DOI: 10.1111/cdev.14129

CHILD DEVELOPMENT

EMPIRICAL ARTICLE

Delay of gratification and adult outcomes: The Marshmallow Test does not reliably predict adult functioning

Jessica F. Sperber[1] | **Deborah Lowe Vandell**[2] | **Greg J. Duncan**[2] | **Tyler W. Watts**[1]

[1]Teachers College, Columbia University, New York, New York, USA

[2]University of California, Irvine, Irvine, California, USA

Correspondence
Tyler W. Watts, Teachers College, Columbia University, New York, NY, USA.
Email: tww2108@tc.columbia.edu

Funding information
Charles Stewart Mott Foundation, Grant/Award Number: G-2017-00786; Eunice Kennedy Shriver National Institute of Child Health and Human Development, Grant/Award Number: 5 U10 HD027040

Abstract

This study extends the analytic approach conducted by Watts et al. (2018) to examine the long-term predictive validity of delay of gratification. Participants (n = 702; 83% White, 46% male) completed the Marshmallow Test at 54 months (1995–1996) and survey measures at age 26 (2017–2018). Using a preregistered analysis, Marshmallow Test performance was not strongly predictive of adult achievement, health, or behavior. Although modest bivariate associations were detected with educational attainment (r = .17) and body mass index (r = −.17), almost all regression-adjusted coefficients were nonsignificant. No clear pattern of moderation was detected between delay of gratification and either socioeconomic status or sex. Results indicate that Marshmallow Test performance does not reliably predict adult outcomes. The predictive and construct validity of the ability to delay of gratification are discussed.

Early childhood is a foundational period for developing behavioral skills that are crucial for success in school (Blair & Raver, 2015) and adult life (Moffitt et al., 2011). Interest in these skills has been fueled by highly cited work on the importance of the ability to delay gratification (Mischel, 2014). Popularized in psychology by Walter Mischel's studies featuring his self-imposed waiting task (commonly referred to as the "Marshmallow Test"), a child's ability to resist temptation in favor of future rewards was shown to predict a host of later outcomes, including higher SAT scores (Shoda et al., 1990), better coping abilities (Mischel et al., 1988), and lower body mass index (BMI; Schlam et al., 2013). Collectively, these correlational studies argue that children who can recruit a higher level of self-control, beyond mere impulse control, may become better-adjusted adults.

Such a theory is highly plausible, as adults with better regulatory skills will be more likely to avoid the failures of self-control that impede success and fulfillment in one's work and personal life (Duckworth et al., 2018). Indeed, longitudinal work with the Marshmallow Test has been foundational to research on self-control, self-regulation, effortful control, and executive function. These related research areas have been reviewed in several recent theoretical articles (Bailey & Jones, 2019; Inzlicht et al., 2021), with commentators noting the need for more conceptual and operational clarity to distinguish between overlapping constructs (Morrison & Grammer, 2016). The current study focuses on the long-term predictive validity of delay of gratification, as measured by performance on the Marshmallow Test. However, we note links to other studies on self-control and self-regulation, given the clear overlap between these research areas.

Importance of early self-regulation

Findings from Mischel's foundational longitudinal studies (e.g., Schlam et al., 2013) continue to inspire new research focused on exploring the long-term importance of developing the early capacity for self-regulation. Developmental theory argues that early skills impact the acquisition of later skills through unfolding developmental cascades. An example of this idea is that early self-regulatory capacities will lead to further cognitive and social–emotional skill development in childhood, creating a positive skill-forming trajectory that will eventually

Abbreviations: BMI, body mass index; EF, executive function; ICPSR, Inter-University Consortium for Political and Social Research; OSF, Open Science Framework; SECCYD, Study of Early Childcare and Youth Development; SES, socioeconomic status.

manifest in measures of adult functioning (e.g., Masten & Cicchetti, 2010).

Indeed, it is not difficult to imagine why early self-regulation could lay the foundation for sustained success throughout development. Blair and Raver (2015) argue that self-regulatory skills are critical for success in school as they promote a child's ability to monitor their attention and emotions, allowing for prolonged engagement in classroom activities and, ultimately, improved academic performance during the schooling years. These theoretical predictions are often supported by longitudinal studies linking early cognitive and social–emotional measures to later indicators of adult health and well-being. For example, previous work from the Study of Early Childcare and Youth Development (SECCYD) has found that higher executive function (EF) at age 4 predicts greater educational attainment in adulthood—an effect partially explained by higher EF in middle childhood (Ahmed et al., 2021). Highly cited research from the Dunedin study in New Zealand also has documented longitudinal correlations between broad measures of behavioral regulation in childhood and social, economic, and behavioral outcomes in adulthood (e.g., Moffitt et al., 2011). The scales used in the Dunedin study contained a comprehensive set of behavioral measures, pulling from scales that tap attention, externalizing, hyperactivity, and social functioning, among other related behavioral constructs. The Dunedin findings were recently replicated in large cohort studies in the USA (using the SECCYD sample) and in the UK (Koepp et al., 2023), providing further confidence to the robustness of the association between early behavioral regulation and adult outcomes.

Self-regulatory skills and early intervention

Not surprisingly, intervention developers have become increasingly focused on educational programs designed to help children build self-regulatory capabilities. Such efforts include interventions promoting broad self-regulatory skills through altering behavioral norms and early childhood curricular practices (e.g., Morris et al., 2014; Nesbitt & Farran, 2021; Raver, 2009) and more narrow intervention approaches that attempt to target delay of gratification directly (e.g., Murray et al., 2016; Rybanska et al., 2018). These programs are typically implemented under the premise that such interventions might lead to important changes in children's long-term trajectories, particularly for children from disadvantaged backgrounds.

It is well documented that children from disadvantaged backgrounds tend to have poorer self-regulatory capacities than their more privileged peers (Duncan et al., 1994; Raver, 2004). These disparities can partially be explained by experiential differences (Blair & Raver, 2015), including the safety, nurturance, and predictability of early caregiving environments (Brotman et al., 2016). Indeed, changes in "noncognitive" skills related to self-regulation have been recently examined as key mediational processes that might explain how early interventions affect adult outcomes. For example, the High-Scope Perry Preschool study, which randomly assigned children from very low-income backgrounds to a 2-year preschool intervention, found long-term impacts on adult earnings, education, and criminality (Heckman et al., 2010). A decomposition analysis of these effects found that reduced externalizing problems were largely responsible for these adult effects (Heckman et al., 2013), further highlighting the importance of "noncognitive" skills in setting the stage for healthy long-term development (see Elango et al., 2015 for greater discussion).

Predictive validity of the Marshmallow Test

Longitudinal examinations of the predictive validity of the Marshmallow Test have also supported the burgeoning interest around the importance of developing early self-regulatory capacities. The bivariate longitudinal correlations Mischel's team observed initially generated substantial excitement in the field of developmental psychology. For example, Mischel found that children who were able to delay gratification at age 4 had higher SAT scores and were rated as more socially competent by their peers in adolescence (Mischel et al., 1989). In adulthood, waiting longer on the task during early childhood was associated with lower BMI, an improved sense of self-worth, and better coping abilities (Mischel, 2014). Interestingly, the team found that the time waited on the Marshmallow Test could be manipulated via small changes to the procedure: providing the children with cognitive strategies (e.g., think of fun/distracting thoughts) improved wait time on the task (Mischel et al., 1989). These findings suggest that delay of gratification might be an important feature of self-regulation (Mischel et al., 1989), and possibly modifiable as well (Mischel, 2014).

Indeed, Mischel argued that the willpower demonstrated in the Marshmallow Test influenced the life course from preschool to retirement planning (Mischel, 2014). In his book *The Marshmallow Test: Mastering Self-Control* (2014), Mischel argues that self-control capacities could be harnessed to make willpower more automatic, improving the chances that one will achieve their goals and long-term success. However, the predictive validity of Marshmallow Test performance has recently been called into question.

Revisiting the Marshmallow Test

A conceptual replication by Watts et al. (2018) found the predictive power of the Marshmallow Test on academic achievement at age 15 to be diminished substantially

when controls for early life factors were considered. These covariates were theoretically motivated as they reflect characteristics that may explain both delay of gratification performance and later academic achievement, including family demographics (e.g., socioeconomic status), the quality of the early home environment, and concurrent behavioral/cognitive functioning at age 4. Furthermore, they found that most of the effect on achievement was driven by variation at the very low end of the waiting time distribution, suggesting that simple impulse control (and not some higher level of control) may have been responsible for any associations between delay of gratification and later outcomes. The authors concluded that interventions targeting delay of gratification in childhood were likely to have only meager effects on adolescent achievement (see also Watts & Duncan, 2020).

Conversely, other studies provide reason to believe Watts and Duncan were too pessimistic in their conclusions. A reanalysis of the SECCYD dataset found links between delay of gratification and several adolescent outcomes, with the strongest association reported for a measure of problem behaviors at age 15 (Michaelson & Munakata, 2020; see also Duckworth et al., 2013; Falk et al., 2020). Moreover, aforementioned findings from Moffitt et al. (2011) and Koepp et al. (2023) provide evidence that measures of self-control across early and middle childhood strongly predict adult measures of health, wealth, and criminality, even when controlling for IQ and socioeconomic status (SES). Indeed, the Koepp et al. (2023) replication effort of the Moffitt and colleagues' study found that this relation held, even when controlling for concurrent measures of academic achievement. If the measures used in these studies of behavioral regulation tap into the same latent capacity for self-control that causes performance on the Marshmallow Test, one would also expect to observe predictive validity for delay of gratification across similar adult outcomes.

Indeed, there are theoretical reasons to expect that early self-regulatory capacities could support adult outcomes through paths not considered by Watts and colleagues. For example, several evaluations of early childhood programs have found null or fading impacts on adolescent measures of academic performance, yet also find significant impacts on high school graduation (Deming, 2009), college attendance, and juvenile incarceration (Gray-Lobe et al., 2021). These findings raise the possibility that changes in early childhood capacities could affect adult outcomes through paths not captured by academic achievement tests. Similarly, a recent follow-up to a randomized control trial evaluating the efficacy of a self-control intervention for boys with substantial behavioral problems found impressive impacts on adult measures of income and reliance on welfare programs (Algan et al., 2022, but see long-term null results reported for an early self-regulation intervention by Watts et al., 2023). Collectively, this evidence suggests that early regulatory skills could explain important variance in adult functioning, even if these same skills do not reliably predict academic achievement.

The ability to delay gratification might lead to long-lasting changes in children's trajectories, even when controls for other child characteristics, such as the home environment and cognitive functioning, are included. To our knowledge, no study to date has pursued a comprehensive examination of associations between delay of gratification ability and broad assessments of adult behavior considered by other key studies in the self-control literature (e.g., Algan et al., 2022; Moffitt et al., 2011). The lone exception can be found in work by Benjamin et al. (2020), which reported links between Marshmallow Test performance and adult attainment for the children who participated in Mischel's original Marshmallow Test studies. Notably, Benjamin and colleagues found no detectable effect of Marshmallow Test performance on human capital formation at age 40. However, as has been previously argued (e.g., Watts et al., 2018), the external validity of this work is limited given the highly selective nature of the sample (the Benjamin et al. follow-up sample reported an average net worth $1.8 million).

Present study

The current study estimates the association between Marshmallow Test performance and a host of early adult outcomes using new data from the NICHD Study of Early Child Care and Youth Development (SECCYD). Two criteria informed the selection of outcomes examined in the present study. Primarily, when available, the authors selected outcomes that directly aligned with previous investigations of the Marshmallow Test (e.g., BMI, educational attainment). Additionally, the authors selected comparable domains that broadly reflected adult functioning used in other highly influential, longitudinal studies of self-control. For example, a seminal study of self-control conducted by Moffitt et al. (2011) examined outcomes categorized under the larger domains of achievement, behavior, and health (e.g., depression, drug use, criminality, annual income). The present study examines these same outcomes, in addition to several outcomes that are products of, in part, one's theorized capacity for self-regulation (e.g., risk-taking, impulsive behaviors, debt). All outcomes are categorized under the achievement, behavior, and health domains specified by Moffitt et al. (2011).

The present study extends the analytic approach used by Watts et al. (2018) and tests the associations between delay of gratification at age 54 months with age-26 outcomes using a series of control variables to adjust for confounding characteristics that may also cause later adult success. Given the equivocal longitudinal evidence from previous literature (e.g., Benjamin et al., 2020; Moffitt et al., 2011), the authors did not have strong a priori

hypotheses, though they expected to observe bivariate associations between Marshmallow Test performance and later adult outcomes. However, the authors expected that any bivariate associations would be heavily attenuated by the inclusion of controls (i.e., Watts et al., 2018). The key analyses in this study were preregistered, and despite our lack of strong a priori hypotheses, we considered our analyses to be confirmatory for questions regarding the long-term predictive validity of performance on the Marshmallow Test.

METHOD

Preregistration

The measures and analytic plan were preregistered with Open Science Framework (OSF; osf.io/67XFN). This OSF page also contains the analytic code used to produce the findings reported here, as well as descriptive statistics and a correlation matrix of all key variables. The childhood waves of the study (including measurement codebooks) are publicly available at the Inter-University Consortium for Political and Social Research (ICPSR IDs: 21940 and 21941), though the adult waves have not yet been released.

As noted in the preregistration, the authors have worked extensively on this data set, including with the key measures involved with this study. However, no prior analyses conducted by the authors had linked the delay of gratification measure to any of the age-26 outcomes prior to posting the preregistration plan. Thus, the preregistration was a commitment to an analytic plan (heavily drawing on the procedures used in Watts et al., 2018) before the authors observed associations between Marshmallow Test performance and adult outcomes.

Data

We relied on data from the SECCYD (NICHD Early Child Care Research Network, 2002). Families were recruited in 1991 in hospitals following their child's birth at 10 U.S. sites across the United States (n=1364). Participants were followed across childhood and into early adulthood, with the most recent assessment occurring in 2017 and 2018 when participants were age 26. Participants were largely White (83%) and middle-class (average income-to-needs across childhood=3.71, SD=2.50), providing a geographically diverse, though not nationally representative, sample. In the current study, the sample is limited to children who had data on the 54-month Marshmallow Test, and at least one measure of adult functioning at age 26 (n=702). Detailed demographic characteristics of the analytic sample are presented in Table S1. All data collection procedures were approved by the Institutional Review Board of the University of California, Irvine, and data analyses for the current study were deemed exempt by the Institutional Review Board of Teachers College, Columbia University.

Measures

Early childhood delay of gratification

A shortened version of Mischel's (1974) Marshmallow Test (i.e., the "self-imposed waiting task") was administered to children at 54 months (n=1038). Children were brought into a laboratory and presented with a treat of their choice by an experimenter (e.g., M&M's, animal crackers, or pretzels). The experimenter presented the child with both a smaller and larger amount of the treat and verified that the child preferred the larger amount. The experimenter told the child that they would leave the room but offered them a choice: they could either ring a bell to end the task early and eat the smaller snack, or they could wait until the experimenter returned and eat the larger snack. The child was left alone with the rewards visible, and the task was ended if the child rang the bell or ate the treat. A small percentage of invalid cases were set aside (n=72; e.g., children who did not understand the directions), and the amount of time the child waited was recorded as the measure of delay of gratification (5 children with valid performance had no wait time recorded and were excluded from our analyses). The task was capped at 7 min, which was much shorter than the task used in the original studies (see Shoda et al., 1990). Approximately half of the sample (56%) waited the full 7 min.

Age-26 measures

Outcomes related to achievement, health, and behavior at age 26 were assessed. Here, we briefly describe the measures used; further information for each measure is provided in the Supplement.

Achievement outcomes

Educational attainment. On the age-26 survey, educational attainment was reported on a Likert scale ranging from 1 (*no high school diploma*) to 9 (*doctoral degree*). These values were recoded to a continuous scale to reflect years of formal schooling.

Annual earnings. Participants reported their annual earnings. Extremely high earners (7 participants) were recoded to match the 99th percentile of reported earnings ($212,000) before all values were log-transformed for analyses.

Debt. Participants responded to categorical bins reflecting the collective debt of themselves and their household, excluding student debt and a mortgage. The distribution fell such that 70% of participants reported

less than $10,000 of collective debt. Therefore, these bins were collapsed to create a dichotomous indicator from 0 (*less than $10,000 of debt*) to 1 (*more than $10,000 of debt*).

Health outcomes

Depression. Participants completed the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977), a well-validated scale of depressive symptoms. Participants responded to a list of 20 symptoms with how frequently they experienced the symptom during the past week. Items were summed together, with higher scores indicating greater depressive symptoms. Incomplete surveys were excluded from analyses (n = 22).

Body mass index (BMI). BMI was calculated through self-reported height and weight (CDC, 2020).

Substance use. Participants responded to three categorical items about the frequency with which they consumed alcohol or marijuana within the last 30 days or other drugs (e.g., cocaine, crystal meth) in the last 12 months. Items were standardized, averaged together, and then re-standardized to form a composite, with higher values indicating greater use.

Behavioral outcomes

Police contact. Police contact was assessed using an index that reflects the frequency and severity of contact with police, including self-reported lifetime arrests, time spent in jail, and the frequency of driving tickets and traffic accidents over the last 12 months. This index reproduced the index previously reported by Vandell et al. (2021) using the SECCYD data set. Items were averaged together and then standardized for analyses, with higher scores indicating greater police contact.

Impulsive behavior. Impulsive behavior was assessed using a subscale of the Weinberger Adjustment Inventory (WAI; Weinberger & Schwartz, 1990). Participants responded to 8 items (7 items were reverse coded; α = .81) such as "I do things without giving them enough thought" and "I should try harder to control myself when I am having fun" using a Likert scale ranging from 1 (*false*) to 5 (*true*). Items were averaged then standardized, with higher values indicating more impulsive behavior.

Risk taking. Risk-taking behaviors were measured through 30 items that index the frequency with which a participant engaged in various behaviors in the past year (e.g., "How many times in the past year have you been in a physical fight?"). Items were averaged and then standardized, with higher scores indicating greater risk-taking behaviors (Cronbach's α = .82; Vandell et al., 2021).

Covariates

We followed the Watts et al. (2018) list of control variables, using data from various points across early childhood. A full list of covariates (Table S1) and how they were measured are available in the supplement. Briefly, covariates grouped under demographics (Panel A) include child race and ethnicity, maternal age and education, average family income across early childhood, the child's actual age at the time of the delay of gratification, site of data collection, and a measure of maternal receptive vocabulary. The second grouping (Panel B) reflects qualities of the child's early background across the first 3 years of life, including birth weight, infant temperament at 6 months, cognitive development at 24 months, and school readiness at 36 months. We also included a measure of the quality of the early home environment assessed at 36 months. Finally, the last grouping (Panel C) reflects the child's concurrent cognitive and behavioral development at the time of the Marshmallow Test at 54 months, assessed via cognitive performance, internalizing problems, and externalizing problems.

Analytic plan

Our analyses were designed to provide a range of estimates regarding the long-term predictive validity of performance on the Marshmallow Test. Here, we begin with a conceptual overview of our modeling approach, before providing our key regression equation with parameter definitions.

We began with a series of simple bivariate models (Model 1), which essentially provide descriptive information regarding correlations between the early ability to delay gratification and later outcomes. With no controls included, the coefficients from these analyses indicate the extent to which performance on the Marshmallow Test predicts later adult functioning, but confounding variables that cause both early delay of gratification ability and a given later adult outcome will heavily bias any observed effects.

Recognizing that researchers and program developers are also interested in the *unique* relations between the early ability to delay gratification and later measures of adult functioning, we also examine models using covariates. We grouped the covariates into three panels and introduced them in a stepwise fashion: (A) demographics and socioeconomic status (Model 2), (B) early childhood background and quality of the early home environment (Model 3), and (C) cognitive and behavioral abilities at 54 months (Model 4). These covariates were theoretically motivated to account for potential confounding characteristics that may cause variation in both delay of gratification performance and adult outcomes. Specifically, we approached these analyses with the thought experiment of conducting a hypothetical intervention that targets children's ability to delay gratification without changing other characteristics of the child or their environment (see Watts & Duncan, 2020 for extended discussion).

With basic demographics controlled (Model 2), we provide the most optimistic view of the possible effects

of an intervention that broadly changed child capabilities related to self-control. Importantly, these models do not control for other psychological constructs related to the ability to delay gratification. Consequently, they provide the broadest assessment of the construct validity of the Marshmallow Test, as any observed prediction could be due to unobserved linkages with other cognitive or social-emotional capabilities (e.g., intelligence, conscientiousness). The next two models (Models 3 and 4) introduce controls that are more related to the construct of gratification delay ability. Here, we were interested in asking whether an intervention that narrowly targeted a child's ability to delay gratification but did not move related psychological or behavioral constructs (e.g., mathematics achievement, externalizing) would have long-term effects on adult outcomes. These results also bear on conversations regarding the construct validity of the Marshmallow Test, as they further clarify whether delay of gratification is uniquely predictive of adult outcomes when variation is separated from constructs such as temperament, general cognitive ability, and behavioral problems. Indeed, such tests have been conducted in similar studies. For example, previous longitudinal work has reported that early self-control abilities predict adult outcomes above and beyond controls for early IQ (Koepp et al., 2023; Moffitt et al., 2011).

Importantly, we do not elevate one statistical model as the "best" model in this analysis. Rather, we consider this an exercise in testing various developmental theories that posit delay of gratification as a uniquely important capability for determining children's long-term outcomes. To concretize our approach, Model 4 (i.e., the most highly controlled model) is presented below:

$$Outcome_i = a_1 + \beta_1 DoG_i + \sum_{j=1}^{n_1} r_j Demographics_i + \sum_{k=1}^{n_2} \Psi_k EarlyBackground/Home_i + \sum_{m=1}^{n_3} \Psi_m Achievement/Behavior_i + e_i.$$

Here, $Outcome_i$ represents a given measure of adult functioning assessed at age 26 for participant i. This model includes a vector of n_1 demographic measures, a vector of n_2 early child background and home environment measures, and a vector of n_3 concurrent achievement and behavioral measures. The key coefficient is represented by β_1, which captures the expected increase in a given adult outcome for a 1-unit increase in the age 54-months delay of gratification measure. Because we standardized both the continuous outcomes and the continuous measure of time waited on the Marshmallow Test (with the exception of earnings, which was log-transformed), this coefficient can be interpreted as the expected SD change in a given outcome for a 1-SD change in Marshmallow Test performance. Returning to our hypothetical intervention, β_1 in this model can be interpreted as the expected adult impact for an exogenous change to delay of gratification at 54 months for an intervention that did not simultaneously change demographics, early child characteristics and the home environment, or concurrent achievement or behavior. Of course, this model is still susceptible to omitted variables bias, as other factors not observed (or perfectly measured) in our analysis could bias β_1.

This modeling approach was then extended using a categorical indicator for delay of gratification, in line with the analyses presented by Watts et al. (2018). For this second model, the amount of time each child spent waiting was split into discrete groups: ≤0.333 min (n = 114, M = 0.11), >0.333 min but ≤2 min (n = 95, M = 0.94), >2 min but <7 min (n = 103, M = 4.12), and 7 full minutes, n = 390. These four mutually exclusive groups were entered into the model simultaneously (with the ≤0.333 min indicator used as a reference group). We employed this categorical approach for several reasons. First, we sought to address the restricted variation at the task's ceiling, since over half (56%) of the children waited the full amount of time on the task (i.e., 7 min). By comparing predictions across these categorical groups, we can observe if this restricted variation results in a large increase in coefficient size for the 7-min group in linear models. This would indicate the presence of nonlinear associations in the data and, consequently, suggest that estimates from the continuous model could be downwardly biased. In addition, we sought to preserve the variation among children who did not wait the full amount of time by creating discrete groups that were near-equal in size and easily interpretable (i.e., ≤20 sc, 2–7 min, etc.). Watts et al.'s (2018) previous investigation found that much of the predictive validity for adolescent achievement was driven by simply waiting longer than 20 s on the measure, rather than completion of the task (i.e., waiting 7 min). Examining this categorical model allowed us to test if extended periods of time spent waiting on the task, rather than mere impulse control, could explain longitudinal effects. In sum, this model provides a more flexible specification that makes no assumptions about the "true" values for the children who had their wait time truncated by the 7-min stop rule, and it allows a test of nonlinear associations throughout the minutes waited distribution.

To examine differences among these categorical groups, we preregistered two post hoc tests and reported the p-values of these tests in Table 3. The first test (labeled *p-value of test of equality of all categories*) examined if the coefficients of all four categorical groups were significantly different from zero, while the second test (labeled *p-value of test of equality of second, third, and fourth categories*) examined if the coefficients of all groups that waited longer than 20 s (i.e., all groups except the reference group) differed significantly from each other.

Notably, some studies using the SECCYD data set employ a binary version of the delay of gratification measure, by which anyone who did not wait 7 min is set

to 0 (i.e., "fail"), and anyone who did wait 7 min is set to 1 (i.e., pass) (e.g., Michaelson & Munakata, 2020). In line with this previous work, we reported results using this version of the measure as part of an exploratory, non-preregistered analysis. Other exploratory tests that were not pre-registered include moderation analyses by SES and sex, which are reported in Tables S4 and S5.

To account for missing data on adult outcomes and early childhood covariates, multiple imputation with chained equations (MICE) with 25 iterations was used in Stata Version 16.0 (StataCorp, 2019). All models were run with robust standard errors, and all continuous variables were standardized using the analytic sample after imputation. As a sensitivity check, all models were also conducted with the robust regression command in Stata to reduce the influence of outliers (Tables S6–S9). Finally, we reported descriptive differences between the full SECCYD sample and those who completed both the 54-month delay of gratification task and the age-26 assessment (i.e., the analytic sample; Table S10).

RESULTS

Descriptive statistics of the analytic sample during early childhood can be found in Table S1. Of the analytic sample, 46% identified as male, 83% identified as White, 8% as Black, and 5% as Hispanic. Participants waited an average of 4.64 (SD = 2.95) minutes on the Marshmallow Test at 54 months.

Table 1 presents descriptive statistics of the analytic sample at age 26 (n = 702). On average, participants obtained 15.16 years (SD = 1.99) of formal education and reported a median income of $37,440 and a mean of $43,595 (SD = $32,929), adjusted for high outliers.

Continuous measure of delay of gratification

Table 2 presents results from Models 1–4 for the continuous measure of delay of gratification (i.e., standardized minutes waited). In bivariate models, better performance on the Marshmallow Test predicted higher educational attainment (β = .17, p < .001) and lower BMI (β = −.17, p < .001), but no other outcomes. After adjusting for child demographic characteristics (i.e., Model 2), the educational attainment coefficient declined to nonsignificance (β = .04, p = .23). Moreover, the addition of early home environment covariates reduced the association with BMI to nonsignificance (β = −.07, p = .11). No other significant associations for the other seven outcomes tested were observed once controls were included.

TABLE 1 Descriptive statistics of the Marshmallow Test and key age-26 outcome variables.

	N	Mean	SD	Min	Max
Marshmallow Test (54 months)	702	4.64	2.95	0.00	7.00
≤0.333 min	180	16%		0.00	1.00
>0.333–2 min	129	14%		0.00	1.00
2–7 min	138	15%		0.00	1.00
7 min	514	56%		0.00	1.00
Adult outcomes (age 26)					
Achievement outcomes					
Educational attainment	702	15.16	1.99	10.00	21.00
Adjusted annual earnings	669	$43,595.51	$32,929.75	$0.00	$212,000.00
Adjusted annual earnings (log)	669	10.36	1.05	6.00	12.00
Other debt	696	0.29	0.45	0.00	1.00
Health outcomes					
Body mass index	691	26.4	6.14	16.00	58.00
Depression	678	14.56	11.72	0.00	58.00
Drug use composite	696	0.00	0.81	−1.00	3.00
Behavioral outcomes					
Risk-taking behaviors	693	1.11	0.13	1.00	2.00
Impulsive behavior	697	3.89	0.77	1.00	5.00
Police contact composite	696	0.23	0.33	0.00	2.00
Observations	702				

Note: Minutes waited (categorical) represents the four discrete groups of time spent waiting on the Marshmallow Test. Adjusted annual earnings reflect reported annual salary with outliers removed, while the log transformation includes an intercept of 500. Other debt reflects a dichotomous variable ranging from 0 (*no debt*) to 1 (*more than $10,000 in debt*). Sample is restricted to those with Marshmallow Test data and at least one adult outcome of interest.

TABLE 2 Minutes waited on the Marshmallow Test at age 54 months (continuous measure) predicting age-26 outcomes.

	No controls	Panel 1: Child demographics	Panel 2: Child background and HOME	Panel 3: Child behavioral and cognitive abilities
Educational attainment	0.17***	0.04	0.01	−0.01
	(0.04)	(0.04)	(0.04)	(0.04)
Adjusted annual earnings	0.07	0.01	0.01	−0.02
	(0.04)	(0.04)	(0.04)	(0.05)
Other debt	−0.03	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.02)
Body mass index	−0.17***	−0.11*	−0.07	−0.07
	(0.04)	(0.04)	(0.04)	(0.04)
Depression	0.01	0.03	0.04	0.05
	(0.04)	(0.04)	(0.04)	(0.04)
Drug use	0.06	0.04	0.04	0.03
	(0.04)	(0.04)	(0.04)	(0.04)
Risk taking	−0.03	0.06	0.05	0.05
	(0.04)	(0.04)	(0.04)	(0.04)
Impulsive behavior	0.02	0.01	−0.01	0.01
	(0.04)	(0.04)	(0.04)	(0.04)
Police contact	−0.03	0.03	0.05	0.05
	(0.04)	(0.04)	(0.04)	(0.04)

Note: Each age-26 outcome was regressed independently onto Marshmallow Test performance at age 54 months. Robust standard errors are in parentheses. All dependent variables (except adjusted annual earnings and other debt) are standardized, so coefficients can be interpreted as effect sizes. Adjusted annual earnings were log-transformed, and other debt was entered as a binary variable (i.e., we used a linear probability model). Estimates shown in the first column contained only delay of gratification measure, the given outcome measure, and the child's actual age when the Marshmallow Test was administered. Panel 1 added controls for child demographics and site fixed effects. Panel 2 added controls for child background characteristics and the quality of the early home environment. Panel 3 added behavioral and cognitive measures at age 54 months. A list of the covariates included in each panel can be found in Table S1, and a detailed description of how each one was measured can be found in the Supplement under "Covariates." $N = 702$.

*$p < .05$. **$p < .01$. ***$p < .001$.

Categorical measure of delay of gratification

Nonlinear associations with Marshmallow Test performance were examined by using four categorical measures capturing discrete groups of wait time (i.e., ≤0.333, 0.33–2, 2–7, and 7 min). These indicator variables were entered into the regression, with the ≤0.333 min (i.e., 20 s) group serving as the reference group. As discussed above, outcome variables were standardized, but these categorical indicators of Marshmallow Test performance were entered as binary variables (i.e., the binary variables were not further standardized). Table 3 displays the results from only the fully adjusted models (Model 4), and results from Models 1–3 are available in Table S2.

No consistent patterns of nonlinearity were detected, suggesting the results shown in Table 2 were not heavily attenuated due to truncating the Marshmallow Test at 7 min. As Table 3 shows, no significant effects for the "7 min" group were detected in the fully adjusted model. This indicates that there were no statistically significant differences between children who waited the maximum time and those who waited less than 20 s on any adult outcome. Surprisingly, several significant effects for the 2–7 min waited group were observed, but many of these effects were in the opposite hypothesized direction compared with the students who waited less than 20 s (educational attainment: $\beta = -.23$, $p = .05$; annual earnings: $\beta = -.44$, $p = .01$; depression: $\beta = .27$, $p = .03$). In the case of earnings, a disproportionately large number of participants in this category reported earning $0 annually, which likely skewed results. We have no ready explanation for the positive coefficient for the 2–7 min indicator variable in the case of education and depression.

Exploratory and supplementary results

Binary pass/fail

Table 4 reports results from a nonpreregistered analysis examining a binary pass/fail version of the delay of gratification measure (i.e., did the child wait 7 min or not) on adult outcomes. In bivariate models, we again observed significant predictions for educational attainment ($\beta = .36$, $p < .001$) and BMI ($\beta = -.35$, $p < .001$). A significant association was observed with annual earnings ($\beta = .23$, $p = .01$), although it should be noted this was largely driven by the unexpected negative effect for

TABLE 3 Minutes waited on the Marshmallow Test at age 54 months (categorical) predicting adult outcomes—fully adjusted models.

	Achievement outcomes			Health outcomes			Behavioral outcomes		
	Educational attainment	Adjusted annual earnings	Other debt	Depression	Body mass index	Drug use	Police contact	Impulsive behavior	Risk-taking behaviors
Delay minutes (categorical)									
≤0.333 min	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
>0.333 to 2 min	−0.19	−0.11	0.05	0.13	0.04	0.05	−0.12	−0.14	−0.22
	(0.12)	(0.14)	(0.07)	(0.13)	(0.15)	(0.14)	(0.13)	(0.14)	(0.12)
2 to 7 min	−0.23*	−0.44**	−0.06	0.27*	0.13	0.16	0.06	−0.25	0.21
	(0.12)	(0.16)	(0.06)	(0.13)	(0.15)	(0.13)	(0.14)	(0.14)	(0.14)
7 min	−0.09	−0.08	0.05	0.19	−0.12	0.11	0.05	−0.07	0.03
	(0.10)	(0.11)	(0.05)	(0.11)	(0.12)	(0.10)	(0.12)	(0.11)	(0.11)
p-value of test of equality of all categories	0.181	0.013*	0.113	0.241	0.143	0.650	0.518	0.316	0.025*
p-value of test of equality of second, third, and fourth categories	0.278	0.012*	0.061	0.621	0.070	0.740	0.324	0.320	0.010*

Note: All models include covariates for all early childhood characteristics (i.e., Model 4). A list of the covariates can be found in Table S1, and a detailed description of how each one was measured can be found in the Supplement under "Covariates." Robust standard errors are in parentheses. All dependent variables (except adjusted annual earnings and other debt) were standardized before being entered into the regression. *p*-values were generated from post hoc *F*-tests to assess whether respective sets of variables were different from one another. *p*-values below .001 have been rounded to 0.001. $N = 702$.

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 4 Nonpreregistered exploratory analysis: completion of the Marshmallow Test predicting adult outcomes.

	No controls	Panel 1: Child demographics	Panel 2: Child background and HOME	Panel 3: Child behavioral and cognitive abilities
Educational attainment				
Waited 7 min	0.36***	0.13	0.08	0.05
	(0.08)	(0.07)	(0.07)	(0.07)
Adjusted annual earnings				
Waited 7 min	0.24**	0.15	0.14	0.11
	(0.08)	(0.09)	(0.09)	(0.09)
Body mass index				
Waited 7 min	−0.35***	−0.24**	−0.16*	−0.17*
	(0.08)	(0.08)	(0.08)	(0.08)
Depression				
Waited 7 min	0.01	0.03	0.03	0.05
	(0.08)	(0.08)	(0.08)	(0.09)
Drug use				
Waited 7 min	0.11	0.06	0.06	0.04
	(0.08)	(0.08)	(0.08)	(0.08)
Risk-taking behaviors				
Waited 7 min	−0.13	0.04	0.02	0.01
	(0.08)	(0.08)	(0.08)	(0.08)
Impulse control				
Waited 7 min	0.09	0.07	0.05	0.06
	(0.08)	(0.08)	(0.08)	(0.08)
Police contact				
Waited 7 min	−0.07	0.03	0.05	0.06
	(0.08)	(0.08)	(0.08)	(0.08)
Other debt				
Waited 7 min	−0.02	0.04	0.05	0.06
	(0.03)	(0.04)	(0.04)	(0.04)

Note: Coefficients reflect the standardized effect of waiting the full time on the Marshmallow Test (i.e., 7 min) compared to all those who did not successfully wait (i.e., binary "pass/fail"). Robust standard errors are in parentheses. All dependent variables (except adjusted annual earnings and other debt) were standardized before being entered into the regression. Estimates shown in the first column contained only the measure of delay of gratification and a given outcome measure. Panel 1 added controls for child demographics and site indicator variables. Panel 2 added controls for child background characteristics and the quality of the early home environment. Panel 3 added behavioral and cognitive measures measured at age 54 months. $N = 702$.

*$p < .05$. **$p < .01$. ***$p < .001$.

the group that waited 2–7 min. However, with the inclusion of covariates, the effect on educational attainment ($\beta = .05$, $p = .46$) and annual earnings ($\beta = .10$, $p = .24$) became nonsignificant. Notably, waiting the full period on the Marshmallow Test significantly predicted lower BMI when using the binary predictor for delay of gratification, even with full covariates ($\beta = -.17$, $p = .03$).

Moderation by SES and sex

Finally, interactions between continuous performance on the Marshmallow Test and two childhood characteristics were evaluated: low SES during early childhood and sex. Low SES was defined as an average reported income-to-needs ratio of less than 2 over the course of early childhood (i.e., 1 to 54 months). Overall, we observed no clear pattern of moderation based on either low SES (Table S4) or sex (Table S5) when accounting for covariates. For boys, *more* time spent waiting on the Marshmallow Test was associated with an increase in risk-taking behaviors ($\beta = .15$, $p = .04$), but this effect did not hold when running sensitivity checks with robust regression (Table S8), indicating the effect was largely driven by outliers.

DISCUSSION

The idea that self-control in early childhood is a critical component to success across the life span has long permeated scientific literature. Indeed, numerous studies have identified early self-control as a robust predictor

of later health, wealth, and achievement (e.g., Koepp et al., 2023; Moffitt et al., 2011). Consequently, numerous interventions have targeted this skill in childhood in order to improve children's developmental outcomes. Recently, numerous studies have called into question whether a popular assessment of this ability (i.e., the Marshmallow Test) predicts these same health, wealth, and achievement outcomes as originally claimed in more diverse samples. In this preregistered analysis, we sought to extend Watts et al's (2018) analysis of the SECCYD sample by examining the association between Marshmallow Test performance in preschool with adult outcomes at age 26. Overall, few bivariate correlations were detected between delay of gratification and adult functioning, apart from educational attainment and BMI. In addition, these associations became statistically nonsignificant when adjusting for demographics, the early home environment, and concurrent cognitive/behavioral abilities. Furthermore, no consistent pattern of moderation by childhood SES or sex was observed. These findings stand in contrast to the robust longitudinal associations between early self-control and adult outcomes, raising skepticism over the long-term predictive and construct validity of the Marshmallow Test. Furthermore, these results suggest that an intervention narrowly targeting delay of gratification abilities are unlikely to produce long-term effects, unless subsequent changes are also made to other aspects of the child's environment and/or characteristics.

Body mass index and Marshmallow Test performance

Although we found largely null effects across most outcomes considered, the measure of adult BMI produced some evidence for associations with age-4 Marshmallow Test performance. Using the binary indicator of "pass/fail" (i.e., 7 min waited versus all other groups), we observed that children who waited the full 7 min had lower BMI scores at age 26 compared with all children who did not reach the 7-min ceiling. This result was consistent even in the fully controlled model, suggesting that the ability to delay gratification in early childhood may be a valuable predictor of a key indicator of later health. Previous studies with the SECCYD data set have also reported that higher self-control across childhood is associated with a lower BMI in adolescence (Datar & Chung, 2018; Tsukayama et al., 2010). Similar findings have been observed in a British cohort, where better self-control in middle childhood (Koepp et al., 2023) and adolescence (Koike et al., 2016) predicted a lower BMI across adulthood, above and beyond earlier cognitive abilities and behavioral problems. Notably, these studies utilized parent and teacher ratings to assess self-control—a measure that does not appear to predictively converge with Marshmallow Test performance (see contrasting results for most outcomes considered here in Koepp et al., 2023).

It is interesting that this finding was most pronounced in the model that included the binary coding (i.e., "pass/fail") for Marshmallow Test performance, and it could better capture children who have reached a milestone in their regulatory development. Previous work finds significant variation in the development of self-regulation during the preschool years (e.g., Montroy et al., 2016), suggesting children who successfully complete the task may fall within a class of "early" or "on-time" developers of self-regulatory capacity. However, we are unaware of any work showing that the 7-min mark on the Marshmallow Test represents a clear developmental threshold for children at age 54 months. Still, it is possible that the food-centric nature of the Marshmallow Test captures some kind of dietary restraint process that is present in early childhood (see Johnson et al., 2012 for discussion). Importantly, these findings should be considered under the context of the largely null results on other outcomes. We did not adjust for multiple comparisons in our analyses, raising the possibility of a Type I error given the large number of statistical tests. Also, the models using the binary measure for Marshmallow Test performance were not preregistered. Future research should examine this association using experimental designs to identify potential causal pathways between childhood delay of gratification and BMI in adulthood.

Predictive validity and construct clarity

Mischel's foundational work suggested that delay of gratification is a critical predictor of adjustment-related outcomes (Mischel et al., 1989), fueling interest in targeting the ability to delay gratification in early interventions (e.g., Murray et al., 2016; Rybanska et al., 2018). However, the present analysis found little evidence for the Marshmallow Test's predictive validity across a wide set of adult outcomes. This lack of association raises doubts regarding the possibility of a causal relation between delay of gratification ability and later functioning, as the few bivariate associations that were observed were largely explained by confounding characteristics. Similarly, Ahmed et al. (2019) found that delay of gratification did not predict adolescent EF independent of other early childhood EF's, including working memory, sustained attention, and impulsivity. This evidence suggests that improving performance on the Marshmallow Test in early childhood will not likely produce effects on adult health, wealth, or behavior, unless boosts in delay of gratification coincide with other changes to broader personal and environmental characteristics. Furthermore, the overall lack of bivariate relations also suggests that

using the Marshmallow Test as a type of "screener" for later adult difficulties may be ill-advised.

Our null findings stand in contrast to studies like Moffitt et al. (2011), which found that a measure of childhood self-control robustly predicted adult functioning, even when controlling for SES and IQ. Given the strength and consistency of Moffit's findings (also see Koepp et al., 2023), it seems possible that the Marshmallow Test captures a different construct than what is captured by these studies. Indeed, Moffitt et al. (2011) measured self-control using a composite of behavioral reports across early childhood (up to age 11), which likely reflect a latent, stable trait. Comparatively, the Marshmallow Test is a point-in-time estimate collected during a lab task in early childhood. Previous meta-analytic work has found that task-based measures of self-regulation (i.e., EF and delay tasks) may only weakly converge with parent and teacher reports of children's behavioral regulation (Duckworth & Kern, 2011). If teacher and parent reports of self-control produce totally distinct predictive relations compared with a direct assessment based on the Marshmallow Task, then the construct validity of the task itself may deserve reconsideration.

It should be noted that this study did detect some important bivariate associations between performance on the Marshmallow Test and measures of adult functioning (i.e., educational attainment and BMI), thus demonstrating that the simple waiting paradigm does have some predictive validity for adult outcomes. However, the fact that these associations were almost entirely explained by covariates for basic demographics and home life suggests that the predictive validity of the Marshmallow Test may have little to do with delay of gratification itself. Rather, the task appears to derive much of its predictive validity from its apparent associations with other important life factors in early childhood (e.g., SES, parenting, etc.). These findings converge with the conclusions of Watts et al. (2018), but they do raise questions about the construct validity of the task (Doebel et al., 2020; Falk et al., 2020). If the ability to delay gratification cannot be easily disentangled in predictive models from other early life factors, such as cognitive functioning and environmental advantage, it is fair to question whether the task truly measures the skill as advertised. In other words, the task may not provide valid inferences regarding the unique ability to delay gratification but may instead be a screener for broader developmental advantages in early childhood.

Early skill development and longitudinal outcomes

The null results reported here should also be understood in the broader context of fading effects that have been observed for many skill-focused early childhood interventions. As argued in the modeling section (see also Watts & Duncan, 2020), these longitudinal models with controls are designed to provide some indication of the likely effects of intervention efforts that might target a child's ability to delay of gratification in early childhood. Although longitudinal studies lacking exogenous variation are certainly limited in their capacity to provide forecasts for intervention (see Bailey et al., 2018), our results are not dissimilar to those reported in a recent meta-analysis of randomized control trials testing a broad set of educational interventions (Hart et al., 2023). Hart et al. (2023) found that most educational intervention impacts on measures of social–emotional skills, many of which captured constructs related to self-regulation, faded in the first few years following the end of the intervention. This fading pattern of intervention effects was also observed for cognitive skills. Thus, it could be the case that early advantages in the ability to delay gratification do not reliably impact skill levels at later periods, as children who lag in this capacity in early development may catch up in later periods.

Limitations

Several study limitations should be noted. Though the SECCYD sample contains more diversity than the sample of children recruited from the Bing Nursery School at Stanford University (see Benjamin et al., 2020), it is not a nationally representative sample (and is primarily White and middle-class). Additionally, the 7-min ceiling on the Marshmallow Test could bias any observed correlation due to the restriction of range (see Falk et al., 2020). However, our analytic adjustment using a categorical measure of minutes waited indicated that the measurement ceiling did not seriously affect the observed relations. Nevertheless, this adjustment is imperfect, and we simply cannot know how these relations would have differed had children been allowed to wait for 15–20 min. Finally, outcome measures were collected in early adulthood, and may only be rough indicators of later adult life (this is certainly the case of employment and finance-based measures).

CONCLUSION

In sum, this study extended the work of Watts et al. (2018) to examine whether the Marshmallow Test predicted key indicators of adult functioning. Consistent with Watts et al.'s analyses during middle adolescence, we found few bivariate associations between performance on the Marshmallow Test in early childhood and measures of adult success, and almost no associations when controls were included. These findings suggest that delay

of gratification as measured by the Marshmallow Test is not an early skill that predicts long-term trajectories. Intervention developers may find more success focusing on broader capacities to produce durable, longitudinal effects.

ACKNOWLEDGMENTS

The authors would like to thank Robert Siegler, Ana Whitaker, Emma Hart, Caroline Botvin, and Xinyu Pan for their feedback on iterations of this manuscript. The authors also wish to thank members of the Consortium on Early Childhood Intervention Impact for providing feedback on this work.

FUNDING INFORMATION

A cooperative agreement (5 U10 HD027040) between the study investigators that included Deborah Lowe Vandell and the Eunice Kennedy Shriver National Institute of Child Health and Human Development supported the design and data collection of the Study of Early Child Care and Youth Development (SECCYD) from birth through age 15 years. The age-26 study design, data collection, and analyses were supported by a grant from the Charles Stewart Mott Foundation (G-2017-00786) to Deborah Lowe Vandell.

DATA AVAILABILITY STATEMENT

This study was preregistered with Open Science Framework (osf.io/67XFN). Because we were not able to provide full access to the data used in the current study (the adult wave of the SECCYD has not been made publicly available), we have instead provided our analytic syntax. We have also provided a correlation matrix and descriptive statistics of all key variables. Earlier waves of data collection for the SECCYD sample are available at the Inter-University Consortium for Polictical and Social Research (icpsr.umich.edu/web/ICPSR/series/233).

ORCID

Jessica F. Sperber https://orcid.org/0000-0002-1636-5560
Tyler W. Watts https://orcid.org/0000-0002-2741-0873

REFERENCES

Ahmed, S. F., Kuhfeld, M., Watts, T. W., Davis-Kean, P. E., & Vandell, D. L. (2021). Preschool executive function and adult outcomes: A developmental cascade model. *Developmental Psychology*, *57*, 2234–2249. https://doi.org/10.1037/dev0001270

Ahmed, S. F., Tang, S., Waters, N. E., & Davis-Kean, P. (2019). Executive function and academic achievement: Longitudinal relations from early childhood to adolescence. *Journal of Educational Psychology*, *111*, 446–458. https://doi.org/10.1037/edu0000296

Algan, Y., Beasley, E., Côté, S., Park, J., Tremblay, R. E., & Vitaro, F. (2022). The impact of childhood social skills and self-control training on economic and noneconomic outcomes: Evidence from a randomized experiment using administrative data. *American Economic Review*, *112*, 2553–2579. https://doi.org/10.1257/aer.20200224

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, *73*, 81–94. https://doi.org/10.1037/amp0000146

Bailey, R., & Jones, S. M. (2019). An integrated model of regulation for applied settings. *Clinical Child and Family Psychology Review*, *22*, 2–23. https://doi.org/10.1007/s10567-019-00288-y

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 & Organization*, *179*, 743–756. https://doi.org/10.1016/j.jebo.2019.08.016

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*, 711–731. https://doi.org/10.1146/annurev-psych-010814-015221

Brotman, L. M., Dawson-McClure, S., Kamboukos, D., Huang, K.-Y., Calzada, E. J., Goldfeld, K., & Petkova, E. (2016). Effects of ParentCorps in prekindergarten on child mental health and academic performance: Follow-up of a randomized clinical trial through 8 years of age. *JAMA Pediatrics*, *170*, 1149–1155. https://doi.org/10.1001/jamapediatrics.2016.1891

CDC. (2020). *All about adult BMI*. Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Datar, A., & Chung, P. J. (2018). Childhood self-control and adolescent obesity: Evidence from longitudinal data on a national cohort. *Childhood Obesity*, *14*, 238–247. https://doi.org/10.1089/chi.2017.0217

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics*, *1*, 111–134. https://doi.org/10.1257/app.1.3.111

Doebel, S., Michaelson, L. E., & Munakata, Y. (2020). Good things come to those who wait: Delaying gratification likely does matter for later achievement (A commentary on Watts, Duncan, & Quan, 2018). *Psychological Science*, *31*, 97–99. https://doi.org/10.1177/0956797619839045

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*, 259–268. https://doi.org/10.1016/j.jrp.2011.02.004

Duckworth, A. L., Milkman, K. L., & Laibson, D. (2018). Beyond willpower: Strategies for reducing failures of self-control. *Psychological Science in the Public Interest*, *19*, 102–129. https://doi.org/10.1177/1529100618821893

Duckworth, A. L., Tsukayama, E., & Kirby, T. A. (2013). Is it really self-control? Examining the predictive power of the delay of gratification task. *Personality and Social Psychology Bulletin*, *39*, 843–855. https://doi.org/10.1177/0146167213482589

Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. *Child Development*, *65*, 296–318. https://doi.org/10.1111/j.1467-8624.1994.tb00752.x

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2015). *Early childhood education* (Working Paper 21766). National Bureau of Economic Research. https://doi.org/10.3386/w21766

Falk, A., Kosse, F., & Pinger, P. (2020). Re-revisiting the marshmallow test: A direct comparison of studies by Shoda, Mischel, and Peake (1990) and Watts, Duncan, and Quan (2018). *Psychological Science*, *31*, 100–104. https://doi.org/10.1177/0956797619861720

Gray-Lobe, G., Pathak, P. A., & Walters, C. R. (2021). *The long-term effects of universal preschool in Boston* (Working Paper 28756).

National Bureau of Economic Research. https://doi.org/10.3386/w28756

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., & Watts, T. W. (2023). Do intervention impacts on social-emotional skills persist at higher rates than impacts on cognitive skills? A meta-analysis of educational rcts with follow-up. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai23-782

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). Analyzing social experiments as implemented: A re-examination of the evidence from the HighScope Perry preschool program. *Quantitative Economics*, *1*, 1–46. https://doi.org/10.3982/QE8

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, *103*, 2052–2086. https://doi.org/10.1257/aer.103.6.2052

Inzlicht, M., Werner, K. M., Briskin, J. L., & Roberts, B. W. (2021). Integrating models of self-regulation. *Annual Review of Psychology*, *72*, 319–345. https://doi.org/10.1146/annurev-psych-061020-105721

Johnson, F., Pratt, M., & Wardle, J. (2012). Dietary restraint and self-regulation in eating behavior. *International Journal of Obesity*, *36*, 665–674. https://doi.org/10.1038/ijo.2011.156

Koepp, A. E., Watts, T. W., Gershoff, E. T., Ahmed, S. F., Davis-Kean, P., Duncan, G. J., Kuhfeld, M., & Vandell, D. L. (2023). Attention and behavior problems in childhood predict adult financial status, health, and criminal activity: A conceptual replication and extension of Moffitt et al. (2011) using cohorts from the United States and the United Kingdom. *Developmental Psychology*, *59*, 1389–1406.

Koike, S., Hardy, R., & Richards, M. (2016). Adolescent self-control behavior predicts body weight through the life course: A prospective birth cohort study. *International Journal of Obesity*, *40*, 71–76. https://doi.org/10.1038/ijo.2015.213

Masten, A. S., & Cicchetti, D. (2010). Developmental cascades. *Development and Psychopathology*, *22*, 491–495. https://doi.org/10.1017/S0954579410000222

Michaelson, L. E., & Munakata, Y. (2020). Same data set, different conclusions: Preschool delay of gratification predicts later behavioral outcomes in a preregistered study. *Psychological Science*, *31*, 193–201. https://doi.org/10.1177/0956797619896270

Mischel, W. (1974). Processes in delay of gratification. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. *7*, pp. 249–292). Academic Press. https://doi.org/10.1016/S0065-2601(08)60039-8

Mischel, W. (2014). *The marshmallow test: Mastering self-control*. Little, Brown and Co.

Mischel, W., Shoda, Y., & Peake, P. K. (1988). The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality and Social Psychology*, *54*, 687–696. https://doi.org/10.1037/0022-3514.54.4.687

Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science*, *244*, 933–938. https://doi.org/10.1126/science.2658056

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 2693–2698. https://doi.org/10.1073/pnas.1010076108

Montroy, J. J., Bowles, R. P., Skibbe, L. E., McClelland, M. M., & Morrison, F. J. (2016). The development of self-regulation across early childhood. *Developmental Psychology*, *52*, 1744–1762. https://doi.org/10.1037/dev0000159

Morris, P. A., Mattera, S. A., Castells, N., Bangser, M., Bierman, K. L., & Raver, C. C. (2014). *Impact findings from the Head Start CARES demonstration: National evaluation of three approaches to improving preschoolers' social and emotional competence (OPRE Report 2014–44)*. Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Morrison, F. J., & Grammer, J. K. (2016). Conceptual clutter and measurement mayhem: Proposals for cross-disciplinary integration in conceptualizing and measuring executive function. In *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (pp. 327–348). American Psychological Association. https://doi.org/10.1037/14797-015

Murray, J., Theakston, A., & Wells, A. (2016). Can the attention training technique turn one marshmallow into two? Improving children's ability to delay gratification. *Behaviour Research and Therapy*, *77*, 34–39. https://doi.org/10.1016/j.brat.2015.11.009

Nesbitt, K. T., & Farran, D. C. (2021). Effects of prekindergarten curricula: Tools of the mind as a case study. *Monographs of the Society for Research in Child Development*, *86*, 7–119. https://doi.org/10.1111/mono.12425

NICHD Early Child Care Research Network. (2002). Early child care and children's development prior to school entry: Results from the NICHD Study of Early Child Care. *American Educational Research Journal*, *39*, 133–164. https://doi.org/10.3102/00028312039001133

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401. https://doi.org/10.1177/014662167700100306

Raver, C. C. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *77*, 302–316. https://doi.org/10.1037/a0015302

Raver, C. C. (2004). Placing emotional self-regulation in sociocultural and socioeconomic contexts. *Child Development*, *75*(2), 346–353. https://doi.org/10.1111/j.1467-8624.2004.00676.x

Rybanska, V., McKay, R., Jong, J., & Whitehouse, H. (2018). Rituals improve children's ability to delay gratification. *Child Development*, *89*, 349–359. https://doi.org/10.1111/cdev.12762

Schlam, T. R., Wilson, N. L., Shoda, Y., Mischel, W., & Ayduk, O. (2013). Preschoolers' delay of gratification predicts their body mass 30 years later. *The Journal of Pediatrics*, *162*, 90–93. https://doi.org/10.1016/j.jpeds.2012.06.049

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology*, *26*, 978–986. https://doi.org/10.1037/0012-1649.26.6.978

StataCorp LLC. (2019). *Stata programming reference manual: Release 16*. Stata Press.

Tsukayama, E., Toomey, S. L., Faith, M. S., & Duckworth, A. L. (2010). Self-control as a protective factor against overweight status in the transition from childhood to adolescence. *Archives of Pediatrics & Adolescent Medicine*, *164*, 631–635. https://doi.org/10.1001/archpediatrics.2010.97

Vandell, D. L., Simpkins, S. D., & Liu, Y. (2021). From early care and education to adult problem behaviors: A prevention pathway through after-school organized activities. *Development and Psychopathology*, *33*, 658–669. https://doi.org/10.1017/S0954579420001376

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, *29*(7), 1159–1177. https://doi.org/10.1177/0956797618761661

Watts, T. W., & Duncan, G. J. (2020). Controlling, confounding, and construct clarity: Responding to criticisms of "Revisiting the Marshmallow Test" by Doebel, Michaelson, and Munakata (2020) and Falk, Kosse, and Pinger (2020). *Psychological Science*, *31*, 105–108. https://doi.org/10.1177/0956797619893606

Watts, T. W., Li, C., Pan, X. S., Gandhi, J., McCoy, D. C., & Raver, C. C. (2023). Impacts of the Chicago School Readiness Project on

measures of achievement, cognitive functioning, and behavioral regulation in late adolescence. *Developmental Psychology*, *59*, 2204–2222. https://doi.org/10.1037/dev0001561

Weinberger, D. A., & Schwartz, G. E. (1990). Distress and restraint as superordinate dimensions of self-reported adjustment: A typological perspective. *Journal of Personality*, *58*, 381–417. https://doi.org/10.1111/j.1467-6494.1990.tb00235.x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sperber, J. F., Vandell, D. L., Duncan, G. J., & Watts, T. W. (2024). Delay of gratification and adult outcomes: The Marshmallow Test does not reliably predict adult functioning. *Child Development*, *00*, 1–15. https://doi.org/10.1111/cdev.14129