffe, 2011; Shaw & Olson, 2012; Zhang & Benozio, 2021), since people may be unwilling to squander such resources.

With adult studies, money is frequently the resource of exchange, often in the form of tokens or points that are subsequently exchanged for money. Money's fungibility means that it works as a powerful incentive for adults; it can be converted into a favored resource of the individual's choosing. However, direct monetary compensation via cash is less suitable for children. It raises ethical concerns as well as methodological problems due to young children's difficulty understanding how money works (Berti & Bombi, 1981) and the potential for long delays between when the money is acquired and when it can be spent (see Ebersbach, Krupa, & Vogelsang, 2022, on delay and decision-making).

With awareness of such concerns, developmental studies in which participants make choices about resources for themselves and others do not typically use money as a resource. Stickers (Benenson, Pascoe, & Radmore, 2007; Blake & Rand, 2010; Cowell, Samek, List, & Decety, 2015; Engelmann, Over, Herrmann, & Tomasello, 2013; Li, Li, Decety, & Lee, 2013; Malti, Gummerum, Ongley, Chaparro, Nola, & Bae, 2016; Samek et al., 2020; Tasimi & Young, 2016) and candy or other sweets (Allgaier, Ścigała, Trautwein, Hilbig, & Zettler, 2020; Amir, Parsons, Ahl, & McAuliffe, 2021; Blake & McAuliffe, 2011; Blake, Corbit, Callaghan, & Warneken, 2016; Fehr, Bernhard, & Rockenbach, 2008; McAuliffe, Raihani, & Dunham, 2017) are especially common resources. Other studies have used erasers (Shaw Olson, 2012), virtual resources in the form of animal videos (Mandalaywala, Benitez, Sagar, & Rhodes, 2021), and tokens to be exchanged for pre-specified goods (stickers; Ebersbach, Krupa, & Vogelsang, 2022; school supplies; Bereby-Meyer & Fiks, 2013) or unspecified prizes (House & Tomasello, 2018; House et al., 2020) at the end of the study. The notion that resource type matters has motivated developmental researchers to vary specific resource types across trials (e.g., diverse candy types; Fehr et al., 2008), allow participants to choose preferred resources (e.g., favorite stickers within a larger set; Blake & Rand, 2010; Li et al., 2013), or pilot resources with a separate group of participants to determine their value to children (Benozio & Diesendruck, 2015).

The relatively few studies that have tested within-category variation in resource value have found that children generally share less of resources they value highly, and more of resources they do not value highly (Benozio & Diesendruck, 2015; Blake & Rand, 2010). In these particular cases, this means sharing less of "favorite" stickers than of "least-favorite" stickers. When choosing to allocate stickers between two individuals when their own resources are not at stake, children tend to give more of the stickers they themselves prefer, and less of the stickers they deem less desirable, to individuals they like more (Chernyak & Sobel, 2016).

Studies that have used different resource categories with the goal of evoking differences in resource value have also found that resource kind influences children's sharing decisions. Children are likelier to share resource kinds they like less (e.g., a pencil) than resource kinds they like more (e.g., a bouncy ball) (Sheskin et al., 2016). We are aware of few studies comparing children's allocations amongst resource kinds that are intended to be similarly valuable. Warneken et al. (2011) found that 3-year-olds were likelier to fairly share gummy bears than stickers after a collaborative task with a partner. A follow-up study, however, found similar rates of sharing between food and non-food items. (It is worth noting that a separate group of children generally preferred the food items). Overall, it is clear that resource value and kind have the potential to affect children's allocation decisions, but whether children's decisions systematically vary between different kinds of desirable resource remain an open question.

1.2. Online testing and recruitment in developmental research

As a result of the COVID-19 pandemic, many developmental researchers have moved their in-person recruitment and data collection methods to an online format. This shift has been met with considerable curiosity, interest, and healthy skepticism (Tsuji, Amso, Cusack, Kirkham, & Oakes, 2022). Given that resource type may affect children's resource allocation decisions, will "virtual" resources work effectively for the online testing of children's prosocial behavior? Tangible resources that can be consumed immediately, which predominate in traditional testing, are not conducive to online use. (For instance, such resources would need to be provided to families prior to testing.) Adapting in-person tasks to web-based platforms may pose challenges for incentivized studies, potentially eclipsing genuine decision-making and instead providing a picture of what children think they *ought* to say. A key question for developmentalists hoping to use incentivized tasks, and the main motivation for our study, is whether online methods using virtual resources replicate in-person results. Next, we will provide an overview of recent advances in online data collection for developmental research.

Online data collection expands the pool of available participants from children who live within the vicinity of a research lab to children across the country, or even other parts of the world. Through web-based studies, families can participate in research at times that are convenient for them, without the need for travel. This ease of participation allows interested families who might not visit research laboratories due to logistical constraints, financial limitations, or scheduling restrictions to participate in studies online. As a result of expanding one's pool of eligible participants, the speed and efficiency of data collection can increase, allowing for the completion of well-powered studies and the targeting of groups that may be difficult to recruit in-person.

Online testing has important implications for sample diversity. The areas surrounding universities are particularly likely to have highly-educated residents; online testing expands participation to a wider range of families, not just to those who live near research labs (see Rhodes et al., 2020,on geographic diversity). While access to computers and high-speed Internet may restrict the socioeconomic diversity of web-based samples (Lourenco & Tasimi, 2020), several research groups have found that children tested online are more geographically and racially diverse, and less likely to come from high-income, college-educated families, than children tested through conventional means (Leshin, Leslie, & Rhodes, 2021; Rhodes et al., 2020; Scott & Schulz, 2017; Sheskin et al., 2020). (Online samples of adult participants are often more diverse than those tested in-person; see Casler, Bickel, & Hackett, 2013; Gosling, Sandy, John, & Potter, 2010.) Thus, online data collection has the potential to address developmental psychology's enduring problem of samples lacking in cultural, racial, geographic, international, and socioeconomic diversity (Henrich, Heine, & Norenzayan; 2010; Nielsen, Haun, Kärtner, & Legare, 2017; Rowley & Camacho, 2015).

With these considerations in mind, efforts to conduct developmental data collection online preceded the COVID-19 pandemic, although the use of online methods has surged since it began (Tsuji et al., 2022). A key distinction between online studies is whether they are "moderated," run by a human experimenter on a live video chat (e.g., Thechildlab.com, Sheskin & Keil, 2018) or "unmoderated," without a live experimenter (e.g., Lookit, Scott & Schulz, 2017; PANDA, Rhodes et al., 2020). Each approach has benefits and drawbacks (Kominsky, Begus, Bass, Colantonio, Leonard, Mackey, & Bonawitz, 2021); since the former approach most closely replicates the experience of a traditional in-person study and allows for comprehension check questions to be posed in real time, it will be our choice in the present study.

Several recent developmental projects have used online samples to address new research questions (Chuey, Lockhart, Sheskin, & Keil, 2020; Leshin, Leslie, Rhodes, 2021[1]; Richardson & Keil, 2022; Thomas, Woo, Nettle, Spelke, & Saxe, 2022). Others have used a mix of online and in-person participants to address new questions, with online participants tested after, and in most cases specifically because, of the COVID-19 pandemic (Aboody, Yousif, Sheskin, & Keil, 2022; Afshordi & Koenig, 2021; Good & Shaw, 2022; Kominsky, Shafto, & Bonawitz, 2021; Shu, Hu, Xu, & Bian, 2022[2]). In these cases, the inclusion of online and in-person data was not for the purpose of comparing these methods but for the sake of completing a research project when in-person data collection was inadvisable.

1.2.1. *Online replications of in-person studies*

Of special interest is the growing body of literature on studies that collect online samples for the sake of comparing them to in-person findings that either stem from new data collection or previously-published papers. In these studies, in-person tasks are converted to an online format and the results of these approaches are compared. In most cases, the success of the replication is evaluated by determining whether a given effect seen in-person (e.g., a significant difference between two experimental conditions of a study) is also seen online, with similar effect sizes. In some cases (e.g., Schidelko, Schünemann, Rakoczy, & Proft, 2021; Lapidow, Tandon, Goddu, & Walker, 2021), values obtained in-person are directly compared to values obtained online, with similarity in values indicating a successful replication.

Consistent with many adult studies that have replicated in-person findings with online methods (Amir, Rand, & Gal, 2012; Buhrmester, Kwang, & Gosling, 2016; Horton, Rand, & Zeckhauser, 2011), several developmental studies have generated similar results in-person and online (Chuey et al., 2021; Schidelko, Schünemann, Rakoczy, & Proft, 2021; Vales et al., 2021), with participants from infancy through middle childhood yielding similar results across formats. A study on decision-making and learning strategies using a computer-based task both online and in-person, for participants in middle childhood through adolescence, also generated similar results across formats (Nussenbaum, Scheuplein, Phaneuf, Evans, & Hartley, 2020[3]).

Other online replications have had more mixed successes. Studies that attempted several separate replications of classic findings in cognitive and social development in childhood, with a moderated format (Sheskin & Keil, 2018), and in infancy, with moderated (Smith-Flores, Perez, Zhang, & Feigenson, 2022) and unmoderated formats (Scott, Chu, & Schulz, 2017), have found close replications on some tasks but not others. A recent online replication of a contemporary cognitive development study with participants in early childhood yielded non-significant effects across multiple online versions, although a final online replication using an older age range than the original sample yielded significant effects (Lapidow et al., 2021). The studies that have not yielded close replications include a mix of weaker yet still significant effects (i.e., smaller effect sizes that are nonetheless significant), and weak, non-significant effects in the online versions.

1.3. *Current study*

Taken as a whole, the use of online methods for conducting developmental research is promising, but not without caveats. A crucial absence in the recent group of online replications is that of tasks using resources to measure costly prosocial behavior. We are unaware of any published studies directly comparing children's online and in-person performance in such tasks. As mentioned previously, the effectiveness of costly prosocial tasks depends on children understanding, and caring about, the resources used in such tasks. Thus, the use of "virtual" resources online, in place of tangible rewards, raises issues that are unique to costly prosocial measures. Aside from this key concern, online studies differ from in-person studies in other ways that could be meaningful for prosocial tasks specifically. For

[1] Leshin et al. (2021), also provided a successful conceptual replication of effects seen with in-person testing.

[2] Aboody, Yousif, Sheskin, & Keil, 2022, and Afshordi & Koenig, 2021 under review, conducted analyses comparing online and in-person data. Aboody et al., 2022, found similar effects, while Afshordi & Koenig, 2021 under review, found some differences in online and in-person results.

[3] Incentives were offered in this study, but the incentives were performance-based and thus differ substantially from the "self vs. other" incentives in the current paper.

instance, the salience of the experimenter may differ across testing formats (e.g., might children feel more "anonymous" online?), and the recipients of children's prosocial actions could seem more "abstract" or less "real" online than in-person, which could reduce children's motivation to share with them.

Thus, we set out to conduct online replications of two tasks measuring children's costly choices regarding trustworthiness (Amir et al., 2021) and fairness (Inequity Game; Blake & McAuliffe, 2011). These tasks had recently been run in our lab in-person as part of a larger cross-cultural study (Amir et al., 2023) . Details about each task are provided in the Method. For Trustworthiness, participants chose how many resources to give to another child who forwent resources in order to allow the participant to receive more resources. Upholding trustworthiness entailed giving resources to the other child, at the expense of one's own resources. For Fairness, participants chose whether to accept or reject allocations that either were equal for themselves and another child or were more advantageous for themselves than another child. Upholding fairness entailed rejecting unequal allocations, which advantaged the participants, resulting in receiving fewer resources. (The decision to accept *equal* allocations was consistent with fairness.) We also ran an exploratory task measuring generosity (Dictator Game; Benenson et al., 2007). This task served as a within-subjects conceptual replication of a between-subjects condition of Amir et al., 2021, which compared trustworthiness and generosity. For Generosity, participants chose how many resources to give to a child with no resources, at the expense of their own resources. Because Generosity was not run in our in-person sample, we cannot directly compare online Generosity results to in-person results.

We designed our online methods to closely replicate our own in-person methods, with substantial overlap in scripts, study personnel, and training procedures, and with full knowledge of the effects obtained with our in-person sample. Although our in-person study was conducted earlier, we will present the online study first in the Method section and then briefly mention ways in which the in-person study differed. We believe this approach is most conducive to highlighting our online testing protocol, which we hope will be useful to other researchers. We note that the in-person study was not designed with an online replication in mind. Our online study, which was designed and run later, was created exclusively for this purpose.

Our in-person study was conducted face-to-face with physical materials (e.g., apparatuses for Trustworthiness and Fairness), while the online study was conducted via computer screens with images and animations instead of physical materials. In piloting, we found that references to the "digital" nature of the images (e.g., "virtual token bowl" rather than "token bowl") were unnecessary. Children had no difficulty accepting the virtual materials as "real," and so the online script generally discussed materials using the exact language as the in-person script. While our in-person sample had an age range of 5 – 13-year-olds, we set our 6—11-year-old age range for the online sample because some younger children struggled with the online procedure during piloting. We chose a priori to not test 12- and 13-year-olds online due to concerns about recruitment for this age group. Our reported analyses include data from 6—11-year-olds only, as described in our pre-registration (https://aspredicted.org/NNF_HJW).

The key conceptual difference between our online and in-person methods is the use of resources. Do physical incentivizes translate well to an online format for children? Can abstract resources be used to study costly prosocial behavior? In our in-person study, the resources were candies (Skittles® for Inequity Game and Starburst® for Trustworthiness). In our online study, the resources were digital tokens described as being exchanged for prizes at the conclusion of the study, with the number collected explained as determining the prize that would ultimately be received. (Ultimately, the tokens were exchanged for Amazon.com gift cards, which varied slightly – $6 or $7 – based on the number collected; thus, they were actually exchanged for prizes, albeit indirectly.) Thus, the resource type differed in a few respects, including whether the reward was tangible or digital, with a clear or ambiguous payoff structure; taken as-is or exchanged for another good; edible candy or of an unclear kind.

We could have more closely controlled the contrast between the in-person and online resources (e.g., online tokens explained as being exchanged for units of candy that exactly matched the in-person resources). However, we believed using tokens that were not directly linked to candy, if yielding a successful replication, would be more useful to our lab and other researchers. Sending candy is logistically difficult, not all children can eat candy, and many labs already use online gift cards as participant compensation. We note that some in-person studies have used tokens to be exchanged for unspecified prizes as compensation (House & Tomasello, 2018; House et al., 2020).

We are aware of only one resource allocation study with children that directly compared tokens to physical rewards. In a task that merged procedures of an Ultimatum Game and Inequity Game, Ebersbach et al. (2022) found that 4—6-year-olds were likelier to reject both unfair and fair offers when the resource was tokens (exchanged for stickers at a 1:1 ratio) than stickers. While this suggests that children are more willing to part with tokens than physical resources, the implications for the current study are limited because tokens in the current study were exchanged for unspecified prizes rather than candy. For adults, a recent study using the Inequity Game found that adults were less likely to reject unfair offers that favored the participants (i.e., advantageous inequity offers) when tokens to be exchanged for money were used as a resource than when candy was used (McAuliffe, Benjamin, & Warneken, 2022).

1.3.1. *Replication criteria*

In order to evaluate the success of our online replication, and therefore the viability of using virtual incentivized resources in behavioral measures with children, we will consider two key criteria. The first criterion, Research Question 1 (RQ1), is whether *effects* observed in-person are also observed online. For instance, in the in-person sample, sharing increases along with age in Trustworthiness; is this also true of the online sample? This will be investigated by conducting analyses with the in-person data and then replicating those analyses with the online data, using separate models. Most recent studies evaluating the effectiveness of online replications take this approach (e.g., Chuey et al., 2021;Smith-Flores et al., 2022). We view RQ1 as crucial for the current study, as it examines whether conclusions about the phenomena of interest (e.g., age-related changes in performance) are common to both samples. The second criterion, Research Question 2 (RQ2), is whether *values* observed in-person are similar to those observed online. For example, is the mean amount of resources shared in Trustworthiness by in-person participants similar to that of online participants? This approach is less commonly used (see Schidelko et al., 2021; Lapidow et al., 2021). It yields important information as well but is less crucial than RQ1. It is possible for a study's replication to succeed according to one criterion but not the other; please see Supplement 1 for examples. We pre-registered our predictions that our replication would succeed according to both criteria for Fairness and Trustworthiness (https://aspredicted.org/NNF_HJW). Our sampling and analytic approach generally aligned with our pre-registration, with a few exceptions; these will be noted in the Method. If our in-person results do not replicate online, this would suggest limitations of testing incentivized tasks with children online, perhaps due to the use of virtual tokens specifically.

2. Method

2.1. Participants

Our in-person sample included 60 participants between the ages of 6 and 11 years ($M = 106.9$ months, $SD = 22.08$, range 72–145 months, 51.7% girls) tested in the summer and fall of 2019, before the start of the COVID-19 pandemic. Further details about the in-person sample are included in Supplement 2.

Our online sample included 87 children between the ages of 6 and 11 years ($M = 107.98$ months, $SD = 20.75$, range 72–143 months, 52.9% girls), tested in the fall of 2020 and winter of 2021, during the COVID-19 pandemic. Participants were contacted from two sources: a new list of families recruited via Facebook advertisements and a pre-existing database of participants who were recruited locally for in-lab studies. An additional 4 children were tested but excluded for all tasks due to chronic comprehension check problems ($n = 2$), family interference ($n = 1$), or a developmental disability ($n = 1$; Autism Spectrum Disorder). Two participants had exclusions for either Fairness or Trustworthiness but not the other (due to family interference or experimenter error[4]), yielding a sample of 86 participants for each task. Our sample size range (72−84) was set a priori and preregistered on aspredicted.org. We slightly exceeded our maximum to accommodate participants' siblings and other interested children who responded to recruitment emails. No analyses were conducted on the online data until the last participant was tested.

Our online sample size was based on the full sample of in-person participants ($n = 80$) and our analysis of the in-person data, which yielded several significant effects (e.g., age-related findings) for Trustworthiness and Fairness. (We note that only a subset of this in-person sample, $n = 60$, matching the ages of our online sample, was used for our analyses.) One limitation of our study is that we did not conduct a formal power analysis to set our online sample size. However, we conducted post-hoc power analyses and sensitivity analyses based on the effect of age on Advantageous Inequity rejections; please see Supplement 4 for details. These analyses revealed that our in-person sample was well-powered to detect this age effect, and our online sample was sufficient to detect a substantially smaller effect than the one we obtained in-person. Our online sample is also broadly consistent with cell sizes for Advantageous Inequity conditions of the Inequity Game in the published literature (e.g., Blake & McAuliffe, 2011; Gonzalez et al., 2020).

Our online sample included participants from across the United States (24.1% Midwest, 20.7% Northeast, 18.4% Mid-Atlantic, 16.1% Southeast, 13.8% West, 5.7% Southwest, 1.1% Northwest). The parents of 82 participants completed an additional demographic form. Since parents chose whether to answer each question based on their comfort level, the number of respondents varied across questions. Overall, participants' parents were highly educated (86.2% held a bachelor's degree or higher, out of 82 respondents), and higher in household income than the 2020 national median of $67,521 (the median bracket selected was $96,000–179,999, out of 67 respondents). Parents' mean subjective socioeconomic status (SES) rating was a 6.6 (out of 76 respondents) on a scale from 1 to 10, with higher ratings corresponding to feeling higher in status relative to other Americans. Our sample was majority-White, but with substantial numbers of non-White participants (60.3% White, 17.9% biracial or multiracial, 16.7% Asian, 2.6% Black and 2.6% Latinx, out of 78 respondents).

2.2. Procedure

2.2.1. Overview

Participants completed the core tasks in one of two orders, with Fairness and Trustworthiness counter-balanced; please see Appendix A for task orders. For the in-person sample, participants also completed two tasks measuring honesty (judged via whether participants returned extra candy on a single trial) and forgiveness (see Amir, Ahl, Parsons, & McAuliffe, 2021, for a task summary), which came before the core tasks for some participants. These tasks did not involve direct tradeoffs in resources between the participant and an anonymous child and will not be discussed here. For the online sample, an honesty task always came after the core tasks, and Generosity was always last. In the online honesty task, participants reported whether a number they were thinking of matched numbers that appeared onscreen for 12 trials, with tokens given for reported matches. In a deviation from the pre-registration, we are not presenting the online honesty task here because of length considerations and the lack of a directly comparable in-person component. (The online and in-person tasks measured honesty in very different ways and were not designed for the sake of direct comparisons). Task order was always used as a covariate in our models; in no case was the effect of task order significant. Online data were collected using Qualtrics. In-person data were collected with live coding sheets recording participants' choices.

[4] Experimenter error was mistakenly left out of the pre-registration as part of the exclusion criteria; the rest of the criteria were pre-registered.

2.2.2. *Pre-study tasks and post-study debriefing*

Parents provided consent (via an online form or a physical form) and participants provided assent before testing began. Participants were told they would do activities in which they would make decisions and ultimately receive prizes; if they ever wanted to stop, they could do so and still receive prizes. For the online and in-person samples, parents were encouraged to give their children space and privacy during the study. More information about the pre-study procedures is provided in Supplement 5. After the study was finished, participants were debriefed and informed that the other children in the study were not real. For online participants, Amazon.com gift cards, which children could use to pick their own prizes, were emailed to families. In-person participants received the candies obtained during the study.

2.2.3. *Introductory tasks*

2.2.3.1. Online. Participants were told that they would do activities in which they could get tokens. The tokens would be added to their token bowls, counted via computer at the end of the activities, and then traded in for a prize. The number of tokens collected would determine the prize's quality: "the more tokens you get, the bigger and better the prize." Participants were asked if they wanted tokens; all responded affirmatively. They were described as being matched up with three "other boys/girls" (gender-matched with the participating child) who could not be present but "really wanted tokens." They could receive tokens in their own bowls and trade them for prizes later. Each of these children would be involved in only one activity. The participant and the other children were depicted with translucent avatars upon which different-colored circles had been placed (brown for the participant; white, gray, and black for the others, sometimes with bowls as well). This approach emphasized that the other children were distinct from each other and "real," like the participant. For the circles, we used neutral colors that were not encountered in the rest of the study and that we believed children would feeling similarly towards; see Fig. 1. Comprehension checks were posed to participants during each section of the study; they are outlined in Supplement 3. For the online and in-person tasks, all included participants gave the correct answer on their first attempt or after a single correction for all questions.

Next, participants were shown an image of a cartoon frog with 6 tokens and a cartoon cat with 0 tokens. The frog was explained as choosing how to divide tokens between himself and the cat. A new image was shown with options ranging from keeping 6 and giving 0 through keeping 0 and giving 6. A letter was displayed next to each option (e.g., "A" for keeping 6 and giving 0). The options were explained to participants, who were then asked which letter corresponded to each of the seven options. All included participants were able to complete this task on their first attempt or with one or more explanations. Although the task was not explained as such, its purpose was to introduce participants to the method they would later use to indicate choices privately in Trustworthiness: saying their answers to the computer without the use of a cursor or assistance from the experimenter.[5]

2.2.3.2. In-person. Instead of tokens, participants were told that they would do activities in which they could get candy (Skittles and Starbursts). Candies received during the activities would be collected in a paper bag and given to participants at the study's end. All included participants expressed an interest in receiving candy. Participants were told that they would be matched up with "other boys/girls" who could not be present but wanted candy, which they could receive later. These children were represented via paper bags with different colored stickers on them, much like the participant's own paper bag, using the same colored circles as the online study's avatars. Since participants' Trustworthiness decisions were made by distributing real Starbursts, no preliminary "frog-cat" task was needed to teach participants how to indicate their choices.

2.2.4. *Trustworthiness*

2.2.4.1. Online. In the Trustworthiness task (modeled after Amir, Parsons, Ahl, & McAuliffe, 2021), participants were introduced to a virtual model of the Trust apparatus; see Fig. 2. The apparatus consisted of one "closer" dish containing two tokens and one "farther" dish containing six tokens. Initially, participants were "given" a virtual short tool that could only reach the closer dish, as shown with an animated video. Next, participants were told that another child (shown with a black circle) who also wanted tokens had been given a choice of using their short tool to reach the closer dish or giving it to the participant, who could combine the tools together to make a long tool capable of reaching the farther dish. This child chose to give their short tool to the participant, with the knowledge that the participant had the ability to share tokens with them. As a result, this child received no tokens.

Next, a long tool for the participants appeared onscreen. This tool was then used by the experimenter to reach the farther dish with six tokens. Participants were told that the six tokens were now theirs. With the aid of an image showing all seven possible resource allocations (ranging from 6 for the participant and 0 for the other child to the opposite), participants decided how many of their tokens to give to the other child and said their choice out loud. To allow participants to make their choice privately, without the experimenter present, participants were told that "the computer would record their decision" while the experimenter stepped away and turned their camera off, ostensibly unable to hear the participants' choice. On comprehension checks, all included participants gave the correct answer on their first attempt or after a single correction, as was the case with the in-person sample.

[5] We had piloted a method of giving participants remote access to the cursor, but many younger participants had difficulty using a mouse or trackpad without help.

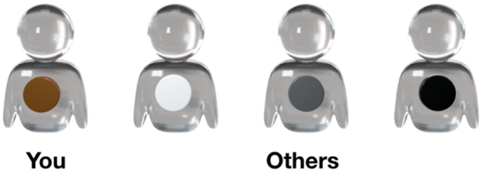

Fig. 1. Images of avatars.

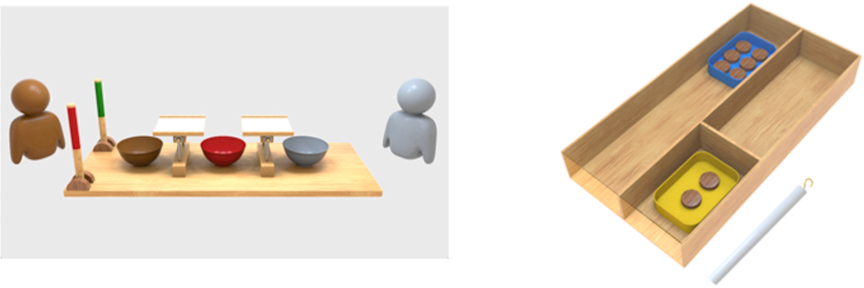

Fig. 2. Sample images of the Fairness and Trustworthiness apparatuses used for online testing.

2.2.4.2. In-person. Instead of a virtual apparatus conveyed with images and videos, a physical apparatus was used, with real tools that were manipulated by the participant and experimenter. Starbursts were used instead of tokens. Participants enacted their decisions by transferring six Starbursts from the "farther" dish to paper bags, one for themselves and one for the other child. Decisions were made privately by having the experimenter turn away and write on a piece of paper while the participant allocated the Starbursts.

2.2.5. Fairness

2.2.5.1. Online. In the Fairness task (modeled after Blake & McAuliffe, 2011; Blake et al., 2015), participants were introduced to a virtual model of the Fairness apparatus, consisting of two trays, three bowls, and two handles; see Fig. 2. The experimenter explained that the tokens on one side of the apparatus were for the other child (shown with a gray circle), and those on the opposite side were for the participant. Using animated videos, the experimenter demonstrated that pulling the green handle caused tokens on the other child's side to fall into that child's bowl and those on the participant's side to fall into their own bowl. Pulling the red handle caused the tokens to fall into the middle bowl, where neither participant could access them ("nobody gets those"). On comprehension checks, all included participants gave the correct answer on their first attempt or after a single correction, as was the case with the in-person sample.

Next, participants were given three trials to familiarize them with the actions of the handles.[6] The first trial always had an Equal distribution (1 for the other child, 1 for the participant). The experimenter explained, "I'm putting one token on his/her tray, and I'm putting one token on your tray. Which color handle do you want to pull?" After participants stated their choice verbally, the experimenter clicked on a handle to enact the distribution. The distribution's outcome was displayed visually and summarized verbally by the experimenter (e.g., "see, you each get one token"). The next two trials were one each of "1–0″ (1 for the other child, 0 for the participant) or "0–1″ distributions, administered in a randomized order. If participants chose all green or all red handles for these three introductory trials, the experimenter asked which handle produced whichever effect was not produced by the previous choices. This question checked for participants' understanding of the unchosen handle. All included participants gave the correct answer on their first attempt.

Participants received seven trials during the main experimental portion of the study. Participants were not informed about the number of upcoming trials, and experimenters refrained from commenting except to straightforwardly narrate the trials' outcomes (e.g., "see, no one gets these tokens"). No record of how many tokens had been accumulated was provided for participants. The first six test trials were administered in a new randomized order for each participant and consisted of two "1–1″ trials (Equal) and four "1–4″ trials (Advantageous Inequity; 1 for the other child, 4 for the participant). The final test trial consisted of a single "0–4″ trial (Extreme Advantageous Inequity; 0 for the other child, 4 for the participant). This trial was included last (as was done in-person) because it was introduced as a unique trial type for this set of studies (the Inequity Game studies we based our current work on had "1–4″ trials but not "0–4″ trials; Blake & McAuliffe, 2011; McAuliffe, Blake, & Warneken, 2020; Tsoi & McAuliffe, 2020). This approach allowed the prior trials to be consistent with literature precedent of trial types while allowing us to test a new trial. Unlike for the regular Advantageous

[6] While these were considered introductory trials from our perspective as researchers and did not constitute key dependent variables of interest, the script did not explain them as such, and the tokens gathered in these trials were distributed to the token bowls.

Inequity trials, Extreme Advantageous Inequity trial rejections reduce the participant's payoff without reducing the other child's payoff (which is 0 tokens regardless of whether the trial is accepted or rejected). Recall that accepting Advantageous Inequity trials and the Extreme Advantageous Inequity trial creates unequal outcomes, whereas rejecting such trials (and accepting or rejecting Equal trials) creates equal outcomes.

2.2.5.2. In-person. Skittles were used instead of tokens. Instead of a virtual apparatus, a physical apparatus was used, with handles that were manipulated by the participant and bowls into which Skittles fell. A paper bag was used to represent the other child. The first six test trials were administered in one of several preset randomized orders, rather than fully randomized for each child (the "0–4″ trial was always last, as was the case in-person). At the conclusion of the task, the experimenter transferred Skittles from the participants' bowls to their paper bags. As was the case online, participants were not told how many tokens had been accumulated, although some participants looked into their own bowls during the test trials.

2.2.6. Generosity

2.2.6.1. Online. In the Dictator Game (modeled after Benenson et al., 2007), participants were given an additional endowment of 10 tokens. They were told that another child (shown with a white circle) did not have a chance to do the activity and had an empty token bowl with no tokens inside. In a procedure similar to the Trustworthiness allocation task, participants were shown all 11 possible resource allocations and chose how many tokens to give to the other child by saying their choice out loud while the experimenter stepped away. On comprehension checks, all included participants gave the correct answer on their first attempt or after a single correction. This task was not conducted with in-person participants.

2.3. General analytic approach

Research Question 1 (RQ1) is concerned with whether the online sample replicates effects seen in-person. Research Question 2 (RQ2) is concerned with whether values are similar in-person and online, i.e., whether there are testing format effects. (A successful replication would be indicated by a *lack* of effect for online versus in-person testing.) The results will be organized by task and research question.

3. Results

3.1. In-person results summary

Our in-person data were collected and analyzed before our online study was designed. To facilitate cross-study comparisons, the in-person results will be presented immediately before the RQ1 and RQ2 results in the rest of this section. We will briefly summarize our main in-person findings here. These main findings, which aligned with the predictions we had before conducting the in-person study, form the in-person component of our RQ1 predictions, namely that our online sample would replicate our in-person effects.

For Trustworthiness, giving increased significantly with age (main effect of age). For Fairness, participants were significantly likelier to reject Advantageous Inequity trials than Equal trials (main effect of trial type); this tendency increased significantly with age (significant age by trial type interaction), in line with prior work on the Inequity Game (Blake & McAuliffe, 2011; McAuliffe, Blake, & Warneken, 2020) Participants were significantly more likely to reject Advantageous Inequity trials with age (main effect of age), in line with prior work showing age-related increases in AI rejections (Blake & McAuliffe, 2011; Blake et al., 2015; Gonzalez et al., 2020). Participants had a non-significant increase in Extreme Advantageous Inequity trial rejections with age; this trial type was not included in the prior studies we used as precedent for the current work.

3.2. Trustworthiness

3.2.1. Analytic approach

We predicted significant age related-increases in sharing online (RQ1) and no significant effect of testing format (RQ2). Linear models were conducted with R statistical software (version 4.1.1, R Core Team, 2021). Figures were produced using ggplot2. The dependent measure was the number of tokens shared, which could range from 0 to 6. Depending on the analyses, the main predictors were age in months (continuous) or testing format, i.e., whether the test was conducted online or in-person (factor; binary). Task order (i.e., whether Trustworthiness or Fairness came first) and participant gender were entered as covariates.

When "model comparisons" are conducted, for this task and all others, they compare full models to null models. The full models include main predictors and covariates; null models include covariates exclusively.

3.2.2. In-person results

A linear model was run with age in months as the main predictor and covariates as described previously (covariates will be included in all the models to follow). This analysis found a significant effect of age ($\beta = .02$, $SE = .007$, $p = .003$), with older children sharing more than younger children, and no significant effects for our covariates ($ps > .12$).

3.2.3. RQ1

The above analysis was re-run on the online data. This analysis also found a significant effect of age ($\beta = .02$, $SE = .006$, $p = .003$), with older children sharing more than younger children, and additionally found a significant effect of gender, ($\beta = -.51$, $SE = .254$, $p = .047$), with boys sharing less than girls.

3.2.4. RQ2

A linear model was run on the in-person and online data, with testing format and age as the main predictors. This analysis found a significant effect of age ($\beta = .02$, $SE = .005$, $p < .001$), with older children sharing more than younger children, and no significant effect of testing format ($\beta = -.15$, $SE = .19$, $p = .44$); please see Table 1. The mean amount shared by the in-person and online samples was 2.60 ($SD = 1.20$) and 2.50 ($SD = 1.24$), respectively. Please see Fig. 3 for the amount shared by age and testing format. For covariates, there was a significant effect of gender ($\beta = .49$, $SE = .19$, $p = .01$) only, with boys sharing less than girls.

Our model comparison found that our full model (including age and testing format) provided a superior fit to the data, $F(2{,}139) = 9.68$, $p < .001$. Next, we examined the significance of our main predictors and covariates by removing them from our full model using 'drop1′ and testing whether their inclusion improved model fit. Age, $F(1{,}139) = 18.87$, $p < .001$, improved model fit, while testing format did not, $F(1{,}139) = .60$, $p = .44$. For covariates, gender improved model fit, $F(1{,}139) = 6.59$, $p = .01$, but task order did not ($p = .89$).

3.3. Fairness

3.3.1. Analytic approach

We predicted a significant age by trial type interaction online (RQ1), with more rejections of Advantageous Inequity trials relative to Equal trials in older children relative to younger children, and significant age effects online (RQ1), with Advantageous Inequity and Extreme Advantageous Inequity rejections increasing with age. We predicted no significant effect of testing format (RQ2). Generalized linear mixed models and generalized linear models were conducted with R statistical software using lme4 (Bates et al., 2014). The dependent measure was the participant's decision to accept or reject a given trial (binary, with a "1" used for rejections). Depending on the analyses, trial type (factor; Equal; Advantageous Inequity; Extreme Advantageous Inequity) was entered as a within-subjects factor, or analyses were run separately for each trial type. Participant ID was entered as a random effect (intercepts) for analyses involving multiple trials (i.e., Equal and Advantageous Inequity), with covariates described previously. Results for all trial types are shown in Fig. 4 and the Supplement (in bar graph form).

3.3.2. Equal and Advantageous Inequity trials

In-person results. A generalized linear mixed model was run for the Equal and Advantageous Inequity (1−4) trials, with age, trial type, and their interaction as the main predictors. Crucially, we found a significant age by trial type interaction ($\beta = .08$, $SE = .02$, $p < .001$), with older children likelier to reject Advantageous Inequity (AI) than Equal trials relative to younger children. Additionally, there was a significant effect for trial type ($\beta = -5.62$, $SE = 2.18$, $p = .01$), with a lower rate of rejections for AI trials. There were no other significant effects (all $ps >= .23$). Our model comparison found that our full model provided a superior fit to the data, $\chi^2_3 = 72.40$, $p < .001$. Using drop1, we found that including the interaction of age and trial type improved model fit (LRT, $\chi^2_1 = 14.40$, $p < .001$), while the covariates did not ($ps >= .24$).

RQ1.. The above analysis was repeated for the online data. This analysis also found a significant age by trial type interaction ($\beta = .06$, $SE = .02$, $p = .003$), with older children likelier to reject AI than Equal trials relative to younger children, and a significant effect for trial type ($B = -4.04$, $SE = 1.86$, $p = .03$), with a lower rate of rejections for AI trials. There was also a significant effect for age ($\beta = -.04$, $SE = .02$, $p = .02$), with overall rejections decreasing with age (note that this effect was driven by the Equal trials). Our full model provided a superior fit to the data, $\chi^2_3 = 39.79$, $p < .001$, and the use of drop1 revealed that including the interaction of age and trial type improved model fit, (LRT, $\chi^2_1 = 10.32$, $p = .001$), while the covariates did not (ps >=.27).

Table 1
Fixed effects in linear models predicting number given in the Trustworthiness task across testing formats.

Statistical models			
	In-person	Online	Combined
(Intercept)	0.73 (0.76)	0.68 (0.69)	0.78 (0.52)
Age (months)	0.02 (0.01)**	0.02 (0.01)**	0.02 (0.00)***
Task order: Trust second	-0.19 (0.29)	0.08 (0.25)	-0.03 (0.19)
Gender: Male	-0.46 (0.29)	-0.51 (0.25)*	-0.49 (0.19)*
Testing format: Online			-0.15 (0.19)
R^2	0.18	0.14	0.16
Adj. R^2	0.14	0.11	0.13
Number of Participants	58	86	144

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Note. Baselines were set as follows: Task order = Trust task first; Gender = Female; Testing format = In-person.

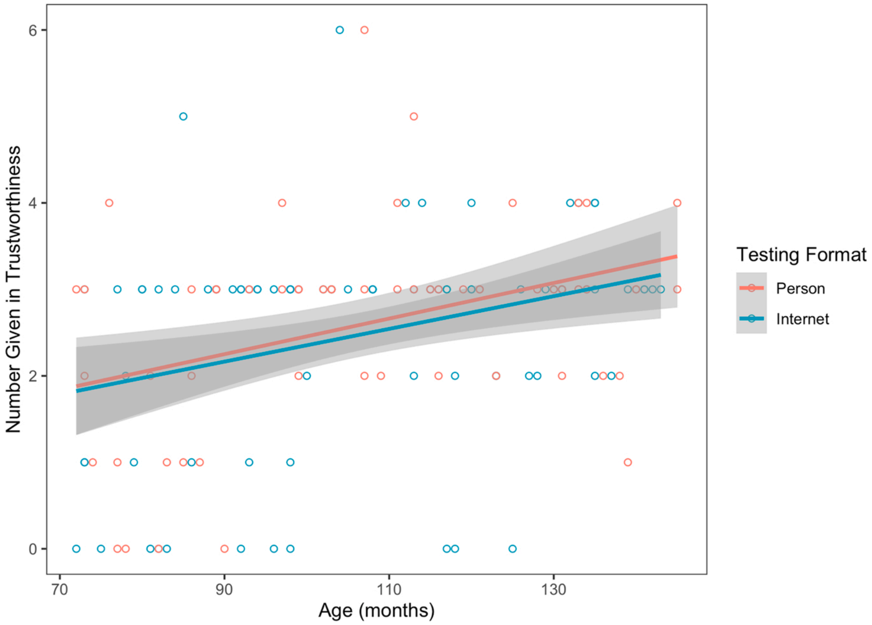

Fig. 3. Number given by testing format and age. Shaded regions show 95% confidence bands.

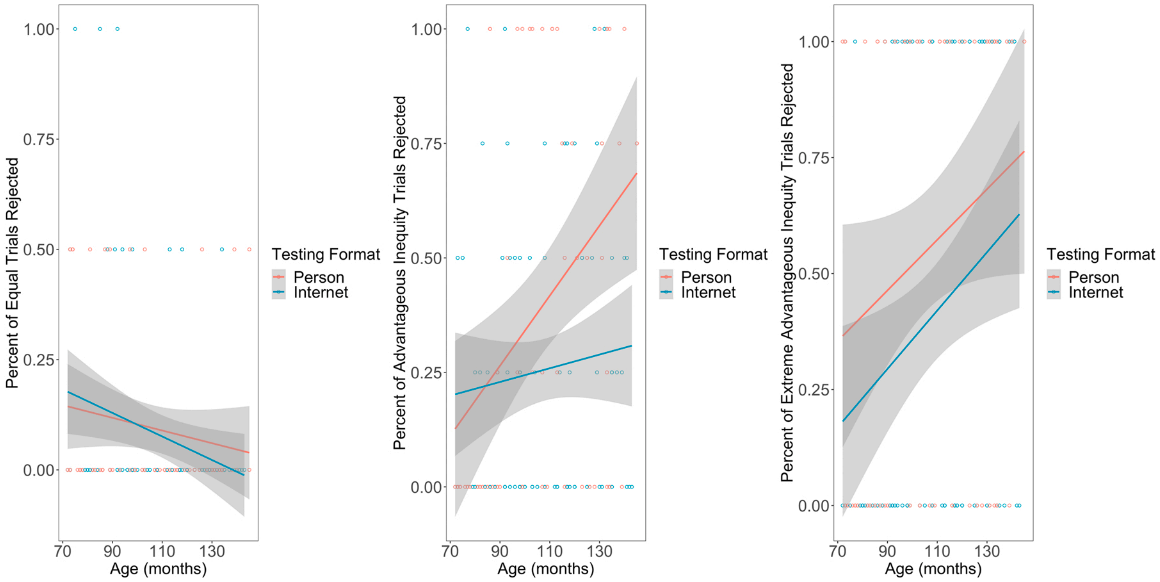

Fig. 4. Rate of rejections across trial type, testing format and age group. Shaded regions show 95% confidence bands.

3.3.3. Advantageous Inequity trials

In-person results. A generalized linear mixed model was run for the Advantageous Inequity trials exclusively, with age as the main predictor. We found a significant effect for age ($\beta = .10$, $SE = .03$, $p = .002$), with rejections increasing with age, and no other significant effects ($ps > =.44$). Our model comparison found that the full model provided a superior fit to the data, $\chi^2_1 = 12.97, p < .001$. Using drop1, we found that age improved model fit (LRT, $\chi^2_1 = 12.97$, $p < .001$), while the covariates did not ($ps > =.44$).

RQ1.. An analysis with the online data found no significant effect of age ($\beta = .01$, $SE = .01$, $p = .35$), and no other significant effects ($ps > =.58$). Our full model including age did not provide a superior fit relative to the null model (LRT, $\chi^2_1 = .89$, p = .35). Thus, our online sample did not replicate a crucial effect in our in-person sample and in prior studies.

RQ2.. A generalized linear mixed model was run on the in-person and online data, with age and testing format as the main predictors; please see Table 2. We found significant effects for age ($\beta = .04$, $SE = .01$, $p = .003$), with more rejections in older children, and testing format ($\beta = -1.11$, $SE = .55$, $p = .04$), with fewer rejections in participants tested online, and no other significant effects ($ps > = .79$). The full model (including age and testing format) provided a superior fit to the data, $\chi^2_2 = 13.58$, $p = .001$. Using drop1, we found that age (LRT; $\chi^2_1 = 9.51$, $p = .002$) and testing format (LRT; $\chi^2_1 = 4.11$, $p = .04$) improved model fit, while the covariates did not ($ps >= .79$).

Based on a visual inspection of the data (Fig. 4), we ran an exploratory generalized linear mixed model to test for an interaction between testing format and age. We found a significant effect of age ($\beta = .07$, $SE = .02$, $p < .001$), with more rejections in older children, and a significant age by testing format interaction ($\beta = .06$, $SE = .03$, $p = .02$). The significant interaction was driven by lower rates of AI rejections in older children tested online. The overall rate of rejections was somewhat lower online, although this effect of testing format was not significant ($\beta = 5.34$, $SE = 2.89$, $p = .06$).

3.3.4. *Extreme Advantageous Inequity trial*

In-person results. A generalized linear model was run for the Extreme Advantageous Inequity trial, with age as the main predictor. We found a non-significant increase in rejections with age ($\beta = .02$, $SE = .01$, $p = .07$), and no other significant effects ($ps > = .61$).

RQ1.. The analysis with the online data found a significant effect of age ($\beta = .03$, $SE = .01$, $p = .02$), with rejections significantly increasing with age, and no other significant effects (ps $> = .75$).

RQ2.. A generalized linear model was run, with age and testing format as the main predictors. There was a significant effect of age ($\beta = .03$, $SE = .01$, $p = .003$), with more rejections in older children. The rate of rejections was somewhat lower online, but this testing format effect fell short of significance ($\beta = -.67$, $SE = .36$, $p = .06$). There were no other significant effects (ps $> = .93$). Our full model provided a superior fit to the data, $\chi^2_2 = 12.46$, $p = .002$. Using drop1, we found that only age significantly improved model fit (LRT, $\chi^2_1 = 9.53$, $p = .002$). An exploratory generalized linear model to test for an interaction between testing format and age for this trial type found no significant interaction ($\beta = .004$, $SE = .02$, $p = .80$); recall that a similar analysis for the standard Advantageous Inequity trials found a significant interaction.

3.4. *Generosity*

3.4.1. *Analytic approach*

This task was always run last due to its role as an exploratory measure that was less important than Trustworthiness and Fairness, and because it was not run in our in-person sample. Thus, instead of comparing online and in-person results, we will consider whether two effects from the previous literature are found in our sample of online participants. We predicted that the number shared would increase with age, as some studies using the Dictator Game have found (Benenson et al., 2007; Blake & Rand, 2010). We also predicted a higher proportion shared in Trustworthiness than the Dictator Game (Amir et al., 2021), an effect seen with a between-subjects design, tested here within-subjects.

Linear models were conducted as described previously. The dependent measure was the number of tokens shared, which could range from 0 to 10. Gender was entered as a covariate (task order was not included because this task always came last). Additional analyses were run after converting the raw numbers shared in Trustworthiness and Dictator Game to the proportion shared (from 0 to 100), since the amount available for sharing differed across tasks. For these linear mixed models, participant ID was entered as a random effect (intercepts). Only participants with usable data from both tasks were included.

3.4.2. ***Generosity only***

A linear model was run with age as the main predictor, the sole covariate of gender, and the dependent variable of tokens shared. This analysis found a significant effect of age ($\beta = .03$, $SE = .011$, $p = .006$), with older children sharing more than younger children, and a significant effect of gender ($\beta = -.89$, $SE = .450$, $p = .049$), with boys sharing less than girls.

3.4.3. *Generosity and Trustworthiness*

A linear mixed model was run with age and task type (within-subjects) as the main predictors, the covariate of gender, and the dependent variable of proportion of tokens shared. This analysis found a significant effect of age ($\beta = .003$, $SE = .001$, $p = .002$), with older children sharing more than younger children, and a significant effect of task type, ($\beta = .09$, $SE = .019$, $p < .001$), with more sharing in Trustworthiness. We also found a significant effect of gender ($\beta = -.09$, $SE = .040$, $p = .02$), with boys sharing less than girls. Please see Table S1.

Our full model provided a superior fit to the data, $\chi^2_2 = 29.85$, $p < .001$. Next, we examined the significance of our main predictors and the covariate using drop1. The inclusion of age (LRT, $\chi^2_1 = 10.27$, $p = .001$) and task (LRT, $\chi^2_1 = 19.57$, $p < .001$) improved model fit. The covariate of gender also improved model fit, (LRT, $\chi^2_1 = 4.65$, $p = .03$). The mean proportion shared in Trustworthiness and the Generosity was 0.42 ($SD = .21$) and 0.33 ($SD = .22$), respectively.

4. Discussion

Our online sample partially replicated the results of our in-person sample. Trustworthiness was an unqualified success, closely replicating the in-person results according to both of our criteria. However, Fairness yielded mixed results, with replications on some dimensions but not others. Below we discuss the results in greater detail. To foreshadow our conclusions, we propose that the partial non-replications of Fairness are best attributed to the use of virtual tokens, rather than candy, with the online sample, and the unique

Table 2
Fixed effects in generalized linear mixed models predicting Advantageous Inequity rejections across testing formats.

Statistical models			
	In-person	Online	Combined
(Intercept)	-12.11 (3.97)**	-3.27 (1.50)*	-5.28 (1.59)* **
Age (months)	0.10 (0.03)**	0.01 (0.01)	0.04 (0.01)**
Task order: Trust second	-0.95 (1.24)	0.16 (0.54)	-0.08 (0.53)
Gender: Male	-0.36 (1.22)	0.30 (0.54)	0.14 (0.53)
Testing format: Online			-1.11 (0.55)*
AIC	211.47	359.86	575.79
BIC	228.69	379.07	601.92
Log Likelihood	-100.74	-174.93	-281.90
Number of Trials	231	344	575
Number of Participants	58	86	144
Variance: Participant ID (Intercept)	15.29	3.45	6.44

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$
Note. Baselines were set as follows: Task order = Trust task first; Gender = Female; Testing format = In-person.

efficiency concerns inherent in this task specifically.

4.1. *Trustworthiness and Generosity*

As we predicted, Trustworthiness succeeded according to both of the criteria we had established. The online data found that giving increases linearly with age, replicating our in-person sample. Additionally, the values generated online and in-person were similar, and no effect of testing format was found. We view the unqualified replication success of the online sample to be very encouraging for the use of online methods and virtual tokens for conducting resource transfer tasks in which all resources get distributed (i.e., none are wasted). One limitation of our study is that the in-person and online samples differ in terms of cohorts, with only the in-person sample tested before COVID-19 began. However, this feature is also a benefit; these results indicate that the COVID-19 pandemic did not lead to substantial differences in children's willingness to share resources with others, at least as measured by this task.

In terms of replicating the effects of Amir et al., 2021,[7] we found that children shared more in a task involving trustworthiness, in which the recipient of the participant's sharing had previously benefited the participant, than one involving straightforward generosity, in which the recipient was neutral. Our within-subjects measures replicated the prior study's finding of more sharing in Trustworthiness than Generosity, which were tested in between-subjects conditions. However, we note the important limitation that Generosity was always run last in the current study; we cannot determine whether the reduced giving in Generosity was because it lacked the special features of Trustworthiness or simply because it came after prior resource allocation tasks. While gender differences were not a focus of our study, boys shared more than girls in our online sample in both Trustworthiness and Generosity (this effect was short of significance for Trustworthiness in our in-person sample); such gender differences were found in Amir et al., 2021, as well as some studies using standard Dictator Games (e.g., Benenson et al., 2007; Gummerum, Hanoch, Keller, Parsons, & Hummel, 2010).

We also found age-related increases in Generosity (i.e., sharing in the Dictator Game) in our online sample, as we predicted based on prior studies (e.g., Blake & Rand, 2010). We did not run Generosity in-person, and comparing the proportion shared in our online sample to various in-person studies is complicated by cross-study differences in scripts, specific procedures, resource types, age distributions, dates of testing, and populations. However, our finding of 33% of tokens shared is broadly consistent with Dictator Game precedent in the literature of roughly 24–41% resources shared within the age range we studied; the values we obtained are similar to, or slightly lower than, those from other studies. For comparison, Benenson et al., (2007) found ranges of sharing from approximately[8] 25% (lower-SES 6-year-olds) to 39% (higher-SES 9-year-olds) of stickers in a sample of 6- and 9-year-olds; Allgaier et al. (2020) found a mean sharing of 40.9% of candies across a sample of 7—11-year-olds; and the control condition of McAuliffe et al., (2017) found candy sharing ranging from 24% (6- and 7-year-olds) to 37.5% (8- and 9-year-olds).

4.2. *Fairness*

For Fairness, our online replication was only partially successful. In terms of successfully replicated effects for RQ1, the most important of our two research questions, the age by trial type interaction for Advantageous Inequity and Equal trials was seen in both samples, with older children rejecting more Advantageous Inequity trials (1 for the other child, 4 for the participant) than Equal trials relative to younger children both online and in-person. (This makes sense, as older children should have more fairness concerns, and accepting Advantageous Inequity trials results in unequal outcomes while accepting Equal trials maintains equality.) For the Extreme Advantageous Inequity trial (0 for the other child, 4 for the participant), the age effect fell short of significance in-person but was significant online, with more rejections in older children (note that the online sample was larger). We interpret this finding as a near-

[7] We note that Amir et al., 2021, did not find a significant effect for age but used a different age range of 5—8-year-olds.

[8] These are estimates based on the article's figures.

replication according to our first criterion, as the effects were in the same direction and of similar magnitudes in both samples. However, the overall rate of rejections for this trial type was non-significantly higher in-person than online; according to RQ2, this “borderline” testing format difference indicates a somewhat unsuccessful replication. As we will discuss later, we believe overall lower rates of rejections online are best attributed to children’s greater reluctance to “waste” tokens than candy.

Our online and in-person results diverged sharply for the Advantageous Inequity trials. For this trial type, our online replication was unsuccessful according to both RQ1 and RQ2. Unlike in our own in-person sample as well as prior published work (Blake & McAuliffe, 2011; Blake et al., 2015), and unlike the Extreme Advantageous Inequity trial, the rate of Advantageous Inequity rejections did not significantly increase with age in our online sample. To reiterate, this means that a key finding in the fairness literature, that of age-related increases in rejections of “1–4″ Advantageous Inequity trials, was not seen in our online sample. Additionally, the overall rate of rejections was significantly lower online than in-person.

Next, we will consider what we deem the most important difference in results across samples: older children were far less likely to reject Advantageous Inequity online than in-person. Why was this trial type uniquely influenced by testing format? First, it is important to acknowledge that many elements varied between the in-person and online studies. As a result of this limitation, it is difficult to determine which specific elements mattered, and which did not, in terms of contributing to differences in results across studies. In-person, our study was conducted face-to-face with a live experimenter, using tangible materials and candy for prizes. Online, our experimenter appeared on a computer screen, with virtual materials and tokens for prizes. Our in-person sample was collected prior to the start of the COVID-19 pandemic; our online sample was collected in the midst of it. At first glance, these differences seem substantial, although their implications for children’s performance are less clear, particularly for this trial type in question relative to the EAI trial, which successfully generated age-related differences in the online sample, and relative to the Trustworthiness task, which generated extremely similar results online and in-person.

The online format did not pose problems for children’s comprehension in the age range we used, with similar rates of comprehension check success across samples. Reputation is a general concern for children (Engelmann et al., 2013), and reputational concerns influence children’s actions in the Inequity Game (McAuliffe, Blake, & Warneken, 2020); whether the participants’ partners in this task know of the participants’ decisions is a particularly powerful influence on their behavior. However, the “other children” in this task were entirely unknown to participants in both our online and in-person studies, represented only by avatars or paper bags. A reputational account for the testing format difference must therefore posit that older children would be uniquely concerned about the opinions of an in-person experimenter relative to an online experimenter, and uniquely so for the Advantageous Inequity trials relative to the Extreme Advantageous Inequity trial. When tested in-person, participants made their choices by grabbing a handle on the Fairness apparatus; online, they said their choices verbally to the experimenter, who clicked on the handle for them. While worth noting as a procedural difference, it is unclear why it would result in a lower rate of rejections (i.e., red-handle choices) for the online sample.

Somewhat disappointingly, our online sample was mostly demographically similar to our in-person sample, whose parents were wealthier and more highly educated than the general population. (We believe our online sample would have been more diverse had only participants recruited via online ads participated.) This means that demographic differences across samples would not be a viable explanation for differences in results, although we note that we have no clear predictions about how any such differences would affect the overall study or Advantageous Inequity rejections specifically.

Our best explanation for the cross-study differences in rates of rejections is that older online participants were more reluctant to waste tokens than older in-person participants were to waste candy. Additionally, for standard Advantageous Inequity trials specifically, we posit that older children were more reluctant to reduce their partner’s payouts for tokens than candy. Such differences in reluctance could emerge if the overall value ascribed to tokens was higher than the value ascribed to candy. Recall that rejecting an Advantageous Inequity trial means denying 1 token for the other child as well as 4 tokens for the participant; both individuals gain fewer tokens as a result of a rejection, even though rejections do reduce the disparity between them. Participants may have been especially reluctant to let tokens go to waste or to deny the other child tokens. This finding parallels a recent study with adults, who were less likely to reject tokens (to be exchanged for money) than to reject candy on 1–4 Advantageous Inequity trials (McAuliffe, Benjamin, & Warneken, 2022; note that a 0–4 trial type was not tested). These results are suggestive of continuity between how resource type affects 1–4 Advantageous Inequity rejections for older children and adults.

Our results do not support the supposition that children online simply wanted more tokens for themselves, independent of other concerns. Recall that sharing in Trustworthiness was similar online and in-person, which speaks against a general reluctance to part with tokens. Additionally, for Extreme Advantageous Inequity trials, rejections increased with age in both samples[9] (although rejection rates were non-significantly higher in-person than online, and the age effect fell short of significance in-person). If older children in the online sample simply did not want to reduce their own payoffs, we should not have seen a significant increase in Extreme Advantageous Inequity rejections with age. Relative to acceptances, Extreme Advantageous Inequity trial rejections only reduce the participants’ payoff; they do not affect the absolute number the other child receives. Thus, a participant who wishes to both reduce the disparity between themselves and the other child while increasing the payoff of the other child may feel conflict about rejecting Advantageous Inequity trials but should feel comfortable rejecting the Extreme Advantageous Inequity trial.At first glance, it may seem surprising that children were more reluctant to reject tokens than a tangible resource, as Ebersbach et al. (2022) found the opposite

[9] Please note that the age-related increase of EAI rejections was significant online and short of significant in-person (RQ1). We note that Beta values were similar in both samples but the in-person sample was smaller and therefore less able to detect an age-based effect for a single-trial measure. There was no significant age by testing format interaction for the EAI trial, unlike for the AI trials.

result. However, in that study, the resources were stickers and tokens exchanged for stickers. In the current study, the tokens were to be exchanged for unspecified (perhaps "mysterious") prizes to which children, who are known for their optimism, likely ascribed higher value. Prior research has indicated that American children are less concerned with waste than Chinese children (Zhang & Benozio, 2021) but are still more reluctant to waste highly valuable resources than less-valuable resources (Choshen-Hillel, Lin, & Shaw, 2020). While American children generally endorse wasting a less-valued resource over distributing it unequally, this preference attenuates and nearly reverses as resource value increases (Choshen-Hillel, Lin, & Shaw, 2020).

4.3. Conclusions and lessons for future work

What do our results mean for future research on incentivized tasks with children in middle childhood? Even with differences in testing modalities and resource types, our Trustworthiness results were similar online and in-person, yielding a close replication. This is a very encouraging result for the prospect of moving incentivized tasks online. Both samples found age-related increases in sharing; the rates of sharing across samples were also very similar. Our Generosity results, while lacking a direct in-person component, are broadly consistent with those in the published literature, with similar base rates across studies. The results indicate that online methods, including with tokens as resources, can be used effectively for tasks involving straightforward resource transfers between a child-age participant and another individual. Potential concerns that children might not give to "anonymous others" online, or might be skeptical of resource exchanges in an online format, are not borne out by the current investigation.

Moving beyond the question of testing format, our results also show that tokens can be a viable alternative to immediately consumable items such as stickers or candy. Abstract rewards work as costly resources that affect real behavior. Beyond their obvious utility for online testing, for which physical rewards are impractical, tokens with unspecified payoffs can be used in-person as well. They enable the single "currency" of tokens to be held constant across the lifespan of a study while allowing the items the tokens get exchanged for at the study's conclusion to vary across testing sites. This approach can yield maximal flexibility in site recruitment (e.g., if a museum testing site does not allow candy, the tokens can be exchanged for a non-candy reward). Additionally, it does not require participants to express liking for a particular prize type during the study itself, since the ultimate reward type is malleable.

Our online methods produced very different results for Fairness. Most crucially, our online sample did not replicate the finding of age-related increases in rejections of 1–4 Advantageous Inequity trials that was obtained in-person, as well as by past studies (e.g., Blake & McAuliffe, 2011). Due to the number of elements that varied across our online and in-person samples, we cannot be certain why this non-replication occurred. However, our best explanation for this discrepancy is that children were reluctant to waste tokens through rejections in general (resulting in lower rates of rejections online for both AI and EAI trials; RQ2), and that older children in particular were reluctant to deny tokens for the other child through rejections (resulting in a lack of age-related increases in AI, but not EAI, rejections in the online sample; RQ1). Such reluctance was less pronounced for the resource type of candy, used in our in-person sample and in most Inequity Game studies. What if the tokens were to be exchanged for something other than "prizes?" It is possible that tokens to be exchanged for candy would function more like candy than the tokens "for prizes" we used in the current study. Using "tokens for candy" could be suitable online, although doing so would eliminate some of the logistical advantages of the fully unspecified prize types we used.

Our results indicate the acceptability of online methods for tasks like Trustworthiness but the need for caution for the Inequity Game and for other tasks in which there is the potential for resources to be "wasted." More broadly, beyond the issue of online testing specifically, our findings suggest that when there is the potential for resources to be wasted, resource type may matter more. In such cases, acceptances of inequality could be motivated by an indifference to unfairness, a desire for one's own resources, or a reluctance to waste resources; the latter motivation becomes especially strong as resource value increases. Different attitudes towards resources and their value could produce very different behavior in the Inequity Game. Given the potential for cross-cultural variation in attitudes towards waste and perhaps different norms surrounding efficiency for groups within the United States, researchers should carefully consider their choice of rewards in such tasks.

Acknowledgments

We are grateful to the teachers, administrators, and staff at Living Laboratory ® at Discovery Museum of Acton, Massachusetts, and all participating children and families. We are grateful to the John Templeton Foundation (#61138) and to the CIFAR Azrieli Global Scholars Program for funding. We are thankful to Tom Coben (tcoben@gmail.com) for the illustrations and animations used in our online study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cogdev.2023.101313.

References

Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. C. (2022). Says who? Children consider informants' sources when deciding whom to believe. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001198

Allgaier, K., Ścigała, K. A., Trautwein, U., Hilbig, B. E., & Zettler, I. (2020). Honesty-humility and dictator and ultimatum game-giving in children. *Journal of Research in Personality, 85*, Article 103907.

Amir, D., Ahl, R. E., Parsons, W. S., & McAuliffe, K. (2021). Children are more forgiving of accidental harms across development. *Journal of Experimental Child Psychology, 205*, 105081.

Amir, D., Ahl, R.E., Bolotin, H., Bogese, M., Gonzalez, G., Callaghan, T., Sugiyama, L.S., & McAuliffe, K. (2023). The development of fairness, honesty, trustworthiness, and forgiveness across five diverse societies. Manuscript in preparation.

Amir, D., Parsons, W. S., Ahl, R. E., & McAuliffe, K. (2021). Trustworthiness is distinct from generosity in children. *Developmental Psychology, 57*(8), 1318–1324.

Amir, O., Rand, D. G., & Gal, Y. A. K. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one, 7*(2), Article e31461.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396–403.

Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children's altruistic behavior in the dictator game. *Evolution and Human Behavior, 28*(3), 168–175.

Benozio, A., & Diesendruck, G. (2015). From effort to value: Preschool children's alternative to effort justification. *Psychological Science, 26*(9), 1423–1429.

Bereby-Meyer, Y., & Fiks, S. (2013). Changes in negative reciprocity as a function of age. *Journal of Behavioral Decision Making, 26*(4), 397–403.

Berti, A. E., & Bombi, A. S. (1981). The development of the concept of money and its value: A longitudinal study. *Child Development*, 1179–1182.

Blake, P. R. (2018). Giving what one should: Explanations for the knowledge-behavior gap for altruistic giving. *Current Opinion in Psychology, 20*, 1–5.

Blake, P. R., Corbit, J., Callaghan, T. C., & Warneken, F. (2016). Give as I give: Adult influence on children's giving in two cultures. In *Journal of Experimental Child Psychology, 152* pp. 149–160).

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.

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., & Warneken, F. (2015b). The ontogeny of fairness in seven societies. *Nature, 528*(7581), 258–261.

Blake, P. R., McAuliffe, K., & Warneken, F. (2014). The developmental origins of fairness: The knowledge–behavior gap. *Trends in Cognitive Sciences, 18*(11), 559–561.

Blake, P. R., & Rand, D. G. (2010). Currency value moderates equity preference among young children. *Evolution and Human Behavior, 31*(3), 210–218.

Blake, P. R., Rand, D. G., Tingley, D., & Warneken, F. (2015a). The shadow of the future promotes cooperation in a repeated prisoner's dilemma for children. *Scientific Reports, 5*(1), 1–9.

Afshordi, N., & Koenig, M. (2021). *Trusting information from friends: Adults expect it but preschoolers do not.*PsyArXiv. 10.31234/osf.io/rsxb2.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?. In A. E. Kazdin (Ed.), Methodological Issues and Strategies in Clinical Research (pp. 133–139). American Psychological Association. https://doi.org/10.1037/14805–009.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156–2160.

Chernyak, N., & Sobel, D. M. (2016). Equal but not always fair: Value-laden sharing in preschool-aged children. *Social Development, 25*(2), 340–351.

Choshen-Hillel, S., Lin, Z., & Shaw, A. (2020). Children weigh equity and efficiency in making allocation decisions: Evidence from the US, Israel, and China. *Journal of Economic Behavior & Organization, 179*, 702–714.

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology*, 4968.

Chuey, A., Lockhart, K., Sheskin, M., & Keil, F. (2020). Children and adults selectively generalize mechanistic knowledge. *Cognition, 199*, Article 104231.

Cowell, J. M., Samek, A., List, J., & Decety, J. (2015). The curious relation between theory of mind and sharing in preschool age children. *PloS one, 10*(2), Article e0117947.

Ebersbach, M., Krupa, J., & Vogelsang, M. (2022). Symbolic distancing in sharing situations restrains children's economic behavior and potentially also their inequity aversion. *Acta Psychologica, 226*, Article 103579.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics, 14*(4), 583–610.

Engelmann, J. M., Over, H., Herrmann, E., & Tomasello, M. (2013). Young children care more about their reputation with ingroup members and potential reciprocators. *Developmental Science, 16*(6), 952–958.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature, 454*(7208), 1079–1083.

Gonzalez, G., Ahl, R. E., Cordes, S., & McAuliffe, K. (2022). Children strategically conceal selfishness. *Child Development, 93*(1), e71–e86.

Gonzalez, G., Blake, P. R., Dunham, Y., & McAuliffe, K. (2020). Ingroup bias does not influence inequity aversion in children. *Developmental Psychology, 56*(6), 1080–1091.

Good, K., & Shaw, A. (2022). Being versus appearing smart: Children's developing intuitions about how reputational motives guide behavior. *Child Development, 93*(2), 418–436.

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*(2–3), 94–95.

Gummerum, M., Hanoch, Y., Keller, M., Parsons, K., & Hummel, A. (2010). Preschoolers' allocations in the dictator game: The role of moral emotions. *Journal of Economic Psychology, 31*(1), 25–34.

Guth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization, 3*(4), 367–388.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399–425.

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., & Silk, J. B. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour, 4*(1), 36–44.

House, B. R., & Tomasello, M. (2018). Modeling social norms increasingly influences costly sharing in middle childhood. *Journal of Experimental Child Psychology, 171*, 84–98.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(8), 589–604.

Kominsky, J. F., Begus, K., Bass, I., Colantonio, J., Leonard, J. A., Mackey, A. P., & Bonawitz, E. (2021). Organizing the methodological toolbox: Lessons learned from implementing developmental methods online. *Frontiers in Psychology, 12*, Article 702710.

Kominsky, J. F., Shafto, P., & Bonawitz, E. (2021). "There's something inside": Children's intuitions about animate agents. *PloS One, 16*(5), Article e0251081.

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*, Article 731404.

Leshin, R. A., Leslie, S. J., & Rhodes, M. (2021). Does it matter how we speak about social kinds? A large, preregistered, online experimental study of how language shapes the development of essentialist beliefs. *Child Development, 92*(4), e531–e547.

Li, Y., Li, H., Decety, J., & Lee, K. (2013). Experiencing a natural disaster alters children's altruistic giving. *Psychological Science, 24*(9), 1686–1695.

Liu, S., Gonzalez, G., & Warneken, F. (2019). Worth the wait: Children trade off delay and reward in self-and other-benefiting decisions. *Developmental Science, 22*(1), Article e12702.

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends in Cognitive Sciences, 24*(8), 583–584.

Malti, T., Gummerum, M., Ongley, S., Chaparro, M., Nola, M., & Bae, N. Y. (2016). "Who is worthy of my generosity?" Recipient characteristics and the development of children's sharing. *International Journal of Behavioral Development, 40*(1), 31–40.

Mandalaywala, T. M., Benitez, J., Sagar, K., & Rhodes, M. (2021). Why do children show racial biases in their resource allocation decisions? *Journal of Experimental Child Psychology, 211*, Article 105224.

McAuliffe, K., Benjamin, N., & Warneken, F. (2022). Reward type influences adults' rejections of inequality in a task designed for children. *PloS one, 17*(8), Article e0272710.

McAuliffe, K., Blake, P. R., & Warneken, F. (2020). Costly fairness in children is influenced by who is watching. *Developmental Psychology, 56*(4), 773–782.

McAuliffe, K., Raihani, N. J., & Dunham, Y. (2017). Children are sensitive to norms of giving. *Cognition, 167*, 151–159.

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

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.

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

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences, 17*(8), 413–425.

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., … Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with children. *Journal of Cognition and Development, 21*(4), 477–493.

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition, 224*, Article 105073.

Rowley, S. J., & Camacho, T. C. (2015). Increasing diversity in cognitive developmental research: Issues and solutions. *Journal of Cognition and Development, 16*(5), 683–692.

Samek, A., Cowell, J. M., Cappelen, A. W., Cheng, Y., Contreras-Ibáñez, C., Gomez-Sicard, N., & Decety, J. (2020). The development of social comparisons and sharing behavior across 12 countries. *Journal of Experimental Child Psychology, 192*, Article 104778.

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*, 4573.

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.

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind, 1*(1), 4–14.

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General, 141*(2), 382.

Sheskin, M., & Keil, F. (2018). TheChildLab. com a video chat platform for developmental research. *PsyArXiv*. https://doi.org/10.31234/osf.io/rn7w5

Sheskin, M., Nadal, A., Croom, A., Mayer, T., Nissel, J., & Bloom, P. (2016). Some equalities are more equal than others: Quality equality emerges later than numerical equality. *Child Development, 87*(5), 1520–1528.

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., & Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences, 24*(9), 675–678.

Shu, Y., Hu, Q., Xu, F., & Bian, L. (2022). Gender stereotypes are racialized: A cross-cultural investigation of gender stereotypes about intellectual talents. *Developmental Psychology. Advance online publication.* https://doi.org/10.1037/dev0001356

Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. *Plos One, 8*(3), Article e59510.

Smith-Flores, A. S., Perez, J., Zhang, M. H., & Feigenson, L. (2022). Online measures of looking and learning in infancy. *Infancy, 27*(1), 4–24.

Tasimi, A., & Young, L. (2016). Memories of good deeds past: The reinforcing power of prosocial behavior in children. *Journal of Experimental Child Psychology, 147*, 159–166.

Thomas, A. J., Woo, B., Nettle, D., Spelke, E., & Saxe, R. (2022). Early concepts of intimacy: Young humans use saliva sharing to infer close relationships. *Science, 375* (6578), 311–315.

Tsoi, L., & McAuliffe, K. (2020). Individual differences in theory of mind predict inequity aversion in children. *Personality and Social Psychology Bulletin, 46*(4), 559–571.

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*, 3011.

Vales, C., Wu, C., Torrance, J., Shannon, H., States, S. L., & Fisher, A. V. (2021). Research at a distance: Replicating semantic differentiation effects using remote data collection with children participants. *Frontiers in Psychology, 12*, Article 697550.

Warneken, F., Lohse, K., Melis, A. P., & Tomasello, M. (2011). Young children share the spoils after collaboration. *Psychological Science, 22*(2), 267–273.

Zhang, Z., & Benozio, A. (2021). Waste aversion reduces inequity aversion among Chinese children. *Child Development, 92*(6), 2465–2477.