A summary of the experimental variations relating to the three annuity decisions is presented at the end of the document in Figure 1. We refer to all treatments as $T1$, $T2$, and so forth, where a "treatment" is a difference between two experimental cells, as seen in the figure. Each difference corresponds to a test of a particular mechanism. We refer to each cell as $G0$, $G1$, and so forth. Because we have both between and within participant variation, each participant generates data points in multiple cells, but not all of the cells.

Unless otherwise stated, all analyses here are run on the full sample in the relevant cells of the given treatment, and not for particular subsamples within a cell. We will not reweight the observations in our primary analysis, although sample weights are provided by AmeriSpeak.

Primary analyses

First, we will report the mean annuity take-up in each of the nine cells in the experiment, together with the standard errors.

Second, we will report the average effects on annuity take-up of the following treatments: $T1$, $T2$, $T3$, $T4$, $T5$, $T10a$, $T10b$, $T20$ and $T35$. Each of these analyses involves two cells, G_a and G_b, where a treatment is the difference in the means between these two cells. Thus, standard errors and hypothesis tests for differences between any two cell means can be derived from a regression, where the covariate is a dummy that equals 0 for the first cell and 1 for the second cell. Specifically, the regression equation is a linear probability model

$$y_{ij} = \beta_0 + \beta_1 \mathbf{1}_{G_b} + \varepsilon_{ij} \tag{1}$$

where $y_{ij} \in \{0, 1\}$ is an indicator that equals 1 if participant i takes up the annuity in cell $j \in \{G_a, G_b\}$. In this and all other regressions, we will compute robust standard errors that are clustered by participant where appropriate (i.e., where some participants appear in multiple cells). We will designate either $T10a$ or $T10b$ as our "preferred estimate" of the "reverse correlation" treatment effect based on whether we find that it is important or not to control for insurance wording. Specifically, to choose our preferred specification, we will run the regression

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\text{ins. wording}} + \varepsilon_i$$

where $\mathbf{1}_{\text{ins. wording}} \in \{0, 1\}$ is an indicator that equals 1 if participant i is randomized into the "insurance wording" condition. We run this regression on all observations in the control cell, $G0$. We will refer to $T10b$ as our preferred specification if and only if the coefficient β_1 is different from 0 at the 10% significance level.

Third, we will report the following four differences between treatments: $T_3 - T_2$ for the entire sample, $T_4 - T_3$ for the entire sample, $T_5 - T_4$ for the subsample of participants for whom we adjust savings in the high-price annuity contingency in the explicit contingency, no context, dominance treatment (i.e., those who are presented a different choice in cell $G4$ than in $G5$), and $T_5 - T_4$ for the subsample of participants for whom choosing an annuity is not a stochastically-dominant decision.

Each of these differences is again based on a comparison of only two cells. Thus, all statistical inference on these differences will be based on estimating the regression in (1), with robust standard errors.

Secondary analyses

First, we will replicate the primary analyses on the subsample of participants who make optimal savings decisions in the following two cases: (i) no annuity and (ii) high-price annuity. For this subsample, cells G_5 and G_4 are identical, and hence will be pooled. Thus, we cannot study the effect of T_5 (we will report the difference between $G0$ and G_4/G_5 as T_4), and we cannot study $T_5 - T_4$. Otherwise, we will conduct identical analyses for this subsample.

Second, we will analyze interaction of selected treatment effects with the following four covariates:

1. Dummy for whether a participant answered all three financial literacy questions correctly
2. Dummy for whether a participant has above median income, where the median income is computed with respect to our sample
3. Dummy for whether a participant has a college degree or higher
4. Dummy for whether a participant is above age 50

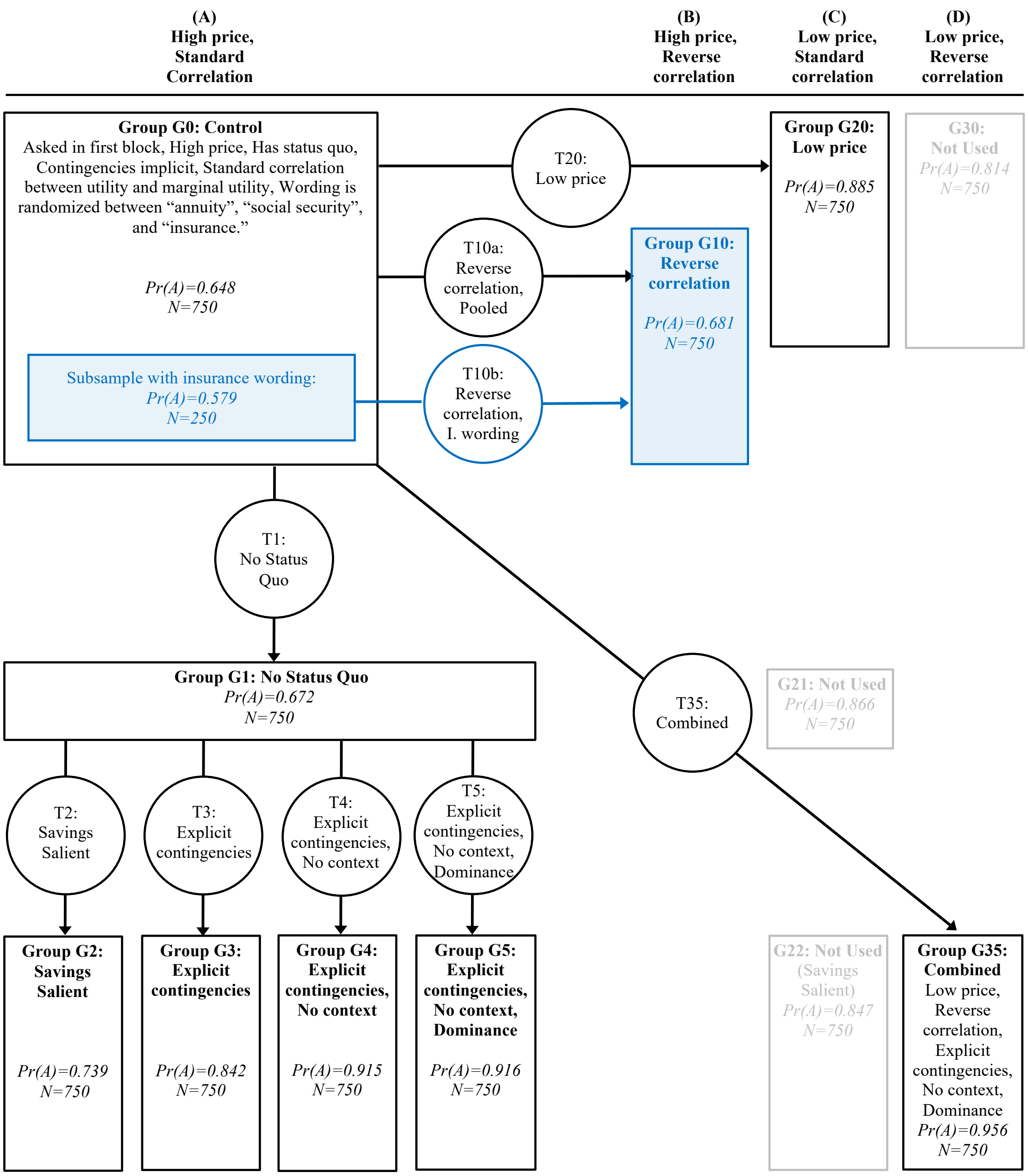
We test whether each of these demographic covariates interacts with the following 6 treatment effects: $T1$, T_2, T_3, the average of T_4 and T_5 , the preferred specification from $T10a$ or $T10b$ (based on the outcome of the test run during the primary analysis), and $T20$. To test for the interaction of a particular treatment with a particular demographic covariate d, we run the regression

$$y_{ij} = \beta_0 + \beta_1 \mathbf{1}_{G_b} + \beta_2 d + \beta_3 \mathbf{1}_{G_b} \cdot d + \varepsilon_{ij} \tag{2}$$

with robust standard errors clustered by participant where appropriate. The coefficient β_3 corresponds to the interaction effect of interest. When testing for the interaction with the average effect of T_4 and T_5, the indicator variable corresponds to the cell being either $G4$ or $G5$.

Because the power of the interaction-effect tests above may be limited, we will perform four tests that pool additional cells. We do these four more powerful tests for each of the four demographic covariates. First, we will also test for joint significance of an interaction with the five non-price treatments (the ones noted above other than $T20$). That is, we will report an F-statistic for the joint significance of the coefficients on $\mathbf{1}_{G_b} \cdot d$ from the five regressions summarized above. Second, will also test for a jointly-significant interaction with the three contingent reasoning treatments: T_2, T_3 , and the average of T_4 and T_5. Third, we will also pool $G1$ and $G2$ when we test whether there is heterogeneity with respect to d in the overall effect of all the treatments aimed at improving contingent reasoning. Specifically, we estimate equation (2) with $G_a = G1 \cup G2$ and $G_b = G4 \cup G5$. Finally, we will test whether there is heterogeneity with respect to d in whether respondents primarily respond to the more minimal contingent-reasoning treatments (primarily to $T2$ or $T3$) or whether they respond mostly to the more comprehensive contingent-reasoning treatments (primarily to $T4$ or $T5$). We implement this test by estimating equation (2) with $G_a = G1 \cup G5$ and $G_b = G2 \cup G3$.

Figure 1: Summary of experimental design, and treatment definitions



Notes: $Pr(A)$ is the fraction of respondents selecting the annuity in the pilot; The number of observations reflect expected numbers in the AmeriSpeak sample of 3000 respondents.