Assessing Hypothetical Scenarios for Estimating Labor Supply Responses: Evidence from a Pension Reform*

Nicola Fuchs-Schündeln[1], Jonas Jessen[2], Pinchuan Ong[3], and Jessica Pan[3]

[1]WZB and Goethe University Frankfurt

[2]WZB and IAB

[3]National University of Singapore

August 19, 2026

ABSTRACT: We assess whether survey responses to hypothetical policy scenarios can recover actual labor supply responses from policy variations. We study the German Müterrente, a pension benefit that credits parents for raising children, benchmarking against the quasi-experimental estimates of Artmann et al. (forthcoming). Surveying mothers and fathers about a hypothetical reform, we find that mothers expect to reduce their labor supply, with an implied marginal propensity to earn of -0.65—close to the quasi-experimental estimate of -0.54. The response is sensible across subgroups, falls mainly on the extensive margin, and spills over to fathers. In the aggregate, spousal predictions match partners' own reports.

JEL: J22, H55, D84, J26, D10.

*We thank conference and seminar participants at the University of Hong Kong, National University of Singapore, and IMEBESS 2026 for helpful comments and suggestions. Zhi Xing Tan provided excellent research assistance. Fuchs-Schündeln and Jessen gratefully acknowledge funding from the German Federal Ministry of Research, Technology and Space (BMFTR). Fuchs-Schündeln: nicola.fuchs@wzb.eu; Jessen: jonas.jessen@wzb.eu; Ong: ong.pinchuan@nus.edu.sg; Pan: jesspan@nus.edu.sg.

1 Introduction

Economists traditionally identify key behavioral parameters of economic models—labor supply elasticities, the marginal propensity to consume, the marginal propensity to earn out of unearned income, and the like—from quasi-experimental policy variation (Imbens et al. 2001; Cesarini et al. 2017; Fetter and Lockwood 2018; Golosov et al. 2024; Boehm et al. 2025; Gromadzki 2024; Artmann et al. forthcoming; Becker et al. 2022). Such variation is compelling where it is available, but it is also scarce: a credible design requires an exogenous reform or shock of the right kind, with the right timing, affecting a population the researcher can follow in the data. These conditions may not align with the precise question at hand, and a researcher who needs an estimate for a particular population or setting often has to extrapolate from estimates produced wherever the necessary variation happened to arise. Policymakers face a sharper version of this problem. They must anticipate how today's population would respond to a reform, yet the available estimates typically come from other settings that may differ in exactly the ways that matter. Moreover, the policies of greatest interest are often not yet enacted, or novel bundles of existing measures, with no variation to exploit.

A complementary approach to estimating such parameters asks respondents directly how they would behave under a hypothetical scenario—a windfall, a tax change, or a policy reform—and infers the parameter of interest from their stated responses (Shapiro and Slemrod 2009; Jappelli and Pistaferri 2020; Christelis et al. 2025; Crossley et al. 2025; Georgarakos et al. 2025). Surveys of this kind are comparatively inexpensive, can be fielded quickly, and can be tailored to policies that have not yet been enacted. Their obvious limitation is that the choices are hypothetical, and there is little direct evidence on how closely stated responses correspond to actual behavior—in part because a convincing test requires a setting that also offers a credible behavioral benchmark to compare against.

This paper assesses the validity of stated labor supply responses by comparing them, within a single setting, to a well-identified quasi-experimental benchmark. We study the

German *Mütterrente* ("mother's pension"), a series of reforms that credit mothers—in principle either parent, but the mother in 98 percent of cases—with additional pension points for time spent raising children. The setting suits our purpose for several reasons. First, a recent literature estimates forward-looking labor supply responses to recent *Mütterrente* expansions using administrative data (Artmann et al. forthcoming; Becker et al. 2022). Our main benchmark, Artmann et al. (forthcoming), exploit rich administrative data to evaluate the effects of the 2014 reform that raised the pension income of mothers of children born before 1992 by 1 pension point (or 3,830 euros in 2013 net present value of total pension accrued in retirement).[1] The authors find that higher future pension wealth reduces mothers' labor supply and, to a lesser extent, their spouses', giving us a yardstick against which to evaluate our survey estimates. Second, the setting is favorable to stated-preference elicitation: the policy environment is stable, Germans are generally well informed about the pension system, such reforms are relatively common, and the reform raises unearned income in a transparent and easily described way. Abstracting from comprehension issues allows us to assess the elicitation method itself, not possible misunderstanding of the policy.

We fielded separate surveys of mothers and fathers. In each survey, we elicit expected labor supply—employment, hours, and earnings—both absent the reform and under a hypothetical reform closely modeled on the *Mütterrente*, with the reform randomly granting one or three additional pension points to mothers. We elicit these expectations at the five-year horizon throughout, and at the one-year horizon for a subset of mothers, pivoting the elicitation around each respondent's own previously stated expectations. We explain the hypothetical reform to respondents using a slideshow with infographics, and include comprehension checks to confirm that they understand the reform.

Each survey yields a direct estimate of that group's own labor supply response: mothers' response to receiving the additional pension points—of which they are the direct recipients—and fathers' response to their partner receiving them, which captures the intra-household

[1] 3,830 € in 2013 ≈ 5,015 € in 2025.

spillover. The two-survey design also lets us elicit each spouse's prediction of their partner's response, allowing us to assess whether household members hold, on average, accurate beliefs about one another's expected behavior.

Our central finding is that the survey recovers a labor supply response close to the quasi-experimental benchmark. Mothers expect to reduce their labor supply in response to the reform—consistent with an income effect, since the additional pension points raise expected retirement income—cutting expected hours and earnings by about 2 percent at the five-year horizon. Translating the earnings response into an implied marginal propensity to earn out of unearned income (MPE), under the same assumptions as Artmann et al. (forthcoming), yields -0.65, close to their quasi-experimental estimate of -0.54. The correspondence is especially striking because the two estimates draw on entirely different data—stated responses to a hypothetical reform on the one hand, and administrative data on a realized reform on the other.

Beyond this headline correspondence, we have four major findings. First, the estimated mothers' response shows a coherent pattern across subsamples: it is negative in every subgroup we examine, larger for the larger pension increase, and concentrated among higher-earning households better placed to afford the implied reduction in labor supply. Second, most of the hours response operates on the extensive margin—movements into and out of employment—the one dimension on which we differ from Artmann et al. (forthcoming), whose response is almost entirely intensive. Georgarakos et al. (2025) likewise recover a largely extensive-margin response from survey-based windfall questions, suggesting that this extensive-margin pattern may partly reflect stated-preference elicitation. Surveying both partners yields our third and fourth contributions, both of which look inside the household. Fathers expect to reduce their own labor supply when their partner receives the points, a within-household spillover that Artmann et al. (forthcoming) also detect but estimate imprecisely. And, comparing our two surveys, each spouse's prediction of the other's response lines up closely in the aggregate with the partner's own report, providing evidence that household

members hold, on average, accurate beliefs about how the other would react.

Our paper is related to three main strands of literature. The first studies income effects in labor supply—the response of earnings or hours to unearned income—using lotteries (Imbens et al. 2001; Cesarini et al. 2017; Picchio et al. 2018; Golosov et al. 2024), inheritances (Joulfaian and Wilhelm 1994), cash transfers (Gromadzki 2024), and old-age support programs (Fetter and Lockwood 2018). Most of this work exploits windfalls to *current* income; Artmann et al. (forthcoming) is distinctive in estimating a forward-looking response to future pension wealth. We contribute a survey-based forward-looking estimate that can be set alongside theirs.

The second literature we build on uses surveys to recover behavioral parameters. One strand asks how studies how accurately survey responses capture actual behavior. While Bhandari et al. (2020) show that survey-reported magnitudes of business finances can diverge substantially from administrative records, Parker and Souleles (2019) find that when households' reported responses to tax rebates are set against revealed-preference estimates from the same setting, the two agree on average, but not in their heterogeneity. A second strand uses hypothetical survey scenarios to estimate parameters such as the marginal propensity to consume or the marginal propensity to earn out of unearned income (Shapiro and Slemrod 2009; Jappelli and Pistaferri 2020; Christelis et al. 2025; Crossley et al. 2025; Georgarakos et al. 2025).Closest to us, Georgarakos et al. (2025) use hypothetical lottery questions to measure labor supply responses to windfalls, finding responses—largely on the extensive margin—only at large gains. We find a response to smaller policy-induced changes in *future* unearned income, in line with quasi-experimental estimates based on administrative data. The key difference in our set-up is that rather than using a generic hypothetical, we mimic an actual policy reform with an explicit policy framing and detailed infographics targeted at the relevant population. Our central contribution is therefore to validate stated responses to a *hypothetical* policy reform directly against a quasi-experimental estimate of a realized version of the *same* reform.

The third literature we contribute to relates to proxy reporting in surveys and information diffusion within the household. Large-scale household surveys routinely rely on proxy reporting, with one member answering for others, so its accuracy directly shapes measured labor-market outcomes. Early work established that self- and proxy reports diverge (Mellow and Sider 1983; Bound et al. 2001), and a series of randomized survey experiments has since shown that whether a person is asked directly or reported on by proxy can shift measured employment, hours, and earnings, with effects that vary systematically by gender and setting (Estrada et al. 2027; Bardasi et al. 2011; Dillon and Mensah 2024; Masselus and Fiala 2024). A more recent literature directly explores informational frictions within the household, often as a test of collective models of the household, which maintain that members are sufficiently well informed about one another to reach efficient allocations (Chiappori 1992). These studies find that spouses' expectations about shared economic facts are often misaligned, and information given to one partner reaches the other only partially and asymmetrically (Fehr et al. 2024; Delavande et al. 2025). Nevertheless, we find that, in the aggregate, spouses hold reasonably accurate expectations of one another's behavioral responses to the reform. To our knowledge, ours is the first survey to elicit both partners' expected responses to the same reform and to document their mutual consistency. More broadly, our results point to the promise of eliciting subjective expectations by proxy, since one respondent's beliefs about a partner track that partner's own expectations in the aggregate.

The rest of the paper proceeds as follows. Section 2 describes the Mütterrente, our surveys, and the samples. Section 3 presents the estimated labor supply responses, the implied marginal propensity to earn, and the spousal-prediction results. Section 4 concludes.

2 Survey

We conducted two separate subjective-expectations surveys of German parents recruited via Bilendi (Appendix B.1). The mothers' survey targeted non-retired German mothers aged

48–60 and was fielded in two waves: a main wave in August–September 2025 which focused on mothers' own labor supply response at the five-year horizon, and a supplemental wave in March–May 2026 which also included shorter-horizon expectations and partner responses. The fathers' survey was conducted in August–September 2025 and targeted non-retired German men aged 50–62 who were partnered and cohabiting with a mother. We did not screen men on their own fertility history to match the information available in the administrative data used by Artmann et al. (forthcoming). Median completion times were 12.4 and 13.7 minutes for the mothers' and men's surveys, respectively.

Each survey proceeded in three parts. First, we collected demographics, pension-system knowledge, and current labor-market outcomes (employment status, hours, wage/salary income, and work expectations if not working) for respondents and, when applicable, their partners. To establish a baseline against which to measure the impact of the hypothetical reform, we elicited respondents' expectations about their work situation five years ahead. Next, respondents viewed a slideshow describing the hypothetical pension reform—our vignette—in four non-skippable, animated slides, before reporting expectations about their own and, where applicable, their partner's work decisions at the relevant post-reform horizons. Finally, we collected additional background characteristics (e.g., education) and expectations about retirement pensions and life expectancy (Appendix B.2).

Appendix Figure A1 shows the vignette slides that mothers saw.[2] The vignette described a reform "that takes child-rearing time into greater account when calculating pensions", granting mothers one or three additional pension points (randomized across respondents). Unlike the actual Mütterrente, which grants more to mothers with more children, ours applied a flat increase: extra points accrue "regardless of how many children [the mother has]". To ensure the response reflects a pure lump-sum shock to future wealth, the increase accrued regardless of labor supply ("...regardless of how much you have worked in total and whether you are currently working or not"). Further, it was given to the mother alone ("Only

[2]The slides were shown in German; Appendix Figure A1 reproduces our English translation. The full instrument is available (in both languages) as an online supplement.

mothers receive the increase. Not your partner"). The slides provided a wealth interpretation in euros: one (three) additional pension point(s) yields "an additional 9,600 (28,800) €" in total pension if collected over 20 years under current conditions. Fathers saw the same vignette framed as their partner receiving the points, otherwise identical in content.

We measured the labor supply response as a change: at the post-reform elicitation, we displayed each respondent's own without-reform answer beside the entry field, so that the object of the report was the adjustment itself (Appendix Figure A2).[3] Prior work finds that asking for a quantity outright (e.g., work hours), without a pre-filtering step of asking whether the respondent would work, can push respondents toward reporting a nonzero number (Crossley et al. 2025); we implemented this "filtering" efficiently by allowing respondents to check a "Do Not Intend to Work" box when reporting hours. Finally, as respondents entered their hours, the screen dynamically updated the implied monthly earnings, computed at their without-reform hourly wage. Importantly, this means the response should be read as a joint decision over hours and earnings.

A natural worry with any hypothetical elicitation is experimenter demand: that respondents infer which direction of response we expect and supply that, rather than reporting what they would actually do. We took several steps to limit this (Appendix B.3). To avoid selection on the topic, the landing page described the study only as research to "better understand human behavior," disclosing a pension focus only after the screening questions. The wording throughout was neutral, and the elicitation was symmetric—respondents could as easily enter more hours than their baseline as fewer, and nothing in the design pointed to a direction. The survey was anonymous and self-administered, so respondents answered privately rather than to an interviewer, mitigating social-desirability bias.

A total of 4,074 mothers and 2,034 fathers completed our surveys (Appendix B.4). We exclude respondents who fail an attention check, spend less than half the median time on the survey pages, or report working more than 80 hours per week—currently or at the five-year

[3] Appendix Figure A3 shows the analogous screens used to elicit labor supply absent the reform, before the vignette slideshow.

horizon, with or without the reform—and we exclude fathers whose partner is younger than 30. This leaves an analysis sample of 3,679 mothers and 1,927 fathers.[4]

Appendix Table A1 describes the two samples, starting with the mothers in Panel A. Relative to the mothers in Artmann et al. (forthcoming), mothers in our survey are slightly older, while still having similar fertility outcomes. 63 percent of mothers cohabit with a male partner; for such respondents in the supplemental wave, we analyze their partners' responses to the hypothetical reform. Compared to Artmann et al. (forthcoming), mothers in our sample are more educated (44 percent hold higher education qualifications, versus 22 percent) and display stronger labor market attachment (86 percent employed, versus 72 percent), with higher earnings even conditional on employment. Panels B and C report summary statistics for fathers and their partners, respectively. Partners of fathers are, on average, 1.9 years younger than mothers in our sample, but are otherwise comparable along demographic and work characteristics.

Our respondents display some familiarity with their pension situation (Appendix Figure A4). 50 percent of mothers and 59 percent of fathers report reading the annual statements from the German Pension Insurance informing about expected future pensions carefully; a further 40 percent and 32 percent, respectively, report having looked at them but not recalling their contents. Furthermore, 72 percent of mothers and 80 percent of fathers have some idea of their own pension entitlements. Unfamiliarity with the pension system is thus unlikely to drive any match or mismatch with the quasi-experimental estimate.

3 Results

We first describe how mothers and fathers expect to adjust their labor supply in response to the hypothetical reform, before translating the earnings response into an implied marginal propensity to earn out of unearned income that we can compare directly with the quasi-

[4] A programming error in the supplemental wave under-sampled some nonworking mothers. We reweight these mothers—2.5 percent of the mothers' sample—to their intended share in all reported estimates (Appendix B.4).

experimental estimate of Artmann et al. (forthcoming).

3.1 The hypothetical reform's effect on labor supply

We start with mothers in Table 1 Panel A. The first row reports changes at the total margin of labor supply: weekly work hours at the five-year horizon, including zeros for those who expect not to work. Columns 1 and 2 give the average expected level without and with the reform, respectively, and columns 3 and 4 the reform's effect in levels and as a percentage of the no-reform level. Mothers expect to work 26.1 hours per week absent the reform and 25.6 under it. The implied reduction of 0.49 hours—1.9 percent—is modest but precisely estimated.[5]

The second row of Panel A reports the extensive margin: mothers' expected probability of employment falls from 0.862 to 0.852, or 1.1 percent. The third row reports the total margin again, using monthly earnings in place of hours. Recall from Section 2 that the survey displayed the earnings implied by each mother's reported hours, holding hourly wage fixed at its no-reform level; the labor supply response is therefore a joint decision over hours and earnings. Reassuringly, using earnings gives a similar total margin: expected earnings fall from 2,547 to 2,496 euros, or 2.0 percent, against 1.9 percent for hours. Since earnings weight each mother's hours change by her wage, the close agreement implies that high-wage outliers are not driving the earnings response.[6]

Figure 1 examines heterogeneity across seven dimensions. The labor supply responses are negative in every subgroup and broadly similar in magnitude, so the income effect is not confined to any particular type of mother. Three sensible gradients nonetheless emerge. First, the response is larger for the larger pension increase, but less than threefold: at three

[5] Appendix Figure A5 plots the distribution of expected changes in hours and earnings, excluding the majority of respondents who report no change. Among those who adjust, changes run in both directions but are net negative.

[6] To see this, suppose half of mothers are high-wage (H) and half are low-wage (L), with $w_H > w_L$. The change in mean expected hours is $(\Delta h_H + \Delta h_L)/2$, which weights both groups equally. The corresponding change in mean expected earnings is $(w_H \Delta h_H + w_L \Delta h_L)/2$. Hence, because $w_H > w_L$, the earnings specification places greater weight on higher-wage mothers.

versus one pension point, expected weekly hours fall by 0.57 versus 0.42 and monthly earnings by 61.8 versus 40.5 euros. This suggests concavity, but the concavity is statistically insignificant: the one-point response is not different from one-third of the three-point response for either hours or earnings (Appendix Table A3). Second, the response is larger for mothers better placed to absorb it. Consistent with liquidity constraints, the reduction is concentrated among higher-earning households—0.67 hours, against a small and insignificant change among lower-earning ones—just as Artmann et al. find responses concentrated among mothers whose partners earn above the median or who have accumulated more non-childcare pension points, two proxies for liquidity constraints. Third, consistent with self-insurance against longevity risk, mothers expecting a longer retirement period and thus a larger pension change appear to also cut back more.

Next, we examine the anatomy of the labor supply response: whether mothers expect to stop working altogether, or to cut back while staying employed. Returning to Table 1, the lower portion of Panel A decomposes the total hours change, in a potential outcomes framework, into an extensive margin—movements into and out of employment—and an intensive margin—changes in hours among mothers who work in both scenarios. Let H_i^0 and H_i^1 denote mother i's expected weekly hours without and with the reform, set to zero if she expects not to work. We assign her entire change in hours to the extensive margin if her employment status differs across scenarios, and to the intensive margin if she works in both:

$$\text{Extensive}_i = \Delta H_i \, \mathbb{1}\{H_i^0 H_i^1 = 0\}, \tag{1}$$

$$\text{Intensive}_i = \Delta H_i \, \mathbb{1}\{H_i^0 H_i^1 > 0\}, \tag{2}$$

where $\Delta H_i = H_i^1 - H_i^0$, and mothers who expect to work in neither scenario ($H_i^0 = H_i^1 = 0$) contribute zero to both terms. Staub (2014) and Golosov et al. (2024, Online Appendix D) construct the same decomposition, imposing monotonicity and functional form or homogeneity assumptions because realized data leave the counterfactual unobserved; our survey

elicits both potential outcomes, and so avoids these assumptions.[7] Averaging Equations 1 and 2 across mothers gives two margins that sum to the total hours effect by construction: 73 percent of the 0.49 hour reduction operates on the extensive margin and the remainder on the intensive, both precisely estimated.

This is where we align least closely with Artmann et al. (forthcoming), who find an almost entirely intensive response: their employment effect is a precise zero. We offer two explanations. First, a stated preference may be easier to express as an exit than as an exact change in hours. Furthermore, our five-year-ahead elicitation compresses uncertainty about future employment into a point response. A mother may check the "do not intend to work" box while remaining open to marginal employment opportunities that become apparent closer to retirement, yielding an expected exit in the survey but lower realized hours in the administrative data. Notably, Georgarakos et al. (2025) recover a survey response that is 82 percent extensive, well above lottery studies that find similar MPEs in administrative data: 57 percent in Golosov et al. (2024), 24 to 40 percent in Cesarini et al. (2017), and no significant extensive margin effect in Picchio et al. (2018). Second, our mothers are closer to retirement when they adjust—they are on average 59 at the five-year horizon, against 54 for Artmann et al.'s mothers five years after the reform. This aligns with Becker et al. (2022), who study the same reform and find extensive-margin responses only from age 56, growing as women approach the early retirement age of 63.[8] That said, the limited response heterogeneity by age in Figure 1 speaks against this channel.

Next, Panel B turns to fathers, who were asked how their own labor supply would respond to their *partner* receiving the additional pension points. Fathers too expect to cut back their labor supply, by 1.7 to 2.0 percent of the without-reform average. Since only the mother receives the points, this is a spillover within the household—one that Artmann et al. (forthcoming) also find, though imprecisely estimated. Our spillover is as large as the

[7] Specifically, Staub uses a censored regression or selection model to recover the mean hours of those who change employment status and of those who work regardless; Golosov et al. assume homogeneity within quartiles of pre-win income to obtain the counterfactual earnings of those leaving employment.

[8] Becker et al. study only employment indicators and hence speak only to the extensive margin.

direct effect: fathers' proportional response is statistically indistinguishable from mothers'. Artmann et al.'s spillover effect is smaller, again consistent with our fathers being older and thus closer to retirement.[9] Similar to what we find for mothers, the decomposition loads on the extensive margin.

3.2 The effect on mothers' labor supply at the one-year horizon

We elicited mothers' expected labor supply responses at the one-year horizon as well as their predictions of their partners' response in the supplemental wave.[10] Table 2 reports the estimated effects at the two horizons side by side, using only the supplemental wave: Panel A at five years, Panel B at one year. This wave is far smaller than the pooled sample, so the estimates are less precise, but they remain informative.

Mothers expect to cut back their labor supply by roughly the same amount at both horizons: hours fall by 2.1 percent at five years versus 1.6 percent at one year, and earnings by 2.1 versus 1.9 percent. None of these differences is statistically significant. Artmann et al., by contrast, find that the response roughly doubles between the first 2.5 years after the reform and years 3 to 5, which they attribute to optimization frictions, since a frictionless model would predict an immediate response. One interpretation of our flatter profile is that such frictions are harder to capture in a survey, though our standard errors are too wide for us to draw a firm conclusion.

[9] Appendix Figure A6 examines heterogeneity across six dimensions; older fathers appear to cut back labor supply roughly twice as much as younger ones.

[10] In the first wave, we had asked mothers about their expected labor supply response one year after the reform; however, the question was posed as a change from their current hours rather than from what they expected in a year absent the reform. Unfortunately, without defining an explicit counterfactual, we were not able to readily isolate the reform's effect from other changes respondents anticipated over the year. This was evident in respondents' open-ended responses. We addressed this by introducing an explicit counterfactual for the one-year horizon (similar to that for the five-year horizon) in the supplemental wave. In addition, the main wave did not ask mothers to predict their partners' responses.

3.3 Spousal predictions of partners' labor supply responses

A distinctive feature of our design is that our separate surveys of mothers and fathers let us compare each spouse's expectation of the *partner's* response with the other group's own reported response—fathers' predictions of mothers' expected labor supply response against mothers' own reports, and mothers' predictions of fathers' against fathers' own reports. Because the predictions and the self-reports come from separate samples rather than the same couples, the comparison is between aggregate distributions rather than within-couple.

This comparison is depicted in Figure 2. For each outcome, the figure plots the mean predicted spousal response and the mean self-reported response—each drawn from the relevant survey of mothers or fathers—together with their 95% confidence intervals, expressed as a percentage of the outcome's pre-reform mean. For mother's labor supply (Panel A), fathers' expectations of their partner's response sit close to, though slightly above, mothers' own reports, with overlapping confidence intervals for hours, employment, and earnings. For father's labor supply (Panel B), the point estimates nearly coincide, though mothers' predictions are less precise. In the aggregate, spouses' expectations of their partners' responses track the partner group's own report quite closely.[11]

This congruence speaks to a literature on beliefs and information within the household that emphasizes frictions, with spouses often holding misaligned expectations about shared facts (Fehr et al. 2024; Delavande et al. 2025). That work concerns how information about such facts is transmitted between spouses; we instead study how one spouse expects the other to respond to a reform. To our knowledge, ours is the first survey to elicit both partners' responses to the same policy and compare them. We find that, on average, spouses anticipate one another's labor supply responses well, suggesting that subjective expectations can be usefully elicited by proxy.

Finally, the same comparison helps address concerns about experimenter demand. Be-

[11] Panel A uses the main wave of mothers, fielded contemporaneously with the fathers' survey. Panel B uses the supplemental wave, the only wave where we asked mothers about their partners.

cause we do not conceal the surveys' purpose, a father told that his partner receives the pension points and then asked what he would do might infer that some response is expected. By contrast, mothers are less subject to this cue when predicting their partners, since they are asked to predict their spouse's behavior and do so only after reporting their own. That mothers' predictions nonetheless match fathers' own reports suggests that the expected spillover to fathers' labor supply is not an artifact of experimenter demand.

3.4 The implied marginal propensity to earn out of unearned income for mothers

The labor supply responses above are specific to our hypothetical reform. To compare them against the quasi-experimental benchmark, or against income effects estimated in other settings, we translate the earnings response into the marginal propensity to earn (MPE)—the change in earnings per additional euro of unearned income. Because it normalizes by the wealth gain that the additional points represent—which in our case depends also on how near a mother is to the pension and how she discounts the future, the MPE is comparable across studies that differ in reform, population, and setting. We can then set our estimate beside Artmann et al.'s (forthcoming) forward-looking estimate, and beside contemporaneous estimates from lotteries, inheritances, and cash transfers (Joulfaian and Wilhelm 1994; Imbens et al. 2001; Cesarini et al. 2017; Picchio et al. 2018; Golosov et al. 2024; Gromadzki 2024).

Table 3 reports the implied MPEs: Panel A reproduces Artmann et al.'s (forthcoming) quasi-experimental benchmark, Panel B our survey-based estimates. Throughout, we follow a similar approach in adopting the annuitization method of Golosov et al. (2024).

In Row 1 of Panel B, we use a straightforward application of Artmann et al.'s (forthcoming) MPE calculation to our earnings estimate from Table 1. First, we compute the present value (PV) gain in lifetime unearned income from an increase of $\Delta P = 2$ pension points—the average of the one- and three-point increases we show respondents. In early

2026, each point is worth $R = 40.79$ euros of monthly pension, paid over a 240-month (20-year) horizon—approximately the average woman's remaining life expectancy at retirement age 67, following the calculation in Artmann et al.. In PV, the extra pension is worth

$$W = \Delta P \cdot R \cdot (\beta(1+g))^J \sum_{m=0}^{239} (\beta(1+g))^m, \tag{3}$$

where J is the number of months until retirement, g is a monthly growth rate of the real value of a pension point computed by Artmann et al., and β is the monthly discount factor. On the right-hand side of Equation 3, the sum is the value at retirement of one euro of monthly pension drawn for 240 months, which the leading $(\beta(1+g))^J$ carries back to the present. Next, following Artmann et al., we spread this PV wealth gain evenly over the J months to retirement and assign our five-year (60-month) window the share $W_5 = W \cdot 60/J$. Finally, we cumulate our estimated monthly earnings response ΔE over the same window, $\Delta E \cdot \sum_{m=1}^{60} \beta^m$, and compute the MPE as the ratio of this cumulated earnings change to the allocated gain W_5.

Strikingly, this yields an MPE of -0.65 (Panel B, row 1), close to Artmann et al.'s (forthcoming) -0.54 (Panel A)—even though our estimated earnings response exceeds theirs, 2.0 percent (Table 1) versus 1.3 percent. Our mothers are nearer retirement, so the same points are discounted less and smoothed over fewer pre-retirement months, delivering a larger windfall in PV terms. Normalizing the response by that windfall returns essentially the same MPE, so the stronger raw response reflects the size of the wealth gain rather than a greater sensitivity to it.

The remaining rows probe robustness to the calculation's main assumptions. Row 2 supposes respondents take the vignette's euro figures at face value: the slides present the additional pension as a flat euro total collected over 20 years, inviting respondents neither to discount the future euros nor to recognize the slowly eroding real value of a pension point.[12]

[12]The pension level's slow erosion relative to wages was a topic of public debate around the time of our survey.

The perceived wealth gain is then larger, so the implied MPE is smaller in magnitude, at −0.42.

Our headline MPE is a ratio of averages—the aggregate earnings response over the aggregate wealth gain—following Artmann et al. (forthcoming) and standard practice for aggregate income effects. Alternatively, we could compute the MPE separately for mothers shown a one- and a three-point increase and average the two, yielding an average MPE of −0.69 (row 3).

Our main estimate applies the five-year earnings response to all five years, as the one- and five-year responses differ little (Table 2). By contrast, Artmann et al. (forthcoming) find the response roughly doubling between the first 2.5 years and years 3 to 5, consistent with optimization frictions that may be difficult to reproduce in a hypothetical survey. To gauge robustness, row 4 of Table 3 imposes Artmann et al.'s profile on our response—the first 2.5 years at about half the year-5 effect, years 3 to 5 at the full effect—and yields an MPE of −0.49.

A survey can supply inputs that administrative data cannot. Here, the MPE's denominator depends on how long the pension is drawn, which we can take from the respondents' own mortality expectations—shown to carry private information beyond that in life tables (Hurd and McGarry 2002; Perozek 2008; Hurd 2009). We asked mothers their chances of surviving to ages 80 and 85 and, following Perozek (2008), fit a Gompertz model to these expectations (Appendix Table A2, column 1; column 2 applies Perozek's cleaning procedure, with similar results but fewer observations).[13] The fit implies a mean life expectancy of 84.6 years; row 5 substitutes this for the 20-year horizon and yields −0.72.

Our headline MPE applies the average two-point increase to the pooled sample. Because Artmann et al. (forthcoming)'s reform corresponds most closely to a one pension point increase, one might instead use the one-point arm alone. However, the MPE should not

[13]The coefficient α puts the subjective mortality rate at the mothers' current age at a sensible 4 per 1,000 person-years, and β implies it rises about 9 percent per year of age, the pace at which adult human mortality is observed to increase.

depend on the increment: we cannot reject proportionality of the one- and three-point responses, as noted in Section 3.1, so the implied MPEs should be similar. Indeed, the two MPEs are statistically indistinguishable ($p = 0.166$), and neither differs significantly from Artmann et al. (forthcoming)'s estimate. In any case, because our mothers earn more than Artmann et al.'s, their one-point reform represents a larger wealth shock relative to earnings than our one-point arm, so matching the increment alone does not make that arm uniquely comparable. We therefore pool the arms to retain all respondents and estimate the common MPE more precisely.

The final row of Table 3 provides the specification most closely aligned with both Artmann et al. (forthcoming)'s sample and treatment. We reweight our sample on earnings to match theirs and restrict it to the one-point arm, yielding an MPE of -0.52 (statistically insignificant). Its point estimate remains close to our headline -0.65. Appendix C extends this comparison by reweighting the full sample to Artmann et al.'s earnings distribution and re-estimating all of our main results. Taken together, the MPE estimates in this section and Appendix Table C3 remain of similar magnitude to Artmann et al. (forthcoming)'s estimate across alternative assumptions and samples.

4 Conclusion

We set out to assess whether stated responses to a hypothetical policy can recover labor supply parameters that are otherwise identified only from realized reforms. Using surveys of German mothers and fathers built around a hypothetical *Mütterrente* reform, we document a close correspondence. Mothers expect to reduce their labor supply in response to the additional pension points, and the implied marginal propensity to earn, computed under the same assumptions as Artmann et al. (forthcoming), is -0.65, close to their quasi-experimental estimate of -0.54. The two estimates draw on very different variation—stated responses to a hypothetical reform and administrative data on a realized one—yet are similar in magnitude.

The survey also recovers features of the response that are hard to obtain from administrative data. The reduction in labor supply is sensible across subgroups and scales with the size of the pension increase. It operates mainly on the extensive margin—the one point on which we depart from Artmann et al. (forthcoming), whose response is almost entirely intensive. We cannot rule out that this is partly an artifact of stated-preference elicitation, a possibility consistent with the largely extensive-margin response that Georgarakos et al. (2025) recover from survey-based windfall questions; at the same time, Becker et al. (2022) document genuine extensive-margin responses among older mothers in administrative data. Distinguishing these explanations is a useful direction for future work.

Finally, our two-survey design lets us look inside the household. Fathers expect to cut their own labor supply when their partner receives the pension points, which is evidence of intra-household spillovers in labor supply decisions. Furthermore, household members hold accurate beliefs about each other's behavioral adjustments. Taken together, these results suggest that hypothetical-scenario surveys—at least when they are context-specific and the underlying policy is easily understood—can approximate the behavioral responses recovered from quasi-experimental variation, and are a promising and flexible tool for forecasting the direction and overall magnitude of labor supply responses to pension and other reforms.

References

Artmann, Elisabeth, Nicola Fuchs-Schündeln, and Giulia Giupponi, "Forward-looking labor supply responses to changes in pension wealth: Evidence from Germany," *American Economic Journal: Applied Economics*, forthcoming.

Bardasi, Elena, Kathleen Beegle, Andrew Dillon, and Pieter Serneels, "Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania," *The World Bank Economic Review*, 2011, *25* (3), 418–447.

Becker, Sebastian, Hermann Buslei, Johannes Geyer, and Peter Haan, "The effect of pension wealth on employment," 2022.

Bhandari, Anmol, Serdar Birinci, Ellen R. McGrattan, and Kurt See, "What Do Survey Data Tell Us about US Businesses?," *American Economic Review: Insights*, 2020, *2* (4), 443–458.

Boehm, Johannes, Etienne Fize, and Xavier Jaravel, "Five facts about MPCs: Evidence from a randomized experiment," *American Economic Review*, 2025, *115* (1), 1–42.

Bound, John, Charles Brown, and Nancy Mathiowetz, "Measurement Error in Survey Data," in James J. Heckman and Edward Leamer, eds., *Handbook of Econometrics*, Vol. 5, Elsevier, 2001, pp. 3705–3843.

Cesarini, David, Erik Lindqvist, Matthew J Notowidigdo, and Robert Östling, "The effect of wealth on individual and household labor supply: evidence from Swedish lotteries," *American Economic Review*, 2017, *107* (12), 3917–3946.

Chiappori, Pierre-André, "Collective Labor Supply and Welfare," *Journal of Political Economy*, 1992, *100* (3), 437–467.

Christelis, Dimitris, Dimitris Georgarakos, Tullio Jappelli, and Geoff Kenny, "Wealth shocks and portfolio choice," *Journal of Monetary Economics*, 2025, *149*, 103632.

Crossley, Thomas F, Paul Fisher, Peter Levell, and Hamish Low, "Eliciting the marginal propensity to consume in surveys," IFS Working Papers W25/25, Institute for Fiscal Studies 2025.

Delavande, Adeline, Gizem Koşar, and Basit Zafar, "Information Spillovers within Couples: Evidence from a Sequential Survey of Spouses," NBER Working Paper 33879, National Bureau of Economic Research 2025.

Dillon, Andrew and Edouard Mensah, "Respondent Biases in Agricultural Household Surveys," *Journal of Development Economics*, 2024, *166.* Add article number from your .bib.

Estrada, Ricardo, Daniela Goyheix, and María Lombardi, "Money and Gender: Proxy Respondents and the Mismeasurement of Income in Surveys," *Journal of Development Economics*, 2027, *184*, 103879.

Fehr, Dietmar, Johanna Mollerstrom, and Ricardo Perez-Truglia, "Listen to Her: Gender Differences in Information Diffusion within the Household," *Journal of Public Economics*, 2024, *239*, 105213.

Fetter, Daniel K and Lee M Lockwood, "Government old-age support and labor supply: Evidence from the old age assistance program," *American Economic Review*, 2018, *108* (8), 2174–2211.

Georgarakos, Dimitris, Tullio Jappelli, Geoff Kenny, and Luigi Pistaferri, "Labor supply response to windfall gains," *Journal of Public Economics*, 2025, *250*, 105476.

Golosov, Mikhail, Michael Graber, Magne Mogstad, and David Novgorodsky, "How Americans respond to idiosyncratic and exogenous changes in household wealth and unearned income," *The Quarterly Journal of Economics*, 2024, *139* (2), 1321–1395.

Gromadzki, Jan, "Labor supply effects of a universal cash transfer," *Journal of Public Economics*, 2024, *239*, 105248.

Hurd, Michael D., "Subjective probabilities in household surveys," *Annual Review of Economics*, 2009, *1* (1), 543–562.

Hurd, Michael D and Kathleen McGarry, "The predictive validity of subjective probabilities of survival," *The Economic Journal*, 2002, *112* (482), 966–985.

Imbens, Guido W, Donald B Rubin, and Bruce I Sacerdote, "Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players," *American economic review*, 2001, *91* (4), 778–794.

Jappelli, Tullio and Luigi Pistaferri, "Reported MPC and unobserved heterogeneity," *American Economic Journal: Economic Policy*, 2020, *12* (4), 275–297.

Joulfaian, David and Mark O Wilhelm, "Inheritance and labor supply," *Journal of Human Resources*, 1994, pp. 1205–1234.

Masselus, Lise and Nathan Fiala, "Whom to Ask? Testing Respondent Effects in Household Surveys," *Journal of Development Economics*, 2024, *168.* Add article number from your .bib.

Mellow, Wesley and Hal S. Sider, "Accuracy of Response in Labor Market Surveys: Evidence and Implications," *Journal of Labor Economics*, 1983, *1* (4), 331–344.

Parker, Jonathan A. and Nicholas S. Souleles, "Reported Effects versus Revealed-Preference Estimates: Evidence from the Propensity to Spend Tax Rebates," *American Economic Review: Insights*, 2019, *1* (3), 273–290.

Perozek, Maria, "Using subjective expectations to forecast longevity: Do survey respondents know something we don't know?," *Demography*, 2008, *45* (1), 95–113.

Picchio, Matteo, Sigrid Suetens, and Jan C van Ours, "Labour supply effects of winning a lottery," *The Economic Journal*, 2018, *128* (611), 1700–1729.

Shapiro, Matthew D and Joel Slemrod, "Did the 2008 tax rebates stimulate spending?," *American Economic Review*, 2009, *99* (2), 374–379.

Staub, Kevin E, "A causal interpretation of extensive and intensive margin effects in generalized Tobit models," *Review of Economics and Statistics*, 2014, *96* (2), 371–375.

Table 1: Effect of hypothetical reform on expected labor supply at the five-year horizon

	In 5 years, expected labor market outcome...			
	Without reform (1)	With reform (2)	Difference (3)	As % of (1) (4)
	Panel A: Mothers (N=3,677)			
Hours worked	26.1	25.6	-0.495***	-1.90***
			(0.099)	(0.38)
Prob. of employment	0.861	0.852	-0.00981***	-1.14***
			(0.00322)	(0.37)
Earnings	2545.7	2494.5	-51.2***	-2.01***
			(9.1)	(0.36)
Decomposition of hours:				
Extensive margin			-0.359***	
			(0.089)	
			[72.6%]	
Intensive margin			-0.136***	
			(0.044)	
			[27.4%]	
	Panel B: Fathers (N=1,927)			
Hours worked	33.1	32.4	-0.655***	-1.98***
			(0.136)	(0.41)
Prob. of employment	0.852	0.837	-0.0145***	-1.71***
			(0.0034)	(0.40)
Earnings	4416.7	4340.5	-76.1***	-1.72***
			(17.5)	(0.40)
Decomposition of hours:				
Extensive margin			-0.522***	
			(0.123)	
			[79.7%]	
Intensive margin			-0.133**	
			(0.059)	
			[20.3%]	

Notes: The sample comprises mothers (Panel A) and fathers (Panel B), excluding respondents who failed the attention check, completed the survey in less than half the median response time, or reported more than 80 weekly hours currently or five years ahead (with or without reform), and excluding fathers whose partners are younger than 30. Supplemental-wave nonworking mothers are reweighted to their intended share (Appendix B.4). The top of each panel reports the average expected outcome at the five-year horizon absent the reform (column 1) and under the hypothetical reform (column 2), together with the reform's effect expressed as a level change (column 3) and as a percentage of the no-reform level (column 4). The lower rows decompose the total hours effect by assigning each respondent's change in hours to the extensive margin if her employment status differs with and without the reform, and to the intensive margin if she works in both, with each margin's percentage share in brackets. Hours are measured per week, and earnings are monthly earnings in current euros. Standard errors in parentheses, robust in column 3 and computed by the delta method in column 4. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 2: Supplemental-wave mothers: Effect of hypothetical reform on expected labor supply at the five- and one-year horizons

	Expected labor market outcome...			
	Without reform (1)	With reform (2)	Difference (3)	As % of (1) (4)
Panel A: 5-year horizon (N=911)				
Hours worked	25.6	25.1	-0.536**	-2.09**
			(0.242)	(0.94)
Prob. of employment	0.829	0.824	-0.00487	-0.587
			(0.00801)	(0.964)
Earnings	2569.3	2514.3	-55.0**	-2.14***
			(21.4)	(0.83)
Decomposition of hours:				
Extensive margin			-0.361	
			(0.225)	
			[67.4%]	
Intensive margin			-0.174*	
			(0.090)	
			[32.6%]	
Panel B: 1-year horizon (N=835)				
Hours worked	28.8	28.4	-0.447*	-1.55*
			(0.239)	(0.82)
Prob. of employment	0.925	0.919	-0.00585	-0.633
			(0.00712)	(0.768)
Earnings	2750.6	2697.7	-52.9**	-1.92**
			(23.4)	(0.84)
Decomposition of hours:				
Extensive margin			-0.248	
			(0.225)	
			[55.6%]	
Intensive margin			-0.198**	
			(0.081)	
			[44.4%]	

Notes: The sample comprises supplemental-wave mothers, excluding respondents who failed the attention check, completed the survey in less than half the median response time, or reported more than 80 weekly hours currently or five years ahead (with or without reform); Panel B additionally excludes respondents reporting more than 80 weekly hours at the one-year horizon. Panel A reports predictions at the five-year horizon and Panel B at the one-year horizon. Nonworking mothers are reweighted to their intended share (Appendix B.4). The top of each panel reports the average expected outcome absent the reform (column 1) and under the hypothetical reform (column 2), together with the reform's effect expressed as a level change (column 3) and as a percentage of the no-reform level (column 4). The lower rows decompose the total hours effect by assigning each respondent's change in hours to the extensive margin if her employment status differs with and without the reform, and to the intensive margin if she works in both, with each margin's percentage share in brackets. Hours are measured per week, and earnings are monthly earnings in current euros. Standard errors in parentheses, robust in column 3 and computed by the delta method in column 4. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$.

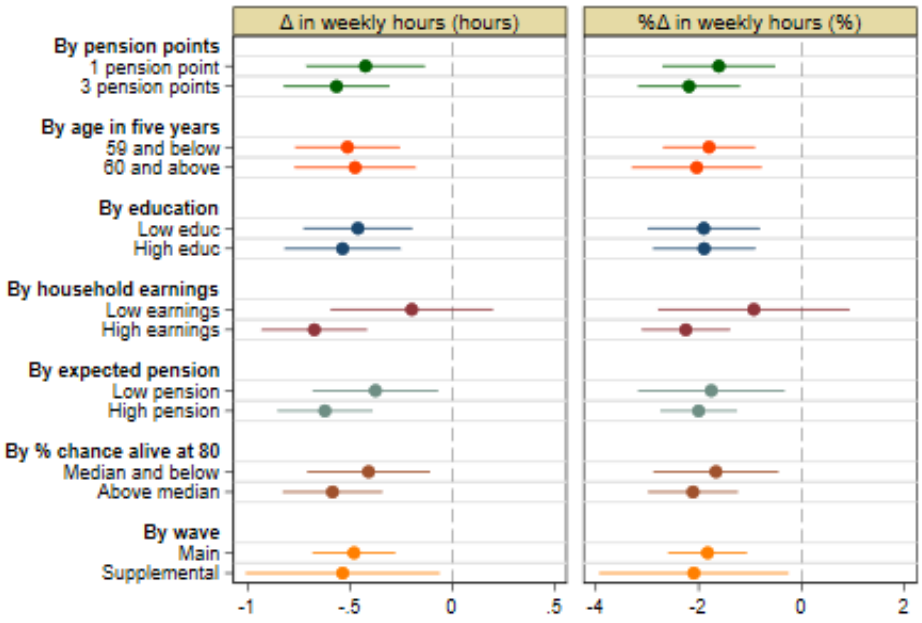
Table 3: Implied marginal propensity to earn under alternative assumptions

S/N	Specification	Description	MPE
		Panel A: Quasi-experimental benchmark	
	Artmann et al. (forthcoming)	Their reported estimate. Main assumptions: about 2 percent annual discount rate; real pension-point value declines slightly over time; collect pension for 20 years.	−0.54
		Panel B: MPE estimates based on our survey	
1	Preferred specification	Artmann's present-value formula, discount rate, and pension point-value change, applied to our main five-year earnings response (Table 1). Main difference: months to retirement based off the sample mean age.	−0.65*** (0.12) [−0.88, −0.42]
2	Face-value framing	Assume that respondents value the additional points at the undiscounted 20-year amount shown on the reform slides—no discounting or pension point-value change over time.	−0.42*** (0.07) [−0.57, −0.27]
3	By-arm average	Compute the MPE separately for the one- and three-pension-point arms and average the two.	−0.78*** (0.18) [−1.14, −0.42]
4	Artmann et al. time profile	Impose Artmann et al.'s labor supply response time profile: scale the first 2.5 years to about half the year-5 effect, holding years 3–5 at the full effect.	−0.49*** (0.09) [−0.66, −0.32]
5	Subjective life expectancy	Replace the 20-year payout horizon with the 17.6 years implied by respondents' subjective mean life expectancy of 84.6 (Appendix Table A2).	−0.72*** (0.13) [−0.98, −0.46]
6	Reweighted on earnings, 1-pension-point arm	Reweight the sample to match Artmann et al.'s earnings distribution, and use only the one-pension-point arm.	−0.52 (0.37) [−1.25, 0.21]

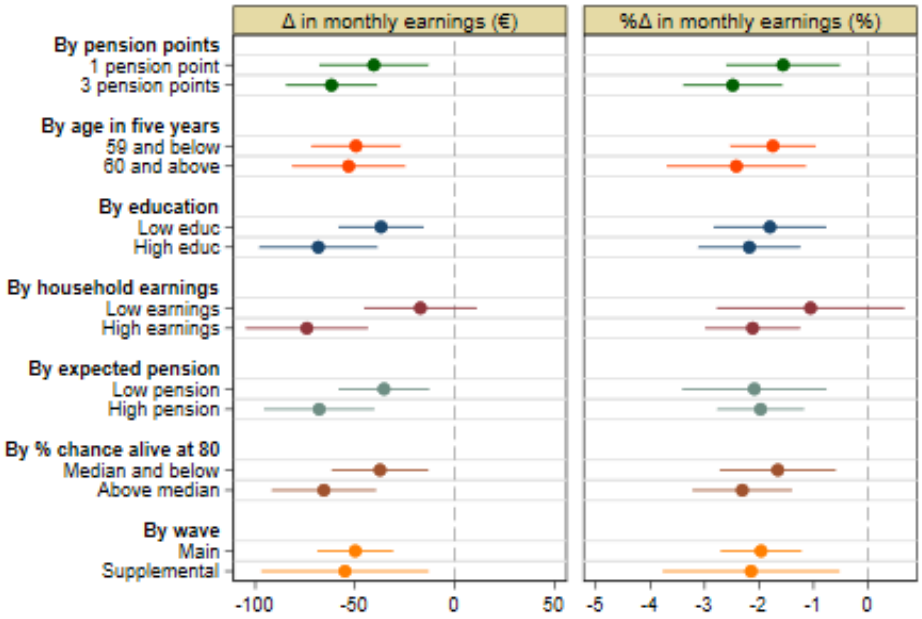
Notes: The table reports the implied marginal propensity to earn out of unearned income (MPE) for mothers. Standard errors computed by the delta method in parentheses unless otherwise stated. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$. Panel A reproduces the quasi-experimental MPE estimate of Artmann et al. (forthcoming). In Panel B, our preferred specification (row 1) uses the same assumptions as Artmann et al. (forthcoming) but updates the pension-point value to its July 2025–June 2026 level (40.79 € per month) and takes the months to retirement from the sample mean age, following the annuitization procedure described in Section 3. The remaining rows of Panel B show robustness specifications with the specified deviation compared to row 1. Panel B, row 2: In the vignette slides, respondents are told that one (three) additional pension point(s) corresponds to 9,600 (28,800) euros if collected over 20 years. This row assumes they take that euro value literally, without adjusting for discounting. Panel B, row 3: We compute the MPE separately for the one- and three-pension-point arms and report their simple average. Panel B, row 4: We rescale our five-year earnings response by Artmann et al.'s short-run/medium-run profile, so that the first 2.5 years carry 52% of the year-5 monthly effect and years 3–5 carry the full effect. Panel B, row 5: We replace the 20-year pension-collection horizon with respondents' subjective mean life expectancy—84.6 years, or 17.6 years past the retirement age of 67 (Appendix Table A2). Standard errors computed based on 2,000 bootstrap replications to capture the added uncertainty in life expectancy.

Figure 1: Heterogeneity in the effect on work hours and earnings, mothers

(a) Weekly work hours



(b) Monthly earnings



Notes: The figure plots the hypothetical reform's effect on mothers' expected weekly work hours (top) and monthly earnings (bottom) at the five-year horizon for each specified subsample, each expressed as a level change (left) and as a percentage of the no-reform level (right). The sample excludes respondents who failed the attention check, completed the survey in less than half the median response time, or reported more than 80 weekly hours currently or five years ahead (with or without reform). Supplemental-wave nonworking mothers are reweighted to their intended share (Appendix B.4). Horizontal bars are 95% confidence intervals computed from robust (left panels) or delta-method (right panels) standard errors. *Lower secondary and below* includes those whose highest schooling degree is none, lower secondary (*Hauptschulabschluss*), or intermediate secondary (*Realschulabschluss*); *upper secondary and above* includes those holding a university of applied sciences entrance qualification (*Fachhochschulreife*), a general university entrance qualification (*Abitur*), or another degree.

Figure 2: Spousal versus own labor supply predictions

(a) Mothers' labor supply



(b) Fathers' labor supply



Notes: Each panel plots the hypothetical reform's effect at the five-year horizon on the labor supply outcomes specified, expressed as a percentage of the pre-reform mean. Panel A compares fathers' expectations of the mother's response with mothers' own reported response; we use the main wave of mothers in this panel because it was fielded contemporaneously with the fathers' survey. Panel B compares mothers' expectations of the father's response with fathers' own reported response, using the supplemental wave of mothers, for whom we additionally elicited expectations about their partners. Horizontal bars are 95% confidence intervals computed by the delta method.

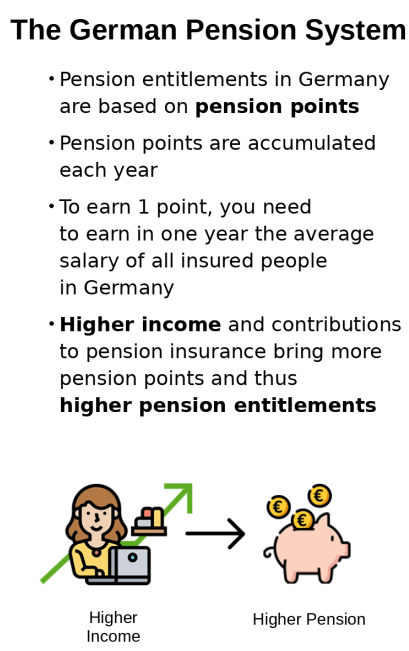
Online Appendix for Assessing the Use of Hypothetical Scenarios to Estimate Labor Supply Responses: Evidence from Pension Reform

By Nicola Fuchs-Schündeln, Jonas Jessen, Pinchuan Ong and Jessica Pan

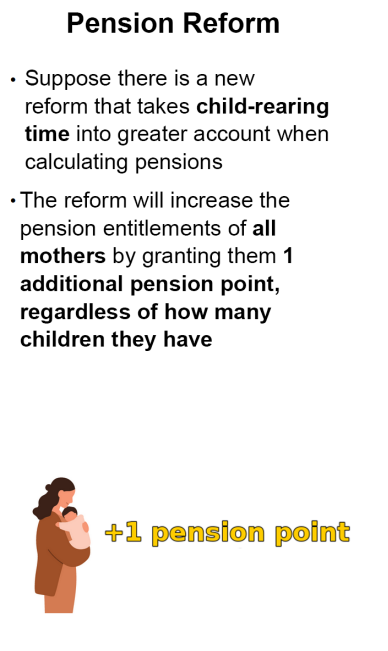
A Appendix tables and figures

Appendix Figure A1: Slides describing the hypothetical reform, shown to mothers

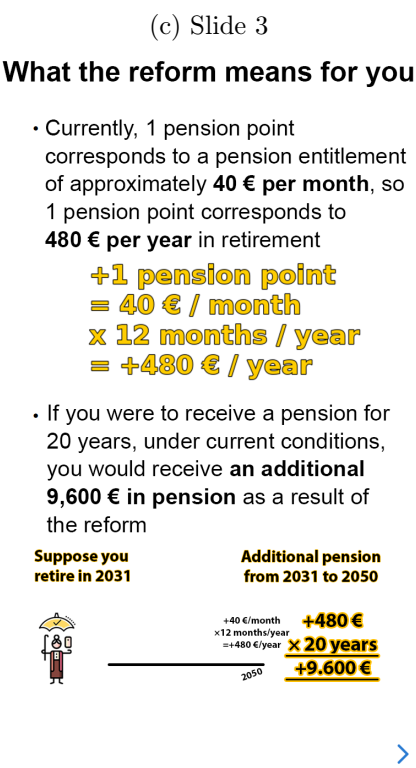
(a) Slide 1



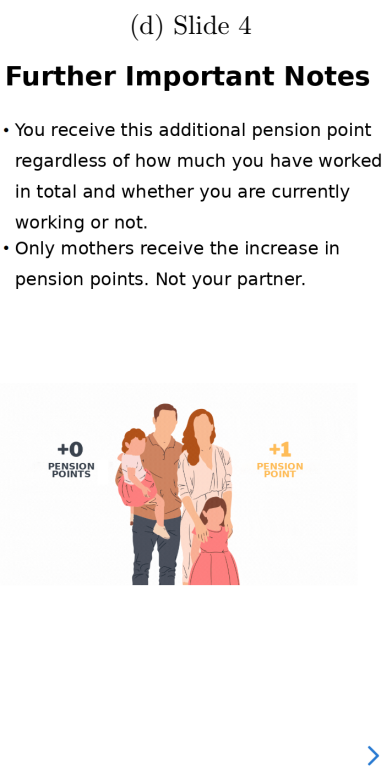
(b) Slide 2



(c) Slide 3

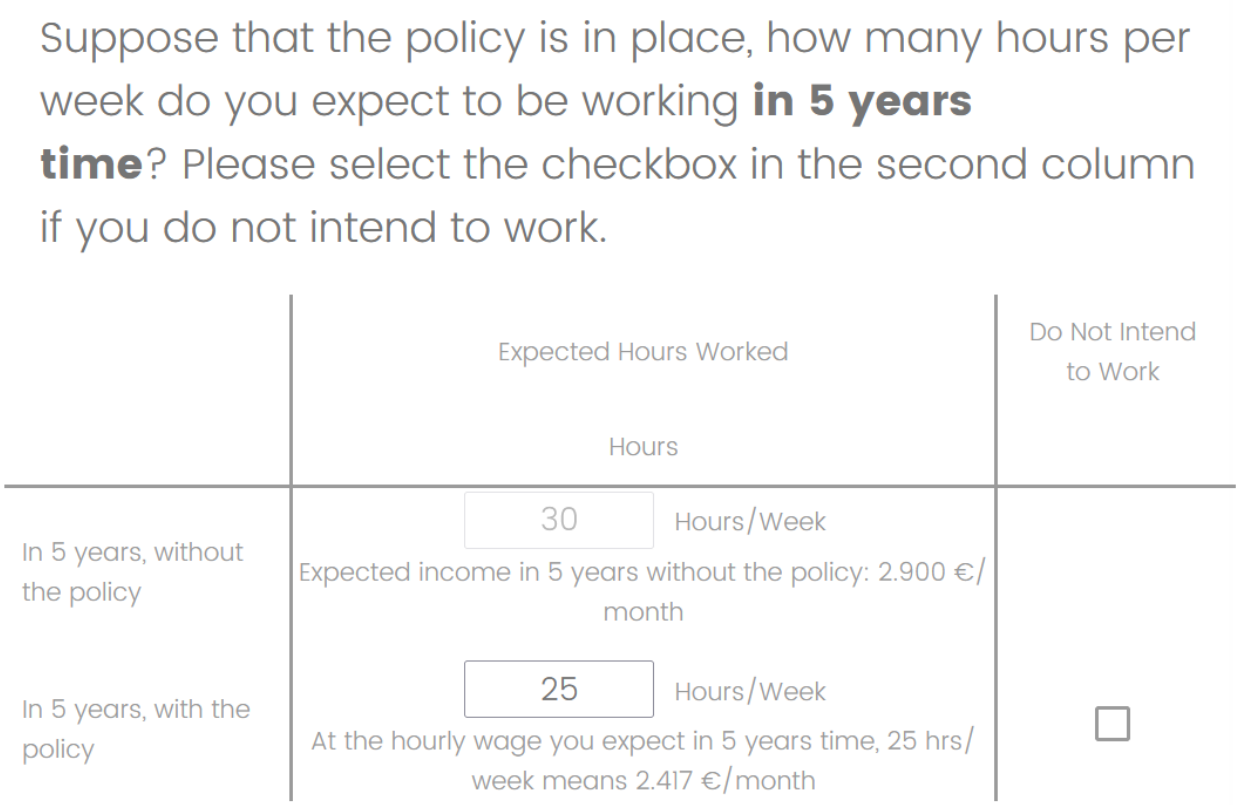


(d) Slide 4



Notes: Figure shows the 4 main vignette slides shown to mothers. A fifth slide informing respondents of the end of the vignette slideshow is omitted. Slides for fathers were phrased from their perspective of their partners receiving the additional pension points but otherwise identical. Respondents were randomly shown 1 or 3 pension point increases (with corresponding monetary value changes). Slides were shown in German; the figure reproduces our English translation.

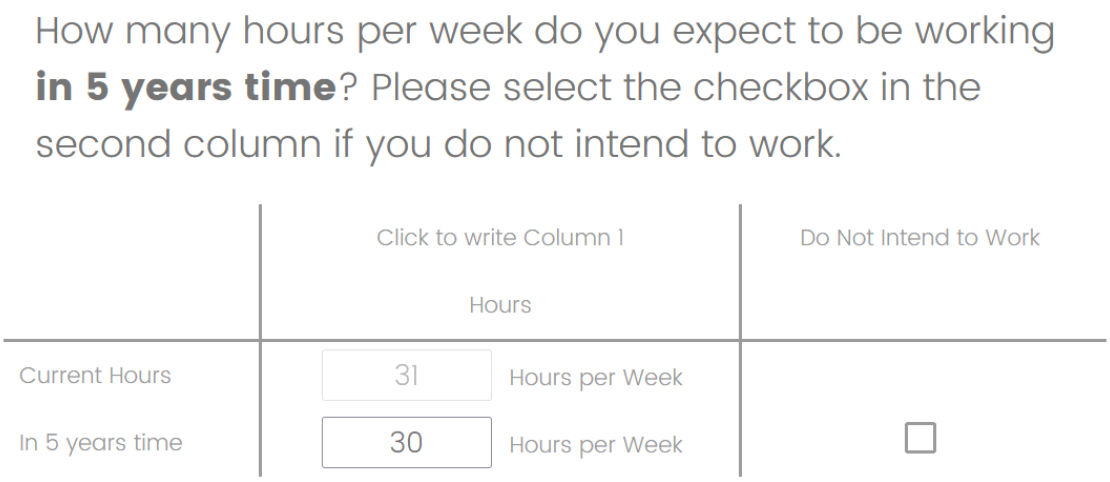
Appendix Figure A2: Elicitation of response to the hypothetical reform



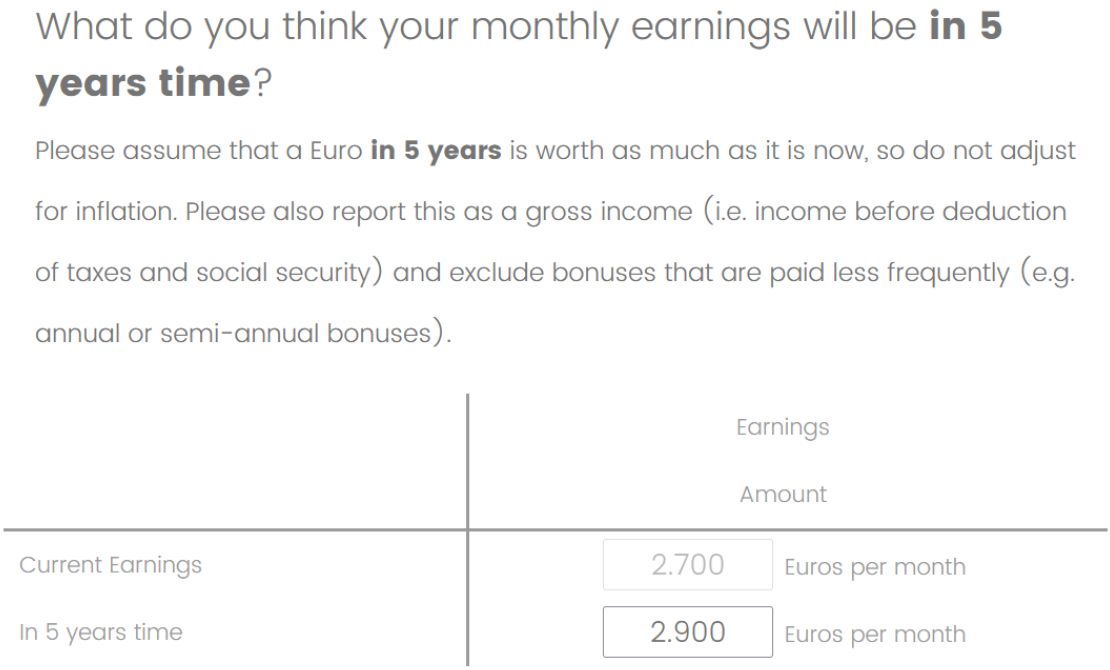
Notes: The figure shows the question used to elicit the labor supply response to the hypothetical reform. The top blank was not editable and showed the respondent's previous report of her 5-years-ahead work expectations, elicited before we described the policy reform. The expected income directly underneath was based on a similar elicitation. Respondents filled in the lower blank. The monthly earnings directly underneath the lower blank changed dynamically with the response, with the earnings computed assuming that the hourly wage was constant. If the respondent checked the "Do Not Intend to Work" box on the right, the blank was set to an uneditable zero.

Appendix Figure A3: Labor-supply elicitation screens absent the reform, shown before the vignette slideshow

Panel A: Weekly hours



Panel B: Monthly earnings

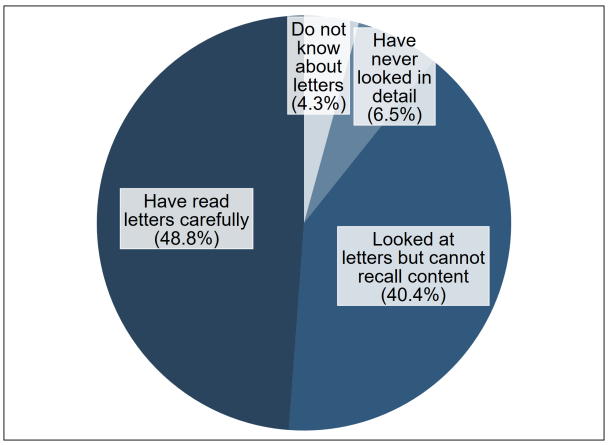


Notes: The figure shows the two screens used to elicit expected labor supply at the five-year horizon, before the hypothetical reform was described: weekly hours (Panel A) and monthly gross earnings (Panel B). In each screen, respondents entered a value in the "In 5 years time" field, with their current value shown above for reference in a non-editable, grayed-out box. On the hours screen (Panel A), respondents who entered positive hours advanced to the earnings screen (Panel B); those who instead checked the "Do Not Intend to Work" box had their hours set to an uneditable zero and, on the next screen, were asked what they expected to do instead. From both follow-up screens, respondents could navigate back to revise their hours or work decision.

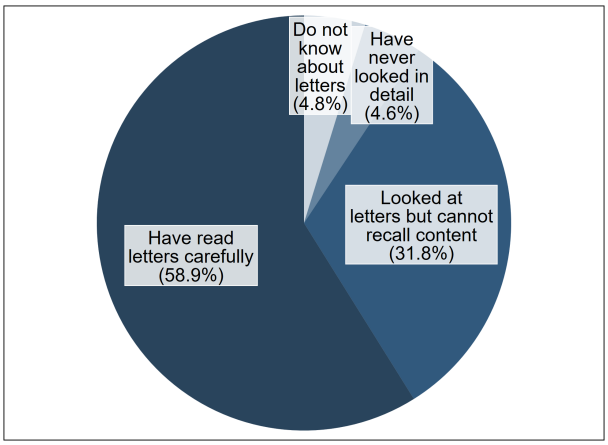
Appendix Figure A4: Pension knowledge

Panel A: Annual letters from the German Pension Insurance

(a) Mothers

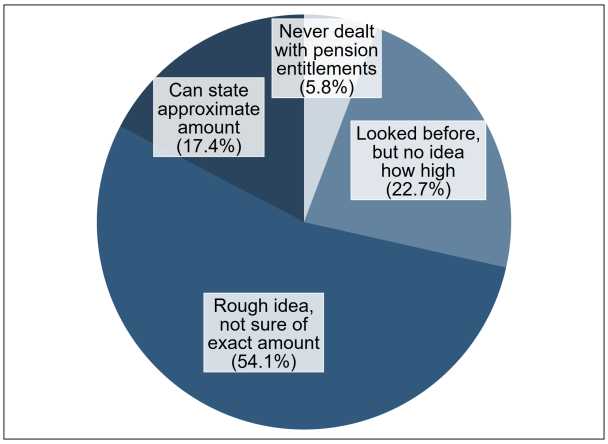


(b) Fathers

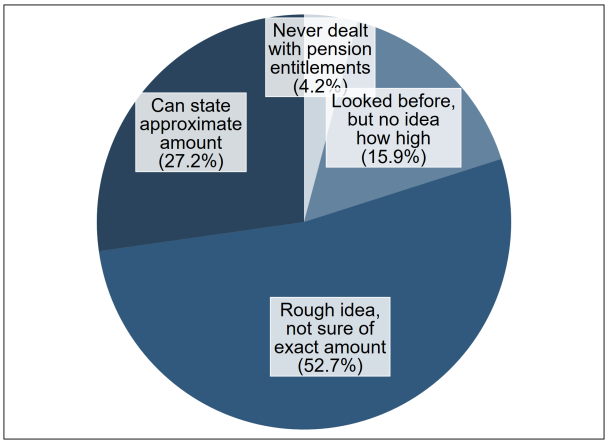


Panel B: Knowledge of own pension entitlements

(c) Mothers



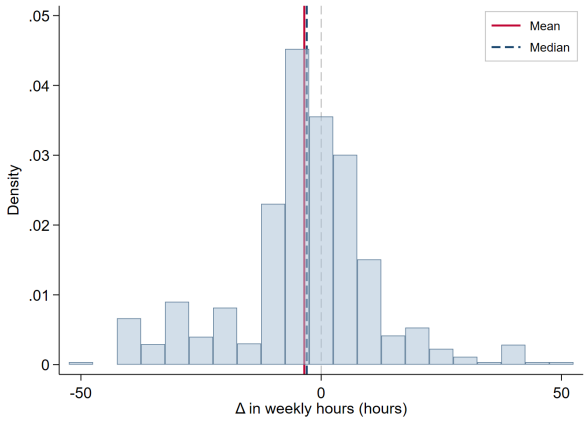
(d) Fathers



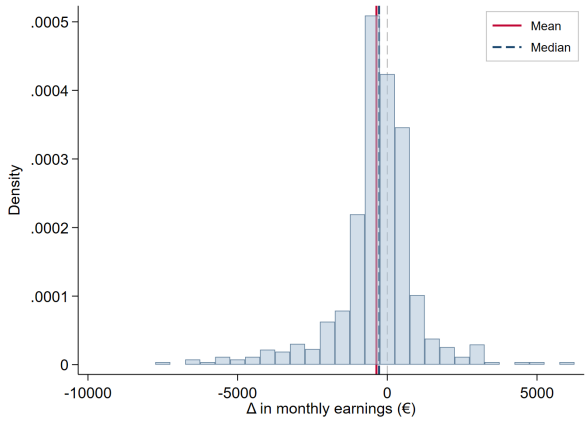
Notes: Panel A shows responses to a question on whether respondents read or knows about the annual pension statements sent by the German Pension Insurance. Panel B shows responses to a question on how well informed respondents are about their own pension entitlements.

Appendix Figure A5: Distribution of expected 5-year-horizon labor supply effects, excluding non-movers

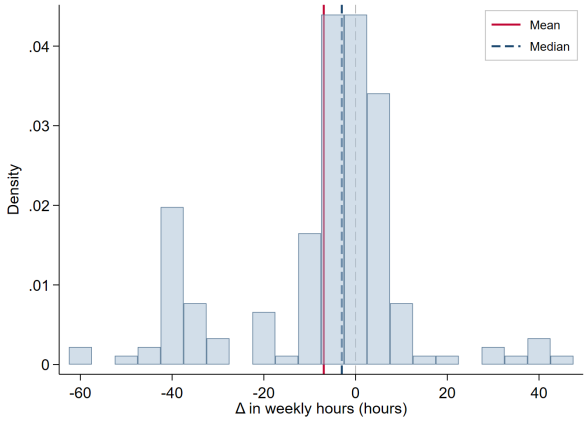
(a) Δ Hours, mothers



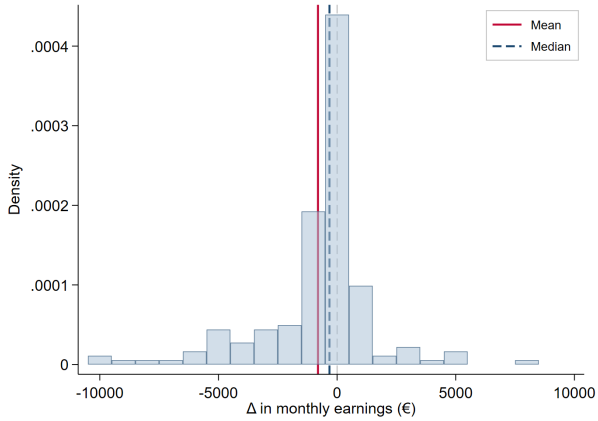
(b) Δ Earnings, mothers



(c) Δ Hours, fathers



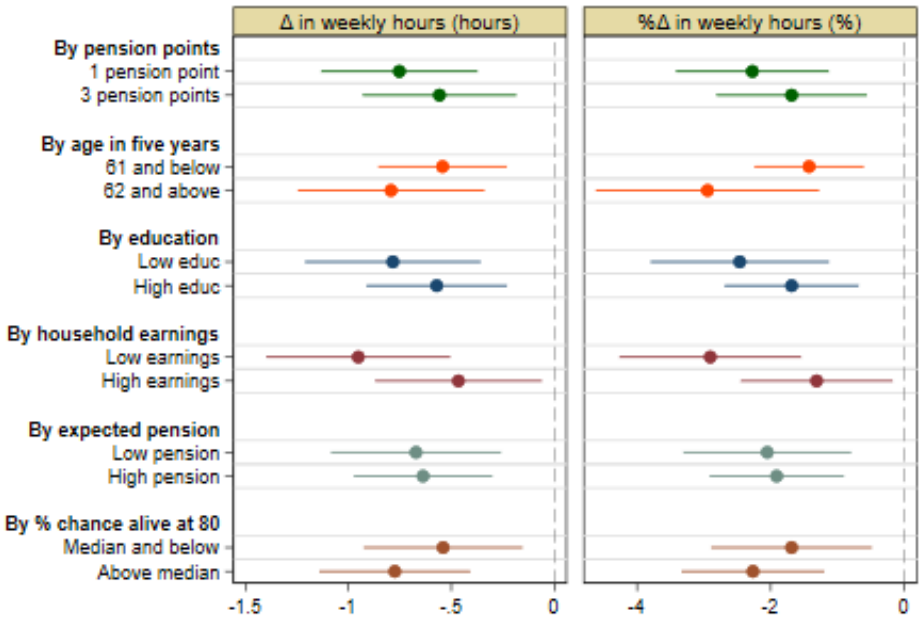
(d) Δ Earnings, fathers



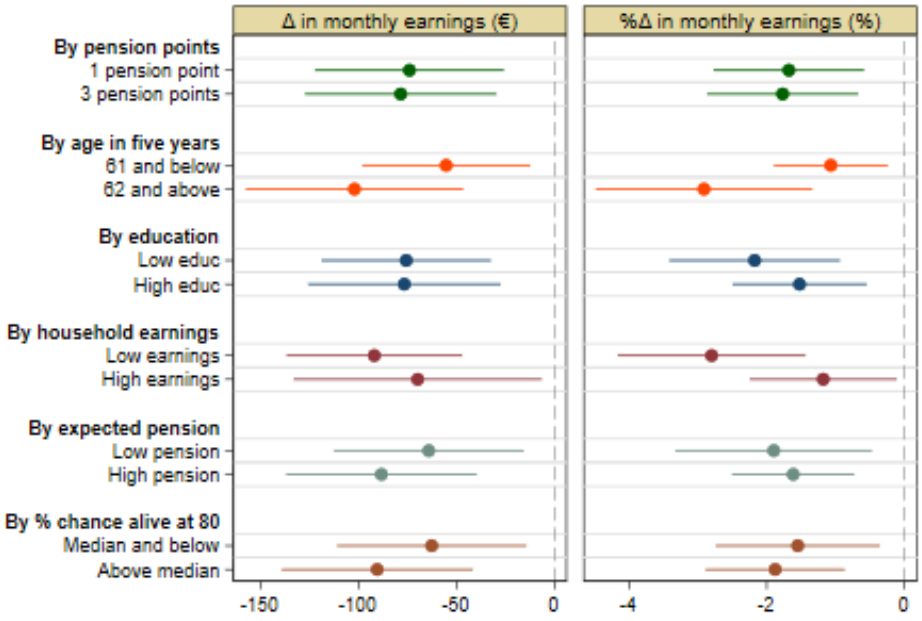
Notes: Each panel plots the distribution across respondents of the hypothetical reform's effect on expected work hours (left) and earnings (right) at the five-year horizon, for mothers (top) and fathers (bottom). Respondents who would not change their work hours or earnings (non-movers) are excluded; these comprise 86% and 91% of mothers and fathers, respectively. Supplemental-wave nonworking mothers are reweighted to their intended share (Appendix B.4). Solid and dashed vertical lines mark the mean and median of the plotted distribution.

Appendix Figure A6: Heterogeneity in the effect on work hours and earnings, fathers

(a) Weekly work hours



(b) Monthly earnings



Notes: The figure plots the hypothetical reform's effect on fathers' expected weekly work hours (top) and monthly earnings (bottom) at the five-year horizon for each specified subsample, each expressed as a level change (left) and as a percentage of the no-reform level (right). The sample excludes respondents who failed the attention check, completed the survey in less than half the median response time, reported more than 80 weekly hours currently or five years ahead (with or without reform), or whose partners are younger than 30. Horizontal bars are 95% confidence intervals computed from robust (left panels) or delta-method (right panels) standard errors. *Lower secondary and below* includes those whose highest schooling degree is none, lower secondary (*Hauptschulabschluss*), or intermediate secondary (*Realschulabschluss*); *upper secondary and above* includes those holding a university of applied sciences entrance qualification (*Fachhochschulreife*), a general university entrance qualification (*Abitur*), or another degree.

Appendix Table A1: Summary statistics

	N	Mean	S.D.	Artmann et al. (forthcoming)
	(1)	(2)	(3)	(4)
Panel A: Mothers				
Demographics				
Age	3,679	53.8	3.82	49.4
Age at first birth	3,679	28.6	6.60	27.4
No. of children ever born	3,679	1.91	0.876	1.95
Is married or in a registered partnership	3,679	0.650	0.477	
Cohabiting with a male partner	3,679	0.626	0.484	
Higher education	3,679	0.453	0.498	0.22
Work				
Currently employed	3,679	0.873	0.333	0.72
Full-time employed	3,679	0.420	0.494	0.21
Part-time employed	3,679	0.420	0.494	0.38
Monthly earnings (unconditional)	3,679	2,432	1,760	1749
Monthly earnings (conditional on working)	3,274	2,785	1,601	2346
Panel B: Fathers				
Demographics				
Age	1,927	55.9	3.60	51.5
Age at partner's first birth	1,927	32.7	7.26	
Higher education	1,927	0.608	0.488	
Work				
Currently employed	1,927	0.968	0.175	0.84
Full-time employed	1,927	0.908	0.289	
Part-time employed	1,927	0.056	0.229	
Monthly earnings (unconditional)	1,927	4,825	2,370	4472
Monthly earnings (conditional on working)	1,866	4,983	2,239	5187
Panel C: Fathers' partners				
Demographics				
Age	1,836	52.0	5.95	
Age at first birth	1,836	29.1	6.73	
No. of children ever born	1,836	1.98	0.866	
Work				
Currently employed	1,836	0.882	0.322	
Full-time employed	1,836	0.429	0.495	
Part-time employed	1,836	0.412	0.492	
Monthly earnings (unconditional)	1,836	2,470	1,871	
Monthly earnings (conditional on working)	1,620	2,799	1,745	

Notes: The table reports demographic and labor-market statistics for mothers (Panel A), fathers (Panel B), and the fathers' partners (Panel C). All samples exclude respondents who failed the attention check, completed the survey in less than half the median response time, or reported more than 80 weekly hours, and the fathers' samples additionally exclude those whose partner is younger than 30. Supplemental-wave nonworking mothers are reweighted to their intended share (Appendix B.4). *Higher education* indicates a university-entrance qualification (*Fachhochschulreife* or *Abitur*) or higher. Monthly earnings in current euros. Column 4 reports corresponding means from Artmann et al. (forthcoming, Appendix Tables A1, A2, and F1, and Appendix F), left blank when unavailable, with euro values inflated to 2025 euros.

Appendix Table A2: Gompertz estimates of mothers' subjective survival and implied life expectancy

	Dependent variable: age-80 and age-85 survival probabilities. Sample:	
	All respondents (1)	Following Perozek (2008) (2)
Baseline mortality hazard (α)	0.00388	0.00421
	(0.00014)	(0.00016)
Mortality hazard growth rate (β)	0.08799	0.08611
	(0.00152)	(0.00157)
Respondents	3,679	3,353
Implied mean life expectancy	84.6	84.2

Notes: The table reports nonlinear least-squares estimates of a Gompertz mortality model fit to mothers' elicited probabilities of surviving to ages 80 and 85. In the model, the mortality hazard rate τ years beyond a respondent's current age is $\alpha\, e^{\beta\tau}$: α is the subjective mortality hazard rate at the current age, and β its growth rate with age. Column 1 uses all respondents; column 2 applies the data cleaning procedures of Perozek (2008), including a sample restriction. Supplemental-wave nonworking mothers are reweighted to their intended share (Appendix B.4). *Implied mean life expectancy* is computed from the fitted parameters. Standard errors in parentheses.

Appendix Table A3: Effect of hypothetical reform on mothers' expected labor supply, by pension-point increase

	Effect of specified pension-point increase on:	
	Weekly hours	Monthly earnings
	(1)	(2)
One pension point	-0.424***	-40.53***
	(0.148)	(13.98)
Three pension points	-0.566***	-61.80***
	(0.132)	(11.67)
Observations	3,677	3,677
Concavity: one-point − three-point/3	-0.235	-19.92
	(0.155)	(14.51)
p-value	0.129	0.170

Notes: The sample comprises mothers, excluding respondents who failed the attention check, completed the survey in less than half the median response time, or reported more than 80 weekly hours currently or five years ahead (with or without reform). Supplemental-wave nonworking mothers are reweighted to their intended share (Appendix B.4). Each column reports the estimated effect of the assigned pension-point increase on the mother's expected labor supply at the five-year horizon, separately for the one- and three-pension-point arms, using weekly hours (column 1) and monthly earnings in current euros (column 2). The *Concavity* row reports the one-point response minus one-third of the three-point response, together with the p-value of the test that it equals zero (a value of zero corresponds to a linear response in pension points). Standard errors in parentheses. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$.

B Survey appendix

This section provides further details on the survey and data cleaning procedures.

B.1 Samples and fieldwork

We recruited respondents from Bilendi's online panel in Germany. Throughout, we screened out respondents who were retired or whose employment status the labor-supply questions could not accommodate: short-time work, apprenticeship, or training. No restriction was placed on the device used to answer the survey; 65 percent of mothers and 50 percent of fathers did so on a mobile device.

The mothers' survey was fielded in two waves, a main wave in August–September 2025 and a supplemental wave in March–May 2026. Both waves targeted German women aged 48–60 with at least one child (screened on their own fertility record). No restrictions were placed on the respondents' partnership status or their partner's gender. The supplemental wave's purpose was to extend the main wave in two respects: to estimate the reform's effect on mothers at the one-year horizon, and to elicit mothers' expectations of how their partner would respond at five years. A programming error early in the supplemental wave initially prevented some nonworking mothers from completing the questionnaire; we describe it and the reweighting correction we apply in Section B.4.

The fathers' survey was fielded contemporaneously with the main mothers' wave, in August–September 2025, and targeted German men aged 50–62—two years older than the mothers' target range. The men had to be "fathers" as well, but unlike the mothers' survey, we required the men to be living with a female partner who had at least one child (screened on the partner's fertility record).[14] This mirrors the matched-couples sample of Artmann et al. (forthcoming), whose data link mothers to their male partners but cannot establish whether the partner fathered the children.

B.2 Questionnaire flow

The mothers' and fathers' surveys were broadly structured as follows (the full instrument is available as a separate online supplement):

1. Demographics and pension knowledge: After a consent screen, respondents gave basic demographics—including gender, birth year, and partnership status—together with their own (in the mothers' survey) or their partner's (in the fathers' survey) fertility

[14] An implication is that the man need not have fathered his partner's children.

history. A short module then gauged their knowledge of the statutory pension system: the regular retirement age, how pension points accrue, the monthly value of a point, how closely the respondent reads the annual statement from the German Pension Insurance, and how well informed they felt about their own entitlements.

2. Current labor-market situation: Workers reported usual weekly hours and gross monthly earnings; from these the survey computed an implied hourly wage and verified it with the respondent. Nonworkers reported whether they were searching for work and, if so, their subjective probability of finding a job along with the hours and earnings they would expect on finding one.

3. Expected outcomes absent the reform: Respondents gave their expected weekly hours and gross monthly earnings at the five-year horizon, with the earnings question instructing them not to adjust for inflation: "Please assume that a Euro in 5 years is worth as much as it is now, so do not adjust for inflation." The supplemental mothers' wave elicited the same expectations at the one-year horizon as well, immediately before the five-year block; the main wave and the fathers' survey asked only about five years.

4. The hypothetical reform: Respondents were randomized to a hypothetical reform granting the mother one or three additional pension points and shown the vignette slideshow explaining it (Appendix Figure A1). They then answered two comprehension questions: whether the extra entitlement depends on the number of children, and by how much the mother's total pension would rise—9,600 or 28,800 euros under the one- and three-point vignettes—if the pension is collected over 20 years. In the fathers' survey the vignette slides framed the mother as the recipient, and a third comprehension question asked whose entitlements would change—the partner's, not the respondent's own.

5. Expected outcomes under the reform: We re-elicited expected weekly hours with the reform in place, at the one- and five-year horizons.[15] As respondents entered their weekly hours, the survey computed and dynamically displayed the implied monthly earnings, holding the hourly wage at its without-reform level.[16] Searchers and nonworkers additionally reported how the reform would change their job search and their updated chance of finding work.

[15] At the one-year horizon, the supplemental wave's elicitation referred to the respondent's without-reform level, while the main wave and the fathers' survey referred to the current level. The latter provide no one-year without-reform baseline, so the one-year analysis uses the supplemental wave alone.

[16] We informed respondents how the implied earnings are computed: at the five-year horizon, for example, they saw "At the hourly wage you expect in 5 years, [entered hours] hours per week means [implied earnings] euros per month."

6. Partner outcomes: Where a cohabiting opposite-sex partner was present, the survey repeated the current, without-reform, and with-reform modules for the partner, reminding the respondent that only the mother's pension points would rise. The partner's blocks always followed the analogous own blocks.

7. Closing background: The survey closed with occupation and primary income source (for the respondent and, where present, the partner), household composition, highest schooling degree, the expected monthly pension at retirement, the subjective probability of living past ages 80 and 85, and state of residence.

B.3 Quality control

We took several steps to ensure the quality of survey responses, summarized in Appendix Table B1.

To ensure respondents understood the hypothetical reform before answering, the vignette slides were interactive but could not be rushed. The text appeared line by line, at about 160 words per minute—slightly below the pace at which German adults read aloud (Trauzettel-Klosinski et al. 2012)—and respondents had to click or tap the button to advance once the whole slide had been shown, demanding active engagement rather than passive reading. A minimum of four clicks or taps are required to proceed; the median respondent clicked or tapped one more than necessary, and 8.7 to 8.8 percent clicked or tapped more than 12 times (more than three times the minimum).[17]

After the vignette slides, we included comprehension questions to check the elements we most needed respondents to understand: that the additional pension points did not depend on the number of children—unlike previous actual Mütterrente reforms, which depended on the number and birth years of children—how much the reform raised total pension wealth, and, in the fathers' survey, that the points accrued to the partner rather than the respondent. Notably, around 40 percent of respondents still thought the points depended on the number of children, as under the actual Mütterrente; by contrast, only 15 percent of fathers mistakenly thought they were the reform's recipient. These questions were meant as much to reinforce these points as to gauge understanding, and so gave immediate feedback whether the respondent got it correct or not.

We also took steps to measure responses accurately. To ease recall, each field that respondents filled in was shown beside a relevant reference value—their current level, or their own without-reform expectation. A second choice, made when we designed the survey,

[17] These also include stray clicks or taps (excluding swipes) elsewhere on the screen, which can happen especially since many respondents did the survey on mobile.

concerned the nature of the response: we treated it as a joint decision over hours and earnings, and so displayed the implied earnings as respondents entered their post-reform hours. That conversion requires an hourly wage. We computed it from respondents' without-reform hours and earnings and, to improve accuracy, asked them to confirm that it "sound[ed] about right," prompting a recheck if not.

To limit sample selection on the survey's topic, the landing page described the study only in general terms as research to "better understand human behavior", and in fact only disclosed its focus on pensions after the screening questions. To avoid cueing respondents on the expected effect, the survey itself used neutral wording throughout. The elicitation of responses was also symmetric: respondents could as easily enter more hours than their no-reform baseline as fewer, and nothing in the design pointed to a particular direction of response. Moreover, the survey was anonymous and self-administered online, so respondents answered privately rather than to an interviewer, mitigating a common source of social-desirability bias. Finally, prior work finds that a question requiring a quantity outright—with no easy way to answer "none"—presumes the behavior and pushes respondents toward reporting it, whereas a filtered question that first lets them opt out is less biased (Crossley et al. 2025). Our hours question worked in the same spirit: respondents could check "will not work" rather than enter a number.

We tracked inattentive respondents with an instructed-response attention check—a question listing the options 10, 20, 30, and 40 that asked respondents instead to select "Others" and enter "25." Overall, 5.3 percent of mothers and 2.7 percent of fathers who completed the survey failed it (Appendix Table B2), within the range reported elsewhere (Shamon and Berning 2020; Gummer et al. 2021). We randomized where the check appeared—before the vignette was introduced for some respondents, after the with-reform questions for others—letting us estimate whether attention decayed over the survey. It did not; if anything, mothers who saw the check after the reform were marginally more likely to pass. We drop respondents who fail from the analysis sample.

Finally, we followed standard practice to verify that respondents were genuine (Stantcheva 2023). We recruit only through an established online panel that vets its members, and use a reCAPTCHA at the beginning of the survey. Overall, 4.4 percent of mothers and 1.2 percent of fathers received reCAPTCHA scores below 0.5, broadly in line with rates reported in other studies using commercial online panels, including another recent Bilendi sample (Al Sayah et al. 2025; Celebi et al. 2026).[18] We additionally included a "honeypot" question—"Check this box if it applies to you" with the options "I am a piece of software" and "I am not a piece of software"—displaced a full screen height downward and shown for one second, so that a

[18] Our results are similar if we exclude these respondents.

human saw what looked like a still-loading blank page, while an automated agent filling in fields programmatically might select any of the two options and reveal itself. In practice, this honeypot only caught and screened out two respondents in the mothers' survey and none in the fathers'.

Appendix Table B1: Survey quality checks

	Mothers (1)	Fathers (2)
Panel A: Engagement and comprehension (analysis sample)		
Clicks or taps while viewing the vignette slideshow		
Median number of clicks/taps	5	5
>8 clicks/taps (%)	18.7	17.6
>12 clicks/taps (%)	8.8	8.7
All comprehension questions correct (%)	51.7	48.0
Correct: extra points do not depend on no. of children (%)	60.9	57.3
Correct: size of total pension increase in euros (%)	79.7	81.7
Correct: reform raises partner's pension, not respondent's (%)	—	85.4
reCAPTCHA score <0.5 (%)	4.4	1.2
Time spent on survey (minutes)		
25th percentile	7.4	8.2
Median	9.5	10.5
75th percentile	12.5	13.8
Respondents	3,679	1,927
Panel B: Attention check (completed responses)		
Failed attention check (%)	5.3	2.7
Failed check shown before the reform (%)	5.9	2.8
Failed check shown after the reform (%)	4.7	2.5
p-value of difference	0.08	0.60
Respondents	4,074	2,034

Notes: Mothers' and fathers' surveys in columns 1 and 2 respectively. Panel A is computed on the analysis sample and Panel B on all completed responses, before the data cleaning exclusions of Section B.4. *Clicks or taps while viewing the vignette slideshow* counts the number of times a respondent clicks or taps the vignette slideshow, a minimum of four was required to complete the slideshow. Comprehension question rows report the percentage correct; the recipient question was asked only in the fathers' survey. *Time spent on survey* is the summed active time across the survey's pages expressed in minutes. The attention check question is an instructed-response item randomly placed either before the vignette or after the with-reform questions; the p-value tests equality of the failure rates at the two locations.

B.4 Data cleaning and sample construction

Appendix Table B2 traces how we go from raw responses to the analysis sample, by survey and—for mothers—by wave, with *Both* pooling the two waves. 6,915 mothers and 3,032 fathers started the survey and were not screened out. Of these, 25 percent of mothers and 21 percent of fathers dropped out before reaching the vignette slideshow, and a further 16 and 12 percent after—leaving 4,074 mothers and 2,034 fathers who completed the survey.

Next, to get to our main sample, we drop respondents who failed the attention check question or whose active time on the survey pages falls below half the sample median—the latter being a standard speeding check for low-effort responding (Zhang and Conrad 2014; Greszki et al. 2015; Leiner 2019). We also drop respondents who report more than 80 work hours per week—currently, or with or without the reform at the five-year horizon—both to reduce the influence of outliers and because the excessive hours likely indicate a low-quality response. In the fathers' survey we additionally drop respondents whose partner is younger than 30. The final analysis sample comprised 3,679 mothers and 1,927 fathers.

Appendix Table B2: Number of respondents and data-cleaning steps

	Number of respondents			
	Mother sample			**Father sample**
	Main	Supplemental	**Both**	
Started the survey and not screened out	4,544	2,371	**6,915**	**3,032**
Reached the vignette slideshow	3,623	1,542	**5,165**	**2,400**
Passed the vignette slideshow	3,194	1,405	**4,599**	**2,154**
Completed the survey	3,038	1,036	**4,074**	**2,034**
Excluded:				
Failed the attention check	147	68	**215**	**54**
Completed in under half the median time	122	51	**173**	**42**
Reported more than 80 weekly hours	11	11	**22**	**7**
Partner younger than 30	N/A	N/A	**N/A**	**7**
Analysis sample	2,766	913	**3,679**	**1,927**

Notes: The table reports sample construction by survey and, for mothers, by wave; *Both* pools the two waves. *Started the survey and not screened out* counts respondents who began the questionnaire and passed screening. *Reached* and *Passed the vignette slideshow* count respondents who reached the pages immediately before and after the slideshow, respectively. Exclusion categories may overlap. The partner-age restriction applies only to fathers.

Early in the supplemental wave's fielding, a programming error prevented nonworking

mothers from completing the questionnaire. We fixed it midway through and extended recruitment, but the final sample still contained too few nonworkers. To correct this, all estimates upweight supplemental-wave nonworking mothers to the share they would have held under normal recruitment. This assumes that nonworking mothers who entered before and after the fix are exchangeable; Appendix Table B3 shows that the two groups are similar on every characteristic we observe. Upweighted mothers are 2.5 percent of the final sample and carry a weight of 1.8.

Appendix Table B3: Nonworking mothers entering the supplemental wave: balance check

Characteristic	Supplemental-wave nonworking mothers who entered the survey			
	Before fix (1)	After fix (2)	Difference (3)	p-value (4)
Age (years)	54.1	53.7	0.4	0.53
Number of children	2.15	2.15	−0.00	0.99
Age of youngest child (years)	22.1	21.2	0.9	0.45
Share partnered (%)	62.4	73.0	−10.6	0.14
Share with partner in household (%)	58.7	68.9	−10.2	0.16
Share currently searching for work (%)	28.4	31.1	−2.6	0.70
Reads pension statement carefully (%)	39.4	32.4	7.0	0.34
Knows own pension entitlements well (%)	18.3	20.3	−1.9	0.75
Respondents (N)	109	74		

Notes: The sample comprises supplemental-wave nonworking mothers who entered the survey before and after we fixed the programming error, in columns (1) and (2) respectively. Mothers in column (1) were unable to complete the survey; those in column (2) completed it. Mothers who were affected by the error but returned to complete the survey after the fix are excluded from both groups. Characteristics are recorded before the point of error and so are observed for both groups. Column (3) is column (1) minus column (2), and column (4) reports the p-value of a two-sample t-test of their equality. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$.

We take without- and with-reform hours directly from each respondent's report, winsorizing the top 1 percent to limit the influence of outliers. For earnings, we apply the same conversion shown to respondents (Section B.3): weekly hours times the without-reform hourly wage times 4. The expected wage distribution exhibits long tails; to limit its influence, we winsorize at the top and bottom 1 percent. The same procedure is used for both five- and one-year horizon outcomes.

A complication is that the wage is missing for a respondent who does not expect to work without the reform. We impute it in such cases, following standard practice in structural labor-supply models, which predict potential wages for nonworkers rather than exclude them from estimation (Van Soest 1995; Blundell et al. 2000). Excluding such observations would be especially problematic here: a respondent who does not work without the reform but

works under it is a labor-market entrant, and dropping such entrants overstates the reform's negative effect on earnings.

When the five-year without-reform wage is missing, we use the respondent's current-wage information if in the labor force—the hourly wage if employed, or expected wage if searching for work—and for respondents currently out of the labor force, we predict using other in-labor-force respondents. The prediction comes from a Mincer regression of the log hourly wage $\ln w_i$,

$$\ln w_i = \alpha + \delta \,\text{Employed}_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{4}$$

where $\mathbf{X}_i$ contains education, age and its square, age interacted with education, number of children, age at first birth, age of the youngest child, state of residence, and an indicator for the supplemental wave. Employed_i is equal to one if the respondent is employed; note that this is always set to zero in prediction, which reflects the plausible similarity between not-in-labor-force and searchers. Exponentiating a fitted log wage understates its level, so we apply a Duan smearing factor—the sample mean of the exponentiated residuals, 1.063—to correct this retransformation bias (Duan 1983). Finally, we cap every imputed wage to the 1st–99th percentile of observed workers' hourly wages. Imputation for the one-year horizon, and fathers' wages, follows a similar procedure. Note that imputation affects only the earnings outcome, and only through these entrants—the hours outcome never uses it.

C Reweighted analysis

Our sample differs from Artmann et al.'s (forthcoming) along two dimensions that could affect the comparison (Appendix Table A1): our mothers are nearer retirement, and they earn more. The MPE calculation already absorbs the main effect of the first: it divides the earnings response by an age-dependent wealth gain, so mothers nearer retirement have both a larger response and a larger wealth gain, leaving the MPE unaffected (Section 3). Earnings has no such built-in adjustment: the wealth gain depends on the size of the pension increment regardless of a mother's earnings, so the MPE does not by construction account for differences in earnings across mothers. And earnings levels do appear to matter—our estimated response is concentrated among higher-earning households (Figure 1), as Artmann et al. also find, and Golosov et al. (2024) document an MPE that increases with income.[19] Hence, this section examines how sensitive our main results are to reweighting the sample to match the earnings distribution in Artmann et al..

We reweight the sample so that the share of respondents in each of three bands of own gross monthly earnings matches the earnings distribution in Artmann et al. (forthcoming), separately within each pension-point arm and survey wave.[20] The reweighted sample's unconditional monthly earnings are 1,914 euros for mothers and 4,707 for fathers. For mothers this is close to the 1,749 in Artmann et al.'s sample, and well below the 2,434 in our unweighted sample.

Reweighting leaves our conclusions intact. On the reweighted sample, mothers expect to cut weekly hours by 1.6 percent and monthly earnings by 1.8 percent, and fathers by about 2.0 and 1.7 percent (Appendix Table C1)—magnitudes close to our headline estimates, and, as before, concentrated on the extensive margin. The heterogeneity pattern for mothers is likewise preserved—the response stays negative in every subgroup and broadly similar in magnitude (Appendix Figure C1)—and it remains similar across the one- and five-year horizons (Appendix Table C2). Applying the same MPE calculation to the reweighted earnings response yields -0.47 in our preferred specification (Appendix Table C3, Panel B row 1)—close both to our headline estimate (Table 3) and to Artmann et al.'s (forthcoming) -0.54—and the robustness rows move much as they do in the main table. Finally, to compare spousal predictions, we reweight both mother and father surveys on own and spousal earnings simultaneously, so that the spouse-prediction and own-report expectations target the same demographics. The reweighted predictions again line up with partners' own reports

[19] Golosov et al. (2024) estimate an MPE that rises from -0.31 for households in the lowest quartile of pre-win income to -0.67 in the highest.

[20] The bands are 0–1,000, 1,001–3,000, and above 3,000 euros for mothers, and 0–2,000, 2,001–4,000, and above 4,000 euros for fathers, defined in 2013 euros and inflation-adjusted to the survey period.

(Appendix Figure C2).

Appendix Table C1: Reweighted sample: effect of hypothetical reform on expected labor supply at the five-year horizon

	In 5 years, expected labor market outcome...			
	Without reform (1)	With reform (2)	Difference (3)	As % of (1) (4)
Panel A: Mothers (N=3,677)				
Hours worked	22.5	22.1	-0.369***	-1.64***
			(0.121)	(0.53)
Prob. of employment	0.806	0.798	-0.00834**	-1.03**
			(0.00423)	(0.52)
Earnings	2092.6	2055.4	-37.3***	-1.78***
			(9.3)	(0.44)
Decomposition of hours:				
Extensive margin			-0.313***	
			(0.109)	
			[84.8%]	
Intensive margin			-0.0561	
			(0.0524)	
			[15.2%]	
Panel B: Fathers (N=1,927)				
Hours worked	32.4	31.7	-0.660***	-2.04***
			(0.143)	(0.44)
Prob. of employment	0.840	0.826	-0.0148***	-1.76***
			(0.0036)	(0.43)
Earnings	4319.2	4246.4	-72.8***	-1.69***
			(17.3)	(0.40)
Decomposition of hours:				
Extensive margin			-0.531***	
			(0.126)	
			[80.3%]	
Intensive margin			-0.130*	
			(0.069)	
			[19.7%]	

Notes: The table reproduces Table 1 on samples reweighted separately by pension-point arm and survey wave to match the gross monthly earnings distribution in Artmann et al. (forthcoming). The matching uses three earnings bands: 0–1,000, 1,001–3,000, 3,001+ for mothers in Panel A; and 0–2,000, 2,001–4,000, 4,001+ for fathers in Panel B. Cutoffs are in 2013 euros (i.e. the units used by Artmann et al.) and inflation-adjusted to our survey period. Sample restrictions, panel layout, column definitions, and the extensive/intensive decomposition are as in Table 1. Standard errors in parentheses, robust in column 3 and computed by the delta method in column 4. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$.

Appendix Table C2: Reweighted sample: supplemental-wave mothers at the five- and one-year horizons

	Expected labor market outcome...			
	Without reform (1)	With reform (2)	Difference (3)	As % of (1) (4)
Panel A: 5-year horizon (N=911)				
Hours worked	22.6	22.0	-0.586*	-2.60*
			(0.323)	(1.41)
Prob. of employment	0.775	0.770	-0.00434	-0.560
			(0.01101)	(1.417)
Earnings	2147.7	2095.4	-52.3**	-2.44**
			(23.8)	(1.10)
Decomposition of hours:				
Extensive margin			-0.390	
			(0.307)	
			[66.5%]	
Intensive margin			-0.197*	
			(0.104)	
			[33.5%]	
Panel B: 1-year horizon (N=835)				
Hours worked	26.7	26.3	-0.361	-1.35
			(0.226)	(0.84)
Prob. of employment	0.920	0.913	-0.00669	-0.728
			(0.00716)	(0.776)
Earnings	2433.7	2391.9	-41.7**	-1.71**
			(20.1)	(0.82)
Decomposition of hours:				
Extensive margin			-0.216	
			(0.211)	
			[59.8%]	
Intensive margin			-0.145*	
			(0.083)	
			[40.2%]	

Notes: The table reproduces Table 2 on the supplemental wave sample reweighted separately by pension-point arm to match the gross monthly earnings distribution in Artmann et al. (forthcoming). The matching uses three earnings bands, 0–1,000, 1,001–3,000, and 3,001+, defined in 2013 euros (i.e. the units used by Artmann et al.) and inflation-adjusted to our survey period. Panel A reports predictions at the five-year horizon and Panel B at the one-year horizon. Sample restrictions, column definitions, and the extensive/intensive decomposition are as in Table 2. Standard errors in parentheses, robust in column 3 and computed by the delta method in column 4. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$.

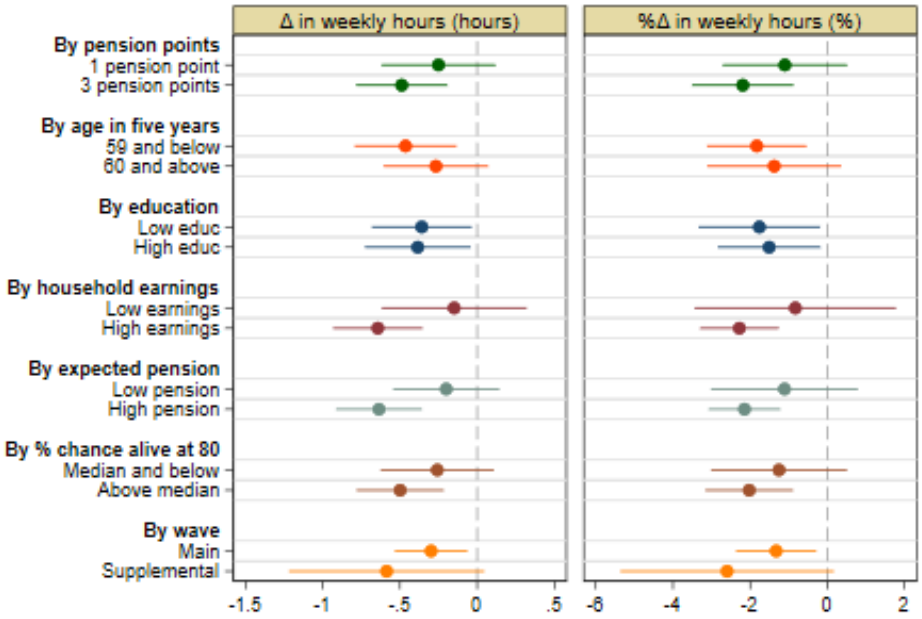
Appendix Table C3: Reweighted sample: implied marginal propensity to earn under alternative assumptions

S/N	Specification	Description	MPE
		Panel A: Quasi-experimental benchmark	
	Artmann et al. (forthcoming)	Their reported estimate. Main assumptions: about 2 percent annual discount rate; real pension-point value declines slightly over time; collect pension for 20 years.	−0.54
		Panel B: MPE estimates based on our survey	
1	Preferred specification	Artmann's present-value formula, discount rate, and pension point-value change, applied to our main five-year earnings response (Table 1). Main difference: months to retirement based off the sample mean age.	−0.47*** (0.12) [−0.69, −0.24]
2	Face-value framing	Assume that respondents value the additional points at the undiscounted 20-year amount shown on the reform slides—no discounting or pension point-value change over time.	−0.30*** (0.08) [−0.45, −0.15]
3	By-arm average	Compute the MPE separately for the one- and three-pension-point arms and average the two.	−0.48** (0.19) [−0.86, −0.11]
4	Artmann et al. time profile	Impose Artmann et al.'s labor supply response time profile: scale the first 2.5 years to about half the year-5 effect, holding years 3–5 at the full effect.	−0.35*** (0.09) [−0.52, −0.18]
5	Subjective life expectancy	Replace the 20-year payout horizon with the 17.1 years implied by respondents' subjective mean life expectancy of 84.1 (Appendix Table A2).	−0.53*** (0.13) [−0.79, −0.27]

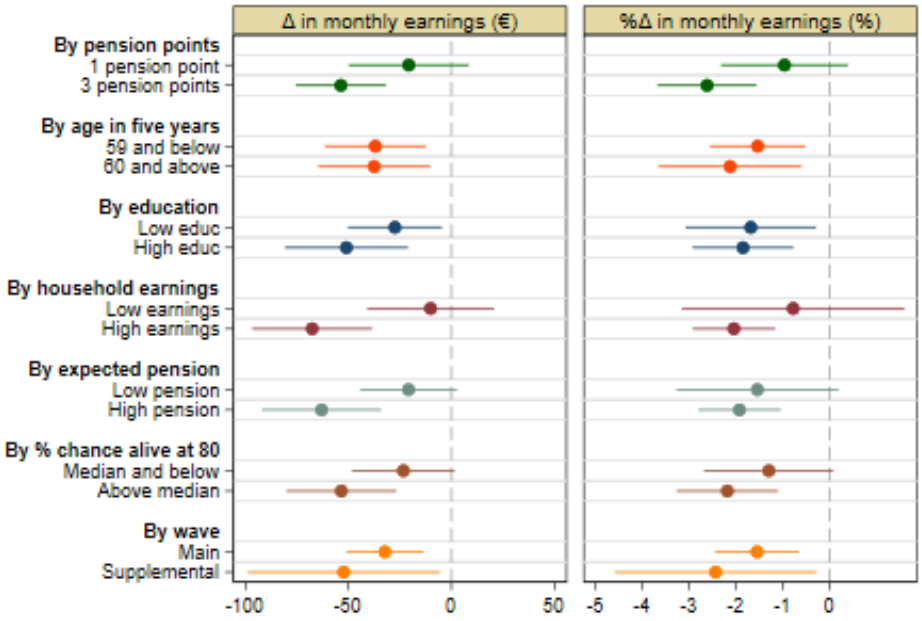
Notes: The table reproduces Table 3, with Panel A reproducing the quasi-experimental benchmark of Artmann et al. (forthcoming) and Panel B computed on the reweighted sample described in Appendix Table C1 Panel A. Standard errors computed by the delta method in parentheses unless otherwise stated. 95% confidence intervals in brackets. Asterisks denote significance: * $p < .10$, ** $p < .05$, *** $p < .01$. Panel B, row 1: We use the same assumptions as Artmann et al. but update the pension-point value to its July 2025–June 2026 level (40.79 € per month) and take the months to retirement from the sample mean age, following the annuitization procedure described in Section 3. Panel B, row 2: We assume that respondents take 9,600 and 28,800 €in the one and three pension-point vignettes literally without adjusting for discounting. Panel B, row 3: We compute the MPE separately for the one- and three-pension-point arms and report their simple average. Panel B, row 4: We rescale our five-year earnings response by Artmann et al.'s short-run/medium-run profile, so that the first 2.5 years carry 52% of the year-5 monthly effect and years 3–5 carry the full effect. Panel B, row 5: We replace the 20-year pension-collection horizon with respondents' subjective mean life expectancy and compute standard errors based on 2,000 bootstrap replications.

Appendix Figure C1: Reweighted sample: heterogeneity in the effect on work hours and earnings, mothers

(a) Weekly work hours



(b) Monthly earnings



Notes: The figure reproduces Figure 1 on the reweighted sample described in Appendix Table C1 Panel A, plotting the hypothetical reform's effect on mothers' expected weekly work hours (top) and monthly earnings (bottom) at the five-year horizon for each subsample, as a level change (left) and as a percentage of the no-reform level (right). Sample restrictions and subsample definitions are as in Figure 1. Horizontal bars are 95% confidence intervals computed from robust (left panels) or delta-method (right panels) standard errors.

Appendix Figure C2: Reweighted sample: spousal versus own labor supply predictions

(a) Mothers' labor supply



(b) Fathers' labor supply



Notes: The figure reproduces Figure 2 on samples reweighted separately by pension-point arm and survey wave to match own and spousal gross monthly earnings distributions in Artmann et al. (forthcoming). The matching uses three earnings bands: 0–1,000, 1,001–3,000, 3,001+ for mothers in Panel A; and 0–2,000, 2,001–4,000, 4,001+ for fathers in Panel B. Cutoffs are in 2013 euros (i.e. the units used by Artmann et al.) and inflation-adjusted to our survey period. Each panel plots the hypothetical reform's effect at the five-year horizon, expressed as a percentage of the pre-reform mean. Panel A compares fathers' expectations of the mother's response with mothers' own reported response; Panel B compares mothers' expectations of the father's response with fathers' own reported response. Horizontal bars are 95% confidence intervals

References

Al Sayah, Fatima, Hilary Short, Juan M Ramos-Goñi, Rosalie Viney, Erica I Lubetkin, Mathieu F Janssen, Jeffrey A Johnson, and the EQ-DAPHNIE Project Team, "Design and implementation of data quality controls in the EQ-DAPHNIE study: insights from the pilot phase and 15-country analysis," *Quality of Life Research*, 2025, *34* (12), 3335–3350.

Artmann, Elisabeth, Nicola Fuchs-Schündeln, and Giulia Giupponi, "Forward-looking labor supply responses to changes in pension wealth: Evidence from Germany," *American Economic Journal: Applied Economics*, forthcoming.

Blundell, Richard, Alan Duncan, Julian McCrae, and Costas Meghir, "The labour market impact of the Working Families' Tax Credit," *Fiscal Studies*, 2000, *21* (1), 75–104.

Celebi, Can, Christine Exley, Soren Harrs, Hannu Kivimaki, Marta Serra-Garcia, and Jeffrey Yusof, "Mission possible: The collection of high-quality data," CESifo Working Paper 12535, CESifo 2026.

Crossley, Thomas F, Paul Fisher, Peter Levell, and Hamish Low, "Eliciting the marginal propensity to consume in surveys," IFS Working Papers W25/25, Institute for Fiscal Studies 2025.

Duan, Naihua, "Smearing estimate: a nonparametric retransformation method," *Journal of the American Statistical Association*, 1983, *78* (383), 605–610.

Golosov, Mikhail, Michael Graber, Magne Mogstad, and David Novgorodsky, "How Americans respond to idiosyncratic and exogenous changes in household wealth and unearned income," *The Quarterly Journal of Economics*, 2024, *139* (2), 1321–1395.

Greszki, Robert, Marco Meyer, and Harald Schoen, "Exploring the effects of removing "too fast" responses and respondents from web surveys," *Public Opinion Quarterly*, 2015, *79* (2), 471–503.

Gummer, Tobias, Joss Roßmann, and Henning Silber, "Using instructed response items as attention checks in web surveys: Properties and implementation," *Sociological Methods & Research*, 2021, *50* (1), 238–264.

Leiner, Dominik Johannes, "Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys," *Survey Research Methods*, 2019, *13* (3), 229–248.

Perozek, Maria, "Using subjective expectations to forecast longevity: Do survey respondents know something we don't know?," *Demography*, 2008, *45* (1), 95–113.

Shamon, Hawal and Carl Berning, "Attention check items and instructions in online surveys with incentivized and non-incentivized samples: Boon or bane for data quality?," *Survey Research Methods*, 2020, *14* (1), 55–77.

Soest, Arthur Van, "Structural models of family labor supply: a discrete choice approach," *Journal of Human Resources*, 1995, *30* (1), 63–88.

Stantcheva, Stefanie, "How to run surveys: A guide to creating your own identifying variation and revealing the invisible," *Annual Review of Economics*, 2023, *15* (1), 205–234.

Trauzettel-Klosinski, Susanne, Klaus Dietz, IReST Study Group et al., "Standardized assessment of reading performance: The new international reading speed texts IReST," *Investigative ophthalmology & visual science*, 2012, *53* (9), 5452–5461.

Zhang, Chan and Frederick Conrad, "Speeding in web surveys: The tendency to answer very fast and its association with straightlining," *Survey Research Methods*, 2014, *8* (2), 127–135.