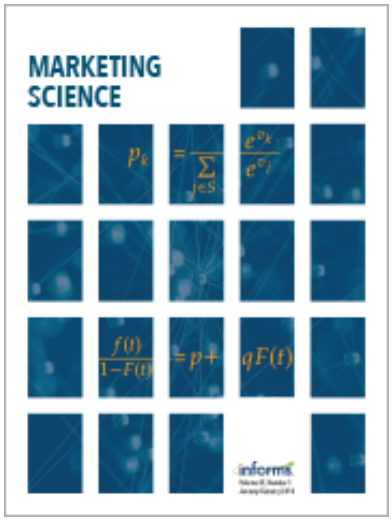


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Database Report: Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, Haozhe Chen

To cite this article:
Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, Haozhe Chen (2025) Database Report: Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions. Marketing Science 44(6):1446-1455. https://doi.org/10.1287/mksc.2025.0262

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 44, No. 6, November–December 2025, pp. 1446–1455
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Database Report: Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions

Olivier Toubia,[a,*] George Z. Gui,[a] Tianyi Peng,[b] Daniel J. Merlau,[a] Ang Li,[c] Haozhe Chen[c]

[a] Marketing Division, Columbia Business School, Columbia University, New York, New York 10027; [b] Decision, Risk & Operations Division, Columbia Business School, Columbia University, New York, New York 10027; [c] Department of Computer Science, Columbia University, New York, New York 10025

*Corresponding author

Contact: ot2107@gsb.columbia.edu, https://orcid.org/0000-0001-7493-9641 (OT); zg2467@gsb.columbia.edu, https://orcid.org/0000-0001-9399-649X (GZG); tianyi.peng@columbia.edu (TP); djm2199@columbia.edu (DJM); al4263@columbia.edu (AL); tonychenxyz@gmail.com (HC)

Received: May 29, 2025
Revised: June 18, 2025
Accepted: June 20, 2025
Published Online in Articles in Advance: August 20, 2025

https://doi.org/10.1287/mksc.2025.0262

Abstract. Large language model (LLM)-based digital twin simulation, where LLMs are used to emulate individual human behavior, holds great promise for research in business, artificial intelligence, social science, and digital experimentation. However, progress in this area has been hindered by the scarcity of real individual-level data sets that are both large and publicly available. To address this gap, we introduce a large-scale public data set designed to capture a rich and holistic view of individual human behavior. We survey a representative sample of $N = 2{,}058$ participants (average 2.42 hours per person) in the United States across four waves with more than 500 questions in total, covering a comprehensive battery of demographic, psychological, economic, personality, and cognitive measures, as well as replications of behavioral economics experiments and a pricing survey. The final wave repeats tasks from earlier waves to establish a test-retest accuracy baseline. Initial analyses suggest the data are of high quality and show promise for constructing digital twins that predict human behavior well at the individual and aggregate levels. Beyond LLM applications, due to its unique breadth and scale, the data set also enables broad social science and business research, including studies of cross-construct correlations and heterogeneous treatment effects.

History: Hema Yoganarasimhan served as the senior editor. This paper was accepted through the *Marketing Science* Database Submission review process.

Funding: This study was funded by the Columbia Business School Digital Future Initiative.

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mksc.2025.0262.

Keywords: generative AI • computational social science • digital twins • LLM-based persona simulation

1. Introduction

The rise of large language models (LLMs) like GPT has sparked interest across disciplines (including marketing, computer science, economics, psychology, and political science) in leveraging these tools to create "silicon samples" that may replicate how these humans would behave in response to any stimuli (Argyle et al. 2023, Brand et al. 2023, Dillion et al. 2023, Horton 2023, Park et al. 2023, Arora et al. 2024, Goli and Singh 2024, Qin et al. 2024, Li et al. 2025). If these LLM simulations can be a faithful substitute for eliciting responses from their human counterparts, the implications for both academics and practitioners are substantial. Academics could use silicon samples for pilot experiments to pinpoint stimuli with significant impact, thus improving the efficiency of theory development and experimental design. Firms could leverage these realistic simulations to explore different ideas and strategies, thereby improving customer insight and product development. Accordingly, in the recent past, we have witnessed a large influx of firms offering services leveraging silicon samples for customer insights (e.g., Synthetic Users, Outset AI, Nexxt, Voxpopme, Evidenza, Expected Parrot, Meaningful, xPolls, Ipsos, CivicSync).

Although silicon samples may be generated using only demographic information or hypothetical "life stories," a promising approach consists in creating silicon samples that are "digital twins" of real people. Notably, Park et al. (2024) use LLMs to create digital twins of more than 1,000 individuals based on transcripts from

qualitative interviews and find that the simulated agents replicated the human participants' responses on the General Social Survey 85% as accurately as participants replicate their own answers two weeks later.

Despite the promise and excitement surrounding digital twins, some uncertainty remains. For example, Brucks and Toubia (2025) show that the answers provided by LLMs may be overly influenced by the architecture of the prompt, such as the labeling or ordering of options in multiple choice questions. Gui and Toubia (2023) show that leveraging LLMs to simulate experiments may introduce unwanted confounding due to the difficulty of clearly instructing the LLM how to draw variables not specified in the prompt. Other research (Santurkar et al. 2023, Motoki et al. 2024, Li et al. 2025) suggests LLMs tend to express opinions that are not representative of the (human) population.

Given this background, it is crucial for the academic and practitioner community to validate digital twins in a transparent, reliable, and replicable manner. However, existing data sets present significant limitations that hinder their effectiveness for this purpose. Some data sets are publicly available (Alattar et al. 2018, Pew Research Center 2023, Santurkar et al. 2023), but they are not well suited for testing the validity of digital twins because they do not contain behavioral data (e.g., experiments) or a test-retest accuracy benchmark. Other data sets have this feature, but they are not publicly available (Park et al. 2024).

In sum, to the best of our knowledge, there is no publicly available data set that combines rich psychological profiles, behavioral data, and demographics from a large, representative sample for the development and testing of digital twin simulations. As a result, researchers often rely on synthetic or proprietary data, which undermines transparency, reliability, and replicability.

To address this gap, we assemble and publicly share an extensive data set from a representative sample of $N = 2{,}058$ people who each answered more than 500 questions covering a wide range of demographic questions, psychological scales, cognitive performance questions, economic preferences questions, as well as replications of a wide range of within- and between-subject experiments on heuristics and biases taken from the behavioral economics literature. The data were collected across four waves of studies lasting on average 2.42 hours per participant in total. Table 1 gives an overview of the measures collected in each wave, and Figure 1 illustrates our overall approach.

We use the responses to the heuristics and biases questions from waves 1–3 as holdout data and train the digital twins based on the rest of the data from waves 1–3. Wave 4 repeated the heuristics and biases experiments, providing us with a measure of test-retest accuracy. Future uses of the data may keep the same split or combine all the data from the four waves to create digital twins.

We report encouraging results regarding the quality of the data: Correlations between measures have good face validity, we replicate almost all known results from the behavioral economics literature, and the test-retest accuracy is robust.

We also report initial tests of the predictive validity of digital twins constructed using the data. At the individual level, we compute the accuracy of the digital twin predictions on holdout questions against the test-retest accuracy benchmark and a random benchmark. At the aggregate level, we test whether the digital twin simulations replicate the average treatment effects observed in human data. Throughout this process, we explore the type of behavior that can be predicted with higher versus lower accuracy by the digital twins to develop insight into the range of potential applications.

The data set and code are publicly available.[1] Given the unique scale and breadth of the data, there is also value in the raw results, irrespective of the application to digital twins. We report descriptive statistics and correlations between the dozens of measures we collected. We encourage others to explore heterogeneous treatment effects (Stanovich and West 2008, Dean and Ortoleva 2019).

2. Methods

We assembled a wide range of measures proposed in the business and social science literature over the past several decades. In addition to 14 demographic questions, we included 19 personality tests that measured 26 constructs over 279 questions, 11 cognitive ability tests (85 questions, 11 measures), and 10 economic preferences tests (34 questions, 10 measures). We also replicated 11 between-subject experiments (16 questions) and 5 within-subject experiments (32 questions) from the behavioral economics literature.[2] Finally, we administered the pricing study from Gui and Toubia (2023), which asks participants to make purchase decisions about 40 different products at randomly selected prices. In total, participants answered 500 questions across the first three waves. Wave 4 repeated all within- and between-subject heuristics and biases experiments from the first three waves, as well as the pricing study from wave 3 ($16 + 32 + 40 = 88$ questions in total). Participants were assigned to the exact same condition in wave 4 as they were in waves 1–3 for each of these experiments, providing us with a clean measure of test-retest accuracy. We programmed the studies on Qualtrics, doing our best to replicate the stimuli and measures from the original papers.[3]

We launched wave 1 on Prolific on January 29, 2025, targeting 2,500 representative U.S. respondents (sampled by age, sex, and ethnicity). Participants received $7 for completing wave 1.[4] They were informed that this was the first of four waves, and they would earn a

Table 1. Complete List of Questions and Related Measures

Task (source)	No. of questions (format)	Extracted measure(s)	Wave(s)
Demographics			
Demographics (Santurkar et al. 2023)	12 (multiple choice)	Region, sex, age, education, race, citizenship, marital status, religion, religious attendance, political party, household income, political ideology (categorical)	1
Additional demographics	2 (multiple choice)	Household size, employment status (categorical)	1
Personality traits			
Big 5 personality test (John and Srivastava 1999)	44 (five-point Likert)	Extraversion, agreeableness, conscientiousness, neuroticism, openness scores (numerical)	1
Need for cognition scale (Cacioppo et al. 1984)	18 (five-point Likert)	Need for cognition score (numerical)	1
Agentic vs. Communal Values scale (Trapnell and Paulhus 2012)	24 (9-point Likert)	Agency score, communion score (numerical)	1
Consumer Minimalism scale (Wilson and Bellezza 2022)	12 (five-point Likert)	Minimalism score (numerical)	1
Empathy scale (Carré et al. 2013)	20 (five-point Likert)	Basic empathy score (numerical)	1
Green values scale (Haws et al. 2014)	6 (five-point Likert)	Green score (numerical)	1
Social Desirability scale (Reynolds 1982)	13 (binary choice)	Social desirability score (numerical)	2
Conscientiousness scale (Johnson et al. 2019)	8 (nine-point Likert)	Conscientiousness score (wave 2) (numerical)	2
Anxiety scale (Beck et al. 1988)	21 (four-point Likert)	Anxiety score (numerical)	2
Individualism vs. Collectivism scale (Triandis and Gelfand 1998)	16 (five-point Likert)	Horizontal/vertical individualism, horizontal/vertical collectivism scores (numerical)	2
Selves questionnaire (Higgins et al. 1985)	3 (open-ended)	NA	2
Regulatory Focus scale (Fellner et al. 2007)	10 (seven-point Likert)	Regulatory focus score (numerical)	3
Tightwads vs. Spendthrift scale (Rick et al. 2008)	4 (multiple choice)	Tightwads vs. spendthrift score (numerical)	3
Depression scale (Beck et al. 1961)	22 (multiple choice)	Depression score (numerical)	3
Need for uniqueness scale (Ruvio et al. 2008)	12 (five-point Likert)	Need for uniqueness score (numerical)	3
Self-monitoring scale (Lennox and Wolfe 1984)	13 (six-point Likert)	Self-monitoring score (numerical)	3
Self-concept clarity scale (Campbell et al. 1996)	12 (five-point Likert)	Self-concept clarity score (numerical)	3
Need for closure scale (Roets and Van Hiel 2011)	15 (five-point Likert)	Need for closure score (numerical)	3
Maximization scale (Nenkov et al. 2008)	6 (five-point Likert)	Maximization score (numerical)	3
Cognitive abilities			
Cognitive Reflection Test (Krefeld-Schwalb et al. 2024)	4 (open-ended)	CRT score (numerical)	1
Fluid intelligence test (Krefeld-Schwalb et al. 2024)	6 (multiple choice)	Fluid intelligence score (numerical)	1
Crystallized intelligence test (Krefeld-Schwalb et al. 2024)	20 (multiple choice)	Crystallized intelligence score (numerical)	1
Syllogisms test (Markovits and Nantel 1989)	12 (multiple choice)	Syllogism score (numerical)	1
Overconfidence (Dean and Ortoleva 2019)	1 (numerical)	Overconfidence score (own predicted-actual score)	1
Overplacement (Dean and Ortoleva 2019)	1 (numerical)	Overplacement score (own predicted score-predicted average)	1
Financial literacy test (Johnson et al. 2019)	7 (multiple choice) + 1 (numerical)	Financial literacy score (numerical)	2
Numeracy test (Johnson et al. 2019)	8 (numerical)	Numeracy score (numerical)	2
Deductive certainty of Modus Ponens test (Stanovich and West 2008)	4 (binary choice)	Deductive certainty score	2
Forward Flow (free associations) (Gray et al. 2019)	20 (open-ended)	Forward flow score (average pairwise semantic distance)	2
Wason Selection Task (Klauer et al. 2007)	1 (multiple choice)	Wason Selection Task score (numerical)	3
Economic preferences			
Ultimatum game (sender) (Güth et al. 1982)	1 (multiple choice)	Ultimatum-send (percentage sent)	1
Ultimatum game (receiver) (Güth et al. 1982)	6 (binary choice)	Ultimatum-receive (acceptance probability)	1
Mental accounting (Thaler 1985)	4 (binary choice)	Mental accounting score (% choices consistent with mental account predictions)	1
Discount (Dean and Ortoleva 2019)	3 (multiple price list)	Discount rate (numerical)	2
Present bias (Dean and Ortoleva 2019)	3 (multiple price list)	Present bias (numerical)	2

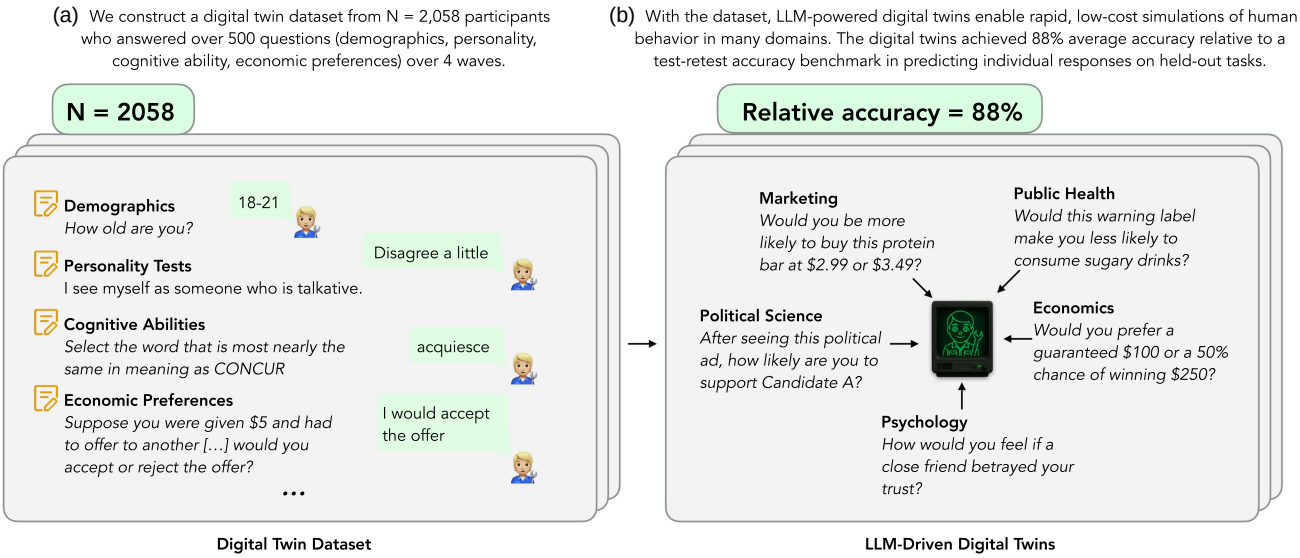
Table 1. (Continued)

Task (source)	No. of questions (format)	Extracted measure(s)	Wave(s)
Risk Aversion (Dean and Ortoleva 2019)	3 (uncertainty equivalence)	Risk aversion coefficient (numerical)	2
Loss Aversion (Dean and Ortoleva 2019)	4 (uncertainty equivalence)	Loss aversion coefficient (numerical)	2
Trust game (sender) (Dean and Ortoleva 2019)	1 (multiple choice)	Trust-send (percentage sent)	2
Trust game (receiver) (Dean and Ortoleva 2019)	5 (multiple choice)	Trust-return (average percentage returned)	2
Trust game (sender) thought listing	1 (open-ended)	NA	2
Trust game (receiver) thought listing	1 (open-ended)	NA	2
Dictator game (Baron and Hershey 1988)	1 (multiple choice)	Dictator-send (percentage sent)	3
Dictator game thought listing	1 (open-ended)	NA	3
Heuristics and biases (between subject)			
Base rate problem (Kahneman and Tversky 1973)	1 (slider scale)	Average prob. Assessment (numerical) in each condition (base rate of 30 vs. 70 engineers)	1, 4
Outcome bias (Baron and Hershey 1988)	1 (seven-point Likert)	Average correctness assessment (numerical) in each condition (success vs. failure)	1, 4
Sunk cost fallacy (Stanovich and West 2008)	1 (numerical)	Average number of purchases (numerical) in each condition (sunk cost yes vs. no)	1, 4
Allais problem (Stanovich and West 2008)	1 (binary choice)	Lottery choice probability in each condition (form 1 vs. 2)	1, 4
Framing problem (Tversky and Kahneman 1981)	1 (six-point Likert)	Average preference for B vs. A (numerical) in each condition (framing gain vs. loss)	2, 4
Conjunction problem (Linda) (Tversky and Kahneman 1983)	3 (six-point Likert)	Average prob. Assessment in each condition (feminist bank teller vs. bank teller)	2,4
Anchoring and adjustment (Tversky and Kahneman 1974, Epley et al. 2004)	2 (numerical)	Average prediction (numerical) in each condition (with high vs. low anchor)	2, 4
Absolute vs. relative savings (Stanovich and West 2008)	1 (binary choice)	Probability of driving to store in each condition (calculator vs. Jacket)	2,4
Myside bias (Stanovich and West 2008)	1 (six-point Likert)	Average ban agreement (numerical) in each condition (German vs. Ford)	2,4
Less is More (Stanovich and West 2008)	3 (five/six-point Likert)	Average attractiveness (numerical) in each condition (Form A vs. B vs. C)	3, 4
WTA/WTP – Thaler problem (Stanovich and West 2008)	1 (multiple choice)	Average in each condition (WTP-certainty, WTA-certainty, WTP-noncertainty)	3, 4
Heuristics and biases (within subject)			
False consensus (Furnas and LaPira 2024)	10 (five-point Likert) + 10 (slider)	Average predicted public support for each level of own support	1,4
Nonseparability of risk and benefits judgments (Stanovich and West 2008)	8 (seven-point Likert)	Correlation between benefits and risks for each item	1,4
Omission bias (Stanovich and West 2008)	1 (four-point Likert)	Likelihood of taking vaccine (numerical)	2,4
Probability matching vs. maximizing (Stanovich and West 2008)	6–10 (binary choice)	Proportion choosing each strategy (Match, Max, other)	3, 4
Dominator neglect (Stanovich and West 2008)	1 (binary choice)	Proportion choosing large tray	3,4
Product preferences			
Pricing study (Gui and Toubia 2023)	40 (binary choice)	Demand curve for each product	3,4

Note. NA, not available.

$10 bonus for completing all waves (with comprehension checks to ensure understanding). We received 2,509 complete responses. The following week (February 4, 2025), we invited these 2,509 participants to wave 2 ($7), receiving 2,263 complete responses. The next week (February 11, 2025), we invited them again for wave 3 ($7), receiving 2,252 complete responses. Waves 2 and 3 were closed the next week. On February 25, 2025, we invited the 2,154 participants who had completed waves 1–3 to wave 4 ($6; two-week delay because wave 4 repeated previous measures), receiving 2,058 complete responses and closing the wave after one week. Those completing all four waves received an additional $10 bonus, totaling $37.

Figure 1. Overview



These 2,058 participants who completed all four waves constitute our final sample. Among our final sample, the average response time was 43.88 minutes for wave 1 (standard deviation (SD) = 19.26), 45.31 minutes for wave 2 (SD = 19.24), 32.66 minutes for wave 3 (SD = 15.68), and 24.09 minutes for wave 4 (SD = 12.51). The average total time across all four waves was 145.47 minutes (SD = 56.10).

3. Data

Section 5 presents initial results of the performance of digital twins created with the data. Those results are subject to change as researchers explore optimal ways to create digital twins, and the current results may be viewed as providing a lower bound on performance. In the current section, we instead focus on exploring the *intrinsic* quality of the data.

Table A.1 in the Online Appendix reports demographic characteristics of our sample. Although our digital twins are created based on the raw responses, it is also informative and potentially useful to extract the measures corresponding to these questions (e.g., the extraversion score is measured by averaging eight questions from the Big Five battery of questions). Across waves 1–3, we collected 47 measures capturing personality traits, cognitive abilities, and economic preferences. Studying the correlations between these measures is of general interest to business and social science scholars and practitioners, above and beyond the question of digital twins. Online Appendix A.3 details the construction of these measures from the raw data, and Table A.2 reports summary statistics for the individual-level measures collected in the study. We compute a total of 1,326 pairwise correlations between the 47 measures listed in Table 1 and five demographic characteristics. We apply the Bonferroni correction and consider a correlation as significant if the p-value is below 0.05/1,326. This gives us 509 pairs of measures with significant correlations. We cannot report them all here, and instead report in Table A.3 in the Online Appendix 10 examples of correlations that are particularly high and/or noteworthy. These correlations all have good face validity, which suggests the data are of good quality, despite the large number of questions.

Next, we test whether our 16 heuristics and biases experiments replicate known results at the aggregate level (Table 3). We see that both in waves 1–3 and wave 4, all between-subject results replicate those in the literature, with the exception of the base rate fallacy. Although Kahneman and Tversky (1973) find that probability assessments are not sensitive to base rate, we find that they are. In terms of within-subject experiments, waves 1–3 and wave 4 also replicate all known results, with the exception of the nonseparability of risks and benefits for one of the items: bicycles. Although Stanovich and West (2008) find a negative correlation between judged benefits and risks for this item, we find no correlation. The fact that our data replicate the vast majority of these known experimental results is again a sign of good data quality.

Finally, we calculate the test-retest accuracy in our data. We use the answers from waves 1–3 to our 88 holdout questions (across 17 tasks) as the ground truth. Given that all holdout questions are either binary or numerical (or transformed into numerical answers), we calculate accuracy as follows. For binary measures, accuracy is simply a binary indicator of whether two answers match. For nonbinary measures, we calculate the absolute deviation between the ground truth and

predicted answer, divided by the range of possible answers.[5] We then compute accuracy as one minus this absolute deviation. This measure generalizes accuracy from binary to numerical questions: It ranges between zero and one, is equal to one when the prediction is equal to the ground truth, and is equal to zero when it is maximally different. When multiple questions are included in the same task, we take the mean accuracy across the questions within each task. Therefore, we are left with one measure of accuracy per respondent for each of the 17 tasks (11 between-subject experiments, 5 within-subject experiments, 1 pricing study). Figure 2 reports the average accuracy across respondents for each task and 95% confidence interval. We see that the average test-retest accuracy across the 17 tasks is 81.72%. This number is aligned with others reported in the literature (Park et al. 2024) and again gives us confidence in the data's quality.

4. Creation of the Digital Twins

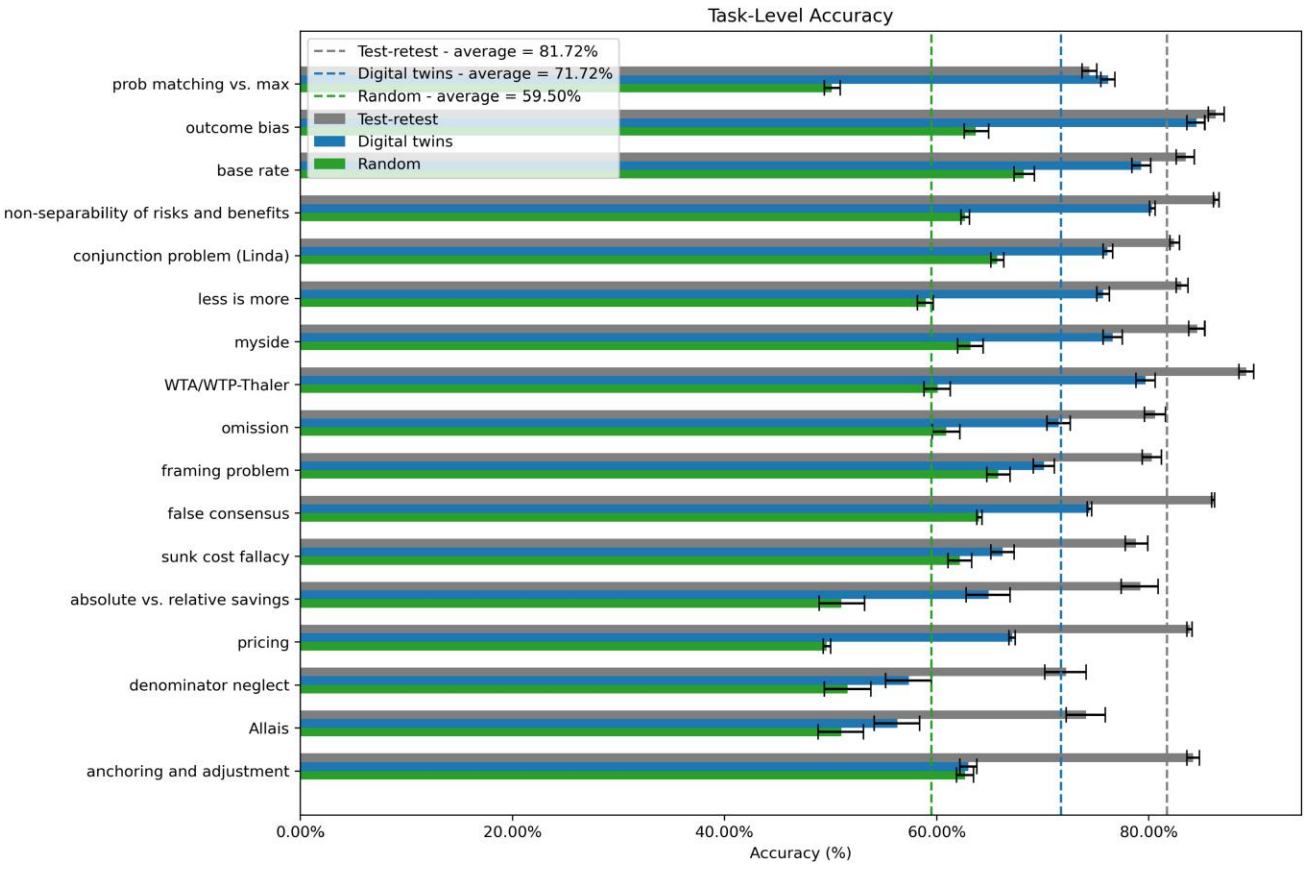
To construct each digital twin, we begin by merging the original Qualtrics survey files (QSFs) with each participant's raw responses, creating a self-contained JavaScript Object Notation (JSON) record for every individual. This record lists, in order, every question the participant actually encountered, the response options shown, and the answers. We then partition this record into three separate files:

- **Persona JSON:** Aggregates all non–hold-out content from waves 1–3, used to define the persona.
- **Evaluation answer-block JSON:** Contains the participant's wave 1–3 responses to hold-out items, providing the ground truth for evaluating simulation accuracy.
- **Retest answer-block JSON:** Stores wave 4 responses to those same hold-out items, used solely to compute the human test–retest accuracy benchmark.

By distributing the data in this modular format, we enable future researchers to experiment with alternative encoding or summarization schemes before presenting the material to an LLM. Future research may also explore alternative ways to split the questions into training and hold-out observations, for example, predicting the responses to cognitive ability questions based on the answers to the other questions in waves 1–3.

In the present work, we use a straightforward text-based approach: The JSON files are converted into text descriptions detailing the questions, options, and participant answers. The prompt instruction is attached in Online Appendix A2.1. The model's completion is

Figure 2. Predictive Accuracy



then postprocessed back into canonical survey coding and compared with the wave 1–3 ground truth, yielding the accuracy statistics reported in Figure 2. Our data set includes all these files for future researchers to train and test LLM-based digital twin simulations.

5. Initial Tests of Digital Twins' Predictive Performance

Section 3 presented evidence that suggests the quality of the data are good, despite the high number of questions. Hence, this data set should be helpful to researchers and practitioners interested in developing, testing, improving, and deploying digital twin simulations. In this section, we present initial tests of the performance of digital twins created using the data, both at the individual and aggregate levels. As mentioned, those initial results may be viewed as providing a lower bound on the predictive performance that may be achieved in the future.

5.1. Individual Level

To systematically evaluate different strategies for LLM-based persona simulation, we experimented with more than a dozen variations in both persona construction and simulation methodology. These include differences in input format (e.g., text versus JSON), model choices, prompting strategies such as proactive reasoning or chain-of-thought, and persona summaries. Full experimental details are provided in Online Appendix A.2.[6] Overall, we find that the predictive accuracy of the answers simulated by the digital twins falls within a similar range across approaches (Table 2). We hope this collection of baseline results will serve as a useful benchmark for future researchers exploring more advanced methods for persona training, such as reinforcement learning with human feedback (RLHF). For the initial analyses reported here, we focus on the text format using GPT4.1-mini.

Figure 2 reports, for each task, the predictive accuracy of the answers simulated by the digital twins, as well as the accuracy of a random benchmark that chooses each answer from a random uniform distribution. On average, across the 17 tasks, the accuracy of the digital twin predictions is 71.72%, and the ratio of the digital twin accuracy to the test-retest accuracy is 87.67%.

Table 2. Various Persona Simulation Approaches and Evaluation Results

Approach	LLM	Accuracy
Text persona	GPT4.1-mini	71.72%
Text persona	Gemini-flash2.5	69.40%
JSON persona	GPT4.1-mini	70.48%
JSON persona	GPT4.1	71.05%
Persona summary	GPT4.1-mini	70.13%
Persona summary + JSON persona	GPT4.1-mini	67.88%
Text persona (reasoning)	GPT4.1-mini	70.39%
Text persona (repeating questions)	GPT4.1-mini	70.45%
Text persona (default temperature)	GPT4.1-mini	71.24%
JSON persona (predicted output)	GPT4.1-mini	69.00%
JSON persona (predicted output)	GPT4.1	71.92%
Random guessing	–	59.17%

5.2. Aggregate Level

We test whether the data simulated from the digital twins replicate the average treatment effects from the 11 classic between-subject studies and the 5 classic within-subject studies included in our experiment. Table 3 shows that for 5 of the 10 between-subject results replicated by waves 1–3 and wave 4, the results from the digital twins also replicate the results. For anchoring and adjustment, the digital twins replicate the effect when asking participants to estimate the height of the highest redwood tree. But when asking participants to estimate the number of African countries in the United Nations, 98.8% of the twins gave the correct answer (54), and no anchoring effect was found. In contrast, only about 10% of humans gave the correct answer (8.79% in wave 2, 10.25% in wave 4). Four other between-subject effects were not replicated. In the outcome bias experiment, participants evaluate a physician's decision to operate on a patient. Humans evaluate the decision more favorably when the operation succeeded than when it failed, despite the risk being greater in the first condition. Overall, about 80% of humans gave a favorable evaluation (78.18% in wave 1, 81.39% in wave 4). In contrast, digital twins all gave a favorable rating ("correct" to "clearly correct"), with no significant different across conditions. In the sunk cost fallacy experiment, the effect was actually reversed with the digital twins versus their human counterparts, which we hope future research can explore. In the Allais problem experiment, which tests for violation of the independence axiom of utility theory, all digital twins chose the lower-risk lower reward option over the higher-risk higher reward one. Humans, on the other hand, were much more split in their decisions and showed systematic differences across conditions (which violates the independence axiom of utility theory). In the Less is More experiment, humans rated the option with no possibility of loss as less appealing than either of the options that contained the possibility of a loss, in all three questions. Twins did so in only one of the questions. Finally, the base rate fallacy, which was replicated neither in wave 2 nor in wave 4, was not replicated by the digital twins either.

Moving to within-subject experiments, we find that the digital twin results match the human results in two of the five within-subject experiments (Table 3). In the nonseparability of risk and benefit judgments study, the digital twins judgments display negative correlation as predicted, but the correlation is significant for only one of the items. For probability matching versus maximizing, the digital twins always selected the

Table 3. Replications of Heuristics and Biases

Task	Source	Prediction	Replicated		
			Waves 1–3	Wave 4	Twins
Between-subject experiments					
Base rate problem	Kahneman and Tversky (1973)	No difference in prob. Assessment when base rate = 30 vs. 70	✗	✗	✗
Outcome bias	Baron and Hershey (1988)	Average correctness assessment higher in success vs. failure condition	✓	✓	✗
Sunk cost fallacy	Stanovich and West (2008)	Average number of purchases higher in sunk cost vs. no sunk cost condition	✓	✓	✗
Allais problem	Stanovich and West (2008)	Violation of independence axiom of utility theory (different choices in Form 1 vs. 2)	✓	✓	✗
Framing problem	Tversky and Kahneman (1981)	Stronger preference for risky option under loss frame vs. gain frame	✓	✓	✓
Conjunction problem (Linda)	Tversky and Kahneman (1983)	Probability assessments higher for feminist bank teller vs. bank teller	✓	✓	✓
Anchoring and adjustment	Tversky and Kahneman (1974), Epley et al. (2004)	Average prediction higher with large vs. small anchor	✓✓	✓✓	✓✗
Absolute vs. relative savings	Stanovich and West (2008)	Probability of driving to store higher when discount is larger vs. smaller % of price	✓	✓	✓
Myside bias	Stanovich and West (2008)	Average agreement higher for ban of German car in United States vs. American car in Germany	✓	✓	✓
Less is More	Stanovich and West (2008)	Average attractiveness higher when possibility of loss vs. no possibility of loss	✓	✓	✗
WTA/WTP – Thaler problem	Stanovich and West (2008)	WTA-certainty>WTP-certainty>WTP-noncertainty	✓	✓	✓
Within-subject experiments					
False consensus	Furnas and LaPira (2024)	Overpredict (underpredict) public support if own support (oppose)	✓	✓	✓
Nonseparability of risk and benefits judgments	Stanovich and West (2008)	Negative correlation between benefits and risks for each item	✓✓✓✗	✓✓✓✗	✓✗✗✗
Omission bias	Stanovich and West (2008)	Significant proportion avoid treatment	✓	✓	✗
Probability matching vs. maximizing	Stanovich and West (2008)	significant proportion choose suboptimal strategy	✓	✓	✗
Dominator neglect	Stanovich and West (2008)	significant proportion choose nonnormative option	✓	✓	✓

normative option, whereas their human counterparts chose the normative option about 30% of the time only. For omission bias, participants were asked whether they would accept a vaccine that prevents catching a flu that has a 10% chance of killing affected patients when the vaccine itself carries a 5% chance of death. Although approximately 45% of the human participants (45.10% in wave 2, 44.80% in wave 4) refused the vaccine, only 4.0% of the twins refused the vaccine. This finding echoes our finding related to outcome bias where digital twins were much more favorable to medical professionals compared with their human counterparts.

Finally, we construct average demand curves from the pricing study. Figure A.1 in the Online Appendix shows the average demand curves from the responses from wave 3 versus wave 4 versus digital twins. We see that the average demand curves from wave 3 versus 4 are practically indistinguishable. We find that the average demand curve obtained from the twin is not fully downward sloping due to the twins' responses to free products. This echoes Gui and Toubia (2023), although

digital twins produce demand curves that are downward sloping for positive prices and that are generally closer to the ground truth compared with the demand curves obtained by Gui and Toubia (2023) without such input data.

In sum, although digital twins replicate about half of between-subject and within-subject effects, there are notable exceptions. Some occur when digital twins fail to mimic the suboptimal or nonnormative behaviors of humans or cannot "unlearn" certain facts (e.g., the number of African countries in the United Nations). In some areas, twins do match suboptimal human responses (e.g., absolute versus relative savings, dominator neglect), raising the broader question of whether digital twins should be seen as "improved" humans or as models that also replicate human deviations from normative behavior and knowledge gaps.

Other deviations appear in the medical domain (e.g., outcome bias, omission bias). Future research should examine whether digital twins are systematically more trusting of the medical profession than humans. Another factor may be that certain topics, such as vaccination, have become highly polarized. This may be considered in light of the result that GPT models tend to struggle to reflect conservative viewpoints (Motoki et al. 2024). For instance, in our false consensus task, although about 45% of humans somewhat or strongly supported increased deportations of those staying in the United States illegally (45.42% in wave 1, 45.09% in wave 4), only 25.85% of digital twins did so, and 74.1% strongly or somewhat *opposed* the measure. More research is needed to systematically study where digital twins diverge from humans, especially in medical and political domains, and to identify other domains where such differences may arise.

6. Conclusion

We present a unique data set spanning more than 500 questions and 2,000 respondents, with high data quality evidenced by sensible correlation patterns, good test-retest accuracy, and replication of known effects. Although this resource can broadly benefit business and social science scholars and practitioners, our primary focus is on using it to build digital twins. In initial tests, these twins predict human behavior with out-of-sample accuracy reaching 88% of the test-retest benchmark. Replication of average treatment effects is generally good, although further research is needed to determine if digital twins can capture nonnormative behaviors and reflect the full diversity of political and domain-specific views. We also hope that future research will explore the full range of potential applications of digital twins in marketing, business, and beyond. Examples include personalization, targeting, product development, professional development and training, advocacy and negotiations, mental health and counseling, and so on. The data set's focus on the United States is a potential limitation. Overall, we hope this resource accelerates LLM research, as well as business and social science applications, while being mindful of societal risks such as dehumanization of research and excessive reliance on AI in decision making.

Endnotes

[1] The data set is publicly available at https://huggingface.co/datasets/LLM-Digital-Twin/Twin-2K-500/, and the LLM simulation code can be found at https://github.com/tianyipeng-lab/Digital-Twin-Simulation.

[2] We included all experiments in Stanovich and West (2008, p. 672) who study "some of the most classic and well-studied biases in the heuristics and biases literature," as well as false consensus that allowed us to both capture participant's opinions on a range of issues and to test another well-known bias.

[3] We made minor adjustments to reflect cultural and societal changes (e.g., in the mental accounting scenarios from Thaler (1985), we replaced "Mr. A" and "Mr. B" with "Person A" and "Person B," and in the sunk cost experiment of Stanovich and West (2008), we replaced video rental stores with coffee shops).

[4] We pretested each wave to estimate response time and adjusted compensation accordingly.

[5] For the anchoring questions that accept unbounded answers, we transform the data into deciles based on the answers from wave 2 before calculating the absolute deviation.

[6] The Online Appendix also examines benchmarks that require some of the holdout data (split into a training and validation subsample), that is, fine-tuning and a traditional machine learning benchmark XG Boost.

References

Alattar L, Messel M, Rogofsky D (2018) An introduction to the understanding America study internet panel. *Soc. Security Bull.* 78(2):13–28.

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D (2023) Out of one, many: Using language models to simulate human samples. *Political Anal. (Oxford)* 31(3):337–351.

Arora N, Chakraborty I, Nishimura Y (2024) Express: AI-human hybrids for marketing research: Leveraging LLMs as collaborators. *J. Marketing* 89(2):43–70.

Baron J, Hershey JC (1988) Outcome bias in decision evaluation. *J. Personality Soc. Psych.* 54(4):569.

Beck AT, Epstein N, Brown G, Steer RA (1988) An inventory for measuring clinical anxiety: Psychometric properties. *J. Consulting Clinical Psych.* 56(6):893.

Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. *Arch. General Psychiatry* 4:561–571.

Brand J, Israeli A, Ngwe D (2023) Using LLMs for market research. Preprint, submitted March 30, https://dx.doi.org/10.2139/ssrn.4395751.

Brucks M, Toubia O (2025) Prompt architecture induces methodological artifacts in large language models. *PLoS One* 20(4):e0319159.

Cacioppo JT, Petty RE, Kao CF (1984) The efficient assessment of need for cognition. *J. Personality Assessment* 48(3):306–307.

Campbell JD, Trapnell PD, Heine SJ, Katz IM, Lavallee LF, Lehman DR (1996) Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *J. Personality Soc. Psych.* 70(1):141.

Carré A, Stefaniak N, d'Ambrosio F, Bensalah L, Besche-Richard C (2013) The basic empathy scale in adults (BES-A): Factor structure of a revised form. *Psych. Assessment* 25(3):679–691.

Dean M, Ortoleva P (2019) The empirical relationship between nonstandard economic behaviors. *Proc. Natl. Acad. Sci. USA* 116(33):16262–16267.

Dillion D, Tandon N, Gu Y, Gray K (2023) Can AI language models replace human participants? *Trends Cognitive Sci.* 27(7):597–600.

Epley N, Keysar B, Van Boven L, Gilovich T (2004) Perspective taking as egocentric anchoring and adjustment. *J. Personality Soc. Psych.* 87(3):327–339.

Fellner B, Holler M, Kirchler E, Schabmann A (2007) Regulatory focus scale (RFS): Development of a scale to record dispositional regulatory focus. *Swiss J. Psych.* 66(2):109–116.

Furnas AC, LaPira TM (2024) The people think what I think: False consensus and unelected elite misperception of public opinion. *Amer. J. Political Sci.* 68(3):958–971.

Goli A, Singh A (2024) Frontiers: Can large language models capture human preferences? *Marketing Sci.* 43(4):709–722.

Gray K, Anderson S, Chen EE, Kelly JM, Christian MS, Patrick J, Huang L, et al. (2019) "Forward flow": A new measure to quantify free thought and predict creativity. *Amer. Psychologist* 74(5):539–554.

Gui G, Toubia O (2023) The challenge of using LLMs to simulate human behavior: A causal inference perspective. Preprint, submitted December 24, https://arxiv.org/abs/2312.15524.

Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. *J. Econom. Behav. Organ.* 3(4): 367–388.

Haws KL, Winterich KP, Naylor RW (2014) Seeing the world through green-tinted glasses: Green consumption values and responses to environmentally friendly products. *J. Consumer Psych.* 24(3):336–354.

Higgins ET, Klein R, Strauman T (1985) Self-concept discrepancy theory: A psychological model for distinguishing among different aspects of depression and anxiety. *Soc. Cognition* 3(1):51–76.

Horton JJ (2023) Large language models as simulated economic agents: What can we learn from Homo silicus? Preprint, submitted January 18, https://arxiv.org/abs/2301.07543.

John OP, Srivastava S (1999) The big-five trait taxonomy: History, measurement, and theoretical perspectives. Pervin LA, John OA, eds. *Handbook of Personality: Theory and Research*, 2nd ed. (Guilford Press, New York), 102–138.

Johnson EJ, Meier S, Toubia O (2019) What's the catch? Suspicion of bank motives and sluggish refinancing. *Rev. Financial Stud.* 32(2):467–495.

Kahneman D, Tversky A (1973) On the psychology of prediction. *Psych. Rev.* 80(4):237.

Klauer KC, Stahl C, Erdfelder E (2007) The abstract selection task: New data and an almost comprehensive model. *J. Experiment. Psych. Learn. Memory Cognition* 33(4):680.

Krefeld-Schwalb A, Sugerman ER, Johnson EJ (2024) Exposing omitted moderators: Explaining why effect sizes differ in the social sciences. *Proc. Natl. Acad. Sci. USA* 121(12):e2306281121.

Lennox RD, Wolfe RN (1984) Revision of the self-monitoring scale. *J. Personality Soc. Psych.* 46(6):1349–1364.

Li A, Chen H, Namkoong H, Peng T (2025) LLM generated persona is a promise with a catch. Preprint, submitted March 18, https://arxiv.org/abs/2503.16527.

Markovits H, Nantel G (1989) The belief-bias effect in the production and evaluation of logical conclusions. *Memory Cognition* 17(1):11–17.

Motoki F, Pinho Neto V, Rodrigues V (2024) More human than human: Measuring ChatGPT political bias. *Public Choice* 198(1):3–23.

Nenkov GY, Morrin M, Ward A, Schwartz B, Hulland J (2008) A short form of the maximization scale: Factor structure, reliability and validity studies. *Judgment Decision Making* 3(5):371–388.

Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative agents: Interactive simulacra of human behavior. Preprint, submitted April 7, https://arxiv.org/abs/2304.03442.

Park JS, Zou CQ, Shaw A, Hill BM, Cai C, Morris MR, Willer R, et al. (2024) Generative agent simulations of 1,000 people. Preprint, submitted November 15, https://arxiv.org/abs/2411.10109.

Pew Research Center (2023) Pew Research Center survey data sets. Accessed May 16, 2024, https://www.pewresearch.org/download-datasets/.

Qin X, Huang M, Ding J (2024) Aiturk: Using ChatGPT for social science research. Preprint, submitted June 7, https://dx.doi.org/10.2139/ssrn.4922861.

Reynolds WM (1982) Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *J. Clinical Psych.* 38(1):119–125.

Rick SI, Cryder CE, Loewenstein G (2008) Tightwads and spendthrifts. *J. Consumer Res.* 34(6):767–782.

Roets A, Van Hiel A (2011) Item selection and validation of a brief, 15-item version of the need for closure scale. *Personality Individual Differences* 50(1):90–94.

Ruvio A, Shoham A, Brenčič MM (2008) Consumers' need for uniqueness: Short-form scale development and cross-cultural validation. *Internat. Marketing Rev.* 25(1):33–53.

Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T (2023) Whose opinions do language models reflect? Lawrence N, ed. *Proc. Internat. Conf. Machine Learn.* (PMLR, New York), 29971–30004.

Stanovich KE, West RF (2008) On the relative independence of thinking biases and cognitive ability. *J. Personality Soc. Psych.* 94(4):672.

Thaler R (1985) Mental accounting and consumer choice. *Marketing Sci.* 4(3):199–214.

Trapnell PD, Paulhus DL (2012) Agentic and communal values: Their scope and measurement. *J. Personality Assessment* 94(1):39–52.

Triandis HC, Gelfand MJ (1998) Converging measurement of horizontal and vertical individualism and collectivism. *J. Personality Soc. Psych.* 74(1):118–128.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185(4157):1124–1131.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.

Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psych. Rev.* 90(4):293–315.

Wilson AV, Bellezza S (2022) Consumer minimalism. *J. Consumer Res.* 48(5):796–816.