Appendix A. Supplementary material

A.1. Additional scripted dialogue

The experimenter used the following scripted dialog to explain each stage of the testing procedure to young study participants.

1. **Onset of experiment** – *"So, today we have a very exciting art project planned for you! Upstairs, we have everything we'll need for you to make your own cup like this one! And you'll be able to take it home with you! Does that sound like something you'd like to do?"*

2. **Art material choice** – "*To decorate your cup, you have a choice of what art supplies to use. You could use these [crayons] right now. Or—if you can wait for me to go get them from another room—you can use our big set of art supplies instead. The big set has markers, pens, colored pencils—a lot of cool stuff. How does that sound? [Response.] Okay, I'm going to go get the big set of art supplies from the other room. You should stay right here in that chair. Can you do that? [Response.] I'll leave these [crayons] right here, and if you haven't used them when I come back, you can use our big set of art supplies instead!"*

3. **Sticker choice** – **"***Would you like to add a sticker to your picture? [Response.] For stickers, you have a choice. You can use this [sticker] right now. Or—if you can wait for me to go get them from the other room—you can have a bunch of stickers to use instead. How does that sound? [Response.] Okay, I'm going to go get more stickers from the other room. You should stay right here in that chair. Can you do that? [Response.] I'll leave this [sticker] here and if you haven't used it when I come back, you can have a bunch of stickers to use instead!"*

A.2. Detailed subject data

Table A.1 contains the wait-time judgments of two video coders for each child participating in the study. Children were randomly assigned to one of two experimental conditions—unreliable and reliable—such that each group was gender and age balanced (nine males, five females, and *M* = 4;6). Two naïve coders watched videos of the children waiting during the final stage of the testing procedure (the marshmallow task). The videos were blinded for condition. The coders measured each child's wait-time until first taste (i.e., lick or bite). The coders' timing judgments were checked against one another to ensure validity, and when timing judgments differed, the later judgment was used (and appears in bold in Table A.1). The judgments of the

Table A.1. Gender, age, and wait-times for each study participant.

				Wait-time (s)			
Condition	**Subject ID**	**Gender**	**Age**	**Coder 1**	**Coder 2**	**Diff**	**Waited 15**
	1	m	3;7	17	17	*0*	n
	3	f	4;0	**21**	19	*2*	n
	5	m	4;0	7	7	*0*	n
	7	m	4;0	15	15	*0*	n
	9	f	4;0	10	10	*0*	n
	11	f	4;1	31	31	*0*	n
Unreliable	13	m	4;4	496	**498**	*-2*	n
	15	f	4;5	900	900	*0*	**y**
	17	m	4;6	457	457	*0*	n
	19	m	4;10	72	**73**	*-1*	n
	21	m	5;3	18	18	*0*	n
	23	m	5;4	150	150	*0*	n
	25	f	5;7	195	195	*0*	n
	27	m	5;7	149	**150**	*-1*	n
		9m, 5f	***M = 4;6***	***M = 181.57***			***7.14%***
	2	m	3;7	900	900	*0*	**y**
	4	m	3;8	785	785	*0*	n
	6	f	3;8	**431**	430	*1*	n
	8	f	4;0	900	900	*0*	**y**
	10	m	4;0	59	59	*0*	n
	12	f	4;1	144	**145**	*-1*	n
Reliable	14	m	4;6	900	900	*0*	**y**
	16	f	4;6	900	900	*0*	**y**
	18	m	4;7	900	900	*0*	**y**
	20	m	4;9	900	900	*0*	**y**
	22	m	4;11	594	594	*0*	n
	24	f	5;5	900	900	*0*	**y**
	26	m	5;9	900	900	*0*	**y**
	28	m	5;10	900	900	*0*	**y**
		9m, 5f	***M = 4;6***	***M = 722.43***			***64.29%***

Table A.2. Mood-variable means and statistical significance tests.

	Group means			
	Unreliable (N = 14)	**Reliable** (N = 14)	**Wilcoxon rank sum test**	**Independent samples t-test**
Contentedness *(z-scores)*	M = 0.03 *(sd = 0.89)*	M = -0.03 *(sd = 0.89)*	W = 106.5, $p > 0.71$	t = 0.178, df = 26, $p > 0.85$
Smiling *(s)*	M = 3.16 *(sd = 3.68)*	M = 4.45 *(sd = 6.53)*	W = 96.5, $p > 0.96$	t = 0.644, df = 26, $p > 0.52$
Fidgeting *(interframe pixel change)*	M = 0.61 *(sd = 0.36)*	M = 0.61 *(sd = 0.39)*	W = 97, $p > 0.98$	t = 0.000, df = 26, $p = 1.00$

two coders were found to differ by at most 2 s on each wait-time. Children's behavior was also coded in terms of a binary outcome measure corresponding to whether or not they waited the entire 15 min without tasting the marshmallow (as indicated in the "Waited 15" column in Table A.1). The percentages in this column reflect the portion of the group that waited the full 15 min: 7.14% in the *unreliable* condition and 64.29% in the *reliable* one.

A.3. Analysis of mood variables

We used three control variables to investigate the potential influence of mood on children's wait times: *contentedness*, *smiling*, and *fidgeting*. Each measurement was based on a portion of each child's video data—the first 30 s of the waiting period.

1. **Contentedness** – Two naïve coders rated each child's apparent contentedness on a scale from 1-9, with 1 indicating *very sad* and 9 indicating *very happy*. We computed *z*-scores for each coder's judgments, and then a mean *z*-score for each child.
2. **Smiling** – Two naïve coders measured for how long each child smiled (s). We use the mean of these two judgments.
3. **Fidgeting** – A Python script automatically calculated an estimate of each child's movement. The script computed the mean number of pixel changes frame-to-frame for each child, above a noise threshold (diff > 50). The threshold served to control for pixel changes caused by the noise inherent in digital frame-to-frame comparisons of this type (caused by, for example, small differences in compression and subtle lighting changes). Thus, the threshold enabled us to measure only changes caused by the body movements of each child.

The mood-variable means for each group and the results of two types of statistical tests appear in **Table A.2**. Wilcoxon rank sum tests indicated that these variables did not significantly differ across conditions in our sample population. Independent samples t-tests ($\alpha_{2\text{-}tail} = 0.05$) also failed to detect significant differences across conditions.