Twice-Told Preferences?

Repeated Elicitation, Response Coding,

and Future Household Wealth

Abstract

Does repeated elicitation improve measurement or mainly expose how it works? Using 56,050 intertemporal-choice elicitations from 6,808 Japanese households, I freeze 2010–2015 responses and predict 2022–2024 wealth for 3,099 households conditional on 2016 wealth. The displayed-rate mean is negatively associated with future wealth rank and split-block ORIV is steeper. But capping the 40% endpoint at 20% makes the negative slope imprecise; saturated first-response and repeated-share models reject displayed-rate linearity. Post-first responses add virtually no out-of-sample prediction beyond the first category. Repetition improves diagnosis—revealing persistence, endpoint use, and coding sensitivity—more than prediction. Estimates are predictive, not causal.

Keywords: household finance, intertemporal choice, repeated elicitation, response coding, survey measurement.

JEL Codes: D14, D15, D91, G11, G51.

1 Introduction

Repeated survey questions are commonly averaged or treated as multiple proxies on the premise that repetition reduces noise. In household-finance applications, that premise matters because preference and belief measures are used to explain wealth and portfolio choices. But repetition need not improve measurement when responses reflect focal categories, persistent response styles, or time-varying circumstances. This paper asks what repeated elicitation actually adds: stronger prediction of later balance sheets, or mainly diagnostics about persistence and coding?

I study an intertemporal-choice item observed as many as fifteen times in two national Japanese household panels. The reference person reports how much money thirteen months from now would be required to forgo 10,000 yen one month from now. Answers occupy eight displayed-rate categories from −5% to 40%. From 2010 through 2024, the panels provide 56,050 valid responses from 6,808 households, with nine responses at the median, linked to deposits, securities, and demographic information.

The design has two parts. I first treat the answer as the ordered response that the questionnaire records. Full-sample transition matrices and focal-category diagnostics describe persistence without imposing cardinal distances. I then compare one response, the household's early mean, alternative order-preserving summaries, and the household's early category distribution. For a repeated-proxy sensitivity, I split 2010–2015 answers into nonoverlapping chronological blocks and instrument either block with the other. Under independent block-error and exclusion restrictions, the split-block estimator can be read as a correction for transitory measurement error on the displayed-rate scale. I

treat it as a sensitivity because persistent response style or financial circumstances can load into both blocks.

The outcome design is prospective. Responses from 2010–2015 form the proxies, 2016 supplies baseline financial position and controls, and outcomes are averaged over 2022–2024. No baseline or future response enters the early score. The design therefore removes the look-ahead embedded in a full-panel household mean and asks whether early responses are prospectively associated with financial position six to eight years after baseline. This timing is a design feature rather than a preregistered holdout.

The headline is that repetition changes the diagnosis more than the prediction. On the displayed-rate scale, a ten-percentage-point higher 2010–2015 mean is associated with a 0.64-percentile lower 2022–2024 wealth rank conditional on 2016 rank; split-block ORIV implies 0.94 percentile and a single response 0.52. ORIV is detectably steeper than the mean, while the mean and ORIV contrasts with the first-response slope are imprecise. But recoding the 40% endpoint as 20% shrinks the negative slope to −2.912 (standard error 4.593), and deleting 40% responses flips it positive on the households with a remaining answer. The mapped scalar is therefore an endpoint-driven linear projection, not an economically invariant marginal effect.

The formal coding tests confirm that diagnosis. Saturated category models reject the mapped-rate linear form using either the first response, $F(6, 3084) = 3.425$ ($p = 0.0023$), or repeated response shares, $F(6, 3084) = 3.796$ ($p = 0.0009$). An order-restricted test rejects weak monotonicity of the full share-coefficient profile ($p = 0.0264$), with the reversal concentrated at the 40% endpoint.

Post-first shares do not add detectably beyond first-response dummies, and repeated cross-validation gives virtually identical out-of-sample gains for the two representations. Cross-panel transported point estimates have the same orientation, but full resampling intervals include zero. Repetition therefore exposes endpoint use, coding sensitivity, and proxy-block behavior, but it does not yield a clearly superior predictor relative to a flexible first-response category.

The outcome decomposition supplies a further boundary. Securities entry is null, retention is suggestive but imprecise, and attrition corrections leave the rank estimate nearly unchanged; none of these analyses rules out stable confounding by planning ability, cognition, earning capacity, financial literacy, financial circumstances, or response style. The estimates are predictive associations.

The contribution is to estimate the empirical return to repetition in a setting where the same household-finance measure is observed for up to fifteen years. Prior work shows both that intertemporal-choice measures covary with wealth (Barsky et al. 1997; Epper et al. 2020; Kureishi et al. 2021; Laibson et al. 2026) and that multiple proxies can correct attenuation in noisy behavioral measures (Gillen et al. 2019; Beauchamp et al. 2017; Goda et al. 2019; Stango and Zinman 2023; Dohmen and Jagelka 2024). The long panel lets me hold the item and outcome design fixed while comparing a flexible first-response category, an early average, a response distribution, and split-block IV. The result is a boundary condition for the usual signal-extraction argument: repetition makes persistence and coding assumptions testable, but adds no detectable predictive content beyond the first category in this application.

Related work documents temporal instability, persistent preference shocks, and person-specific

response noise (Meier and Sprenger 2015; Chuang and Schechter 2015; Salamanca 2018; Arts et al. 2026; Belzil and Jagelka 2025). Money-based intertemporal responses can also combine time preference with utility curvature, liquidity, outside returns, attention, and response conventions (Cubitt and Read 2007; Dean and Sautmann 2021; Fulford and Shupe 2026; Belot et al. 2025); household-finance choices reflect rich preference heterogeneity and portfolio frictions (Calvet et al. 2026; Choukhmane and de Silva 2026). These results therefore do not recover a structural discount parameter; I use the neutral term *elicited required return*. Their broader implication is methodological: with coarse repeated survey measures, averaging and multiple-proxy correction should not be assumed to dominate a flexible first-response encoding. Sections 2–3 develop the data and measurement design, Section 4 reports the prospective wealth results, Section 5 evaluates robustness and interpretation, and Section 6 concludes.

2 Data, Elicitation, and Outcomes

2.1 The two household panels

The data come from the Japan Household Panel Survey (JHPS) and the Keio Household Panel Survey (KHPS), administered by the Panel Data Research Center at Keio University. KHPS began in 2004 and JHPS in 2009; common questionnaires have been administered jointly since 2014. Each follows a national, population-based sample of Japanese adults and their households and records household deposits, securities, income, expenditure, employment, and demographic characteristics.

The samples are distinct. I prefix household identifiers by survey, never link a respondent across surveys, and include survey indicators in pooled specifications.

The unit of observation in the underlying panel is a household-year. Financial variables refer to the household; the intertemporal-choice item is answered by the reference person. The reference person is female in 51.1% of observations and has mean age 56.6. Sex and birth-year checks indicate that the respondent is stable within nearly every household; two households show an apparent change and remain in the descriptive counts.

The analysis begins in 2010, the first year of the repeated item in JHPS. KHPS contributes the same item from 2014. Both run through 2024. The pooled measurement sample contains 56,050 valid elicitation-year observations from 6,808 households: 29,729 observations from 3,463 JHPS households and 26,321 from 3,345 KHPS households. A household contributes 8.23 answers on average and nine at the median; the interquartile range is four to eleven and the maximum is fifteen. Thus the identifying asset is not merely a large cross-section, but repeated observation of the same item and respondent.

2.2 The elicitation

The questionnaire asks:

> Instead of receiving 10,000 yen one month from now, at least how much would you require to receive thirteen months from now?

Respondents choose one of eight amounts. The questionnaire displays the associated annual

interest rates, which are −5, 0, 2, 4, 6, 10, 20, and 40 percent. A higher category therefore records a higher required return to wait. Both payment dates are in the future, so the item is not an immediate-versus-delayed present-bias task. The choice is hypothetical, coarse, and stated in money. These features make it a useful survey proxy but not a direct observation of a structural discount factor.

Panel A of Table 1 shows why treating the item as ordered is important. The categories are not equally spaced, and responses cluster at 0, 10, 20, and 40 percent. The two most popular answers alone account for more than half of observations. Assigning the displayed rates is transparent and useful for economic magnitudes, but it imposes distances that the questionnaire labels alone do not identify. I therefore pair displayed-rate block estimates with full-sample category transitions, four scalar category encodings, and a flexible category-composition test.

The answer can reflect several economic objects. In addition to willingness to delay consumption, it may incorporate utility curvature, borrowing or saving opportunities, attention, misunderstanding, and a stable convention for using the response scale (Andersen et al. 2008; Andreoni and Sprenger 2012; Cohen et al. 2020; Dean and Sautmann 2021; Fulford and Shupe 2026). I use the neutral term *elicited required return*. The empirical objective is to characterize repeated responding and to study predictive content under explicit codings; whether an answer reflects pure patience is left open.

Table 1: Sample, elicitation, and financial outcomes

Panel A. Response categories								
Displayed annual rate	−5%	0%	2%	4%	6%	10%	20%	40%
Share of responses (%)	0.3	15.3	4.7	4.2	6.2	27.8	16.8	24.6

Panel B. Repetition and early-to-late sample	
Valid elicitation-year observations	56,050
Households with a valid elicitation	6,808
Elicitations per household: mean / median / IQR / maximum	8.23 / 9 / 4–11 / 15
At least two valid responses in 2010–2015	6,096
Early score and a 2016 household record	4,831
Primary later-wealth-rank sample	3,099

Panel C. Selected variables in the valid-elicitation sample				
Variable	*N*	Mean	SD	Median
Elicited required return	56,050	0.166	0.147	0.100
Securities participation	52,954	0.247	0.431	0
Financial wealth (10,000 yen)	54,296	1,123.1	1,818.1	450.0
log(1 + financial wealth)	54,296	4.950	3.041	6.111
Age	56,050	56.57	14.45	57

Panel A uses all valid JHPS and KHPS responses from 2010–2024. Panel B reports household counts; the early-to-late design uses 2010–2015 responses, a 2016 baseline, and 2022–2024 outcomes. Panel C uses available observations within the valid-elicitation sample for each variable. Financial wealth is deposits plus securities, not net worth, in units of 10,000 yen. Deposits and securities are each winsorized at their 99th percentiles. When exactly one component is observed, the primary construction treats the missing component as zero; observations with both components missing remain missing. Securities include stocks, bonds, and investment trusts. Monetary nonresponse codes are set to missing before transformation.

2.3 Financial outcomes

Financial wealth is deposits plus securities. It excludes housing, other real assets, pensions, and debt, so throughout the paper I say *financial wealth*, not net worth or total wealth. Monetary amounts are expressed in units of 10,000 yen. After survey nonresponse codes are set to missing, deposits and securities are each winsorized at their panel-wide 99th percentiles. The primary construction treats a single missing component as zero when the other is observed and leaves wealth missing when both

components are missing; the Internet Appendix requires both components in a robustness check. Because 24.1% of observed household-years report zero financial wealth, no single transformation is sufficient. I use four complementary outcomes: the household's within-survey-year wealth rank; a Poisson pseudo-maximum-likelihood model for the level of W; a two-part decomposition into positive wealth and log wealth conditional on positive wealth; and $\log(1 + W)$ as a unit-sensitive sensitivity specification.

For survey-year s, t with n_{st} finite wealth observations, the rank is $100\{\text{midrank}(W_{ist}) - 0.5\}/n_{st}$. Ties, including the mass at zero, receive their average rank. Ranks use the full contemporaneous estimation-panel survey-year cross-section and are computed before prospective-sample selection; separately distributed refresh-cohort files are not part of that rank universe. This definition makes the outcome relative to the observed population in each survey-year; the two-part outcomes separately expose the zero mass and the positive-wealth amount.

Securities participation equals one when the household reports positive stocks, bonds, or investment trusts. The prospective analysis averages this indicator across available outcome waves, so it can be read as the share of 2022–2024 waves in which a household participates. I separately examine entry among 2016 nonparticipants and retention among 2016 participants. This baseline-status split prevents a pooled participation coefficient from hiding offsetting transitions. Conditional portfolio shares are secondary because the survey categories do not identify equity exposure precisely enough to support a broad "risky asset" label.

Wealth rank is calculated within survey and calendar year using all households with observed

financial wealth before imposing the prospective sample. This outcome is robust to currency units, skewness, and common nominal changes. PPML retains zeros and is invariant to the denomination of yen; because the conditional variance is large, I use it as a scale check. The log-one-plus specification combines the extensive and intensive margins and is not invariant to whether the same wealth is recorded in yen or units of 10,000 yen. I therefore do not interpret its coefficient as a denomination-free percentage effect.

2.4 Temporally separated sample

For each household I freeze the response information set at 2015 by averaging valid responses in 2010–2015, requiring at least two. I record the chronological odd and even responses as nonoverlapping proxy blocks. Baseline outcomes and controls are measured in 2016. The outcome is the household's mean over every observed 2022–2024 wave. Thus the score is temporally separated from baseline and outcome by construction. The long gap reduces the chance that a single transitory survey condition drives both the response and the outcome, while averaging the outcome reduces year-specific balance-sheet noise.

JHPS and KHPS contribute different amounts of pre-period information. An eligible JHPS household has 5.89 early responses on average; an eligible KHPS household has exactly two, because the common item enters KHPS in 2014. Their odd/even correlations are 0.70 and 0.41, respectively. Pooled, the early-score standard deviation is 11.9 percentage points and the odd/even correlation is 0.50. All pooled models therefore include survey indicators, control for the number of early

responses, and are repeated by survey. The cross-survey comparison is substantively useful: it asks whether the finding depends on the unusually long JHPS measurement block.

Of 10,101 households appearing anywhere in the constructed panel, 6,758 provide at least one early response and 6,096 provide at least two. Separately, 4,993 households have a 2016 record. Intersecting the repeated-response and baseline-record samples leaves 4,831 households before complete-case baseline checks; 4,716 satisfy the wealth observation model's pre-outcome requirements, and later wealth rank is observed for 3,099. Section 5 models this observation probability using only information available by 2016 and reports an inverse-probability-weighted robustness estimate.

3 Repeated-Response Measurement

3.1 Full-sample ordered-response diagnostics

Let $Y_{it} \in \{1, \ldots, 8\}$ denote respondent i's recorded option in year t. I first characterize persistence without assigning cardinal distances to these labels. For every adjacent pair of valid annual responses, I calculate the Spearman rank correlation, the probability of repeating the identical option, and the row-normalized transition matrix

$$T_{jk} = \Pr(Y_{it} = k \mid Y_{i,t-1} = j).$$

Table 2: Full-sample ordered-response persistence

Sample	Respondents	Pairs	ρ_S	Exact repeat	40% stay	Focal stay
Main panel	6,808	48,389	0.513	0.444	0.578	0.893
Continuous JHPS, 2019-24	969	4,845	0.655	0.521	0.648	0.907
Refresh B (2019 entry)	2,175	6,108	0.450	0.415	0.589	0.896
Refresh C (2023 entry)	1,686	1,161	0.427	0.394	0.596	0.899

Entries use observed adjacent annual response pairs and impose no cardinal spacing between categories. ρ_S is the Spearman correlation; exact repeat is the unconditional share selecting the same option in both years. Endpoint and focal statistics are conditional on the previous response. The low-endpoint cell is suppressed because it has insufficient support under the prespecified disclosure rule. Refresh cohorts are not part of the main panel.

This exercise uses all 56,050 responses from 6,808 households. It is descriptive: repeated categories can reflect a persistent economic signal, recall, focal response, or a stable convention for using the questionnaire. No transition statistic labels the remaining movement as measurement error.

Table 2 documents substantial but incomplete repetition. In the main panel, adjacent-category Spearman correlation is 0.513 and exact repetition is 0.444. Conditional on selecting one of the four salient values (0, 10, 20, or 40 percent), the next response remains in that set with probability 0.893. Two further patterns in the table matter later. The 40 percent endpoint retains 0.578 of its occupants from one year to the next, and exact repetition rises with panel tenure at similar calendar dates: 0.394 in the 2023 refresh cohort, 0.415 in the 2019 cohort, and 0.521 among JHPS respondents observed continuously over 2019–2024. Section 5 returns to both patterns, the first for salient responding and the second for panel conditioning. These quantities establish persistence of questionnaire use, not a monotone relationship between category location and later wealth.

Figure 1 displays the observed row-normalized transitions for the main panel and the 2019 JHPS refresh cohort. The diagonal is visible in both samples, but movement is not simply local: many

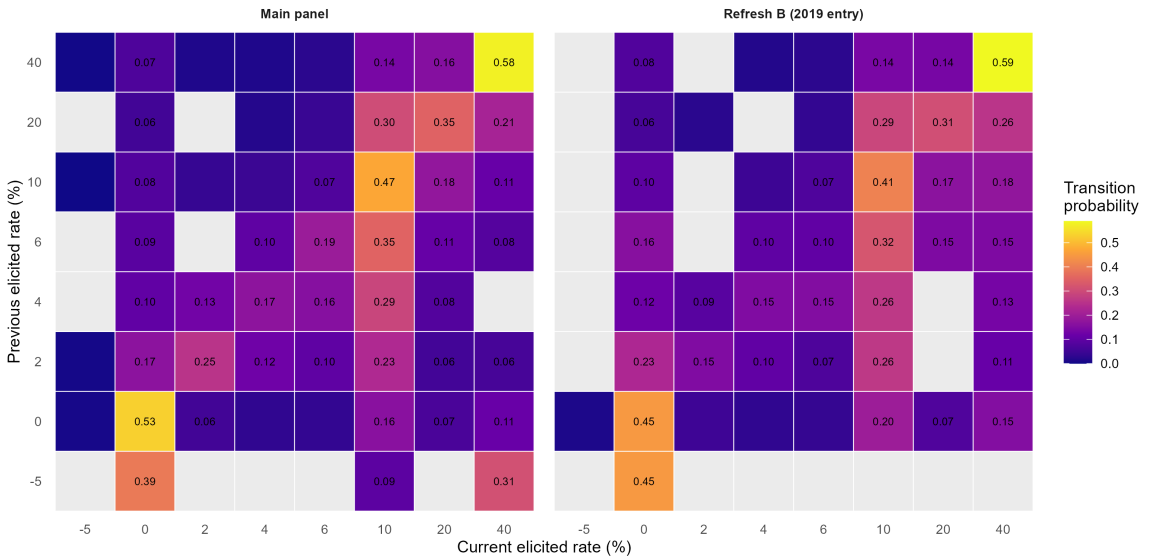

Figure 1: Observed annual transitions in elicited-rate categories

Alt text: Two side-by-side heat maps compare one-year transitions among eight elicited-rate categories. Both concentrate on repeated and focal responses; the main panel and 2019 refresh cohort show similar focal retention.

Notes: Cells are row-normalized probabilities of the current category conditional on the previous category. The left panel uses the 56,050-response main sample; the right panel uses JHPS refresh cohort B, which entered in 2019. Cells with fewer than ten adjacent pairs, plus a complementary cell when needed for disclosure protection, are suppressed. The figure is nonparametric and contains no fitted-model probabilities.

responses flow toward the salient 0, 10, and 40 percent categories. Cells supported by fewer than ten adjacent pairs are suppressed. The refresh comparison is informative about whether persistence is unique to long-tenured respondents, although cohort composition and the common questionnaire prevent a causal interpretation of panel tenure.

I also classify respondents descriptively as always, sometimes, or never using the four salient categories and report response counts, adjacent rank correlation, and exact repetition in the Internet Appendix. These classes are defined ex post and are not preference types. Stable reuse of a salient label is especially important for repeated-proxy designs because averaging does not remove it and both proxy blocks can inherit it.

3.2 Prospective category encodings

The displayed rates impose unequal distances on the eight labels. I therefore use four transparent scalar alternatives constructed only from 2010–2015 responses: the mean option code, the mean within-survey-year ridit, the lower household median category, and the share of early responses at or above the 10 percent option. Each score is standardized on the same prospective wealth-rank sample. These tests ask whether a simple monotone ordering, without the displayed-rate spacing, reproduces the benchmark slope.

A scalar summary can miss a nonlinear category profile. For household i, let q_{ic} be the share of early responses in category c. I regress later wealth rank on seven shares, omitting category 6—the most common response in the prospective sample—and test the seven terms jointly. A complementary specification treats the lower household median as an eight-level factor and uses the same reference category. Both specifications retain the 2016 baseline, demographic, response-count, and survey controls used in the main prospective design. The global tests, not selected category coefficients, are the relevant diagnostics.

I separate scale form from information added by repetition. A saturated first-response model tests whether the seven category coefficients lie on the one-dimensional direction implied by displayed-rate differences. For repeated shares, a companion Wald test imposes the analogous linear direction. I also test whether adjacent coefficients in the linear share model are weakly nonincreasing, weakly nondecreasing, or monotone in either direction. The test statistic is the HC1-weighted distance to the relevant order cone; p-values use 9,999 household-level Rademacher multiplier draws at the

least-favorable equal-profile boundary. Endpoint-exclusion results expose whether a rejection is concentrated in a boundary category. These are restrictions on the linear category-share coefficient profile, not on every possible monotone household aggregator.

Because the shape diagnostic isolates the 40% endpoint, I follow it with a coding audit. Let $\bar{q}_i^{20}$ recode every 40% answer as 20% and let s_{i8} be the household's share of 40% answers. The identity $\bar{q}_i = \bar{q}_i^{20} + 0.20 s_{i8}$ nests two scalar codings in one regression: the 20% cap imposes a zero coefficient on s_{i8}, while the original mapping imposes that its coefficient equal 0.20 times the coefficient on $\bar{q}_i^{20}$. I test both restrictions and report direct 40%-versus-20% contrasts from the saturated first-response and repeated-share models. Deleting endpoint responses, restricting to households that never use the endpoint, and leave-one-category-out scores are influence diagnostics with response deletion and, where necessary, altered samples, not independent confirmatory tests.

To isolate information collected after a one-shot survey would have ended, I add seven post-first category shares to saturated first-response dummies and test them jointly. Repeated stratified cross-validation compares predictive performance without treating its repetitions as independent inference. A cross-panel exercise estimates the repeated-share profile in one survey, freezes it, and calibrates a single score in the other; a two-sample household bootstrap re-estimates both stages to account for generated-score uncertainty. Every comparison retains the repeat-eligible common sample; "first response" is a one-answer representation, not a separate one-shot sampling frame.

3.3 A nonoverlapping-block estimator

For economic magnitudes and repeated-proxy sensitivity, I use the questionnaire's displayed rates. Within 2010–2015, sort each household's valid answers chronologically and form means for the odd-positioned and even-positioned observations:

$$r_i^o = s_i + u_i^o, \qquad r_i^e = s_i + u_i^e. \tag{1}$$

Here s_i is the persistent component on the displayed-rate scale and u_i^o, u_i^e are block-specific deviations. I stack two copies of each household, using r_i^e as an instrument for r_i^o in one copy and reversing their roles in the other. The stacked regression includes a proxy-block indicator, duplicates the same outcome and controls, and clusters standard errors by household. This symmetric obviously-related-instruments construction avoids choosing one block as privileged and uses no response in both proxy and instrument (Gillen et al. 2019).

The block estimate has a narrow interpretation. It corrects the predictive slope for transitory block-specific error if (i) both blocks measure the same stable component over 2010–2015; (ii) their deviations are mean independent across blocks, conditional on controls; and (iii) the instrument is unrelated to the financial-outcome residual except through the persistent component and any stable confounders it contains. Together, these conditions give the slope a measurement-error-correction interpretation, not a causal interpretation. A stable planning disposition, cognitive trait, or focal response style enters both blocks and remains in the block-IV slope.

The odd/even split makes overlap impossible and keeps all measurements before the 2016 baseline. It does not make proxy errors independent: a slowly moving state or remembered response can enter both chronological blocks. The cost is especially visible in KHPS, which contributes exactly one response to each block, whereas JHPS usually contributes three. Survey-specific estimates and the response-count control expose that imbalance. Strong first stages rule out weak-instrument instability but not endogeneity from stable confounders, correlated proxy errors, or failure of the exclusion restriction. I therefore present ORIV as an assumption-dependent measurement sensitivity rather than the recovered "true" slope.

3.4 Estimands and threats

The diagnostics answer three distinct questions. Observed transitions describe how consistently households reuse ordered response labels. Alternative scalar and flexible category encodings test whether later-outcome content is robust to the spacing assigned to those labels. The block estimator estimates an outcome slope per unit of the displayed annual-rate coding under independent-proxy restrictions. Persistence in the first exercise cannot validate the cardinal scale or the exclusion restriction in the third.

Four threats remain. First, stable reporting style is observationally equivalent to a stable economic component unless another construct separates them. Focal-response and flexible-category diagnostics can bound, not eliminate, this concern. Second, slow economic states can enter both proxy blocks. Third, panel conditioning may change response behavior with survey tenure (Cernat

et al. 2026); tenure comparisons and refresh-cohort diagnostics address its observable component. Fourth, an early required return may proxy for liquidity, sophistication, or outside investment opportunities (Dean and Sautmann 2021; Fulford and Shupe 2026). Baseline conditioning and auxiliary controls narrow those interpretations, but only exogenous variation or a richer structural design could separate them.

These limitations motivate the prospective outcome design. If early responses have no relationship with a later balance sheet once baseline wealth is held fixed, repeated measurement would be descriptively interesting but financially unimportant. The next section asks that question while keeping the score, baseline, and outcome windows separate.

4 Early Responses and Later Financial Outcomes

4.1 Specification

The primary outcome design uses one observation per household and a horizon fixed independently of each household's response history. Let q_i^{pre} be the mean displayed rate among valid 2010–2015 answers, requiring at least two. For outcome j I estimate

$$y_{ij}^{2022:24} = \alpha_j + \beta_j q_i^{\text{pre}} + \theta_j y_{ij}^{2016} + Z_i^{2016\prime}\delta_j + \lambda_{s(i)} + e_{ij}, \tag{2}$$

where $y_{ij}^{2022:24}$ is the household's mean outcome across available 2022–2024 waves, y_{ij}^{2016} is the corresponding baseline outcome, and $\lambda_{s(i)}$ is a survey fixed effect. The vector Z_i^{2016} contains age and age squared, sex, marital status, and the number of valid pre-period responses. Heteroskedasticity-robust standard errors are reported. For participation I additionally condition on baseline wealth rank; entry and retention are estimated separately by 2016 participation status. For PPML in wealth levels, the baseline balance-sheet control is 2016 log-one-plus wealth rather than the level, which avoids allowing a few upper-tail observations to dominate both sides of the model.

The estimand pertains to households that satisfy the stated early-response and 2016 requirements and have at least one observed outcome in 2022–2024. It is therefore a selected-survivor conditional association, not a population-representative effect. Section 5 reports observation models, attrition weighting, and official-weight sensitivities without claiming that they identify selection on unobservables.

This specification is intentionally predictive. Baseline conditioning absorbs the level of the earlier balance sheet, including the accumulated consequences of preferences and constraints before 2016. The coefficient β_j therefore describes conditional future financial position rather than an unconditional level gap. With a noisy single-wave baseline, it need not equal a change-score or accumulation effect. Conditioning on baseline changes the estimand and may reduce its magnitude relative to a contemporaneous cross-section; its timing is more credible and its economic question clearer.

The harmonized panel does not contain a common 2016 household-income measure for

both surveys. Equation (2) therefore holds baseline wealth and demographics fixed but not contemporaneous income. This is a material limit: persistent earning capacity can correlate with both the early response and later wealth. Income-observed JHPS specifications are reported as construct-validity sensitivities in Section 5.

The outcome hierarchy is deliberately narrow but was not preregistered. Financial-wealth rank is the primary outcome because it is invariant to currency denomination and common nominal shifts. PPML and the two-part margins test whether the same sign survives a level model and the zero mass; log-one-plus wealth is a unit-sensitive sensitivity. Securities participation, entry, and retention mark a distinct portfolio boundary. I therefore do not count agreement across transformations as independent discoveries.

I report three measurement choices on exactly the same repeat-eligible outcome-specific sample. "First response" uses only the first available answer in 2010–2015 and mimics a one-answer representation, not the sampling frame of a genuine one-shot survey. "Early mean" uses every eligible pre-period answer. "Split-block ORIV" is the symmetric estimator of Section 3. Comparing the columns does not require the first response to be unbiased. It shows how the estimated predictive gradient changes with repetition and, for ORIV, with the independent-block restriction; paired household-bootstrap contrasts account for covariance among estimators fit to the same households.

Table 3: Core prospective household-finance estimates

Outcome	First response	Early mean	Split-block ORIV	Dep. mean	R^2	N
Wealth rank (0–100)	-5.24^{***} (1.89)	-6.36^{***} (2.45)	-9.40^{***} (3.61)	49.6	0.678	3,099
Wealth level (PPML)	-0.242^{**} (0.113)	-0.282^{*} (0.145)	–	1,263	–	3,099
Positive-wealth incidence	−0.032 (0.037)	−0.069 (0.048)	−0.102 (0.070)	0.744	0.454	3,099
Conditional log wealth	−0.226 (0.139)	-0.348^{**} (0.174)	-0.509^{**} (0.253)	6.600	0.448	2,498
Securities participation	−0.043 (0.033)	−0.054 (0.042)	−0.080 (0.062)	0.243	0.496	3,054
Entry (2016 nonholders)	−0.008 (0.030)	−0.011 (0.039)	−0.017 (0.059)	0.082	0.047	2,284
Retention (2016 holders)	-0.179^{*} (0.106)	-0.244^{*} (0.128)	-0.334^{*} (0.173)	0.719	0.097	770

Notes: Rate-score coefficients are per unit (100 percentage points); multiplying by 0.1 gives a ten-percentage-point contrast. The first response and early mean use heteroskedasticity-robust standard errors. Symmetric ORIV stacks nonoverlapping odd/even pre-period proxies and clusters by household. All rows condition on the corresponding 2016 baseline outcome, age and age squared, sex, marital status, pre-period response count, and survey fixed effects; participation also conditions on baseline wealth rank. The dependent-variable mean and R^2 are from each row's early-mean estimation sample and specification; the wealth-level mean is in units of 10,000 yen. PPML is nonlinear, so no ORIV entry or OLS R^2 is reported. Raw slope ratios are intentionally omitted for null or near-zero first-response denominators. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.2 Financial wealth

Table 3 begins with financial-wealth rank. With the early mean, the coefficient is −6.357 (standard error 2.454). On the questionnaire's displayed-rate coding, a ten-percentage-point increase is associated with a rank 0.636 percentile points lower (standard error 0.245). The split-block estimate is −9.395 (standard error 3.610), or 0.940 percentile points lower per ten percentage points under the block restrictions. The complementary block strongly predicts the included block; the first-stage F is about 980. These are coding-dependent projections for conditional future rank, not percentage changes in wealth or local structural marginal effects.

The first-response rank coefficient is −5.237 (standard error 1.895), so one response already reveals a statistically detectable gradient. The early-mean and ORIV point estimates are larger in absolute value, but paired same-household percentile intervals for their differences from the

first-response slope do not exclude zero for rank. ORIV is more negative than the early mean, a distinction reported below. Averaging can retain residual proxy error, while ORIV can amplify violations of the independent-block restriction; none of these scalar comparisons validates the displayed-rate mapping.

The comparisons resample households 999 times and re-estimate all three slopes on identical draws. For rank, the early-mean-minus-first contrast is -1.120 with a percentile interval of $[-3.999, 1.717]$; the ORIV-minus-first contrast is -4.158 with interval $[-8.941, 0.614]$ and a descriptive two-sided bootstrap tail probability of 0.090. The ORIV-minus-mean contrast is -3.039, with interval $[-5.561, -0.796]$ and tail probability 0.004. For log-one-plus wealth, ORIV-minus-first is -0.634 with interval $[-1.299, -0.017]$ and tail probability 0.038, while ORIV-minus-mean is -0.361 with interval $[-0.678, -0.080]$ and tail probability 0.018. Throughout the paper, bootstrap tail areas are descriptive ordinary-bootstrap quantities; they are never null-imposed p-values. The block restriction therefore produces a detectably steeper scalar projection than averaging, but the rank contrast with one response remains imprecise and the flexible category evidence below rejects the common scalar mapping.

The distributional alternatives locate the relationship. A ten-percentage-point increase in the early mean predicts a 0.69-percentage-point lower share of later waves with positive wealth, but the interval includes zero. Conditional on positive later wealth, it predicts 0.035 lower log wealth (standard error 0.017); the split-block counterpart is 0.051 (standard error 0.025). Requiring strictly positive 2016 wealth gives a coefficient of -0.216 (standard error 0.143, $N = 2{,}243$), or 2.14%

lower conditional wealth per ten percentage points ($p = 0.131$). The conditional-amount evidence is therefore suggestive rather than a stand-alone intensive-margin result.

PPML provides a denomination-invariant scale check. Its coefficient of -0.282 (standard error 0.145) implies 2.8% lower expected wealth for a ten-percentage-point higher early mean, with a confidence interval that narrowly includes zero. The PPML conditional mean gives more influence to the noisy upper tail, whereas rank compresses it. I treat the agreement in sign as the robust evidence; the magnitudes are transformation-specific.

The log-one-plus specification is reported only as a sensitivity because 24.1% of observed household-years are at zero. In units of 10,000 yen, a ten-percentage-point higher early mean predicts 0.077 lower $\log(1 + W)$ (standard error 0.031), and ORIV predicts 0.113 lower (standard error 0.045). Re-denominating the identical wealth variable in yen changes these coefficients because the added constant does not scale with the units. An inverse-hyperbolic-sine outcome gives the same sign without supporting a unit-invariant percentage interpretation. The log ORIV-minus-first contrast excludes zero while its rank counterpart does not, but that evidence concerns a unit-sensitive outcome.

The harmonized wealth measure sums deposits and securities. When exactly one component is missing, the primary construction treats it as zero; both missing components produce a missing total. On the locked sample, 460 households (14.8%) invoke this convention at baseline or in at least one post-period wave. A strict per-wave construction instead sets wealth missing unless both components are observed, then requires a strict 2016 rank and at least one strict 2022–2024 rank.

Recomputing survey-year ranks and otherwise preserving the design yields −6.210 (standard error 2.447, N = 2,953), close to the primary −6.357 estimate.

These magnitudes are modest. On the locked rank sample, the early score's standard deviation is 12.05 percentage points, so a one-standard-deviation higher score predicts a later rank 0.77 percentile points lower. Applying the conditional-positive point estimate to the relevant positive-wealth median of about 8 million yen gives roughly 330,000 yen, but the stricter baseline-positive estimate is imprecise. The design concerns conditional future position six to eight years after baseline rather than the level gap accumulated before it.

4.3 Participation, entry, and retention

The prospective participation result is weak. The early-mean coefficient is −0.054 (standard error 0.042), or only −0.54 percentage point for a ten-percentage-point higher response. The split-block estimate is −0.80 percentage point with a standard error of 0.62 percentage point. Both confidence intervals include economically small effects and moderately negative effects.

Splitting transitions explains why. Among 2,284 households not participating in 2016, the early response has essentially no relationship with later entry: the ten-percentage-point effect is −0.11 percentage point (standard error 0.39), and the block estimate is similarly close to zero. The 95 percent interval runs from about −0.9 to +0.7 percentage points per ten percentage points. That is narrow in absolute terms but not relative to the 8.24% entry mean, so the null does not rule out every economically relevant proportional association. Among the 770 baseline participants, the

corresponding estimate for later retention is −2.44 percentage points (standard error 1.28); the block estimate is −3.34 points (standard error 1.73). The retention estimates are suggestive but imprecise, and the subsample is small. The defensible conclusion is therefore asymmetric: early responses are associated with later financial wealth, but the data do not establish a broad entry effect.

This boundary is consistent with the wider household-finance evidence that participation reflects defaults, fixed costs, attention, trust, and access as well as preferences (Gomes et al. 2021; Choukhmane and de Silva 2026). It also guards against a tempting but unwarranted mechanism claim. A stable required-return response can predict later liquid-wealth position without being the margin that determines whether a household opens a securities account.

4.4 Linear specification and category-distribution content

The endpoint audit makes the coding failure economically concrete. Category 8, the displayed 40% response, accounts for 25.3% of eligible early answers in the locked sample; 44.9% of households use it at least once. Recoding 40% as 20% on the same 3,099 households changes the mapped coefficient from −6.357 to −2.912 (standard error 4.593, $p = 0.526$). Deleting 40% responses before recomputing the household mean flips the slope to +10.205 (standard error 4.954) on the 2,780 households with a remaining response; the never-40% subpopulation is likewise positive. Those altered-sample rows are influence diagnostics, not the same estimand.

Table 4 reports the nested test. In a model with the 20%-capped mean and the 40% share, the capped-score coefficient is +18.231 (standard error 6.185) and the share coefficient is −5.769

Table 4: Robustness to the 40-percent response endpoint

Specification	Estimate	HC1 SE	Effect per 10 pp	*N*
Panel A. Scalar-score specifications				
Original mapped mean	−6.357	(2.454)	−0.636	3,099
Cap 40% at 20%	−2.912	(4.593)	−0.291	3,099
Delete 40% responses	10.205	(4.954)	1.021	2,780
Never selected 40%	17.761	(7.081)	1.776	1,708
Panel B. Endpoint restriction tests				
Test	Contrast	HC1 SE	p-value	*N*
Cap coding: endpoint-share term = 0	−5.769	(1.186)	< 0.0001	3,099
Original coding: endpoint term = 0.20β	−9.416	(2.216)	< 0.0001	3,099
Any-40% indicator = 0	−4.528	(0.748)	< 0.0001	3,099
Any/share endpoint terms jointly zero	–	–	< 0.0001	3,099
Repeated shares: 40% − 20%	−5.149	(1.330)	0.0001	3,099
First response: 40% − 20%	−3.182	(0.877)	0.0003	3,099

Notes: The outcome is mean 2022–2024 financial-wealth rank. Full-sample models use the locked 3,099 households and condition on 2016 wealth rank, age and age squared, sex, marital status, the original number of early responses, and survey fixed effects. Inference is HC1.

The capped score recodes 40% as 20%. The delete-response row recomputes the mean from categories 1–7 and excludes only households with no remaining response. The never-40% row restricts to households that never selected category 8 and therefore describes a different subpopulation.

The cap-plus-share model nests the cap and original codings because original mean equals capped mean plus 0.20 times the category-8 share. Saturated contrasts compare category 8 directly with category 7 without imposing a scalar mapping. These are predictive conditional associations, not causal effects.

Panel B contrast estimates are in wealth-rank points for a one-unit change in the relevant share or indicator; a 0.10 change in the category-8 share is one tenth of the reported share contrast.

(1.186). The data reject both the simple cap, which sets the share term to zero ($p < 0.0001$), and the original mapping, which constrains that term to 0.20 times the capped-score slope ($p < 0.0001$). Saturated models also reject equality of the 40% and 20% coefficients using either repeated shares ($p = 0.0001$) or the first response ($p = 0.0003$). The negative mapped slope is therefore driven by a discrete endpoint-use pattern rather than a portable local required-return gradient.

The broader category evidence reaches the same conclusion. Simple scalar summaries that retain category order but discard the displayed-rate distances are individually imprecise, while the flexible

Table 5: Formal inference on the prospective linear specification

Test or paired contrast	Difference or F	Boot. SE	95% CI	p-value / tail area
Panel A: Displayed-rate linear restriction				
Unrestricted shares vs. mapped rate	$F(6, 3084) = 3.80$	–	–	$p = 0.0009$
Panel B: Mapped-rate slope minus alternative slope				
Mean option code (1–8)	−0.670	0.140	[−0.922, −0.376]	0.002
Survey-year ridit mean	−0.337	0.075	[−0.476, −0.178]	0.002
Lower household median	−0.985	0.177	[−1.314, −0.600]	0.002
Share responses ≥ 10%	−0.833	0.219	[−1.235, −0.374]	0.002

Notes: The outcome is mean 2022–2024 wealth rank. All tests use the locked sample of 3,099 households and condition on 2016 wealth rank, age and age squared, sex, marital status, pre-period response count, and survey fixed effects. Panel A is an HC1 Wald/F test of the six restrictions $\beta_c = \gamma(r_c - r_6)$ in the unrestricted seven-share model, with category 6 omitted.

Panel B uses 999 successful paired ordinary household-bootstrap draws (seed 20260831). All scores are standardized once on the locked sample and use identical household resamples within each draw. Entries are mapped-rate minus alternative slopes, percentile intervals, and descriptive two-sided ordinary-bootstrap tail probabilities. The latter are not null-imposed p-values.

repeated-share model rejects the joint null. A direct HC1 Wald test rejects the six restrictions imposed by the displayed-rate mapping, $F(6, 3084) = 3.80$ ($p = 0.0009$), and a saturated first-response model rejects its analogous restriction, $F(6, 3084) = 3.43$ ($p = 0.0023$). A cone-distance test rejects monotonicity of the full repeated-share profile ($p = 0.0264$), but not after excluding the 40% endpoint ($p = 0.8633$); excluding the sparse −5% endpoint still rejects ($p = 0.0188$). The shape tests use 9,999 household-level multiplier draws at the least-favorable equal-profile boundary. Thus first-response data already expose functional-form failure; repetition uniquely supplies the response distribution, persistence diagnostics, and nonoverlapping proxies.

Repetition nevertheless adds little incremental rank prediction after the first category is observed. Seven post-first response shares are jointly insignificant conditional on saturated first-response

dummies, $F(7, 3077) = 1.015$ ($p = 0.418$), with partial $R^2 = 0.0022$. In 20 repetitions of survey-stratified 10-fold cross-validation, incremental out-of-sample R^2 relative to baseline controls is 0.00525 for first-response dummies, 0.00534 for repeated shares, and 0.00185 for the mapped mean. The first two gains are small and virtually identical.

Cross-panel transport gives the same qualified message. Estimating the seven-share profile in JHPS, freezing it, and applying it to KHPS gives a same-direction holdout calibration of 1.312 wealth-rank percentiles per score standard deviation; reversing training and holdout panels gives 0.877. Conditional HC1 standard errors are 0.390 and 0.450, but they hold the source-panel weights fixed. A 1,999-draw two-sample household bootstrap that re-estimates both source weights and holdout calibration gives intervals of $[-0.207, 1.985]$ and $[-0.149, 1.754]$. The transported point orientation is encouraging, but full uncertainty includes zero and both panels share the same questionnaire and institutional setting.

4.5 Baseline diagnostics

A repeated-baseline check replaces the single 2016 control with averages over 2013–2016, 2014–2016, or 2015–2016 while leaving the 2022–2024 outcome unchanged. The ten-percentage-point estimates imply ranks 0.59–0.65 percentile points lower and remain statistically distinguishable from zero. The log estimate falls from −0.077 with the single-wave baseline to −0.053–−0.067 with averaged baselines and is less precise in the longer windows. Samples vary slightly with baseline availability, and the multi-year baselines overlap the early response period; this is a

regression-to-the-mean sensitivity, not a cleaner timing design.

Flexible baseline adjustment leaves the mapped projection nearly unchanged. On the locked 3,099 households, replacing linear 2016 rank with a four-degree-of- freedom natural spline gives −6.287 (standard error 2.436); allowing the entire nuisance profile to differ by survey gives a common score coefficient of −6.255 (2.435). In the 1,197-household JHPS income-complete subsample, adding the verified 2016-wave log report of prior-year household income changes the same-sample coefficient from −9.800 (4.430) to −10.344 (4.393). These checks reduce concern about a linear baseline control or the available income field, but they do not resolve endpoint coding or omitted stable traits.

A common 2014–2015 two-response design removes the mechanical difference between the nearly six-response JHPS mean and the two-response KHPS mean. On 3,062 households, its pooled mean coefficient is −5.771 (standard error 2.322) and symmetric two-wave ORIV is −8.846 (3.568; first-stage $F = 823$), with the same point direction in both surveys. A separately timed JHPS replication using 2010 baseline rank, 2011–2012 responses, and 2013–2015 outcome ranks also gives a negative mapped coefficient, −8.437 (2.502, $N = 2{,}415$). This is an earlier prospective replication, not a pretrend placebo, and it inherits the displayed-rate limitation.

4.6 Why the magnitudes differ from contemporaneous estimates

The prospective estimates should not be compared mechanically with coefficients from a pooled contemporaneous regression. There are three differences. First, equation (2) conditions on 2016

wealth, removing the portion of the cross-sectional level gap already present by baseline, subject to baseline measurement error. Second, it uses a selected but auditable long-horizon survivor sample. Third, it freezes the response information set in 2015; a full-panel household mean uses responses collected alongside and after the outcomes. The prospective design consequently sacrifices power and estimates conditional future position, not a clean accumulation effect.

The gap is quantitatively large. In the pooled contemporaneous panel behind the construct-validity sensitivities (Internet Appendix), one response predicts about 0.21 lower log-one-plus wealth per ten percentage points; the prospective early-mean counterpart in the same units is 0.077. The two designs use different samples and controls, so the comparison conveys scale; the three differences above, jointly, produce the gap.

That sacrifice disciplines the main claim. A large contemporaneous block-IV coefficient can arise because persistent financial circumstances influence both long-run answers and long-run wealth, because a full-panel score leaks future response information, or because true response heterogeneity accumulated into wealth before the sample began. The design here cannot distinguish historical channels, but it rules out future-score leakage and makes the baseline gap explicit.

5 Robustness, External Validity, and Interpretation

The empirical design can fail in several distinct ways. A first-response slope may be attenuated by transitory error; a repeated score may instead capture stable response style; the apparent persistent

component may move with financial circumstances; conditioning on long-horizon survival may select particular households; or a common questionnaire convention may drive both panels. No single robustness check resolves all of these concerns. This section organizes the evidence by the threat it addresses and separates tests of measurement from tests of economic interpretation.

5.1 Two independently sampled surveys

JHPS and KHPS provide a demanding design comparison because they use the same item but offer different pre-period measurement intensity. Eligible JHPS households contribute 5.89 answers on average from 2010–2015, and their odd/even block correlation is 0.70. Eligible KHPS households contribute exactly two answers, with a correlation of 0.41. If the pooled result were an artifact of the long JHPS average, the wealth gradient should collapse in KHPS. I estimate equation (2) separately by survey and formally test the interaction between the early score and a KHPS indicator.

The wealth coefficients have the same sign in both panels, while their precision and raw scale reflect the different information in a nearly six-response and a two-response mean. The JHPS and KHPS rank coefficients are -9.27 and -4.92; their equality-test p-value is 0.329, while the KHPS estimate is individually imprecise. For log wealth the corresponding coefficients are -1.08 and -0.61 and the equality-test p-value is 0.403; the participation equality-test p-value is 0.382. Failure to reject equality is not evidence that the latent effects are equal. I treat the estimates as directionally consistent across independently sampled households answering the same questionnaire; replication across instruments or countries would be a stronger claim than these panels can support. Because

questionnaire wording and institutional setting are common, the two surveys cannot rule out a stable response convention specific to this item.

Survey-specific split-block estimates provide a sharper measurement check. The rank coefficient is -11.33 (standard error 5.15) in JHPS and -8.15 (standard error 4.98) in KHPS; the corresponding first-stage F statistics are 1,966 and 380. The estimates overlap and remain negative despite sharply different response counts and proxy correlations. This is not a formal cross-survey equality test, and the KHPS coefficient remains imprecise, but the matching point-estimate direction makes it less plausible that the pooled block result is generated only by the longer JHPS response history.

The JHPS 2019 refresh cohort supplies a further check on panel tenure. It is not part of the main 56,050-response panel. Figure 1 compares its observed adjacent transitions with those in the main panel, without fitted probabilities. The refresh cohort's Spearman correlation is 0.450, exact-repeat rate is 0.415, and focal-to-focal retention is 0.896, compared with 0.513, 0.444, and 0.893 in the main panel. Similar focal retention makes a convention unique to long-tenured respondents less plausible, but the cohorts differ in composition and share the same questionnaire.

5.2 Focal responses and panel conditioning

The four especially salient displayed values—0, 10, 20, and 40 percent—account for about 85 percent of responses. A respondent who always chooses 10 percent may be economically stable, may round a more precise answer, or may simply reuse a salient label. All three generate persistence, but only the first two plausibly contain preference information. I therefore report, for respondents who always,

sometimes, or never use those values, the number of respondents, adjacent-pair rank correlation, and exact repetition. The full-sample transition table separately reports observed diagonal, endpoint, and focal retention. These are descriptions of scale use.

Panel conditioning remains possible even with tenure controls. Repeated exposure could teach respondents the scale, encourage recall, or change attention (Cernat et al. 2026). The refresh-cohort comparison is useful because it varies tenure at similar calendar dates, but it is not randomized. I therefore interpret a tenure pattern as a feature of measurement quality; it says nothing about whether the underlying economic preference changes with survey experience.

5.3 Attrition

The primary later-wealth sample retains 3,099 of 4,716 households eligible for the wealth observation model, a retention rate of 65.7%. For participation the corresponding rate is 65.2%. I first regress an indicator for observing the 2022–2024 outcome on the early score and all 2016 controls. A ten-percentage-point higher response predicts only a 0.55-percentage-point lower probability of observing wealth (standard error 0.56) and a 0.53-point lower probability of observing participation (standard error 0.57). Neither relationship is distinguishable from zero.

I then estimate stabilized observation probabilities using only the early score, baseline outcome, demographics, response count, and survey. Probabilities are bounded between 0.02 and 0.98 and retained-sample weights are capped at the 99th percentile. Inverse-probability weighting changes the early-mean rank coefficient from -6.357 to -6.222 and its standard error from 2.454 to 2.448.

The participation estimate remains small. This exercise addresses selection on observed baseline variables; attrition driven by unobserved future shocks is outside its reach. Within that limit, the stability of the rank estimate indicates that the observed selection channel produces no large correction.

Official 2016 integrated-household cross-sectional weights address baseline representativeness rather than longitudinal attrition. On the 3,051 households with a valid weight and observed wealth, the weighted rank coefficient is −8.835 (standard error 5.466), compared with −6.555 (2.467) unweighted on those same households; the weighted PPML coefficient is −0.126 (0.430) on the same 3,051 households. The weighted participation coefficient is −0.102 (0.080) on its distinct 3,007-household sample. Point estimates therefore remain directionally consistent, but the weights substantially reduce precision; they do not turn this selected long-horizon panel into a representative survivor sample.

Varying the number of waves entering the later outcome also leaves the result intact. Requiring at least two of the three 2022–2024 wealth observations gives coefficients of −6.64 (standard error 2.45) for rank and −0.788 (standard error 0.312) for log wealth, on 2,826 households. Requiring all three gives −7.19 (2.66) and −0.819 (0.337), respectively, on 2,382 households. Participation becomes somewhat more negative in the balanced samples but remains imprecise. The main design retains all available later observations to avoid selecting on complete survival.

5.4 Construct validity and the accumulation margin

Two auxiliary results help interpret the persistent response without turning it into a structural discount factor. First, contemporaneous specifications in the income-observed JHPS subsample show that controlling for household income reduces but does not eliminate the wealth gradient. A financial-literacy-battery sensitivity reaches a similar qualitative conclusion. These controls are measured imperfectly and sometimes after the earliest responses, so I report them as sensitivities and claim no selection-on-observables identification.

Second, the four JHPS waves with both household income and expenditure allow a saving-rate check. A ten-percentage-point higher response is associated with a 0.64-percentage-point lower saving rate using one response, 1.37 points using the household mean, and 1.64 points using a lagged-response instrument. This pattern is consistent with an accumulation mechanism, but it covers only 2010–2013, relies on noisy self-reported income and expenditure, and remains observational. It cannot establish that the saving-rate difference compounds into the full wealth gradient.

The outcome decomposition reinforces this restrained reading. Later wealth rank, PPML wealth, and conditional positive wealth move in the same direction. Entry into securities does not. This is compatible with patient or financially organized households accumulating larger liquid balances, but it is also compatible with persistent cognition, planning, or resource access affecting both the elicitation and saving. The evidence establishes predictive content and estimator sensitivity; adjudicating among such mechanisms would require variation this design lacks.

Outcome measurement adds a final limitation. Recent evidence linking experimental choice

inconsistency to both administrative and self-reported wealth finds a stronger association in self-reports, consistent with decision quality also predicting reporting error (Arts et al. 2026). JHPS/KHPS do not provide an administrative balance-sheet benchmark. Wealth ranks, PPML, the two-part decomposition, and the prospective baseline reduce sensitivity to any one transformation, but they cannot remove response error shared between the preference item and the balance sheet.

Taken together, the identified object is a coding-explicit predictive association. The complementary response block is an instrument for a noisy regressor, not an exogenous shock to intertemporal preference; a strong first stage rules out weak-instrument instability but not endogeneity or correlated proxy error. The evidence documents persistence and joint category content without identifying every within-person movement as error, a causal effect, or a structural primitive.

6 Conclusion

Repeated elicitation changes what can be diagnosed more than what can be predicted. Observing the same item up to fifteen times documents substantial but incomplete label persistence and permits transparent comparisons among one response, averages, response distributions, and split-block IV. Those comparisons do not justify interpreting the item as a structural discount factor.

In the prospective wealth design, early displayed-rate averages are negatively associated with 2022–2024 wealth rank conditional on 2016 rank, and split-block ORIV is steeper than the average under its maintained block restrictions. But capping 40% at 20% makes the negative scalar slope

imprecise, deleting 40% responses flips its sign on an altered sample, and the data reject both scalar codings in a nested endpoint model. Flexible category models likewise reject displayed-rate linearity, while post-first responses add little beyond the first category. The distinctive value of repetition is therefore diagnostic, not a demonstrable improvement in wealth prediction.

That conclusion is deliberately limited. Stable cognition, planning, financial circumstances, earning capacity, and response style can affect both repeated answers and wealth, and securities entry is not detectably predicted. The general lesson is to report what repetition reveals—persistence, transitions, and sensitivity to aggregation—without equating observed persistence with a latent preference or causal mechanism. The design can be applied to other preference and belief measures to test whether repetition improves prediction without inferring that observed persistence identifies the structural object.

Data Availability

JHPS and KHPS microdata are distributed under a research-use agreement by the Panel Data Research Center at Keio University and cannot be redistributed. Researchers can apply to the Center. The replication code reconstructs the analysis panel from licensed files and regenerates all tracked aggregate tables and figures. No respondent identifiers or restricted microdata are included.

References

Steffen Andersen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. Eliciting risk and time preferences. *Econometrica*, 76(3):583–618, 2008.

James Andreoni and Charles Sprenger. Estimating time preferences from convex budgets. *American Economic Review*, 102(7):3333–3356, 2012.

Sara Arts, Qiyan Ong, Jianying Qiu, and Jana Vyrastekova. Choice (in)consistency and household wealth. *Journal of Risk and Uncertainty*, 2026. doi: 10.1007/s11166-026-09499-5. Published online June 23, 2026.

Robert B. Barsky, F. Thomas Juster, Miles S. Kimball, and Matthew D. Shapiro. Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *Quarterly Journal of Economics*, 112(2):537–579, 1997.

Jonathan P. Beauchamp, David Cesarini, and Magnus Johannesson. The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and Uncertainty*, 54(3):203–237, 2017.

Michèle Belot, Philipp Kircher, and Paul Muller. Eliciting time preferences when income and consumption vary: Theory, validation, and application to job search. *American Economic Journal: Microeconomics*, 17(1):130–170, 2025.

Christian Belzil and Tomáš Jagelka. Separating preferences from endogenous effort and cognitive

noise in observed decisions. IZA Discussion Paper 18315, IZA Institute of Labor Economics, December 2025.

Laurent E. Calvet, John Y. Campbell, Francisco Gomes, and Paolo Sodini. The cross-section of household preferences. *Journal of Finance*, 2026. doi: 10.1111/jofi.70067. Published online July 23, 2026.

Alexandru Cernat, Joseph W. Sakshaug, Bella Struminskaya, Susanne Helmschrott, and Tobias Schmidt. Panel conditioning in measuring financial behavior and expectations. *Journal of Economic Psychology*, 116:102936, 2026. doi: 10.1016/j.joep.2026.102936.

Taha Choukhmane and Tim de Silva. What drives investors' portfolio choices? separating risk preferences from frictions. *Journal of Finance*, 81(1):5–48, 2026. doi: 10.1111/jofi.70013.

Yating Chuang and Laura Schechter. Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. *Journal of Development Economics*, 117: 151–170, 2015.

Jonathan Cohen, Keith Marzilli Ericson, David Laibson, and John Myles White. Measuring time preferences. *Journal of Economic Literature*, 58(2):299–347, 2020.

Robin P. Cubitt and Daniel Read. Can intertemporal choice experiments elicit time preferences for consumption? *Experimental Economics*, 10(4):369–389, 2007.

Mark Dean and Anja Sautmann. Credit constraints and the measurement of time preferences. *Review of Economics and Statistics*, 103(1):119–135, 2021.

Thomas Dohmen and Tomáš Jagelka. Accounting for individual-specific reliability of self-assessed measures of economic preferences and personality traits. *Journal of Political Economy Microeconomics*, 2(3):399–462, 2024. doi: 10.1086/727559.

Thomas Epper, Ernst Fehr, Helga Fehr-Duda, Claus Thustrup Kreiner, David Dreyer Lassen, Søren Leth-Petersen, and Gregers Nytoft Rasmussen. Time discounting and wealth inequality. *American Economic Review*, 110(4):1177–1205, 2020.

Scott L. Fulford and Cortnie Shupe. Time preferences, rates of return, and real-world investment decisions. *Journal of Economic Behavior & Organization*, 241:107396, 2026. doi: 10.1016/j.jebo.2025.107396.

Ben Gillen, Erik Snowberg, and Leeat Yariv. Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863, 2019.

Gopi Shah Goda, Matthew R. Levy, Colleen Flaherty Manchester, Aaron Sojourner, and Joshua Tasoff. Predicting retirement savings using survey measures of exponential-growth bias and present bias. *Economic Inquiry*, 57(3):1636–1658, 2019.

Francisco Gomes, Michael Haliassos, and Tarun Ramadorai. Household finance. *Journal of Economic Literature*, 59(3):919–1000, 2021.

Wataru Kureishi, Hannah Paule-Paludkiewicz, Hitoshi Tsujiyama, and Midori Wakabayashi. Time preferences over the life cycle and household saving puzzles. *Journal of Monetary Economics*, 124:123–139, 2021.

David Laibson, Sean Chanwook Lee, Peter Maxted, Andrea Repetto, and Jeremy Tobacman. Estimating discount functions with consumption choices over the lifecycle. *Review of Financial Studies*, 39(6):1580–1610, 2026. doi: 10.1093/rfs/hhae035.

Stephan Meier and Charles D. Sprenger. Temporal stability of time preferences. *Review of Economics and Statistics*, 97(2):273–286, 2015.

Nicolás Salamanca. The dynamic properties of economic preferences. Working Paper 4/18, Melbourne Institute of Applied Economic and Social Research, 2018.

Victor Stango and Jonathan Zinman. We are all behavioural, more, or less: A taxonomy of consumer decision-making. *Review of Economic Studies*, 90(3):1470–1498, 2023.