Cognitive Development 65 (2023) 101271

Contents lists available at ScienceDirect

Cognitive Development

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

Who is running our experiments? The influence of experimenter identity in the marshmallow task ☆

Thomas St. Pierre [a,*], Katherine S. White [b], Elizabeth K. Johnson [a,c]

[a] University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada
[b] University of Waterloo, Waterloo, ON N2L 3G1, Canada
[c] University of Toronto, Toronto, ON M5S 3GS, Canada

A R T I C L E I N F O

Keywords:
Experimenter effects
Delay of gratification
Marshmallow task
Social cognition
Experimental design
Implicit bias

A B S T R A C T

While developmental researchers take great care to report on the characteristics of their participants, they rarely report on the characteristics of their experimenter(s). This is surprising, given the real potential for experimenter identity (e.g., gender, race, age, etc.), especially as it relates to children's identities, to influence children's behavior in experiments. In the current study, we investigate how experimenter identity (as signaled by language and race cues) influences 3- to 5-year-old children's (N = 159) behavior in the famous marshmallow task. Results show that experimenter identity indeed influenced children's wait times in the marshmallow task; specifically, we found that racial mismatch between experimenter and child led to longer wait times, and in an exploratory analysis, we found that this effect was exaggerated by an additional mismatch in accent. We thus reveal a previously overlooked factor that may influence children's behavior in a delayed gratification task—experimenter identity—and discuss the important implications of these findings for developmental research more broadly.

1. Introduction

In recent years, it has become increasingly recognized how important it is for researchers to include detailed demographic information about participants, such as their gender, race, SES, etc., to "give readers a more complete understanding of the sample and the generalizability of results" (American Psychological Association, 2020, p. 83). Many journals, in fact, now require researchers (rightly so!) to include detailed descriptions of their samples in manuscripts. For most experiments, however, there is an additional 'participant' that researchers rarely consider and report demographic information on, namely, the experimenter(s) (St. Pierre et al., 2022). This lack of information about experimenters is particularly striking given the wealth of studies demonstrating that children are highly sensitive to the social properties of others (Rhodes & Baron, 2019), and may behave differently depending on whom they are interacting with. Using the well-known marshmallow test, in which the influence of experimenter identity might seem negligible (at least at first blush), we investigated whether the identity of an experimenter (especially *as it relates* to the identity of the children they test) can indeed influence children's behavior. In doing so, we illustrate how important it is for researchers to consider experimenter

☆ This research was supported by grants awarded to the second and third authors from the Social Sciences and Humanities Research Council (SSHRC), and the Natural Sciences and Engineering Research Council (NSERC), as well as additional financial support from MITACS Global Links. We have no conflicts of interest to disclose.

* Correspondence to: Department of Psychology, University of Toronto Mississauga, 3359 Mississauga Rd., Mississauga, ON L5L 1C6, Canada.
E-mail addresses: thomas.stpierre@utoronto.ca (T.St. Pierre), white@uwaterloo.ca (K.S. White), elizabeth.johnson@utoronto.ca (E.K. Johnson).

https://doi.org/10.1016/j.cogdev.2022.101271
Received 9 May 2022; Received in revised form 4 October 2022; Accepted 7 November 2022
0885-2014/© 2022 Elsevier Inc. All rights reserved.

identity in combination with participant identity.

The idea that the identity of the experimenter might influence participant behavior is not a new idea in psychology (see Innes & Fraser, 1971; McGuigan, 1963; Rosenthal, 1966; Silverman, 1974 for discussions about this issue). Work with adults, for example, has shown how characteristics of experimenters—including their race (e.g., Marx & Goff, 2005), gender (see Chapman et al., 2018 for a review), accent (e.g., Hay et al., 2009), age (e.g., Sindi et al., 2013), and sexual orientation (e.g., Berry, 2015)—can influence participants' behavior in a variety of domains, with some evidence of similar effects with children (e.g., Clark et al., 1980; Corenblum & Annis, 1987; Greenberg & Gordon, 1983; Katz et al., 1975; Kwong See et al., 2012; Shatz & Gelman, 1973). However, despite the widespread recognition that experimenter identity can potentially influence participants' behavior, and despite over a half century of efforts to bring attention to experimenters as neglected variables in research (McGuigan, 1963; Silverman, 1974), only a fraction of developmental studies actually report the attributes of the experimenters running their studies, which is all the more concerning given that interactions with experimenters (of some sort) are especially common in developmental work (St. Pierre et al., 2022).

How exactly might experimenter identity influence children's behavior? One way might be through the different social inferences that children form about others based on category membership (see St. Pierre et al., 2022 for other possibilities). Like adults, children make inferences about the thoughts, beliefs, and intentions of others based on the social categories they belong to, and critically, these inferences can shape how they subsequently behave around those individuals. For example, children may alter how they produce and process language depending on who their interlocutor is, as when they use simpler sentence structures when interacting with younger siblings compared to adults (Shatz & Gelman, 1973). And even in cases where children do not necessarily consider the mental states of others, they may nevertheless respond differently to different individuals based on more superficial characteristics like how they look (Charlesworth et al., 2019), and what their perceived status is (McGuigan, 2013). For example, based on facial features alone, children may make inferences about others' personalities (e.g., how trustworthy and/or intelligent they appear), and later use those inferences to decide how to respond to those individuals (e.g., choosing whom they will opt to give a treat to; Charlesworth et al., 2019). Thus, we might expect experimenters from different social categories to elicit different responses from children based on the inferences they form about those individuals.

Importantly, the specific inferences that children form about others—and their responses to those individuals—will depend on children's own background/experiences, and how they situate themselves in relation to others (Dunham, 2018; Liberman et al., 2017). For example, children are more likely to positively evaluate in-group members compared to out-group members (Dunham, 2018), and conversely, are more likely to attribute negative traits and social behavior to out-group members than in-group members (Aboud, 2003; Dunham et al., 2011; Kuhn et al., 1978). As a result of these varying inferences, children may, among other things, be more anxious and/or inhibited around out-group members compared to in-group members (Yu et al., 2021; see MacInnis & Page-Gould, 2015), exhibit more dehumanizing behavior towards them (McLoughlin & Over, 2017), and be less credulous and trusting of them (Ma & Woolley, 2013; McDonald & Ma, 2016). Therefore, researchers need to consider not just the identity of the experimenter (as an objective property of the environment), but the identity of the experimenter *in relation* to that of the children they test (as something that is subjectively perceived/interpreted by children; Bronfenbrenner, 1977).

In the current study, we ask whether the identity of an experimenter (as it relates to children's identities) can influence children's behavior, using the well-known marshmallow task as a test case. In this task, children are tested on their abilities to forgo a small, immediately available reward (e.g., one marshmallow) in order to later obtain a larger reward (e.g., two marshmallows; see Mischel et al., 2011; Tobin & Graziano, 2010 for reviews). Traditionally, children's ability to wait for a second treat was thought to arise mainly from their executive functioning skills (i.e., their ability to resist temptation), but in recent years, researchers have discovered the additional role that rational decision-making processes play in children's waiting behavior, showing that children strategically weigh the costs and benefits of waiting for a second treat, and act accordingly. Interestingly (for the current study), many of these studies have focused on how children's perceptions of others (including the experimenter) influence their decisions to wait or not (Kidd et al., 2013; Leonard et al., 2014; Ma et al., 2020; Moffett et al., 2020; Michaelson & Munakata, 2016; Pesch & Koenig, 2018; see Michaelson et al., 2013 for work with adults). For example, in Kidd et al. (2013), before participating in the marshmallow task, children gained firsthand knowledge of the experimenter's general reliability, having seen the experimenter either renege on or successfully fulfill two promises (to return with a fancy art set and some cool stickers, respectively). During the experiment, children who interacted with a previously untrustworthy or unreliable experimenter were less likely to wait for a second treat compared to children who interacted with a trustworthy/reliable counterpart, presumably because they had inferred (for good reason) that the former would be unlikely to follow through and bring back an additional treat. While this study (and similar ones) focused on children's evaluations of the experimenter's *previous behavior*, in our study, we instead investigated whether experimenter identity alone (and children's perception of it)—in the absence of direct evidence about an experimenter's previous behavior—influences children's behavior in the marshmallow task (see Strickland, 1972).

To test this idea, we used linguistic and race cues to signal different aspects of experimenter identity, recruiting four female experimenters that varied with respect to race and accent. In addition, we manipulated whether the experimenters spoke with grammatical errors or not. With this set of experimenters, we tested a locally representative group of 3- to 5-year-old Canadian English-learning children on whether their wait times in the marshmallow task depended on the identity of the experimenter. The decision to focus on these features was based on a large body of previous literature showing that preschool-aged children are highly sensitive to linguistic and race differences, using them to categorize and evaluate others (Rhodes & Baron, 2019). For example, with respect to accent cues, children are more likely to endorse information from (Corriveau et al., 2013; Kinzler et al., 2011; Wagner et al., 2014), trust (Kinzler et al., 2007; Shutts et al., 2009), and prefer to be friends with (Kinzler et al., 2007; Paquette-Smith et al., 2019; St. Pierre & Johnson, 2020) fellow native-accented speakers over foreign-accented speakers; similar results have been found with respect to race, with children—in particular, racial majority children—favoring same-race individuals over different-race individuals (Chen et al., 2011; Glaeser et al., 2000; Kinzler et al., 2009; Simpson et al., 2007; see Aboud, 1988 and Raabe & Beelmann, 2011 for reviews/metanalysis). While much less work has studied how children evaluate ungrammatical (vs. grammatical) speech, the work that does exist suggests that children may prefer grammatical speakers over ungrammatical speakers (Corriveau et al., 2011; Rett & White, 2022; Sobel & Macris, 2013). Thus, we hypothesized that children's perceptions of the experimenter would vary as a function of experimenter identity, and would lead to differences in wait times in a delayed gratification task.

Specifically, we expected that experimenter identity might influence children's decisions to wait for the second treat or not in one of two ways. First, children might vary in how much they trust specific experimenters, which might then influence their willingness to delay gratification. Children, for example, tend to be less trusting of out-group members compared to in-group members (e.g., Corriveau et al., 2013; Ma & Woolley, 2013; Plötner et al., 2015; Rotenberg & Cerda, 1994; Shutts et al., 2010; Xiao et al., 2018). Therefore, we might expect children to wait less for a larger reward when interacting with ungrammatical, foreign-accented, and/or other-race experimenters compared to their grammatical, native-accented, same-race counterparts, if they assume that the former are less likely to follow through on their promises (see Strickland, 1972). Thus, in the same way that children decide to wait less for previously unreliable experimenters (e.g., Kidd et al., 2013), so too might they decide to wait less for experimenters they perceive as less trustworthy (based on their identity). Alternatively, children might feel more anxious around some experimenters compared to others, which could influence their behavior in the marshmallow task. In general, children tend to feel more anxious in unfamiliar laboratory settings (compared to their home; Ross et al., 1975), which may be amplified or reduced depending on whom they are interacting with (Feinman, 1980; see MacInnis & Page-Gould, 2015 for a discussion on intergroup anxiety). As a result, we might instead find that children interacting with experimenters who are more dissimilar from them (e.g., ungrammatical, foreign-accented, other-race experimenters) actually decide to wait *more* for the second treat, if, for example, they feel less at ease after their interaction, and as a result, less at liberty to eat the immediately available treat (should they be inclined to do so). Either outcome (that children wait less or more for more dissimilar experimenters) would show that the identity of the experimenter alone influences children's behavior in the marshmallow task, which would have important implications for how experimenters in many fields of developmental psychology might influence children's behavior.

2. Methods

2.1. Participants

As mentioned previously, children interacted with one of four experimenters during the experiment (two native speakers and two non-native speakers, one who spoke English with a Ukrainian accent, and one who spoke with a Mandarin accent). Since our two non-native experimenters were interns from abroad who were able to test in our lab for only two months, we aimed to recruit as many children as we reasonably could for each experimenter in the time allotted to us (40 per experimenter: 20 in the grammatical condition, and 20 in the ungrammatical condition, for a total of 160 participants). Altogether we tested 216 children from the Greater Toronto Area who were between 3 and 5 years of age and exposed to English at least 70% percent of the time. After excluding 57 children (see details below), our final sample contained 159 children ($M_{\text{age}} = 4.43$ years, range = 3;10 – 5;2; $M_{\text{English exposure}} = 90.71\%$, $SD = 9.98$). No children were regularly exposed to Ukrainian, and only one child was exposed to Mandarin and interacted with the Mandarin-accented experimenter (the exclusion of this participant does not influence our results). While all children were L1 speakers of Canadian English, they were racially diverse and representative of the local community: Of the 150 caregivers who provided at least some race information about themselves and their child, 61% had at least one caregiver who identified as White, 15% South Asian, 13% Black, 9% Chinese, 9% Latin American, 7% Arab/West Indian, 6% South East Asian, 4% Filipino, 2% Japanese, 2% Aboriginal, and 1% Greek, Korean, Punjabi, Tamil, Portuguese, Uyghur, Trinidadian, and Maylay.

The majority of children came from mid- to high-SES families. Of the 129 families who provided information about parental education, 72% reported that both caregivers had received a college/university degree, 23% reported one caregiver completing a college/university degree, and the remaining 5% reported that no caregivers had obtained a college/university degree. Ninety-four families also provided information about family income: 36% reported family incomes over $140,000 (CAD) a year, 38% reported a family income between $90,000 and $140,000, and 25% reported family incomes of less than $90,000.

Participants were recruited from a lab database and were randomly placed into one of four language exposure conditions, which were defined by the identity of the experimenter (their accent and grammaticality; see Procedure for details): (1) Native grammatical ($n = 40$), (2) Native ungrammatical ($n = 39$), (3) Non-native grammatical ($n = 38$), (4) Non-native ungrammatical ($n = 42$). Within each of the four conditions, children interacted either with a Chinese/Chinese-Canadian or White experimenter.

The random assignment to conditions resulted in an even distribution of children from different backgrounds. There were no significant differences between the conditions with respect to age, English exposure, and exposure to non-local accents ($M = 3.44$ on a 7-point scale, with 1 indicating exposure only to the locally dominant accent, and 7 exposure only to non-local accents). Children's morphosyntactic knowledge, as measured by the Structured Photographic Expressive Language Test Third Edition (SPELT-3) (Dawson et al., 2003), also did not significantly differ between conditions.

The additional 57 children were excluded from the analysis prior to examining the data due to children getting upset/not wanting their caregiver(s) to leave the room (17), needing to use the restroom during the experiment (4), not wanting to eat any (or two) treats (4), eating the treat before the experimenter finished with the instructions (3), explicitly indicating lack of understanding of the task (e.g., asking for clarification as the experimenter was exiting the room) (5), exiting the room during the experiment (6), interacting with parents through the door (7), and experimenter error (e.g., deviations from the script; 11). The exclusions were distributed relatively equally across experimenters ($n_{\text{Non-native, Chinese}} = 16$, $n_{\text{Non-native, White}} = 12$, $n_{\text{Native, Chinese-Canadian}} = 17$, $n_{\text{Native, White}} = 12$).

2.2. Procedure

After obtaining consent, children were seated at a centrally located, small round table in a room that was otherwise empty, save for some decorations on the walls (see Fig. 1). Cameras positioned in two of the corners of the room allowed caregivers to monitor their children on a screen outside the testing room as children completed the delayed gratification task.

During the delayed gratification task, children interacted with one of the four female, undergraduate experimenters, who wore identical gray t-shirts during the experiment (see Fig. 1). All experimenters were trained together to deliver the script with similar affect, timing, and body movements throughout. As mentioned previously, two of the experimenters were native-accented speakers from the local area, and the other two were non-native speakers of English, one of whom spoke with a Mandarin accent and the other with a Ukrainian accent (see Johnson et al., 2022, for a review of children's abilities to identify and comprehend other-accented speech). Within each accent condition, one experimenter was White and the other was Chinese(-Canadian).

The experimenter began by presenting children with two plates, one containing a gummy shark and the other a marshmallow, and asked them which one they wanted to eat, thereby helping to ensure that the children found the treat to be of high value. After children selected their preferred treat, the experimenter then proceeded to explain the delayed gratification task based on a memorized script adapted from Kidd et al. (2013), indicating that the child could eat their preferred treat at any time, or, if they could wait for the experimenter to return, they could eat two instead. Unlike most studies, which signal linguistic identity via accent only, we additionally utilized grammaticality, another cue of linguistic identity often coinciding with non-native accentedness. Specifically, the script was either error-free or contained a number of grammatical errors in the form of missing verbs (copula, auxiliary) and determiners, tense errors, incorrect auxiliary verbs, and word order errors (see Table 1 for the grammatical and ungrammatical scripts).[1] Despite the presence of salient errors in the ungrammatical scripts, children indicated their comprehension by readily responding to the questions posed by the experimenter during their interaction. As in Kidd et al. (2013), children were then left alone in the room until they consumed the treat or until 15 min had elapsed, at which point a research assistant entered the room and presented children with a second treat (regardless of whether they waited the entire 15 min or not). Although they were not explicitly told to do so, most children remained seated at the table while waiting for the experimenter to return. Before leaving the lab, children's morphosyntactic knowledge of English (e.g., use of pronouns, negation, modal verbs, etc.) was assessed with the SPELT-3 (Dawson et al., 2003).

Caregivers also filled out a language/demographic questionnaire during their visit, where they provided information about their child's linguistic background (English and accent exposure) and race. To determine each child's race, we asked a caregiver to check all of the categories that applied to them, the other caregiver (if relevant), and their child, using the following categories taken from a scale previously used by Statistics Canada (Statistics Canada, 2017): Aboriginal (Inuit, Métis, North American Indian), Arab/West Asian (e.g., Armenian, Egyptian, Iranian, Lebanese, Moroccan), Black (e.g., African, Haitian, Jamaican, Somali), Chinese, Filipino, Japanese, Korean, Latin American, South Asian, South East Asian, White (Caucasian), Other (specify), or prefer not to say.

2.3. Coding

Two research assistants coded video recordings of the children, noting (1) when the experimenter released her hand from the plate containing the treat (*I'll leave this here*), and (2) when children bit/licked the treat, or when 15 min had elapsed (whichever came first). Wait times for each coder were then calculated by taking the difference in time (seconds) between the start and end times for each child. Calculated wait times that were within 3 s of one another between the two coders (80% of cases) were considered to be in agreement; in these cases, the larger of the two wait times was selected as the final wait time, as in Kidd et al. (2013). A third rater was asked to code the videos for the remaining cases (20%) in which there was not agreement. In all cases, the third coder's calculated wait times were within 3 s from that of one of the first two coders, and as before, the larger of the two agreeing values was selected for the final analysis.

3. Results

3.1. Planned analyses

Linguistic cues. First, we were interested in whether an experimenter's linguistic cues predicted the likelihood of successfully holding out for the second treat. To test this, we ran a mixed-effects logistic regression using the *glmer* function in the lme4 package in R (Bates et al., 2015; R Core Team, 2019), predicting the likelihood of waiting the entire 15 min. We supplemented this analysis by running the same model using the *brm* function in the brms package (Bürkner, 2017) with weak priors in order to calculate credible intervals for our estimates (see Appendix for a table showing results of all models). Given that all of the children tested were native speakers of Canadian English, we expected the linguistic cues (accent and grammaticality) to have a similar influence across all children. To this end, we included fixed effects of Experimenter Accent (Non-native, coded −0.5, Native, coded 0.5), Grammaticality (Ungrammatical, −0.5, Grammatical, 0.5), and their interaction into the model, but no factors relating to children's linguistic

[1] Since infants and toddlers are already sensitive to these kinds of morphosyntactic errors in speech (e.g., Kedar et al., 2006; Shady et al., 1995; van Heugten & Johnson, 2011), we expected 3- to 5-year-old children to have little difficulty in perceiving the morphosyntactic errors in the ungrammatical condition, especially given previous work showing that preschool-aged children respond differently to grammatical and ungrammatical informants (e.g., Sobel & Macris, 2013).

Fig. 1. An image of the testing room, showing an experimenter interacting with a participant.

Table 1
Scripts used by experimenters in the grammatical and ungrammatical conditions.

Grammatical	Ungrammatical
Hello, how are you doing today? What is your name? Hi [child's name]. Today we are going to start off with snack time! Look, I brought some goodies [Kneel down. Set treats on marks placed on table.]. A marshmallow and a gummy shark. Which would you like to eat? [child chooses. Put the other treat under the table.]. You can eat this one [preferred treat] right now. Or—if you can wait for me to get another [preferred treat] from the other room—you can have a second one. Does that sound good? [Response.] Okay, I'm going to go get another [preferred treat] from the other room! You should stay right here in that chair. Can you do that? [Response. Starting to get up. Take other treat out with you.] So I'll leave this here [pushes plate so that it is approximately 12 in. from the child], and if you haven't eaten it when I come back, you can have two [preferred treat] instead.	Hello, how **are** you doing today? What **is** your name? Hi [child's name]. Today we **are going to** start**ing off** with snack time! Look, I **bring some** goodies [Kneel down. Set treats on marks placed on table.]. A marshmallow and **a** gummy shark. Which **snack would** you like to eat? [child chooses. Put the other treat under the table.]. You can eat**ing** this one [preferred treat] right now. Or—if you can wait for me to get another [preferred treat] from **the** other room—you can have **a** second one. **Is** that sound good? [Response.] Okay, **I'm go other room** to **go** get **another more** [preferred treat] **from the other room**! You should stay right here in **that** chair. Can you do**ing** that? [Response. Starting to get up. Take other treat out with you.] So I'**ll** leav**ing** this here [pushes plate so that it is approximately 12 in. from the child], and if you haven't eat**en** it when I come back, you can have two [preferred treat] instead.

backgrounds (since all were native speakers of Canadian English). Although it was not an original variable of interest, since children matched or mismatched the experimenter with respect to gender, and previous research has shown that children (at least older children) evaluate same-gender individuals more favorably than different-gender individuals (e.g., Powlishta, 1995; Susskind & Hodges, 2007), we included a fixed effect of Child Gender (Male, −0.5, Female, 0.5). Finally, the model included mean-centered Age, English Exposure, and Accent Exposure as co-variates, along with random intercepts for each experimenter. Since 9 parents did not complete our language/race questionnaire, the reported results are for 150 instead of 159 participants, though a model run on the full data set (leaving out English and Accent Exposure as co-variates) produced similar results. Results revealed a significant, negative intercept ($\beta = -1.43$, $SE = 0.29$, $z = -5.01$, $p < .001$), indicating that children were more likely to eat or lick the treat than to wait for the experimenter to return. Indeed, only 22% of children waited the entire 15 min, which is consistent with previous literature, where children (especially those under the age of 5) are more likely (than not) to forgo the second treat (e.g., out of 543 children tested in the Bing cohort, about one third waited the entire 15 min; Benjamin et al., 2020). Although children were numerically more likely to wait in the Non-native and Ungrammatical conditions, the differences were not found to be statistically significant ($\beta_{Accent} = -0.46$, $SE = 0.55$, $z = -0.83$, $p = .41$ and $\beta_{Grammaticality} = -0.40$, $SE = 0.42$, $z = -0.95$, $p = .34$), nor was the interaction between Experimenter Accent and Grammaticality ($\beta = 0.01$, $SE = 0.87$, $z = 0.02$, $p = .99$), providing little evidence that language cues alone influenced the degree to which children waited for the experimenter to bring back a second treat.

Next, we examined whether the amount of time that children waited differed based on Accent and Grammaticality. Since children's wait times were right-censored (i.e., the true wait times of children who waited the entire 15 min were unknown) and not normally distributed, we conducted a survival analysis, estimating whether the probability of an event occurring across time (biting/licking the treat) differed as a function of Accent and Grammaticality (for other examples of survival analysis applied to delayed gratification data, see Doebel & Munakata, 2018; Ma et al., 2018 and Michaelson & Munakata, 2016). Specifically, we conducted a mixed-effects Cox proportional hazards regression model using the coxme package (Therneau, 2020) in R, predicting the probability of resisting the treat across time from Accent (Non-native, −0.5, Native, 0.5), Grammaticality (Ungrammatical, −0.5, Grammatical, 0.5), and their interaction. As in the previous analysis, children's mean-centered Age, English Exposure, and Accent Exposure, as well as Gender (Male, −0.5, Female, 0.5) were included, along with random intercepts for each experimenter. As in the previous analysis, children were numerically more likely to hold out for non-native speakers compared to native speakers, though neither the effect of Experimenter Accent ($\beta = 0.19$, $SE = 0.21$, $z = 0.89$, $p = .37$) nor its interaction with Grammaticality ($\beta = 0.21$, $SE = 0.38$, $z = 0.55$, $p = .58$)

were found to be significant. Similarly, there was no effect of Grammaticality ($\beta = 0.08$, $SE = 0.19$, $z = 0.42$, $p = .67$), suggesting that children resisted licking or biting the treat at similar rates across the conditions.

Race cues. We additionally investigated whether the race of the experimenter (White or Chinese/Chinese-Canadian) influenced children's wait times. In contrast to children's linguistic background, which was uniform across all participants (in terms of them being native speakers of Canadian English), children's racial/ethnic backgrounds were quite varied. As a result, we could not assume that experimenter race would influence all children uniformly, but needed to consider the specific race of each child in relation to that of their experimenter. Of primary interest was how children behaved when interacting with a same-race vs. a different-race experimenter. Given our heterogeneous sample, we had many children from backgrounds that matched neither of the experimenters, as well as a number of biracial children whose race in relation to their experimenter's was difficult to determine. Furthermore, we had very few participants who were identified as being exclusively of Chinese ancestry (only 6, or 4 %), making it impossible to analyze how children responded to same-race, Chinese(-Canadian) experimenters. Therefore, we decided to focus our analysis on the children who were categorized as White (only), who represented the largest demographic of children obtained in our sample (62, or 39%), and who also represented the demographic for which we had data in both the same-race and different-race conditions ($n_{\text{White Experimenter}} = 31$, $n_{\text{Chinese(-Canadian) Experimenter}} = 31$). Towards this analytical goal, we categorized children whose caregivers both identified as White in one group ($n = 62$), and the remaining children in another group ($n = 86$).

To determine each child's race, we asked caregivers on the questionnaire they completed before the start of the experiment to check all of the categories that applied to them, the other caregiver, and their child (see Procedure above for list of categories). We did not obtain complete race information from 11 families, reducing the sample size to 148 participants (note that each experimenter had approximately the same number of children in each race category, both in the Grammatical and Ungrammatical conditions).

We then investigated the effect of Experimenter Race (White vs. Chinese/Chinese-Canadian) on both the likelihood of waiting the entire 15 min and on children's wait times. For the first analysis, we conducted a mixed-effects logistic regression (supplemented by Bayesian credible intervals) predicting the log odds of successfully waiting 15 min from Experimenter Race (Chinese/Chinese-Canadian, −0.5, White, 0.5), Child Group (Diverse, −0.5, White, 0.5), and their interaction. Child Gender (Male, −0.5, Female, 0.5), and mean-centered Age, English Exposure, and Accent Exposure were included as fixed effects, along with random intercepts for each experimenter. Interestingly, we found a significant interaction between Child Group and Experimenter Race ($\beta = -2.08$, $SE = 0.93$, $z = -2.23$, $p = .03$) (see Fig. 2). To understand the nature of this interaction, we re-ran the model twice, once where the White Group served as the reference level (White, coded 0, Diverse, 1), and once with the Diverse Group as the reference level (Diverse, 0, White, 1). Interestingly, we observed a significant effect of Experimenter Race in the White Group ($\beta = -2.01$, $SE = 0.72$, $z = -2.81$, $p < .01$), such that White children were significantly less likely to wait the entire 15 min when interacting with a same-race (i.e., White) experimenter compared to when they were interacting with a different-race (Chinese/Chinese-Canadian) experimenter, but found no such effect in the Diverse Group ($\beta = 0.06$, $SE = 0.59$, $z = 0.10$, $p = .92$). Given that this group of children was quite heterogeneous, it is difficult to interpret what this result might mean. However, this apparent lack of an effect helps to underscore the fact that the difference observed in the White group is indeed about the relationship between the experimenter and the participant, and not just an artifact of individual experimenters (who were identical between the White and Diverse groups).

With respect to children's wait times, we similarly see a significant interaction between Experimenter Race and Child Group ($\beta = 0.86$, $SE = 0.40$, $z = 2.18$, $p = .03$), as revealed by a mixed-effects Cox proportional hazards regression predicting the probability of resisting the treat across time, with Child Group (Diverse, −0.5, White, 0.5), Experimenter Race (Chinese(-Canadian), − 0.5, White, 0.5), Age (mean-centered), English Exposure (mean-centered), Accent Exposure (mean-centered), Gender (Male, −0.5, Female, 0.5), and the interaction between Child Group and Experimenter Race included as fixed effects. Random by-experimenter intercepts were also included. Consistent with the previous analysis, White children interacting with a Chinese(-Canadian) experimenter were significantly more likely to resist eating/licking the treat across time compared to White children interacting with a White experimenter ($\beta = 0.82$, $SE = 0.31$, $z = 2.61$, $p < .01$), while no differences were observed between experimenters in the Diverse group of children ($\beta = -0.04$, $SE = 0.24$, $z = -0.18$, $p = .86$) (see Fig. 2).

3.2. Exploratory analyses

Considering accent cues together with race cues. Despite the wealth of studies showing that children have strong preferences for native-accented speakers, which has been sometimes shown to trump preferences for same-race individuals (Kinzler et al., 2009), we found no evidence of experimenter accent influencing children's behavior in our primary analyses. It should be noted, however, that most studies investigating children's evaluations of speaker accent explicitly contrast non-native speakers with native speakers by pitting them against one another (e.g., *Which one do you want to be friends with?*), which might serve to make accent particularly salient (Vanderbilt et al., 2014). In our experiment, however, children interacted *either* with a native or non-native speaker, but not both. In other studies, where children are exposed to only one individual (e.g., either a native or non-native speaker), researchers have observed differences in how children evaluate them only when individuals mismatch (or match) children on *multiple* cues (e.g., mismatched accent + mismatched race), and not when they mismatch on a single cue (e.g., matched accent and mismatched race). For example, McDonald and Ma (2016)—who exposed children either to a native or non-native speaker of English—found that 3- to 4-year-old White monolinguals were significantly less likely to believe the non-native speaker compared to the native speaker, but *only* when racial cues were present as well (the native speaker was White, while the non-native speaker was Asian). In contrast, children were equally credulous when interacting with a native-accented Asian or a foreign-accented White speaker (Experiment 2) (see also Spence & Imuta, 2020 for a similar example).

In light of these findings, we conducted one final set of analyses exploring the potentially added influence of accent beyond that of

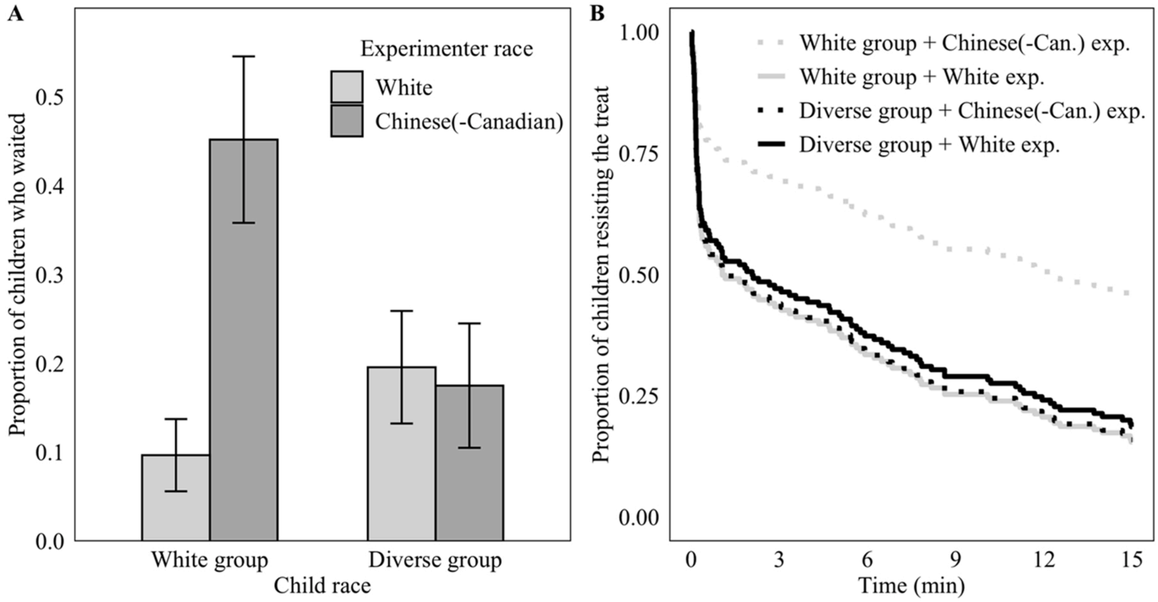

Fig. 2. Children's waiting behavior based on the race of the experimenters and the race of children. (a) shows the proportion of children who waited the entire 15 min before eating the treat, with error bars representing + /- 1 standard error, and (b), the survival functions reflecting the risk of biting/licking the treat across time for each race combination.

race, comparing children's behavior when the experimenter matched both children's native accent and their race to cases where the experimenter mismatched children on either one or both dimensions.[2] As before, we focused our analyses on the White children in our sample, and whether they matched (or mismatched) the experimenter's race and/or accent ($n_{\text{MatchedBoth}} = 17$, $n_{\text{MismatchedAccent}} = 14$, $n_{\text{MismatchedRace}} = 15$, $n_{\text{MismatchedBoth}} = 16$). Given the results of McDonald and Ma (2016), we predicted that children's behavior when the experimenter mismatched children on either race or accent alone might not differ considerably from children's behavior when the experimenter matched children on both dimensions, but that children's behavior when the experimenter differed in both race *and* accent would. First, we examined the likelihood of children waiting the entire 15 min between the four conditions (Matched Race and Accent, Mismatched Race, Mismatched Accent, and Mismatched Race and Accent) using a logistic regression. Since we were interested in how children's behavior differed between the completely matched condition (Matched Race and Accent) and each of the mismatched conditions, we treatment coded our independent variable, specifying the Matched condition as the reference level, and created three dummy-coded variables, each of which contrasted the reference level with one of the mismatched conditions (Mismatched Race, Mismatched Accent, and Mismatched Race and Accent). As in previous analyses, Child Gender, mean-centered Age, mean-centered English Exposure, and mean-centered Accent Exposure were included as well. Consistent with McDonald and Ma, children's behavior in the Mismatched Accent condition did not differ from children's behavior in the completely Matched condition ($\beta = -0.55$, $SE = 1.30$, $z = -0.43$, $p = .67$), and while children appeared more likely to wait in the Mismatched Race condition, this difference was also not found to be significant ($\beta = 1.40$, $SE = 0.97$, $z = 1.44$, $p = .15$). However, children's behavior in the completely Mismatched condition *did* significantly differ from that in the completely Matched condition ($\beta = 2.22$, $SE = 0.93$, $z = 2.39$, $p = .02$). Using the survival package in R (Therneau, 2021), a Cox proportional hazards regression model on wait times—using the same predictors—found comparable results. Specifically, children were significantly more likely to resist biting/licking the marshmallow (compared to the reference level) when interacting with an experimenter who mismatched on both Race and Accent ($\beta = -1.18$, $SE = 0.48$, $z = -2.49$, $p = .01$), as opposed to an experimenter who mismatched on only Accent ($\beta = 0.04$, $SE = 0.40$, $z = 0.09$, $p = .93$) or Race ($\beta = -0.56$, $SE = 0.43$, $z = -1.31$, $p = .19$), providing some evidence that accent information, in conjunction with race information, influenced children's interpretation of the experimenter (see Fig. 3). However, given the exploratory nature of this analysis, the fact that there was only one experimenter in each of the four conditions, and the fact that we have collapsed the results across different levels of grammaticality, we readily acknowledge that this issue needs to be explored further in future work with more experimenters.

4. General discussion

The goal of the present study was to investigate whether the identity of an experimenter, a factor rarely considered or reported on in

[2] Since previous literature has focused primarily on the relationship between race and accent, we chose to focus our exploration on accent, and therefore, do not include an analysis examining grammaticality as well.

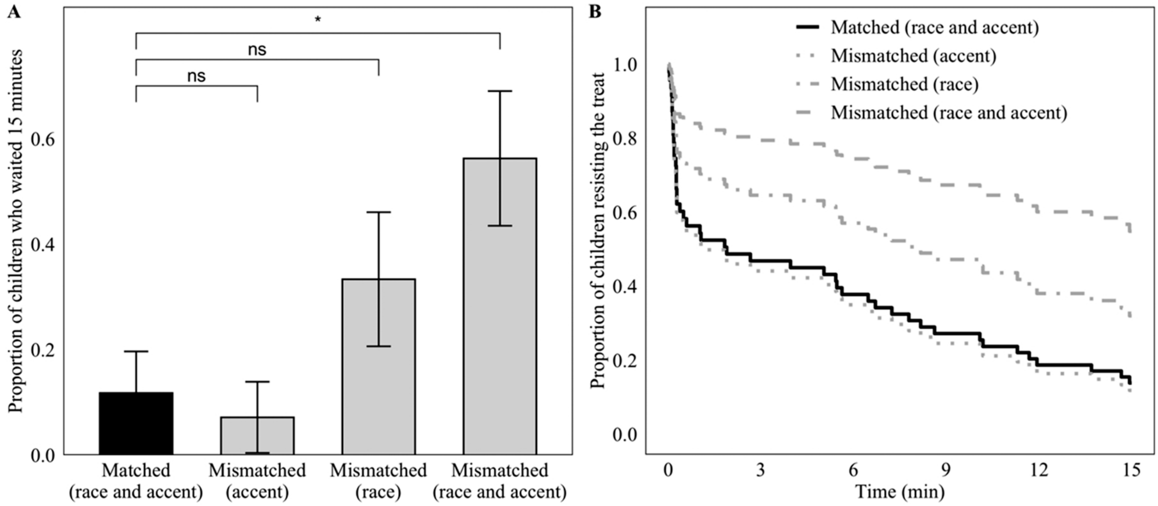

Fig. 3. White children's waiting behavior with each of the four experimenters, who either matched children with respect to race and accent, or mismatched along one or both dimensions. (a) shows the proportion of children who waited the entire 15 min before eating the treat, with error bars representing + /- 1 standard error, and (b) shows the survival functions reflecting the risk of biting/licking the treat across time for each match/mismatch condition.

developmental work, might influence children's performance in an experiment where experimenter identity (at least at first blush) might not be expected to matter. As a case in point, we ran a version of the classic marshmallow task and investigated whether children would differ in their willingness to wait for a second treat based on the experimenter's identity (in relation to that of the children they tested), focusing on linguistic and race cues. Although we did not find convincing evidence that the experimenter's accent (by itself), or their propensity to make grammatical errors, influenced children's wait times, we *did* find a robust effect of race, such that the White children in our experiment were more likely to wait for the second treat when interacting with a Chinese(-Canadian) experimenter who mismatched their race compared to one who matched their race (i.e., was White). In line with previous studies (McDonald & Ma, 2016; Spence & Imuta, 2020), we also found *some* evidence in an exploratory analysis that accent information—in conjunction with race information—influenced children's waiting behavior. Specifically, while White children who interacted with a same-race and same-accent experimenter did not behave differently from White children who interacted with an experimenter sharing only one of those attributes, they did wait significantly less than White children who interacted with an experimenter who mismatched children along both dimensions, suggesting that two cues together were enough to affect children's behavior. These results have implications across a variety of areas in developmental psychology, as they demonstrate that experimenter identity influences children's behavior even in a task that has not traditionally focussed on children's social judgments.

Although it is clear that the identity of the experimenter influenced children's behavior in the marshmallow task, what remains less clear is why exactly White children waited more for other-race experimenters (particularly the experimenter who mismatched children in both race *and* accent) than same-race experimenters. Earlier, we outlined two possibilities, a reliability-/trust-based account, in which children might wait less for an out-group experimenter, assuming that the experimenter would not follow through on their promise, and a comfort-based account, in which children might instead wait *more* for out-group experimenters, if they felt more uncomfortable and less at liberty to eat the immediately available treat (should they have wanted to). Our results are consistent with the latter possibility. Although children were theoretically given a free choice between eating the immediately available treat or waiting for an additional one ("You can eat this one marshmallow right now. Or—if you can wait for me to get another marshmallow from the other room—you can have a second one."), they may have been more anxious/wary when interacting with other-race experimenters (Feinman, 1980; Yu et al., 2021), and as a result, felt more deterred from eating the immediately available treat, should they have wanted to. Given that White adults report trying to avoid intergroup contact because of their intergroup anxiety (Anicich et al., 2021), it is not unreasonable to think that White children too might be more uncomfortable interacting with other-race individuals. At the same time, children's tendency to view in-group members more positively (e.g., Dunham et al., 2011; Patterson & Bigler, 2006; Plötner et al., 2015) may have made them feel more at ease to do as they pleased after interacting with same-race individuals. Future research is needed to better understand how exactly experimenter identity influences children's behavior in the marshmallow task, especially through more targeted recruitment of participants (e.g., in our case, recruiting both White European and Chinese descended participants in equal numbers), and greater recruitment of experimenters. Indeed, while we made use of a sizeable number of experimenters, who were carefully selected for their race and linguistic background, future work should include even greater numbers of experimenters, in order to be more confident that differences between conditions are due to the specific characteristics under investigation (e.g., accent/race) and not to idiosyncrasies of the experimenters assigned in each group (though we find it unlikely that the logically interpretable pattern of results we found as a function of the number of matching cues would have occurred by chance). Future work might additionally benefit from exploring the potential effects of other experimenter cues in the marshmallow task. For example, although we did not observe an effect of child gender in our study, it would be important in future work to examine the role of gender more carefully, specifically through the addition of male experimenters. Finally, given that this is the first study to suggest that children's behavior in the marshmallow task is influenced by their level of comfort, it will be necessary in future research to confirm this hypothesis by correlating children's behavior in the marshmallow task with physiological measures of stress/anxiety (see e.g., Page-Gould & Akinola, 2015; Wei et al., 2018).

Interestingly, the waiting behavior of the diverse group of children for both the White and Chinese(-Canadian) experimenters was similar to the waiting behavior of White children who interacted with a same-race experimenter, perhaps suggesting that the diverse group of children felt comfortable with experimenters of both races (see Fig. 2). Given the high degree of heterogeneity in the group, it is difficult to determine why that would be the case, though the finding is consistent with a large body of work showing that children from minority racial and ethnic groups tend to show weaker in-group favoritism than children from the majority racial group (in this case, White) (e.g., Aboud & Skerry, 1984; Alexandre et al., 2007; Corenblum & Annis, 1987; Griffiths & Nesdale, 2006; Milner, 1973; Teplin, 1976).

This study is the first study to propose that the identity of the experimenter influences children's comfort level and decision-making in a delayed gratification task. Beyond the marshmallow task, we expect that experimenter identity has the potential to influence a variety of other mental processes as well, including children's attention and learning (van Rooijen et al., 2019), motivation (Labov, 1972), and perception (Niedzielski, 1999), among other things (see St. Pierre et al., 2022 for a more in-depth discussion). Given this potential for experimenter identity to influence a variety of cognitive processes, we expect that the influence of experimenter identity is not specific to the marshmallow task (and self-control), but might be present in a variety of areas of developmental psychology which include interactions between children and experimenters. More research is needed to understand the different ways in which experimenter identity can influence children's behavior. Importantly, given that children around the world are exposed to highly variable cultural contexts, the inferences that children form about different experimenters will be highly dependent on their local socio-cultural environment, for example, their neighborhood demographics (Hwang et al., 2021), the social structure (Shutts et al., 2011), and how children are socialized (Lamm et al., 2018). We expect, therefore, that the specific ways in which the identity of the experimenter influences children's behavior will likely differ from place to place, but identifying patterns across various contexts will aid in understanding experimenter effects and interpersonal/intergroup interactions across development more generally (Bigler et al., 2016).

Despite the field of developmental psychology being aware of the potential influence of experimenter identity on participants' behavior, most researchers nevertheless fail to consider (or a least report on) the identity of the experimenter in their studies. In the current study, using the well-known marshmallow task, where experimenter identity might not be obviously expected to influence children's behavior, we find that experimenter identity indeed influences children's willingness to wait for a second treat. We expect that experimenter identity effects are not specific to the marshmallow task or self control, but are present across a wide range of domains in developmental psychology, and may be influencing the results of our studies. As we have outlined elsewhere (St. Pierre et al., 2022), we propose that, going forward, researchers should report demographic information about their experimenters in their studies, especially those attributes which are routinely reported about participants (e.g., age, gender, and race or ethnicity). This will help to provide a more complete understanding of the experimental context of studies and the generalizability of findings, and may help researchers better understand divergent findings between labs. We may even find, in some situations, that an effect that appears to be caused by one variable, is in fact due to the identity of the experimenter(s) instead; for example, in samples where SES correlates with race, an apparent effect of SES could actually be due to experimenter identity (race) and its relation to that of the children tested. In sum, we hope that this study will help bring attention to this *still* neglected variable in experimental work (McGuigan, 1963; Silverman, 1974), and encourage researchers in a variety of domains to consider the potential impact of experimenter identity in their own research, especially as we seek to diversify not only the participants we test, but also the students that we recruit to work in our labs.

Data Availability

The data that support the findings of this study are openly available in OSF at https://osf.io/tjbru/?view_only=37a835e9c05b4a8abb68aa325e7bff2a.

Acknowledgements

We would like to thank Ala Refai, Runyi Yao, Khrystyna Mandziy, Mariam Galytskyy, Grace Wang, Lisa Hotson, and the rest of the Child Language and Speech Studies Lab, as well as the participating families in our community, for their assistance in the completion of this project.

Appendix. . Regression analyses

Linguistic Cues

Variable	Logistic regression	Bayesian Est.	Survival analysis

(continued on next page)

(*continued*)

Variable	Logistic regression				Bayesian Est.	Survival analysis			
	Est.	*SE*	*z*-value	*p*	Est. [Cred. interval]	Est.	*SE*	*z*-value	*p*
	Est.	*SE*	*z*-value	*p*	Est. [Cred. interval]	Est.	*SE*	*z*-value	*p*
Intercept	-1.43	0.29	-5.01	< .001	-1.46 [− 2.75, 0.02]				
Accent (Non-native −.5, Native, .5)	-0.46	0.55	-0.83	.41	-0.33 [− 2.41, 1.84]	0.19	0.21	0.89	.37
Grammaticality (Ungram., −.5, Gram., .5)	-0.40	0.42	-0.95	.34	-0.40 [− 1.24, 0.42]	0.08	0.19	0.42	.67
English Exposure (mean centered)	0.07	0.03	2.77	< .01	0.08 [0.03, 0.14]	-0.03	0.01	-2.78	< .01
Accent Exposure (mean centered)	0.08	0.12	0.69	.49	0.08 [− 0.16, 0.34]	-0.06	0.06	-1.13	.26
Age (months) (mean centered)	0.08	0.04	1.78	.08	0.08 [− 0.01, 0.18]	-0.04	0.02	-1.90	.06
Child Gender (Male, −.5, Female, .5)	-0.06	.42	-0.15	.88	-0.05 [− 0.88, 0.78]	-0.01	0.19	-0.07	.94
Accent x Grammaticality	0.01	0.87	0.02	.99	0.01 [− 1.55, 1.59]	0.21	0.38	0.55	.58

Race Cues

Variable	Logistic regression				Bayesian Est.	Survival analysis			
	Est.	*SE*	*z*-value	*p*	Est. [Cred. interval]	Est.	*SE*	*z*-value	*p*
Intercept	-1.48	0.20	-6.03	< .001	-1.51 [− 2.60, − 0.25]				
Child Race (Diverse, −.5, White, .5)	-0.21	0.50	0.43	.67	-0.20 [− 1.16, 0.74]	-0.17	0.22	-0.76	.45
Experimenter Race (Chinese, −.5, White, .5)	-0.98	0.46	-2.11	.03	-0.80 [− 2.63, 1.17]	0.39	0.20	1.95	.05
English Exposure (mean centered)	0.07	0.03	2.47	.01	0.08 [0.02, 0.14]	-0.02	0.01	-2.26	.02
Accent Exposure (mean centered)	0.08	0.12	0.69	.49	0.09 [− 0.16, 0.34]	-0.05	0.06	-0.84	.40
Age (months) (mean centered)	0.07	0.04	1.56	.12	0.08 [− 0.01, 0.17]	-0.03	0.02	-1.52	.13
Child Gender (Male, −.5, Female, .5)	-0.10	0.43	-0.23	.82	-0.03 [− 0.87, 0.83]	-0.03	0.19	-0.16	.87
Child x Experimenter Race	-2.08	0.93	-2.23	.03	-1.85 [− 3.76, − 0.15]	0.86	0.40	2.18	.03

Exploratory analyses

Variable	Logistic Regression				Bayesian Est.	Survival analysis			
	Est.	*SE*	*z*-value	*p*	Est. [Cred. interval]	Est.	*SE*	*z*-value	*p*
Intercept	-1.94	0.79	-2.45	.01	-1.88 [− 3.40, − 0.59]				
Matched vs. Mismatched (Accent)	-0.55	1.30	-0.43	.67	-0.80 [− 3.25, 1.19]	0.04	0.40	0.09	.93
Matched vs. Mismatched (Race)	1.40	0.97	1.44	.15	1.12 [− 0.52, 2.95]	-0.56	0.43	-1.31	.19
Matched vs. Mismatched (Accent and Race)	2.22	0.93	2.39	.02	2.08 [0.48, 3.88]	-1.18	0.48	-2.49	.01
Age (months) (mean centered)	0.03	0.07	0.44	.66	0.04 [− 0.10, 0.18]	-0.02	0.03	-0.47	.64
Child Gender (Male, −.5, Female, .5)	-0.39	0.68	-0.57	.57	-0.35 [− 1.62, 0.89]	0.20	0.32	0.62	.53
English Exposure (mean centered)	-0.02	0.05	-0.43	.66	-0.03 [− 0.14, 0.08]	-0.01	0.03	-0.19	.85
Accent Exposure (mean centered)	0.002	0.20	0.01	.99	-0.03 [− 0.44, 0.37]	-0.03	0.09	-0.30	.76

References

Aboud, F. E. (1988). *Children and prejudice*. Basil Blackwell.

Aboud, F. E. (2003). The formation of in-group favoritism and out-group prejudice in young children: Are they distinct attitudes? *Developmental Psychology, 39*(1), 48–60. https://doi.org/10.1037/0012-1649.39.1.48

Aboud, F. E., & Skerry, S. A. (1984). The development of ethnic attitudes: A critical review. *Journal of Cross-Cultural Psychology, 15*(1), 3–34. https://doi.org/10.1177/0022002184015001001

Alexandre, J. D., Monteiro, M. B., & Waldzus, S. (2007). More than comparing with majorities: The importance of alternative comparisons between children from different minority groups. *International Journal of Psychology and Psychological Therapy, 7*(2), 201–212. ⟨https://www.redalyc.org/articulo.oa?id=56070206⟩.

American Psychological Association. (2020). *Publication manual of the American Psychological Association 2020: The official guide to APA style* (7th ed.). American Psychological Association.

Anicich, E. M., Jachimowicz, J. M., Osborne, M. R., & Phillips, L. T. (2021). Structuring local environments to avoid racial diversity: Anxiety drives Whites' geographical and institutional self-segregation preferences. *Journal of Experimental Social Psychology, 95*, Article 104117. https://doi.org/10.1016/j.jesp.2021.104117

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

Benjamin, D. J., Laibson, D., Mischel, W., Peake, P. K., Shoda, Y., Wellsjo, A. S., & Wilson, N. L. (2020). Predicting mid-life capital formation with pre-school delay of gratification and life-course measures of self-regulation. *Journal of Economic Behavior and Organization, 179*, 743–756. https://doi.org/10.1016/j.jebo.2019.08.016

Berry, B. A. (2015). Experimenter characteristics, social desirability, and the Implicit Association Test. *Psi Chi Journal of Psychological Research, 20*(4), 247–257.

Bigler, R. S., Rohrbach, J. M., & Sanchez, K. L. (2016). Chapter five – Children's intergroup relations and attitudes. *Advances in Child Development and Behavior, 51*, 131–169. https://doi.org/10.1016/bs.acdb.2016.05.005

Bürker, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*(7), 513–531. https://doi.org/10.1037/0003-066X.32.7.513

Chapman, C. D., Benedict, C., & Schiöth, H. B. (2018). Experimenter gender and replicability in science. *Science Advances, 4*(1), Article e1701427. https://doi.org/10.1126/sciadv.1701427

Charlesworth, T. E. S., Hudson, S.-k. T. J., Cogsdill, E. J., Spelke, E. S., & Banaji, M. R. (2019). Children use targets' facial appearance to guide and predict social behavior. *Developmental Psychology, 55*(7), 1400–1413. https://doi.org/10.1037/dev0000734

Chen, E. E., Corriveau, K. H., & Harris, P. L. (2011). Children are sociologists. *Anales Ďelelőtt Psicología / Annals of Psychology, 27*(3), 625–630. ⟨https://revistas.um.es/analesps/article/view/135181⟩.

Clark, A., Hocevar, D., & Dembo, M. H. (1980). The role of cognitive development in children's explanations and preferences for skin color. *Developmental Psychology, 16*(4), 332–339. https://doi.org/10.1037/0012-1649.16.4.332

Corenblum, B., & Annis, R. C. (1987). Racial identity and preference in native and white Canadian children. *Canadian Journal of Behavioural Science / Revue Canadienne des Sciences Ďelután Comportement, 19*(3), 254–265. https://doi.org/10.1037/h0079985

Corriveau, K. H., Kinzler, K. D., & Harris, P. L. (2013). Accuracy trumps accent in children's endorsement of object labels. *Developmental Psychology, 49*(3), 470–479. https://doi.org/10.1037/a0030604

Corriveau, K. H., Pickard, K., & Harris, P. L. (2011). Preschoolers trust particular informants when learning new names and new morphological forms. *British Journal of Developmental Psychology, 29*(1), 46–63. https://doi.org/10.1348/2044-835X.002009

Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). *SPELT-3: Structured photographic expressive language test* (Third ed.). Janelle Publications.

Doebel, S., & Munakata, Y. (2018). Group influences on engaging self-control: Children delay gratification and value it more when their in-group delays and their out-group doesn't. *Psychological Science, 29*(5), 738–748. https://doi.org/10.1177/0956797617747367

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences, 22*(9), 780–793. https://doi.org/10.1016/j.tics.2018.06.004

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development, 82*(3), 793–811. https://doi.org/10.1111/j.1467-8624.2011.01577.x

Feinman, S. (1980). Infant response to race, size, proximity, and movement of strangers. *Infant Behavior and Development, 3*, 187–204. https://doi.org/10.1016/S0163-6383(80)80025-7

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics, 115*(3), 811–846. https://doi.org/10.1162/003355300554926

Greenberg, R. P., & Gordon, M. (1983). Examiner's sex and children's Rorschach productivity. *Psychological Reports, 53*(2), 355–357. https://doi.org/10.2466/pr0.1983.53.2.355

Griffiths, J. A., & Nesdale, D. (2006). In-group and out-group attitudes of ethnic majority and minority children. *International Journal of Intercultural Relations, 30*(6), 735–749. https://doi.org/10.1016/j.ijintrel.2006.05.001

Hay, J., Drager, K., & Warren, P. (2009). Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics, 29*(2), 269–285. https://doi.org/10.1080/07268600902823128

Hwang, H. G., Debnath, R., Meyer, M., Salo, V. C., Fox, N. A., & Woodward, A. (2021). Neighborhood racial demographics predict infants' neural responses to people of different races. *Developmental Science, 24*(4), Article e13070. https://doi.org/10.1111/desc.13070

Innes, J. M., & Fraser, C. (1971). Experimenter bias and other possible biases in psychological research. *European Journal of Social Psychology, 1*(3), 297–310. https://doi.org/10.1002/ejsp.2420010302

Johnson, E. K., van Heugten, M., & Buckler, H. (2022). Navigating accent variation: A developmental perspective. *Annual Review of Linguistics, 8*, 365–387. https://doi.org/10.1146/annurev-linguistics-032521-053717

Katz, P. A., Sohn, M., & Zalk, S. R. (1975). Perceptual concommitants of racial attitudes in urban grade-school children. *Developmental Psychology, 11*(2), 135–144. https://doi.org/10.1037/h0076446

Kedar, Y., Casasola, M., & Lust, B. (2006). Getting there faster: 18- and 24-month-old infants' use of function words to determine reference. *Child Development, 77*(2), 325–338. https://doi.org/10.1111/j.1467-8624.2006.00873.x

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition, 126*(1), 109–114. https://doi.org/10.1016/j.cognition.2012.08.004

Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011). Children's selective trust in native-accented speakers. *Developmental Science, 14*(1), 106–111. https://doi.org/10.1111/j.1467-7687.2010.00965.x

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences of the United States of America, 104*(30), 12577–12580. https://doi.org/10.1073/pnas.0705345104

Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social Cognition, 27*(4), 623–634. https://doi.org/10.1521/soco.2009.27.4.623

Kuhn, D., Nash, S. C., & Brucken, L. (1978). Sex role concepts of two- and three-year-olds. *Child Development, 49*(2), 445–451. https://doi.org/10.1111/j.1467-8624.1978.tb02334.x

Kwong See, S. T., Rasmussen, C., & Pertman, S. Q. (2012). Measuring children's age stereotyping using a modified Piagetian conservation task. *Educational Gerontology, 38*(3), 149–165. https://doi.org/10.1080/03601277.2010.515891

Labov, W. (1972). Academic ignorance and black intelligence. *The Atlantic Monthly, 229*(6), 59–67.

Lamm, B., Keller, H., Teiser, J., Gudi, H., Yovsi, R. D., Freitag, C., Poloczek, S., Fassbender, I., Suhrke, J., Teubert, M., Vöhringer, I., Knopf, M., Schwarzer, G., & Lohaus, A. (2018). Waiting for the second treat: Developing culture-specific modes of self-regulation. *Child Development, 89*(3), e261–e277. https://doi.org/10.1111/cdev.12847

Leonard, J. A., Berkowitz, T., & Shusterman, A. (2014). The effect of friendly touch on delay-of-gratification in preschool children. *Quarterly Journal of Experimental Psychology, 67*(11), 2123–2133. https://doi.org/10.1080/17470218.2014.907325

Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The origins of social categorization. *Trends in Cognitive Sciences, 21*(7), 556–568. https://doi.org/10.1016/j.tics.2017.04.004

Ma, F., Chen, B., Xu, F., Lee, K., & Heyman, G. D. (2018). Generalized trust predicts young children's willingness to delay gratification. *Journal of Experimental Child Psychology, 169*, 118–125. https://doi.org/10.1016/j.jecp.2017.12.015

Ma, L., & Woolley, J. D. (2013). Young children's sensitivity to speaker gender when learning from others. *Journal of Cognition and Development, 14*(1), 100–119. https://doi.org/10.1080/15248372.2011.638687

Ma, F., Zeng, D., Xu, F., Compton, B. J., & Heyman, G. D. (2020). Delay of gratification as reputation management. *Psychological Science, 31*(9), 1174–1182. https://doi.org/10.1177/0956797620939940

MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science, 10*(3), 307–327. https://doi.org/10.1177/1745691614568482

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology, 44*(4), 645–657. https://doi.org/10.1348/014466604×17948

McDonald, K. P., & Ma, L. (2016). Preschoolers' credulity toward misinformation from ingroup versus outgroup speakers. *Journal of Experimental Child Psychology, 148*, 87–100. https://doi.org/10.1016/j.jecp.2016.03.011

McGuigan, F. J. (1963). The experimenter: A neglected stimulus object. *Psychological Bulletin, 60*(4), 421–428. https://doi.org/10.1037/h0039936

McGuigan, N. (2013). The influence of model status on the tendency of young children to over-imitate. *Journal of Experimental Child Psychology, 116*(4), 962–969. https://doi.org/10.1016/j.jecp.2013.05.004

McLoughlin, N., & Over, H. (2017). Young children are more likely to spontaneously attribute mental states to members of their own group. *Psychological Science, 28*(10), 1503–1509. https://doi.org/10.1177/0956797617710724

Michaelson, L., de la Vega, A., Chatham, C. H., & Munakata, Y. (2013). Delaying gratification depends on social trust. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00355. Article 355.

Michaelson, L. E., & Munakata, Y. (2016). Trust matters: Seeing how an adult treats another person influences preschoolers' willingness to delay gratification. *Developmental Science, 19*(6), 1011–1019. https://doi.org/10.1111/desc.12388

Milner, D. (1973). Racial identification and preferences in 'black' British children. *European Journal of Social Psychology, 3*(3), 281–296. https://doi.org/10.1002/ejsp.2420030307

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., Kross, E., Teslovich, T., Wilson, N. L., Zayas, V., & Shoda, Y. (2011). 'Willpower' over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience, 6*(2), 252–256. https://doi.org/10.1093/scan/nsq081

Moffett, L., Flannagan, C., & Shah, P. (2020). The influence of environmental reliability in the marshmallow task: An extension study. *Journal of Experimental Child Psychology, 194*, Article 104821. https://doi.org/10.1016/j.jecp.2020.104821

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*(1), 62–85. https://doi.org/10.1177/0261927×99018001005

Page-Gould, E., & Akinola, M. (2015). Incorporating neuroendocrine methods into intergroup relations research. *Group Processes & Intergroup Relations, 18*(3), 366–383. https://doi.org/10.1177/1368430214556371

Paquette-Smith, M., Buckler, H., White, K. S., Choi, J., & Johnson, E. K. (2019). The effect of accent exposure on children's sociolinguistic evaluation of peers. *Developmental Psychology, 55*(4), 809–822. https://doi.org/10.1037/dev0000659

Patterson, M. M., & Bigler, R. S. (2006). Preschool children's attention to environmental messages about groups: Social categorization and the origins of intergroup bias. *Child Development, 77*(4), 847–860. https://doi.org/10.1111/j.1467-8624.2006.00906.x

Pesch, A., & Koenig, M. A. (2018). Varieties of trust in preschoolers' learning and practical decisions. *PLoS One, 13*(8), Article e0202506. https://doi.org/10.1371/journal.pone.0202506

Plötner, M., Over, H., Carpenter, M., & Tomasello, M. (2015). The effects of collaboration and minimal-group membership on children's prosocial behavior, liking, affiliation, and trust. *Journal of Experimental Child Psychology, 139*, 161–173. https://doi.org/10.1016/j.jecp.2015.05.008

Powlishta, K. K. (1995). Gender bias in children's perceptions of personality traits. *Sex Roles, 32*, 17–28. https://doi.org/10.1007/bf01544755

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨http://www.R-project.org/⟩.

Raabe, T., & Beelmann, A. (2011). Development of ethnic, racial, and national prejudice in childhood and adolescence: A multinational meta-analysis of age differences. *Child Development, 82*(6), 1715–1737. https://doi.org/10.1111/j.1467-8624.2011.01668.x

Rett, A., & White, K. S. (2022). Children treat grammatical errors differently for native and non-native speakers. *Frontiers in Psychology, 13*, Article 855130. https://doi.org/10.3389/fpsyg.2022.855130

Rhodes, M., & Baron, A. (2019). The development of social categorization. *Annual Review of Developmental Psychology, 1*(1), 359–386. https://doi.org/10.1146/annurev-devpsych-121318-084824

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton-Century-Crofts.

Ross, G., Kagan, J., Zelazo, P., & Kotelchuck, M. (1975). Separation protest in infants in home and laboratory. *Developmental Psychology, 11*(2), 256–257. https://doi.org/10.1037/h0076465

Rotenberg, K. J., & Cerda, C. (1994). Racially based trust expectancies of Native American and Caucasian children. *The Journal of Social Psychology, 134*(5), 621–631. https://doi.org/10.1080/00224545.1994.9922992

Shady, M.E., Gerken, L.A., Jusczyk, P.W. , 1995. Some evidence of sensitivity to prosody and word order in ten-month-olds. In D. MacLaughlin & S. McEwen (Eds.), *Proceedings of the 19th Annual Boston University Conference on Language Development*, Vol. 2 (pp. 553–562). Cascadilla Press.

Shatz, M., & Gelman, R. (1973). The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the Society for Research in Child Development, 38*(5), 1–38. ⟨https://www.jstor.org/stable/1165783⟩.

Shutts, K., Banaji, M. R., & Spelke, E. S. (2010). Social categories guide young children's preferences for novel objects. *Developmental Science, 13*(4), 599–610. https://doi.org/10.1111/j.1467-7687.2009.00913.x

Shutts, K., Kinzler, K. D., Katz, R. C., Tredoux, C., & Spelke, E. S. (2011). Race preferences in children: Insights from South Africa. *Developmental Science, 14*(6), 1283–1291. https://doi.org/10.1111/j.1467-7687.2011.01072.x

Shutts, K., Kinzler, K. D., McKee, C. B., & Spelke, E. S. (2009). Social information guides infants' selection of foods. *Journal of Cognition and Development, 10*(1–2), 1–17. https://doi.org/10.1080/15248370902966636

Silverman, I. (1974). The experimenter: A (still) neglected stimulus object. *Canadian Psychologist/Psychologie canadienne, 15*(3), 258–270. https://doi.org/10.1037/h0081759

Simpson, B., McGrimmon, T., & Irwin, K. (2007). Are Blacks really less trusting than Whites? Revisiting the race and trust question. *Social Forces, 86*(2), 525–552. https://doi.org/10.1093/sf/86.2.525

Sindi, S., Fiocco, A. J., Juster, R.-P., Pruessner, J., & Lupien, S. J. (2013). When we test, do we stress? Impact of the testing environment on cortisol secretion and memory performance in older adults. *Psychoneuroendocrinology, 38*(8), 1388–1396. https://doi.org/10.1016/j.psyneuen.2012.12.004

Sobel, D. M., & Macris, D. M. (2013). Children's understanding of speaker reliability between lexical and syntactic knowledge. *Developmental Psychology, 49*(3), 523–532. https://doi.org/10.1037/a0029658

Spence, J. L., & Imuta, K. (2020). Age-related changes in children's accent-based resource distribution. *Journal of Experimental Child Psychology, 193*, Article 104807. https://doi.org/10.1016/j.jecp.2020.104807

Statistics Canada , 2017. Previous standard – Race (ethnicity). ⟨https://www.statcan.gc.ca/en/concepts/definitions/previous/preethnicity⟩.

St. Pierre, T., & Johnson, E. K. (2020). The development of accent-based friendship preferences: Age and language exposure matter. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 2635–2641). Cognitive Science Society. ⟨https://cogsci.mindmodeling.org/2020/papers/0639/0639.pdf⟩.

St. Pierre, T., White, K. S., & Johnson, E. K. (2022). Experimenter identity: An invisible, lurking variable in developmental research. *Infant and Child Development.* , Article e2357. https://doi.org/10.1002/icd.2357

Strickland, B. R. (1972). Delay of gratification as a function of race of the experimenter. *Journal of Personality and Social Psychology, 22*(1), 108–112. https://doi.org/10.1037/h0032385

Susskind, J. E., & Hodges, C. (2007). Decoupling children's gender-based in-group positivity from out-group negativity. *Sex Roles, 56*, 707–716. https://doi.org/10.1007/s11199-007-9235-z

Teplin, L. A. (1976). A comparison of racial/ethnic preferences among Anglo, Black and Latino children. *American Journal of Orthopsychiatry, 46*(4), 702–709. https://doi.org/10.1111/j.1939-0025.1976.tb00968.x

Therneau, R.M. , 2020. coxme: Mixed Effects Cox Models. R package version 2.2–16, https://CRAN.R-project.org/package=coxme.

Therneau, R.M. , 2021. A Package for Survival Analysis in R. R package version 3.2–10, https://CRAN.R-project.org/package=survival.

Tobin, R. M., & Graziano, W. G. (2010). Delay of gratification: A review of fifty years of regulation research. In R. H. Hoyle (Ed.), *Handbook of personality and self-regulation* (pp. 47–63). Wiley-Blackwell. ⟨https://doi.org/10.1002/9781444318111.ch3⟩.

Vanderbilt, K. E., Heyman, G. D., & Liu, D. (2014). In the absence of conflicting testimony young children trust inaccurate informants. *Developmental Science, 17*(3), 443–451. https://doi.org/10.1111/desc.12134

van Heugten, M., & Johnson, E. K. (2011). Gender-marked determiners help Dutch learners' word recognition when gender information itself does not. *Journal of Child Language, 38*(1), 87–100. https://doi.org/10.1017/S0305000909990146

van Rooijen, R., Bekkers, E., & Junge, C. (2019). Beneficial effects of the mother's voice on infants' novel word learning. *Infancy, 24*(6), 838–856. https://doi.org/10.1111/infa.12312

Wagner, L., Dunfield, K. A., & Rohrbeck, K. L. (2014). Children's use of social cues when learning conventions. *Journal of Cognition and Development, 15*(3), 479–494. https://doi.org/10.1080/15248372.2013.782459

Wei, J., Luo, H., Wu, S. J., Zheng, P. P., Fu, G., & Lee, K. (2018). Transdermal optical imaging reveal basal stress via heart rate variability analysis: A novel methodology comparable to electrocardiography. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00098. Article 98.

Xiao, N. G., Wu, R., Quinn, P. C., Liu, S., Tummeltshammer, K. S., Kirkham, N. Z., Ge, L., Pascalis, O., & Lee, K. (2018). Infants rely more on gaze cues from own-race than other-race adults for learning under uncertainty. *Child Development, 89*(3), e229–e244. https://doi.org/10.1111/cdev.12798

Yu, C., Qian, M., Amemiya, J., Fu, G., Lee, K., & Heyman, G. D. (2021). Young children form generalized attitudes based on a single encounter with an outgroup member. *Developmental Science, 25*(3), Article e13191. https://doi.org/10.1111/desc.13191