# VI. Experimental Instructions and Related Information

## A.0 Recruitment language for Mechanical Turk studies

For all Mechanical Turk studies except for the one that eliminated any mention of politics (Study C2), the recruitment language was as follows:

Title: Answer questions about the economy under US presidents! Usually takes less than 4 minutes.

|  |  |
| --- | --- |
| **Description:** | Contribute to research! |
| **Keywords:** | survey, economy, quick, easy |
| **Qualification Requirement:** | Location is UNITED STATES (Required for preview)  HIT approval rate (%) greater than or equal to 95 (Required for preview) |

Answer questions about the economy under US presidents! Usually takes less than 4 minutes.

You can find the survey here:

Click here to take a survey.

At the end of the survey, you'll find a code. To get paid, please enter the code below:

For Study C2, the recruitment language was slightly modified in the sentence following the qualification requirement. The language for that study is below:

Title: Answer questions about the economy! Usually takes less than 4 minutes.

|  |  |
| --- | --- |
| **Description:** | Contribute to research! |
| **Keywords:** | survey, economy, quick, easy |
| **Qualification Requirement:** | Location is UNITED STATES (Required for preview)  HIT approval rate (%) greater than or equal to 95 (Required for preview) |

Answer questions about the economy! Usually takes less than 4 minutes.

You can find the survey here:

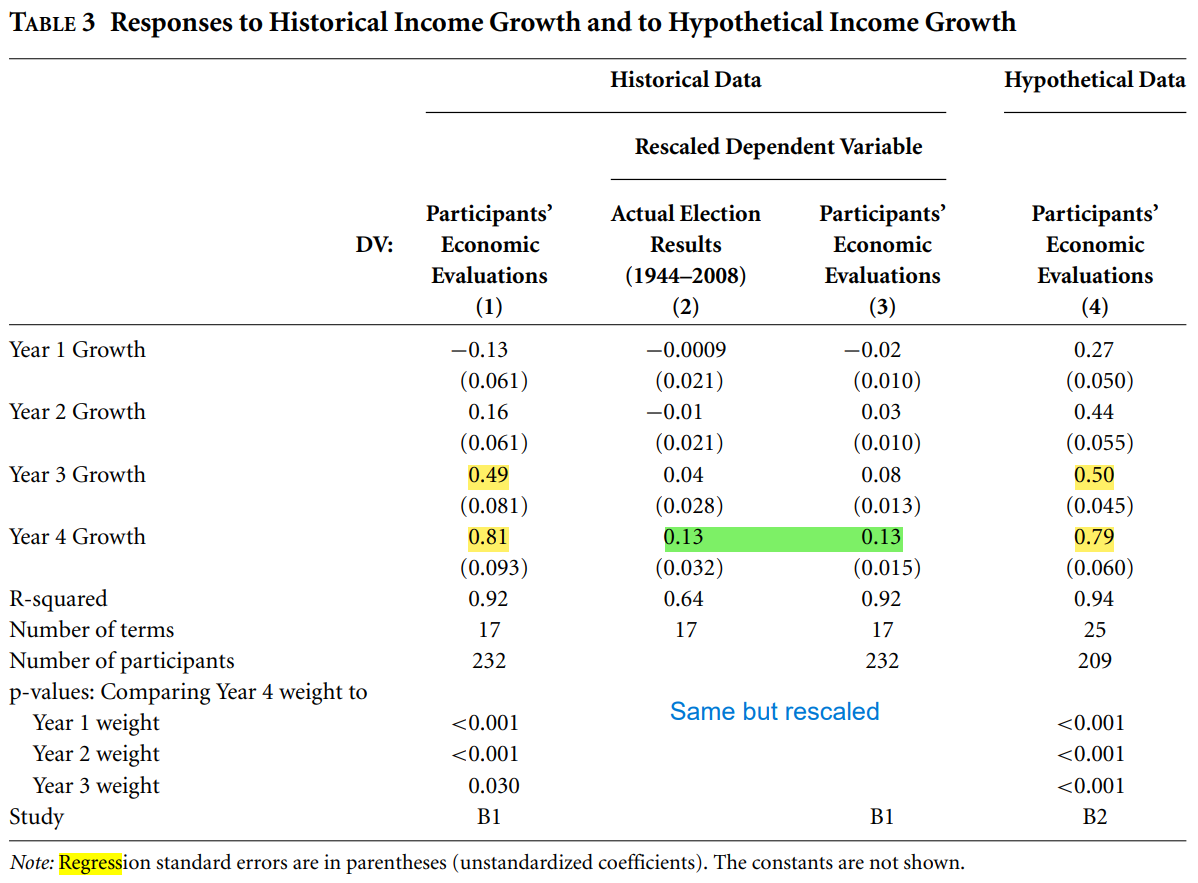
Click here to take a survey.

At the end of the survey, you'll find a code. To get paid, please enter the code below:

## A.1 Economic Evaluations and Regression Logic

Below each plot, we asked, “How would you rate the condition of the national economy during this period? Is it very good, fairly good, fairly bad, or very bad?” To put income growth and these evaluations on a similar scale, we recoded responses to vary from 0 to 10, with 10 corresponding to “very good.” We take the average of the responses for each of the 17 terms and call this variable Economic Evaluations.

In Table 3, we show that this pattern holds more generally. To do so, we estimate a regression similar to the retrospective voting model in Table 2, but with our participants’ ratings of the 17 historical economies as the dependent variable (instead of incumbent party vote margin). That is, to determine the influence that each year had on participants’ economic evaluations, we regress participants’ average ratings of these economies on the percentage change in income growth in each of the four years (so the N is 17, not 232).10 As shown in the first column, participants did indeed overweigh the last year, putting an estimated weight of 0.81 on Year 4 growth, compared to −0.13, 0.16, and 0.49, for Years 1, 2, and 3, respectively.



## A.3 WORDING 4 CONDITIONS

**Wording Yearly** **(Blind Control)**

Evaluate the economy under 17 hypothetical presidents.

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms.

The measurements shown are percent change in personal income growth, calculated on a yearly basis. They provide a good measure of the strength of the national economy.

After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?

Thank you so much for your time. Your responses are very important to us.

**Wording Yearly + Identified (Identified Control)**

Evaluate the economy under 17 presidents.

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms.

The measurements shown are percent change in personal income growth, calculated on a yearly basis. They provide a good measure of the strength of the national economy.

After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?

Thank you so much for your time. Your responses are very important to us.

**Wording Yearly Cumulative (Blind Treatment)**

Evaluate the economy under 17 hypothetical presidents.

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms.

The measurements shown are percent change in personal income growth, which provides a good measure of the strength of the national economy.

We show percent change in personal income growth in two ways: yearly and cumulatively. The cumulative growth figures simply add together the yearly income growth. For example, if personal income growth grew 2% in each year of a presidents four-year term, cumulative growth would be (4\*2% = 8%).

After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?

Thank you so much for your time, your responses are very important to us.

**Wording Yearly Cumulative Identified (Identified Treatment)**

Evaluate the economy under 17 presidents.

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms.

The measurements shown are percent change in personal income growth, which provides a good measure of the strength of the national economy.

We show percent change in personal income growth in two ways: yearly and cumulatively. The cumulative growth figures simply add together the yearly income growth. For example, if personal income growth grew 2% in each year of a presidents four-year term, cumulative growth would be (4\*2% = 8%).

After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?

Thank you so much for your time, your responses are very important to us.

## B. Description of task for Mechanical Turk studies

Description of the task in historical income growth study (B1):

Evaluate the economy under 17 hypothetical presidents who are serving in their second terms.  
  
On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' terms. The figures only show the economy for presidents in their second terms.  
  
After looking at each figure, evaluate the economy.  
  
Thank you so much for your time. Your responses are very important to us.

(To remove the hypothetical and second terms we use the replication studies without those mentions)

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms.   
  
The measurements shown are percent change in personal income growth, calculated on a yearly basis. They provide a good measure of the strength of the national economy.   
  
After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?    
  
Thank you so much for your time. Your responses are very important to us.

Description of the task in hypothetical income growth study (B2):

Evaluate the economy under 25 hypothetical presidents who are serving in their second terms.  
  
On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' terms. The figures only show the economy for presidents in their second terms.  
  
After looking at each figure, evaluate the economy.  
  
Thank you so much for your time. Your responses are very important to us.

Description of task when there was no mention of politics (Study C2):

Now, we would like you to evaluate the economy during four-year periods.

On the following page, you will see figures showing measurements of how much better off people were during 25 four-year periods. The measurements shown are percent change in personal income growth for each year. They provide a good measure of the strength of the national economy.

After looking at the figure, evaluate the economy during this period. Would you say it is very good, fairly good, fairly bad, or very bad?

Description of task for the percent change and cumulative percent change condition (Study D1):

*In the percent change condition (the control condition), the language was the same as on the one shown two pages later where presidents were labeled as being in their second terms.*

Evaluate the economy under hypothetical presidents who are serving in their second terms.

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms. The figures only show the economy for presidents in their second terms.

The measurements shown are percent change in personal income growth, which provides a good measure of the strength of the national economy.

We show percent change in personal income growth in two ways: yearly and cumulatively. The cumulative growth figures simply add together the yearly income growth. For example, if personal income growth grew 2% in each year of a presidents four-year term, cumulative growth would be (4\*2% = 8%).[[1]](#footnote-1)

After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?

Thank you so much for your time, your responses are very important to us.

Please note: As we describe earlier in the SI, we replicated this study with more instructions for yearly income growth and with minimal instructions (no explanation of yearly or cumulative growth). The replication study also asked about the president's handling of the economy, not “evaluating the economy during this period.”

Description of task for the percent change and level condition (Study D2):

*In the percent change condition (the control condition), the language was the same as on the one shown on the next page where presidents were labeled as being in their second terms.*

Evaluate the economy under hypothetical presidents who are serving in their second terms.

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms. The figures only show the economy for presidents in their second terms.

Each figure presents measurements of personal income, which provides a good measure of the strength of the national economy.  It presents the information on personal income in two ways.

First, it shows percent change in personal income growth, calculated on a yearly basis.

Next, it shows per capita personal income at the end of each year (after taxes and government transfers, and adjusting for inflation). During periods shown, yearly personal income ranges from about $31,000 to about $37,000.

After looking at each figure, evaluate the economy during this period.  Would you say it was very good, fairly good, fairly bad, or very bad?

Thank you so much for your time, your responses are very important to us.

Please note: As we describe earlier in the SI, we replicated this study with minimal instructions. The replication study also asked about the president's handling of the economy, not “evaluating the economy during this period.”

Description of the task with the president’s term unspecified (for replication studies):

On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms.   
  
The measurements shown are percent change in personal income growth, calculated on a yearly basis. They provide a good measure of the strength of the national economy.   
  
After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?    
  
Thank you so much for your time. Your responses are very important to us.

Wording was changed for studies about governors and about crime.

Description of task for studies where presidents were labeled as second term (for replication studies):

Evaluate the economy under hypothetical presidents who are serving in their second terms.  
  
On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' four-year terms. The figures only show the economy for presidents in their second terms.  
  
The measurements shown are percent change in personal income growth, calculated on a yearly basis. They provide a good measure of the strength of the national economy.   
  
After looking at each figure, evaluate the economy during this period. Would you say it was very good, fairly good, fairly bad, or very bad?    
  
Thank you so much for your time, your responses are very important to us.

## C. Addressing concerns about demand with the level experiment

As we describe in the text, there are several reasons we might be concerned that experimenter demand could lead people to weigh all years equally in either the cumulative income treatments or the income-in-levels treatments. It is important to note that most of these concerns would also operate in the yearly income growth conditions, where we found not that people put equal weights on all years, but instead a substantially greater weight on the last year.

One possible concern is that, since we tell participants that growth “provides a good measure of the strength of the national economy” and that they were asked to “evaluate the economy during this period,” we signaled them to put more weight on measures of the economy stretching over the whole period. (Again, however, we used the same language in the growth treatments where people put more weight on Year 4 than earlier years.)

To determine whether the instructions could be driving the equal weights that subjects assigned to the different years, we replicated Study D1, where we showed participants either yearly income growth or income in levels. We made two key changes. First, we eliminated the language about income growth providing a good measure of the national economy. Second, more importantly, we asked participants to evaluate, not the economy during the period, but the president's handling of the economy.

Here are the new instructions used for this study:

Evaluate the economy under 25 hypothetical presidents who are serving in their second terms.  
  
On the following pages, you will see figures with measurements of how much better off people are during hypothetical U.S. presidents' terms. The figures only show the economy for presidents in their second terms.  
  
After looking at each figure, evaluate the president's handling of the economy.  
  
Thank you so much for your time. Your responses are very important to us.

Here is the new question respondents answered below each four-year term, which we adapted from the ANES:

Do you approve or disapprove of the way the president handled the economy?

Respondents could answer: disapprove strongly, disapprove not strongly, approve not strongly, and approve strongly (the ANES response options)

231 respondents passed the attention screener and constitute our sample in this study.

These changes to the instructions in question wordings leave the main findings unchanged (see the table on the following page). When individuals saw yearly income growth (control condition), they placed a great deal more weight on the end. In contrast, when they saw a level growth (treatment condition), they did not.

Please note: As we describe earlier in the SI, we replicated this study with more instructions for yearly income growth and with minimal instructions (no explanation of yearly or cumulative growth). The replication study also asked about the president's handling of the economy, not “evaluating the economy during this period.”

## D. Mechanical Turk validation

### 1. Published validation studies

Mechanical Turk (MTurk) allows researchers to recruit and pay participants for participating in web-based studies. Although initially set up by Amazon.com to collect data for basic tasks such as recognizing products in pictures, it has expanded and people now post jobs of numerous sorts, including social science experiments. As we describe below, MTurk samples are more diverse than typical experimental samples, not that different on many demographic and political variables from nationally representative samples, and appear to be more attentive than other samples. Researchers have successfully replicated many experiments and published MTurk validation studies in several fields, including political science.

Berinsky, Huber, and Lenz (published in *Political Analysis*)*,* attempts to validate MTurk samples on external and internal validity grounds. To assess external validity, it compares MTurk samples to typical convenience samples used in experiments (other Internet panels, undergrad volunteers, recruits off the street) as well as nationally representative Internet surveys (e.g. Knowledge Networks) and face-to-face surveys (ANES and CPS). On demographics and political attitudes, we find that MTurk samples are often more representative than the convenience samples. Although obviously not as nationally representative as the ANES or CPS, MTurk samples are not that different from Knowledge Networks on these variables (the principal exception is age and variables related to age such as income). To further assess external validity, Berinsky, Huber, and Lenz also replicated three well-known experiments with MTurk samples. Finally, they examined the potential problem of "habitual customers"—participants who take part in numerous social science studies and so could become unrepresentative of citizens. Examining seven large MTurk samples conducted over five months, we found that only about 2% of participants fell in this category.

Berinsky, Huber, and Lenz also examines attentiveness, which is a concern not just with MTurk but with most experimental subject pools. Compared with other samples, however, MTurk participants appear to be more attentive, not less so. Since individuals requesting workers ("Requesters") often specify at least a 95% prior "approval rate" – that is, Requesters accepted 95% or more of the work previously submitted by an individual—MTurk workers have an incentive to read instructions carefully and consider their responses. As Berinsky, Huber, and Lenz report, our own experiences are consistent with this expectation. In one study, the authors asked participants to identify the political office held by a person mentioned in a story they had just read. The format of this question was multiple choice with five possible responses. On the MTurk study, 60% of the respondents answered the question correctly. An identical question concerning the same article was also included on experiments run through Polimetrix/YouGov and Survey Sampling International (SSI). The correct answer rates on these platforms were markedly lower than in the MTurk sample – 49% on Polimetrix/YouGov and 46% on SSI.

The attentiveness of MTurk participants is also evident in direct tests of attention. To ensure that participants were reading instructions, we included trick questions in all our studies. Here is a typical example:

We are interested in learning about your preferences on a variety of topics, including colors. To demonstrate that you've read this much,  
just go ahead and select both green and yellow among the alternatives below, no matter what your favorite color is. Yes, ignore the question below and select both of those options.  
  
What is your favorite color?

* pink
* red
* green
* white
* yellow
* blue

MTurk participants pass these tests at substantially higher rates than do participants in other samples. In an MTurk experiment just conducted by the authors, 93% passed the color test. On a harder attention test, which asked about which of many news media sources individuals read (but deep in the question told them to demonstrate their attention by selecting two in particular) 81% passed. When the same questions were included on a recent SSI survey by Berinsky, Margolis, and Sances, however, the proportion passing was more than 20 percentage points lower: only 69% passed the color test and only 59% passed the news test.

In addition to the Berinsky, Huber, and Lenz article, several articles now validate MTurk in other fields, reaching similar conclusions. In a paper published in *Perspectives on Psychological Science*, Buhrmester, Kwang, and Gosling ([2011](#_ENREF_4)) conclude:

(a) MTurk participants are slightly more representative of the U.S. population than are standard Internet samples and are significantly more diverse than typical American college samples;

(b) participation is affected by compensation rate and task length but participants can still be recruited rapidly and inexpensively;

(c) realistic compensation rates do not affect data quality; and

(d) the data obtained are at least as reliable as those obtained via traditional methods.

Largely the same conclusions are reached in a paper recently published in *Judgment and Decision Making* ([Paolacci, Chandler, and Ipeirotis 2010](#_ENREF_10)). Finally, the journal of *Experimental Economics* has published an evaluation of MTurk for economic experiments ([Horton, Rand, and Zeckhauser 2010](#_ENREF_7)) that also successfully replicates previous studies and reaches similar conclusions about the strengths of MTurk.

In part because four validation studies have now been published or are forthcoming, MTurk appears to be rapidly gaining acceptance. Quantifying the number of published social sciences articles using MTurk is hard, but Google scholar now lists 981 social science papers that refer to Mechanical Turk. Political science journals publishing with MTurk samples include *World Politics* ([Lawson et al. 2010](#_ENREF_8)) and *Political Psychology* ([Fausey and Matlock 2011](#_ENREF_5)). In psychology, where time to publication is quicker, *JPSP* and *Psych Science* (arguably the top two journals) have now published 15 article using MTurk (e.g.,[Alter, Oppenheimer, and Zemla 2010](#_ENREF_1); [Brady and Alvarez 2011](#_ENREF_3); [Gómez et al. 2011](#_ENREF_6)). The prestigious journal *PNAS* has just published two papers with experiments using MTurk samples ([Mason and Watts 2011](#_ENREF_9); [Rand, Arbesman, and Christakis 2011](#_ENREF_11)). .

### 2. Citations for MTurk validation studies and selected MTurk publications

Alter, Adam L., Daniel M. Oppenheimer, and Jeffrey C. Zemla. 2010. "Missing the Trees for the Forest: A Construal Level Account of the Illusion of Explanatory Depth." *Journal of Personality and Social Psychology* 99(3): 436-51.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.Com's Mechanical Turk." *Political Analysis*. 20(3): 351-68.

Brady, Timothy F., and George A. Alvarez. 2011. "Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items." *Psychological Science* 22(3): 384-92.

Buhrmester, Michael D., Tracy Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?" *Perspectives on Psychological Science* 6(1): 3-5.

Fausey, Caitlin M., and Teenie Matlock. 2011. "Can Grammar Win Elections?" *Political Psychology* 32(4): 563-74.

Gómez, Ángel, Matthew L. Brooks, Michael D. Buhrmester, Alexandra Vázquez, Jolanda Jetten, and William B. Swann, Jr. 2011. "On the Nature of Identity Fusion: Insights into the Construct and a New Measure." *Journal of Personality and Social Psychology* 100(5): 918-33.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2010. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14(3): 399-425.

Lawson, Chappell, Gabriel S. Lenz, Michael Myers, and Andy Baker. 2010. "Looking Like a Winner: Candidate Appearance and Electoral Success in New Democracies." *World Politics* 62(4): 561-93.

Mason, Winter, and Duncan J. Watts. 2011. "Collaborative Learning in Networks." *Proceedings of the National Academy of Sciences*.

Norton, Michael I., Lalin Anik, Lara B. Aknin, and Elizabeth W. Dunn. 2011. "Is Life Nasty, Brutish, and Short? Philosophies of Life and Well-Being." *Social Psychological and Personality Science* 2(6): 570-75.

Paolacci, G., J. Chandler, and P.G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5).

Rand, David G., Samuel Arbesman, and Nicholas A. Christakis. 2011. "Dynamic Social Networks Promote Cooperation in Experiments with Humans." *Proceedings of the National Academy of Sciences* 108(48): 19193-98.

### 3. Mechanical Turk demographics

We did not collect demographics for our studies on Mechanical Turk (MTurk), but have done so for other studies. As we noted above, Berinsky, Huber, and Lenz (forthcoming) compare MTurk demographics and standard political questions to the Current Population Study (CPS) and the American National Election Study (ANES). They summarize the results as follows:

All told, these comparisons reinforce the conclusion that the MTurk sample does not perfectly match the demographic and attitudinal characteristics of the U.S. population, but does not present a wildly distorted view of the U.S. population either. Statistically significant differences exist between the MTurk sample and the benchmark surveys, but these differences are substantively small. MTurk samples will often be more diverse than convenience samples and will always be more diverse than student samples ([Berinsky, Huber, and Lenz 2012](#_ENREF_13)).

### 4. Attention to task and screener questions

Since Mechanical Turk workers become increasingly ineligible for tasks when their work is rejected, they are incentivized and selected to be attentive, which is evident in their high rates of success on manipulation checks ([Berinsky, Huber, and Lenz 2011](#_ENREF_13)). Neither the students used in experimental work, nor respondents in Internet panels face similar incentives or selection processes.

To screen out Mechanical Turk workers who were not paying attention, we asked one of the following questions during the studies. About 10% of participants failed to pass these tests. We excluded these respondents before conducting the analysis (though the results are essentially unchanged if we include them).

**Favorite number test**

I understand that this survey will test my reading comprehension (really whether I'm reading at all). I understand that the questions will be a bit tricky. To demonstrate that you've read this much, just go ahead and select both one and two among the alternatives below, no matter what the question asks. Yes, ignore the question below and select both of those options.

Which number is the largest?

1.

2.

3.

**Favorite color test (already noted above)**

We are interested in learning about your preferences on a variety of topics, including colors. To demonstrate that you've read this much, just go ahead and select both green and yellow among the alternatives below, no matter what your favorite color is. Yes, ignore the question below and select both of those options.

What is your favorite color?

pink

red

green

white

yellow

blue

1. As we mentioned in the article, we calculate growth with natural logs, so that cumulative growth is the sum of yearly growth. Of course, we did not explain this technicality to participants. One might be concerned that mathematically sophisticated participants would have been confused on this point. In open-ended responses at the end of the study, however, no participants mentioned this point and almost all passed a test where they had to calculate cumulative growth at the end of the study. [↑](#footnote-ref-1)