Things Take the Time They Take

Lessons on measurement error from high-frequency time use data

Tamara McGavock
with Ellen McCullough (UGA) & Thomas Assefa (IFPRI)

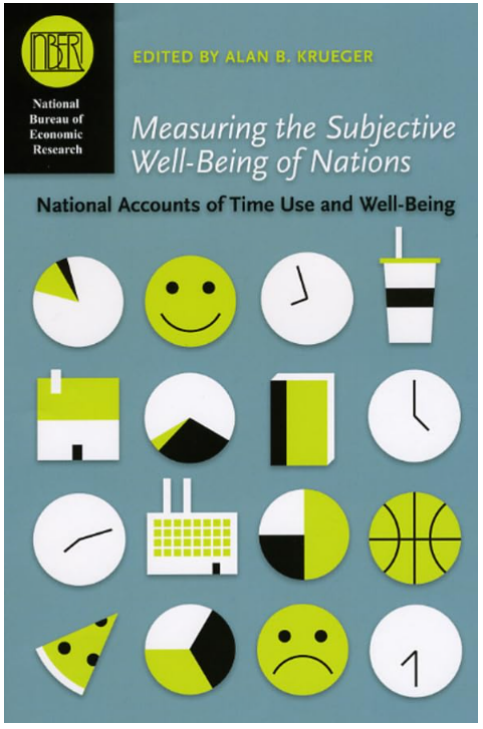
Grinnell College & University of Chicago

March 27, 2026
NBER Gender in the Economy

Implications of time use measurement extend beyond labor

Without good measures of time use, we cannot:

- link time allocation to subjective wellbeing
- quantify welfare and living standards
- evaluate programs and policies credibly
- measure unpaid domestic and care work
- reconcile household production with national accounts



NBER Conference Report (2008)

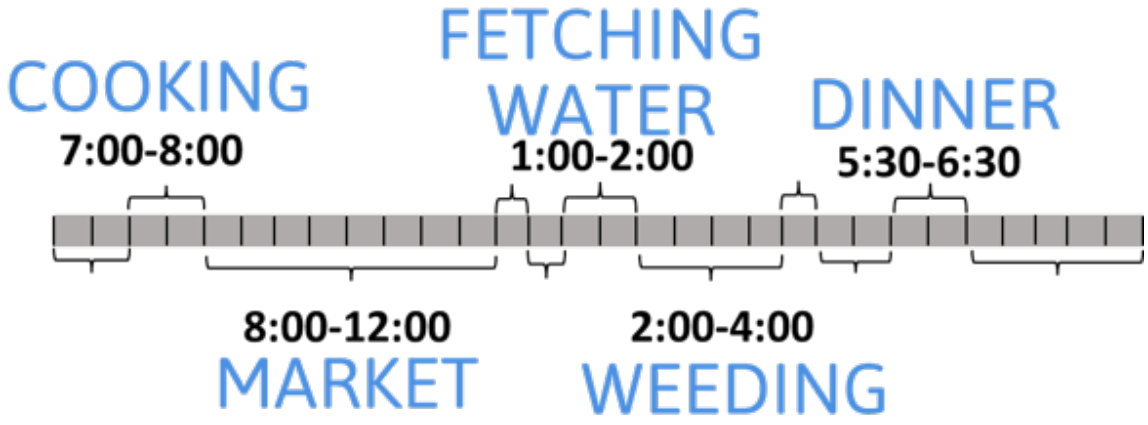
Measurement errors in time use/labor are often overlooked

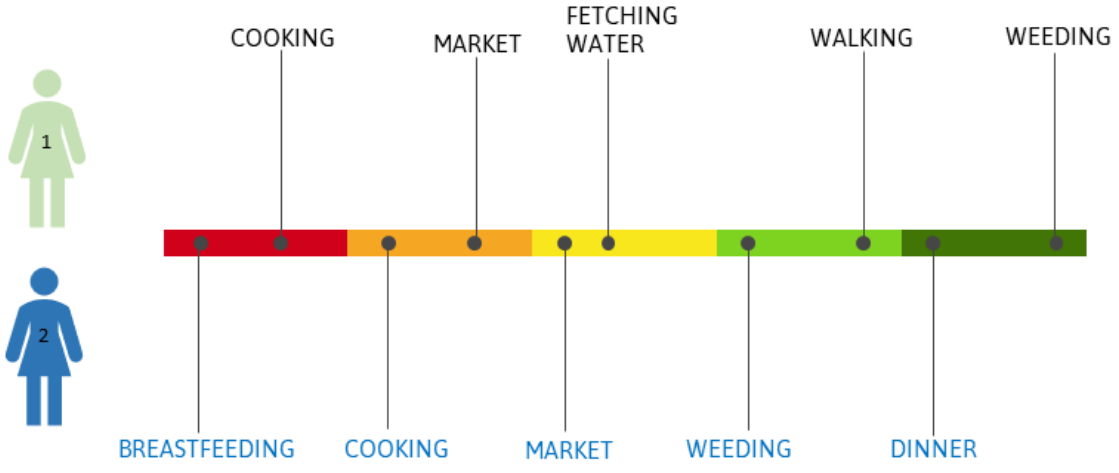
Could measurement errors help explain ... ?

- Sluggish labor supply and proliferation of self-employment in low-income countries
  e.g., Breza and Kaur, 2025; Agness et al., 2025
- Gender and other wage gaps e.g., Borjas and Hamermesh, 2024; Gottlieb et al., 2026
- Sectoral productivity gaps
  e.g., Gollin, Lagakos, and Waugh, 2014; Bick, Fuchs-Schündeln, and Lagakos, 2018; Gollin and Udry, 2021
- Inverse farm size productivity relationship
  e.g., Foster and Rosenzweig, 2022

Our Method: At the Time of the Call (ATOC)

- Day Reconstruction
- Krueger Proposal (2007)





Adaptation of "beeper method"

Experience Sampling Method (ESM)/Ecological Momentary Assessment (EMA)
Csikszentmihalyi, 2014

1. *in-the-moment:* avoid some recall problems (Dabalen et al., 2016)

Adaptation of "beeper method"

Experience Sampling Method (ESM)/Ecological Momentary Assessment (EMA)
Csikszentmihalyi, 2014

1. *in-the-moment:* avoid some recall problems (Dabalen et al., 2016)
2. *quick:* additional info (Kahneman et al., 2004; Krueger and Schkade, 2008)

Adaptation of "beeper method"

Experience Sampling Method (ESM)/Ecological Momentary Assessment (EMA)
Csikszentmihalyi, 2014

1. *in-the-moment:* avoid some recall problems (Dabalen et al., 2016)
2. *quick:* additional info (Kahneman et al., 2004; Krueger and Schkade, 2008)
3. *by phone:* easy, cheaper to measure outcomes repeatedly

Adaptation of "beeper method"

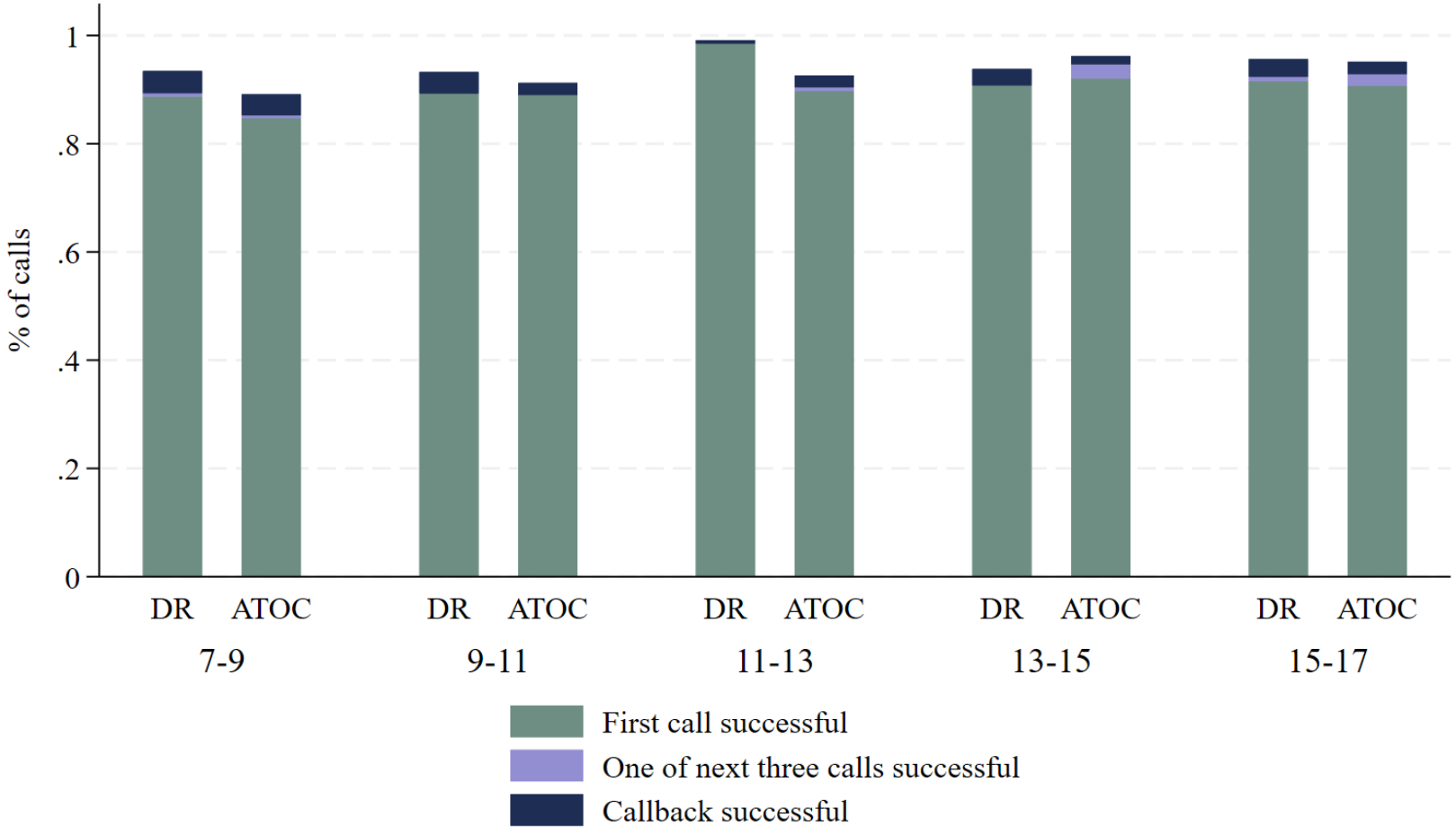
Experience Sampling Method (ESM)/Ecological Momentary Assessment (EMA)
Csikszentmihalyi, 2014

1. *in-the-moment:* avoid some recall problems (Dabalen et al., 2016)
2. *quick:* additional info (Kahneman et al., 2004; Krueger and Schkade, 2008)
3. *by phone:* easy, cheaper to measure outcomes repeatedly
4. *refer to previous call:* "bounded recall" (FAO and The World Bank, 2018)

High Compliance!

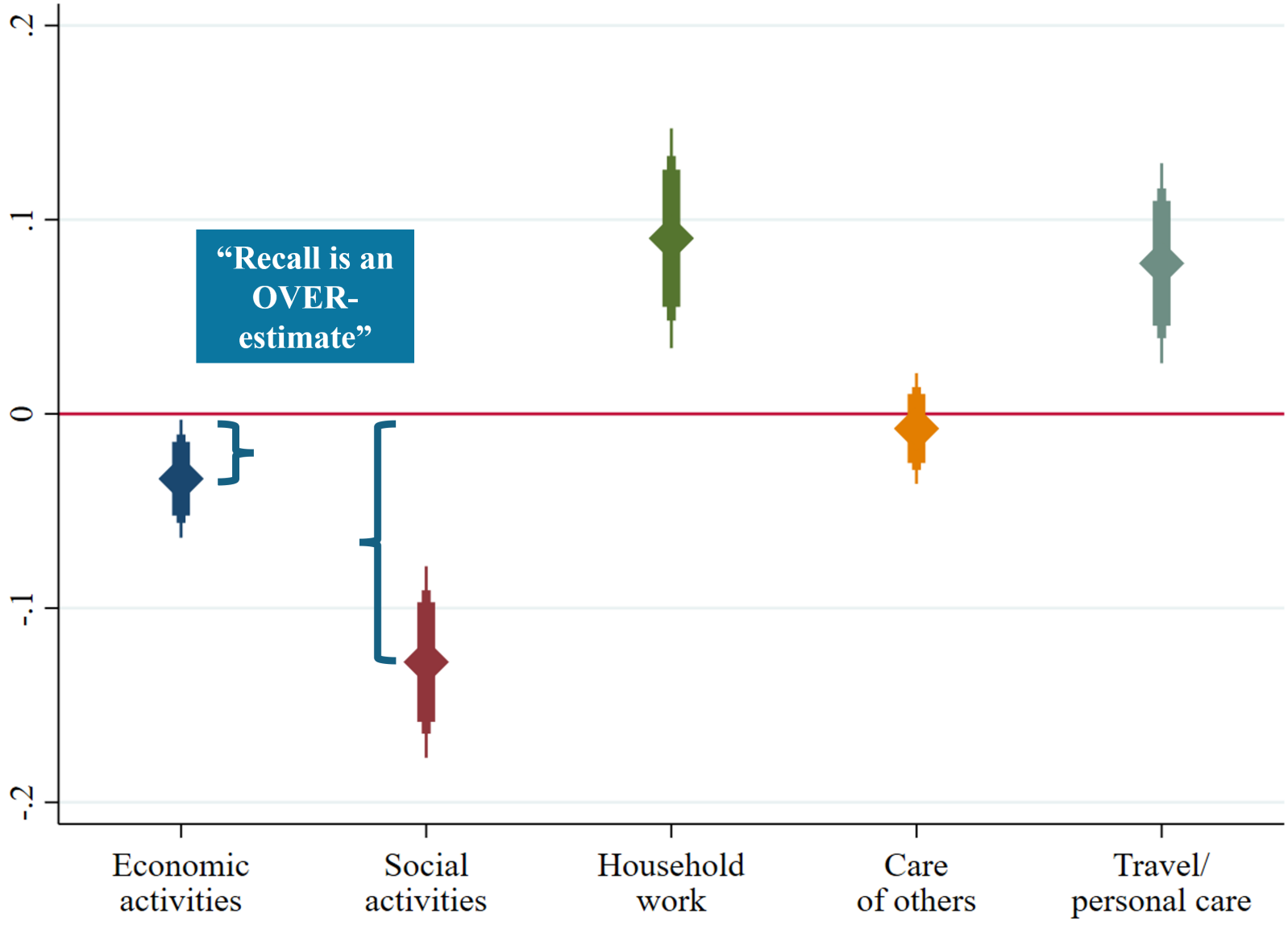
More than 17,000 calls

▸ Individual-level compliance



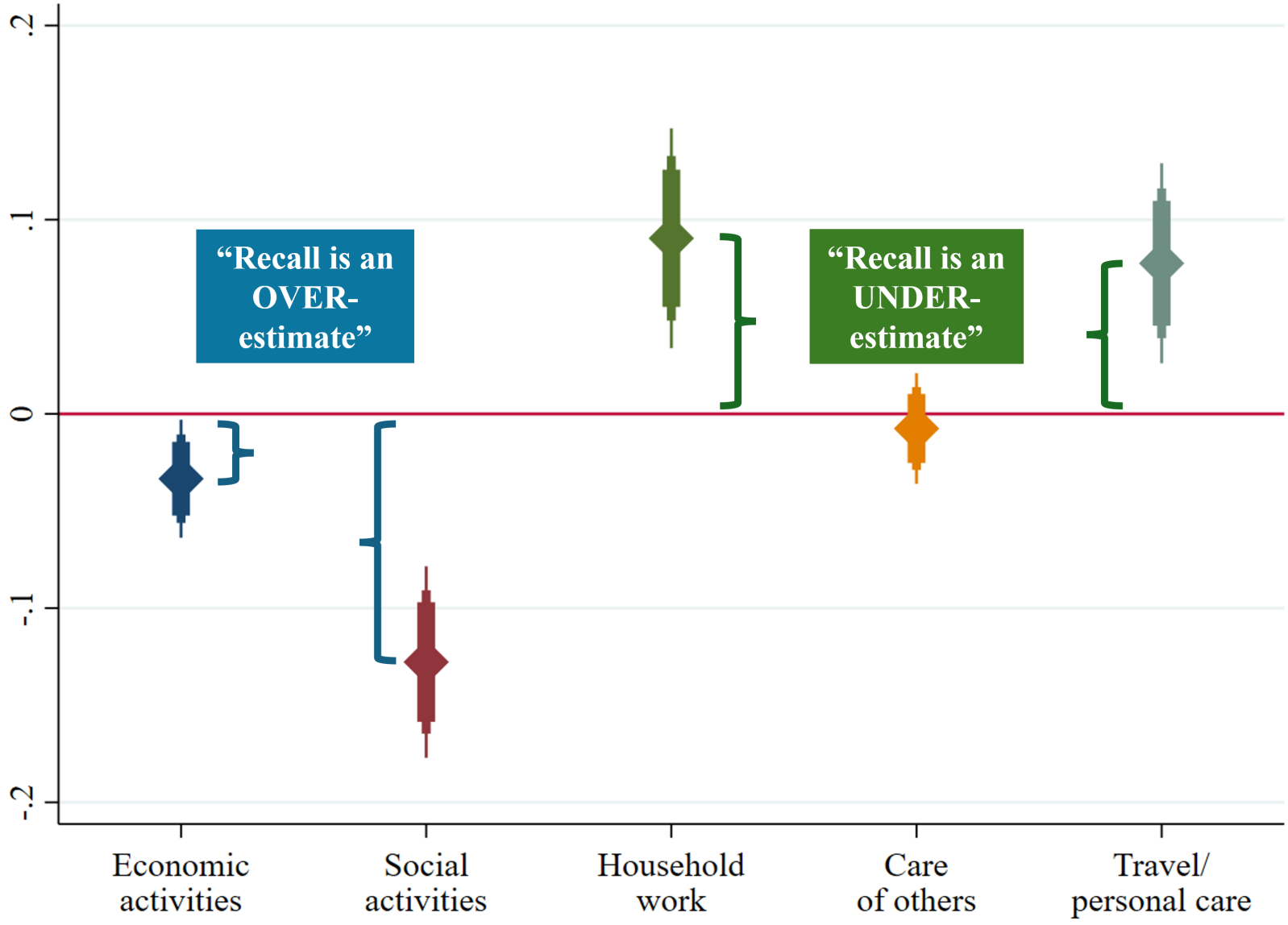
Average activity share differences

▸ Average shares



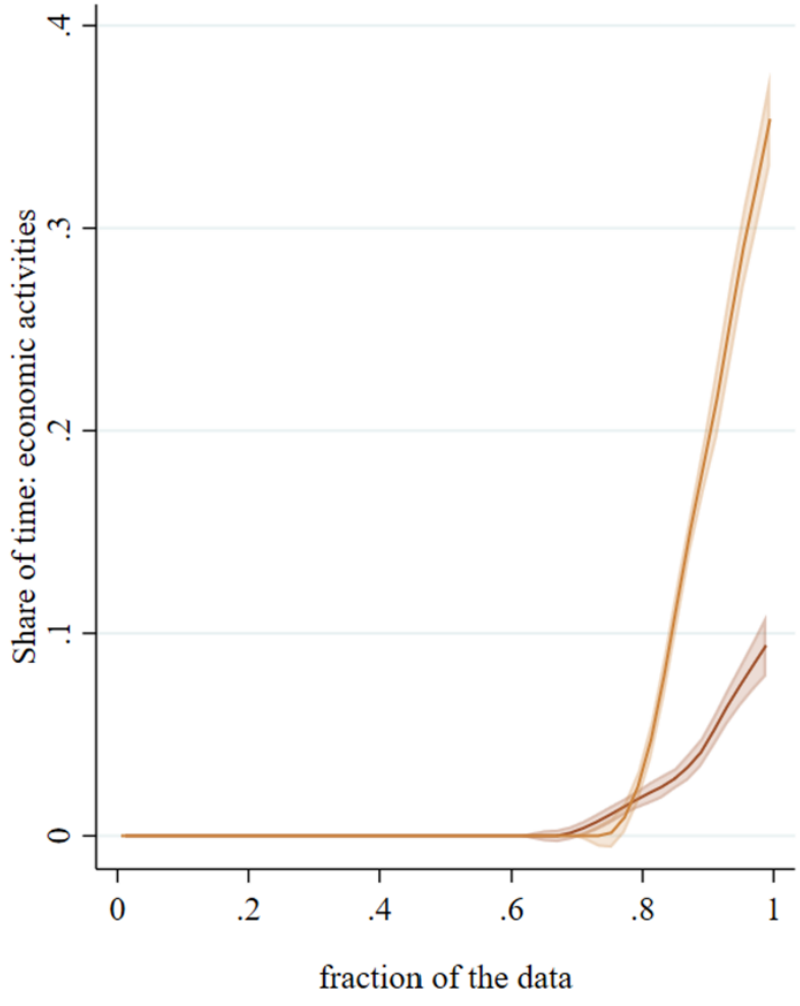
Average activity share differences

▸ Average shares



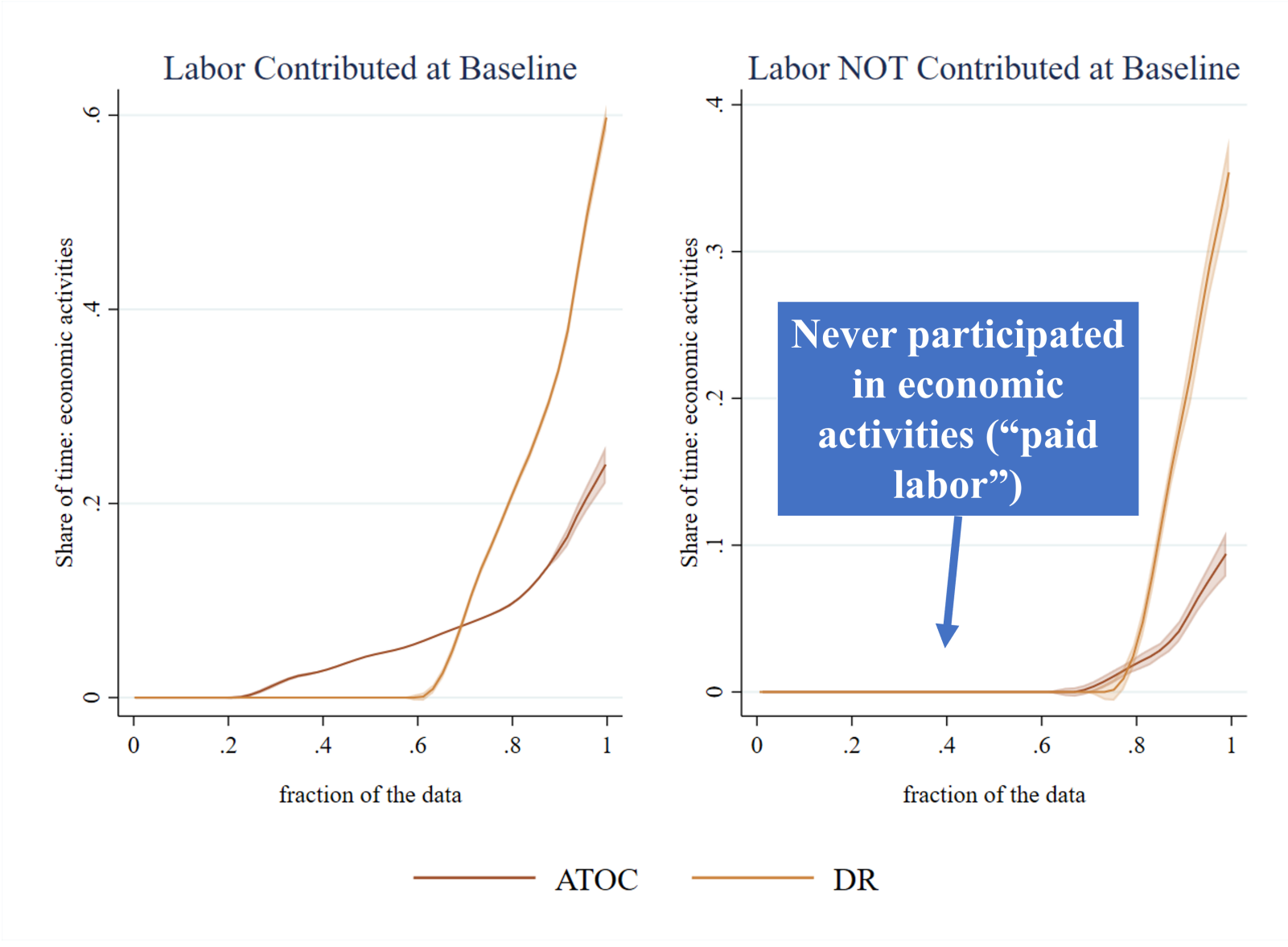
Comparing the distribution of economic activity shares

subsample: did not participate during 7 days prior to baseline



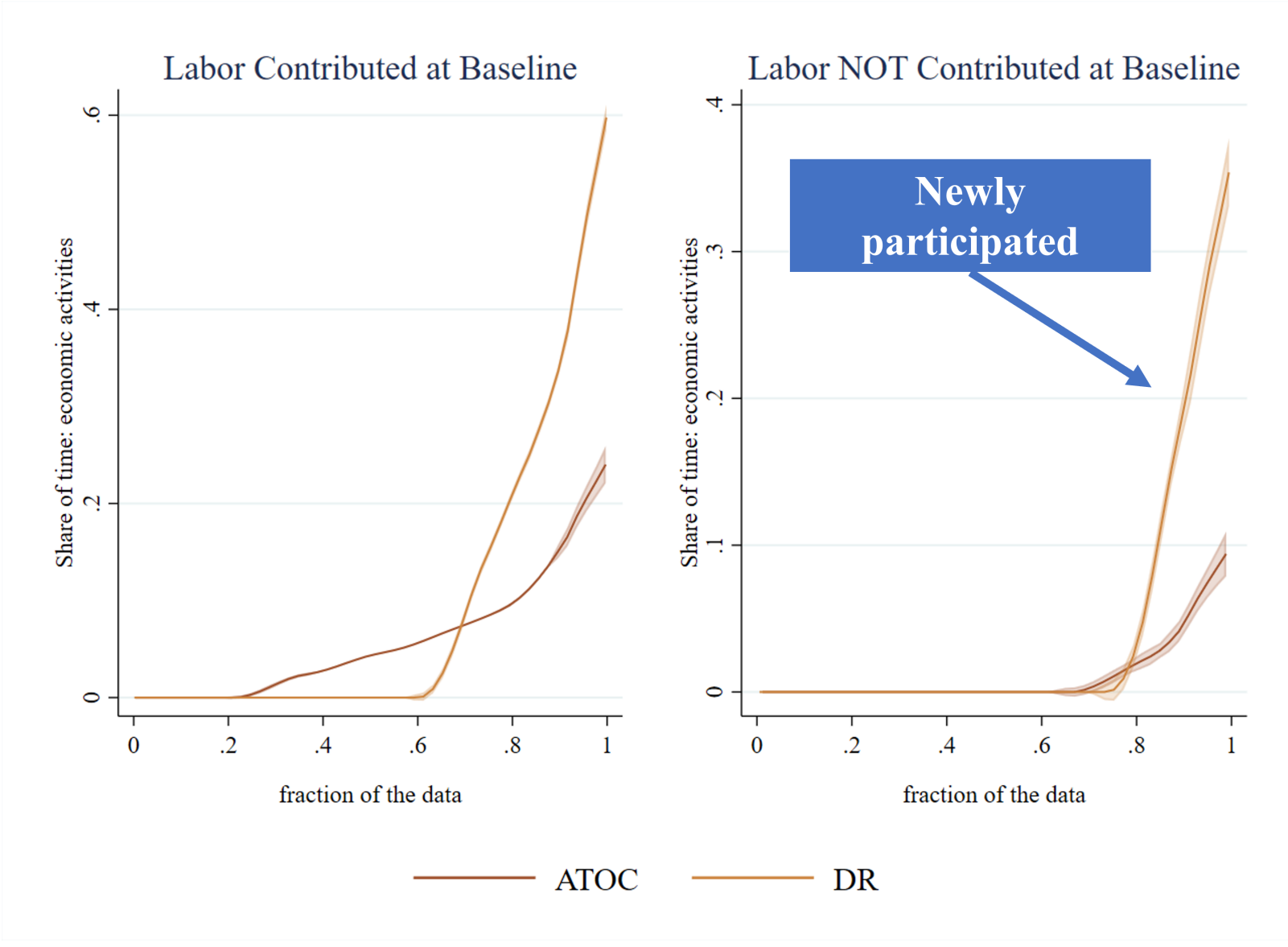
Conditional Distribution of Economic Activity Shares

by respondent's economic participation at baseline ▸ Decomposition



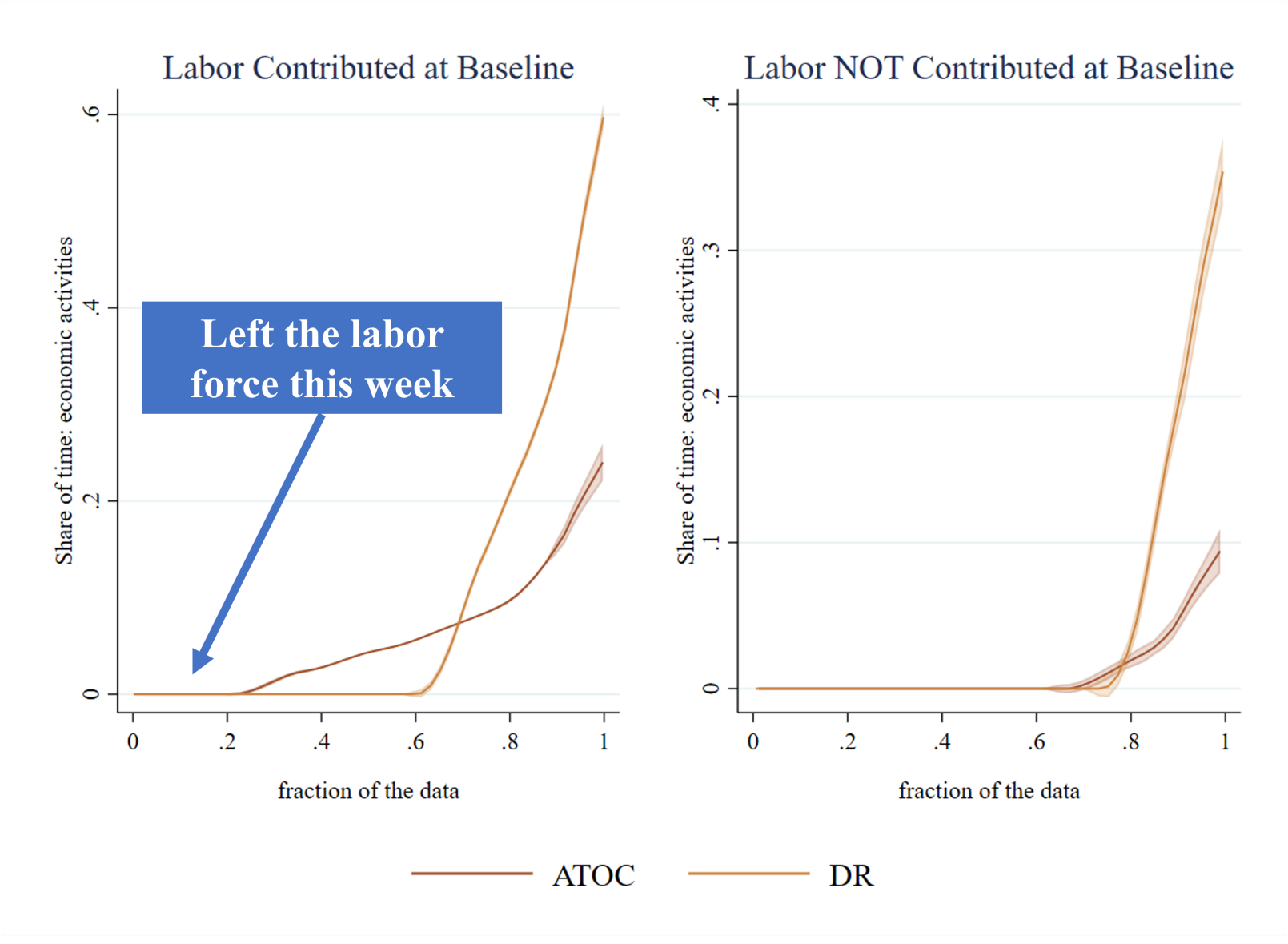
Conditional Distribution of Economic Activity Shares

by respondent's economic participation at baseline ▸ Decomposition



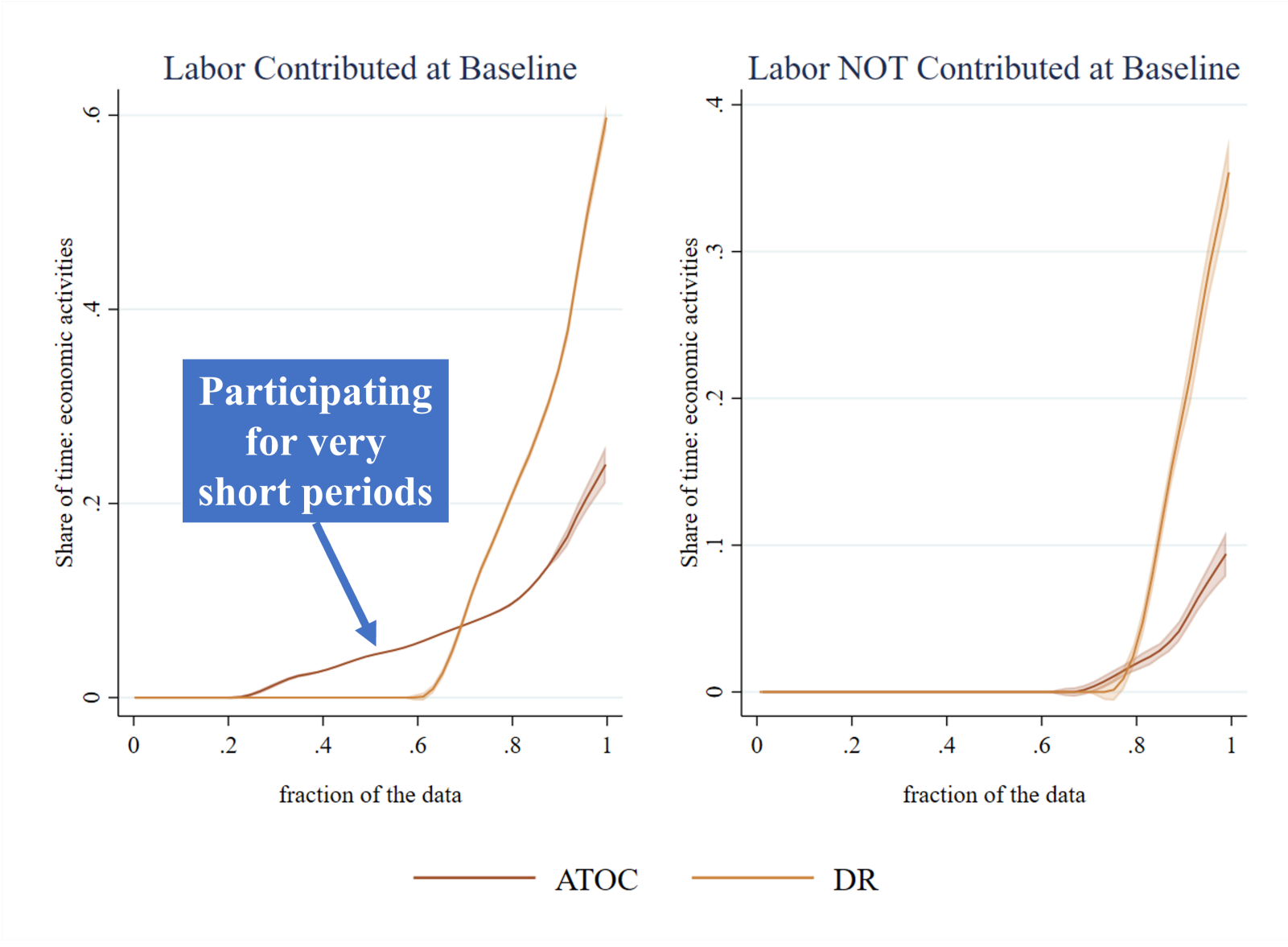
Conditional Distribution of Economic Activity Shares

by respondent's economic participation at baseline ▸ Decomposition



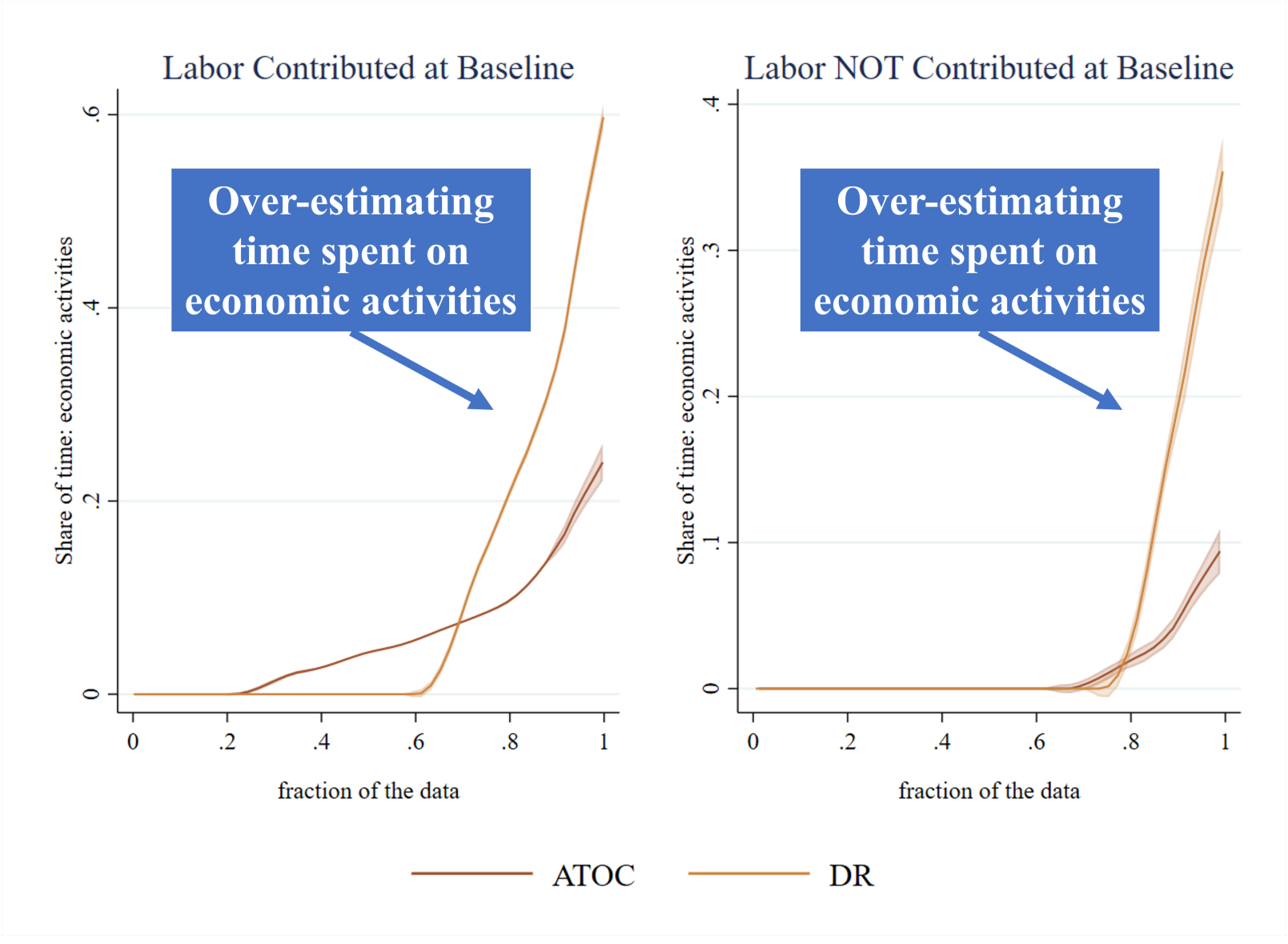
Conditional Distribution of Economic Activity Shares

by respondent's economic participation at baseline ▸ Decomposition



Conditional Distribution of Economic Activity Shares

by respondent's economic participation at baseline ▸ Decomposition



Measurement error mechanisms we can set aside

Our results include evidence dismissing:

- Hawthorne/social desirability effects - reports are consistent over days in survey
- Endowment/survey burden effects - frequent calls and phone provision have no impact on DR reports

Measurement error mechanisms we can set aside

Our results include evidence dismissing:

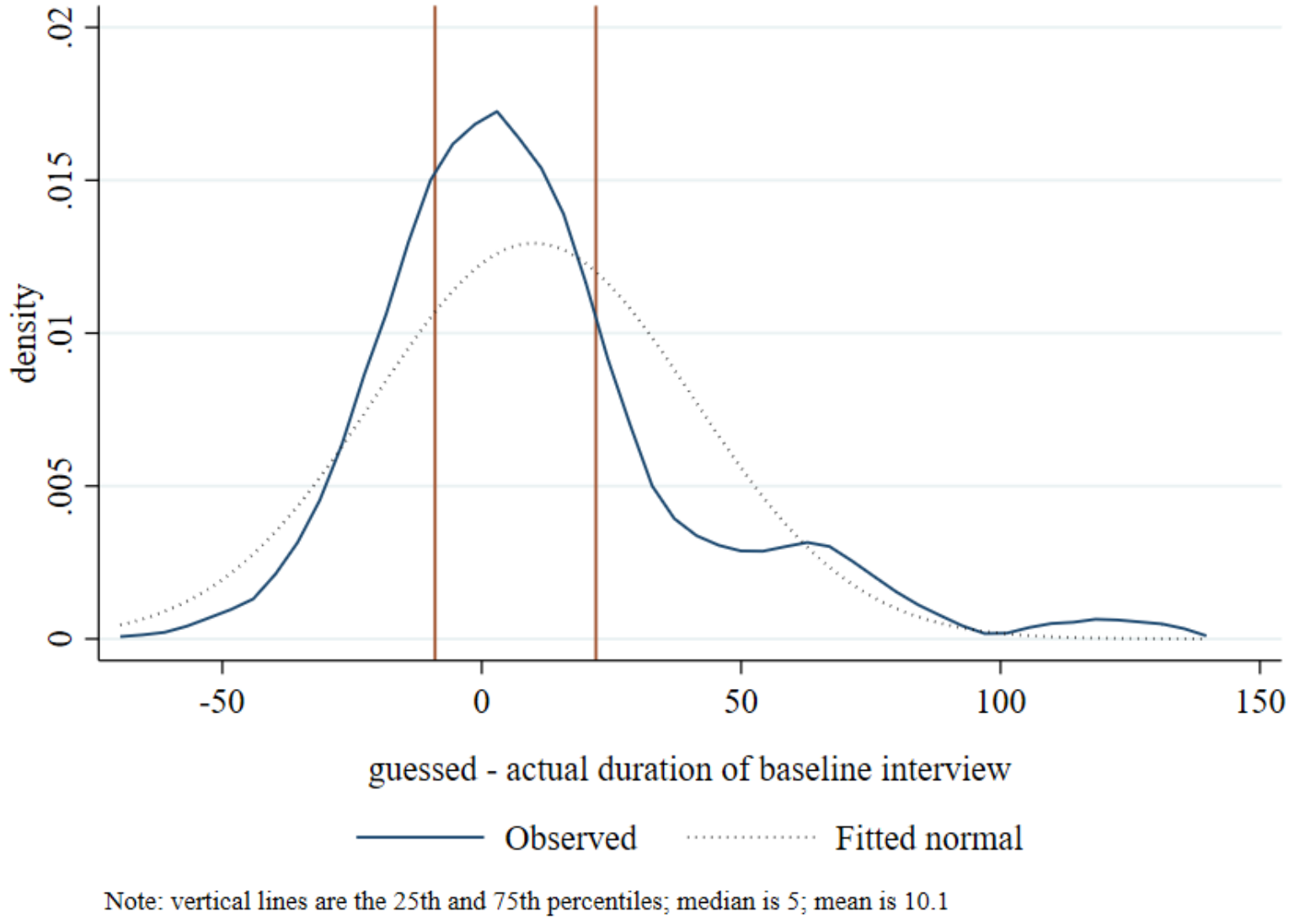
- Hawthorne/social desirability effects - reports are consistent over days in survey
- Endowment/survey burden effects - frequent calls and phone provision have no impact on DR reports

We show impacts of these are minimal:

- Selection effects (which calls answered) - bounded and minimal
- Person-specific time literacy - results not explained by characteristics, e.g., watch ownership
- Survey design effects (e.g., 24-hr DR vs 7 days ATOC reference periods) - do not drive results, though hard to isolate from recall error

Ability to report baseline interview duration varies

IQR is [-5,+20]



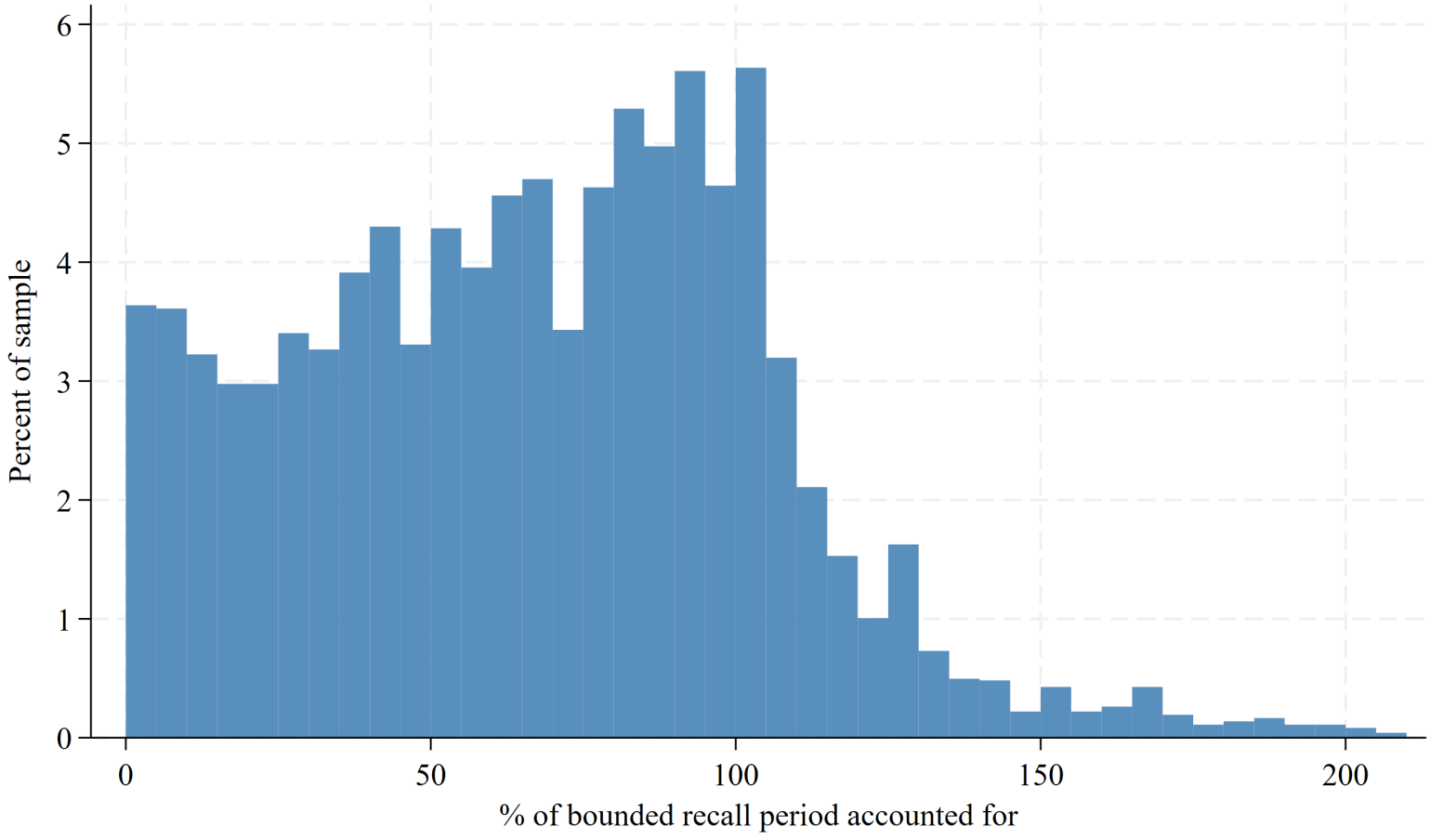
Note: vertical lines are the 25th and 75th percentiles; median is 5; mean is 10.1

Explanations for remaining differences: recall error

Recall error mechanisms in DR:

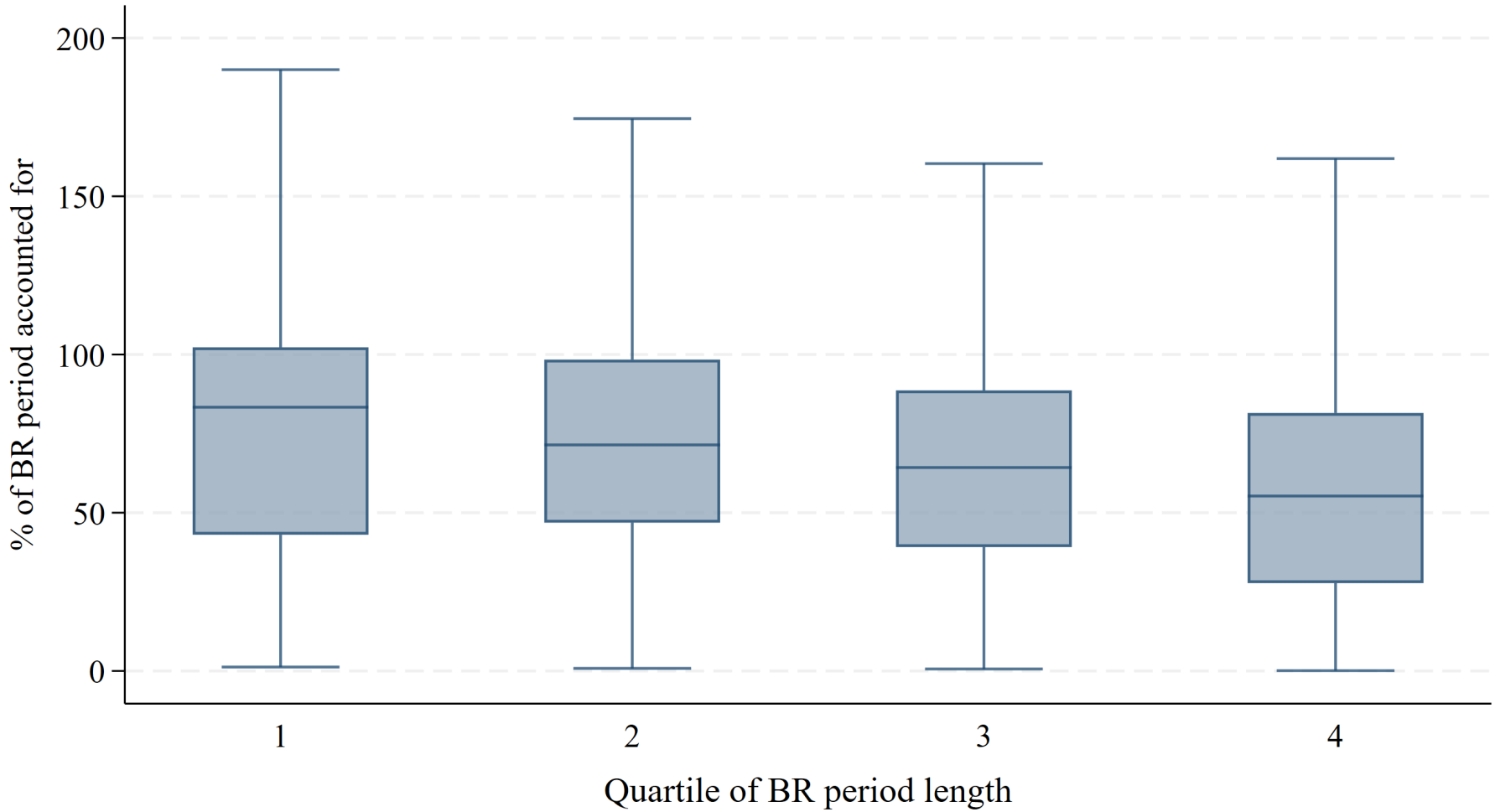
- **Activity switching & durations:** days are busy sequences of activities, brief ones are omitted and long ones over-estimated
- **Activity classification:** e.g., commuting may be coded as work in DR
- **Reference period length:** longer bounded recall windows $\Rightarrow$ lower share of time accounted for

Respondents cannot account for all time between calls (bounded recall periods)



Recall error mechanisms: Longer reference periods are harder to explain

▸ Ref period length by accuracy

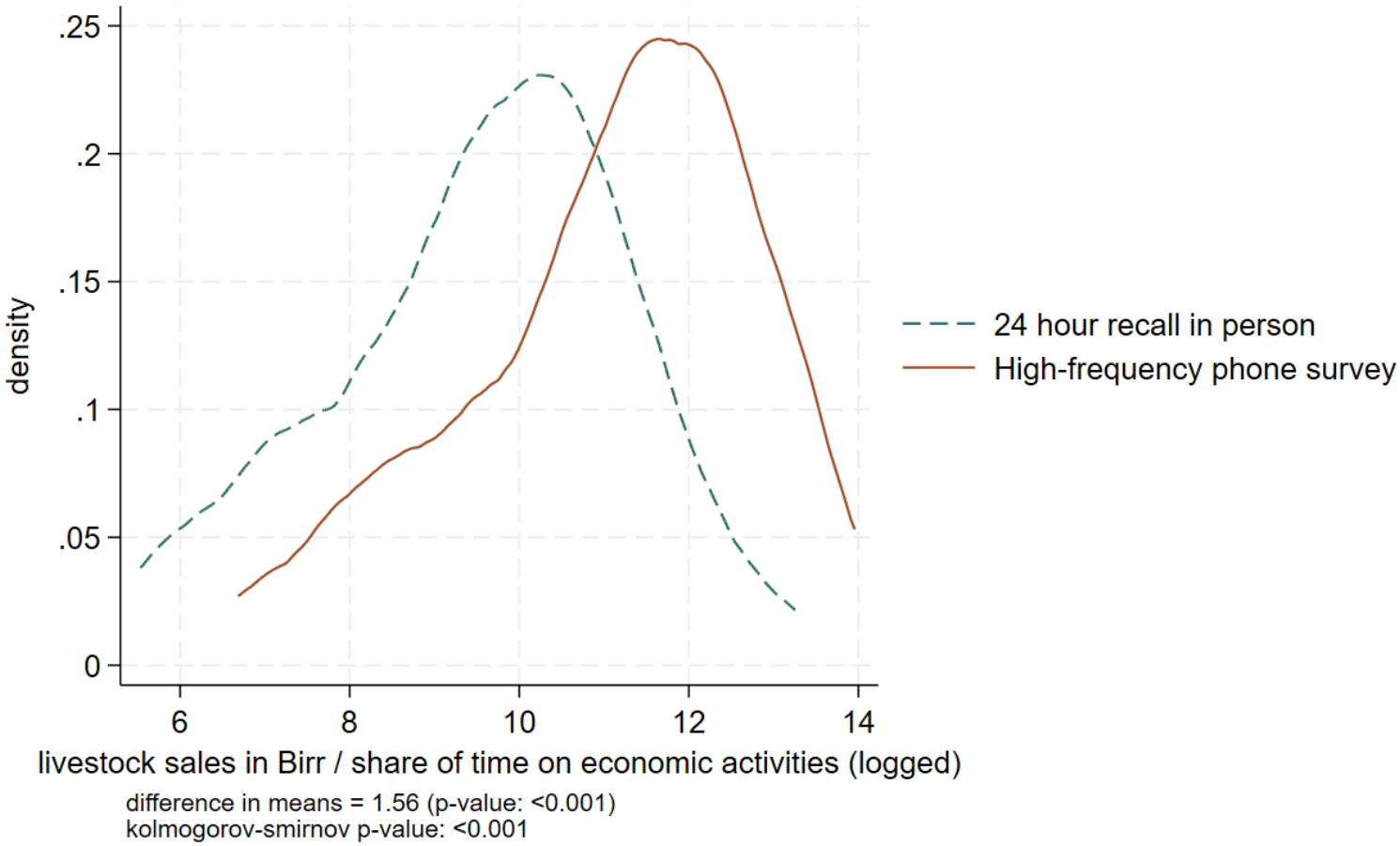


Top and bottom of boxes are 75th and 25th percentiles; horizontal line within box is the median value for respondents in each decile; whiskers extend past the box by 1.5*IQR; omits outside values beyond this

Implied Labor Productivity

if livestock sales and economic activities reported

▸ Labor supply is much more elastic with ATOC



difference in means = 1.56 (p-value: <0.001)
kolmogorov-smirnov p-value: <0.001

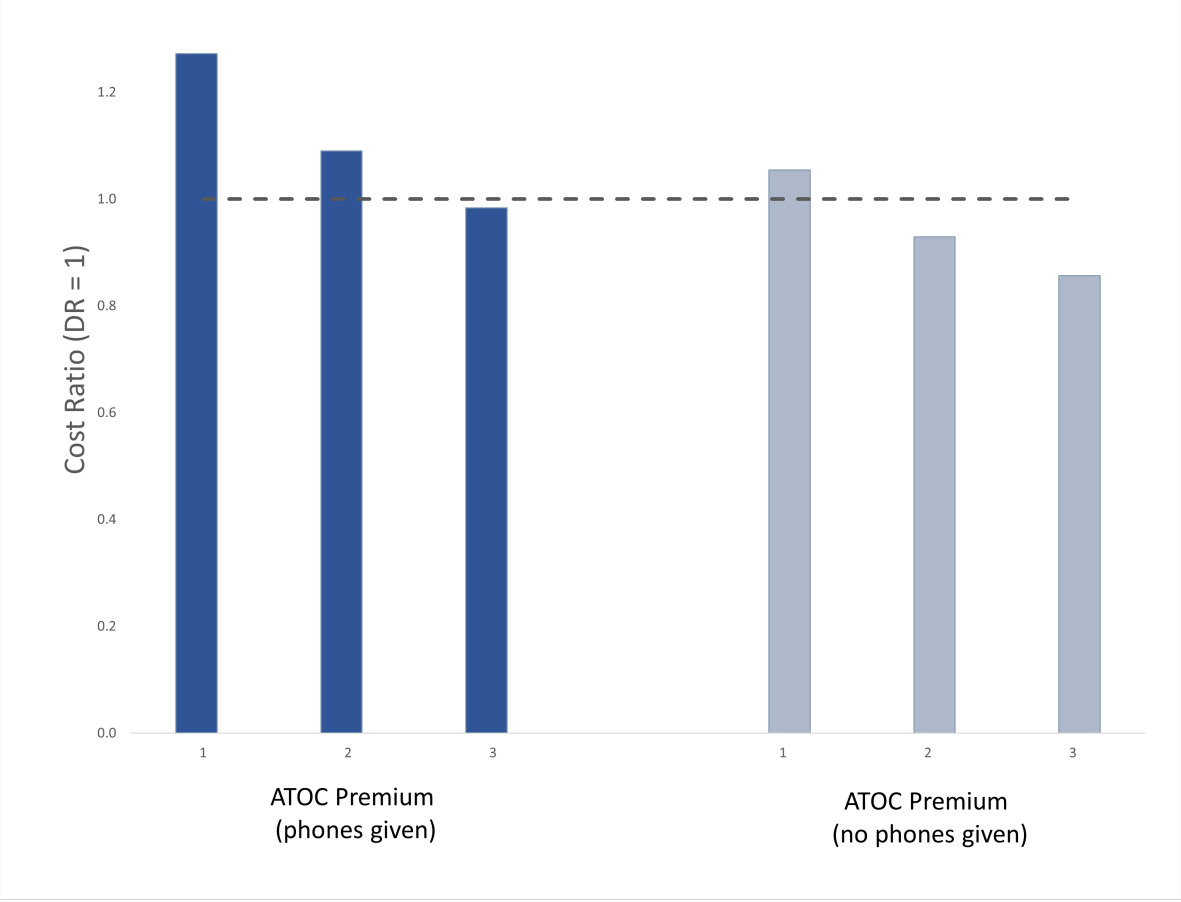
Discussion

- Labor supply errors lead to under-estimation of productivity:
  - LF *participation* is underestimated and dynamic
  - Labor *hours* are overestimated, "chores" under-estimated
  - Labor supply elasticity is **negative!** for ATOC; null for DR
- Errors are common and non-classical:
  - Errors relate to participation. Working more $\rightarrow$ greater over-estimation
  - Errors not driven by respondent characteristics, appear with all types of respondents
- Opportunities to better understand intra-household dynamics:
  - Spouses misestimate each other's activities
  - Childcare likely under-reported with both methods

Scaling ATOC is feasible

ATOC has lower implementation costs than DR for repeated rounds

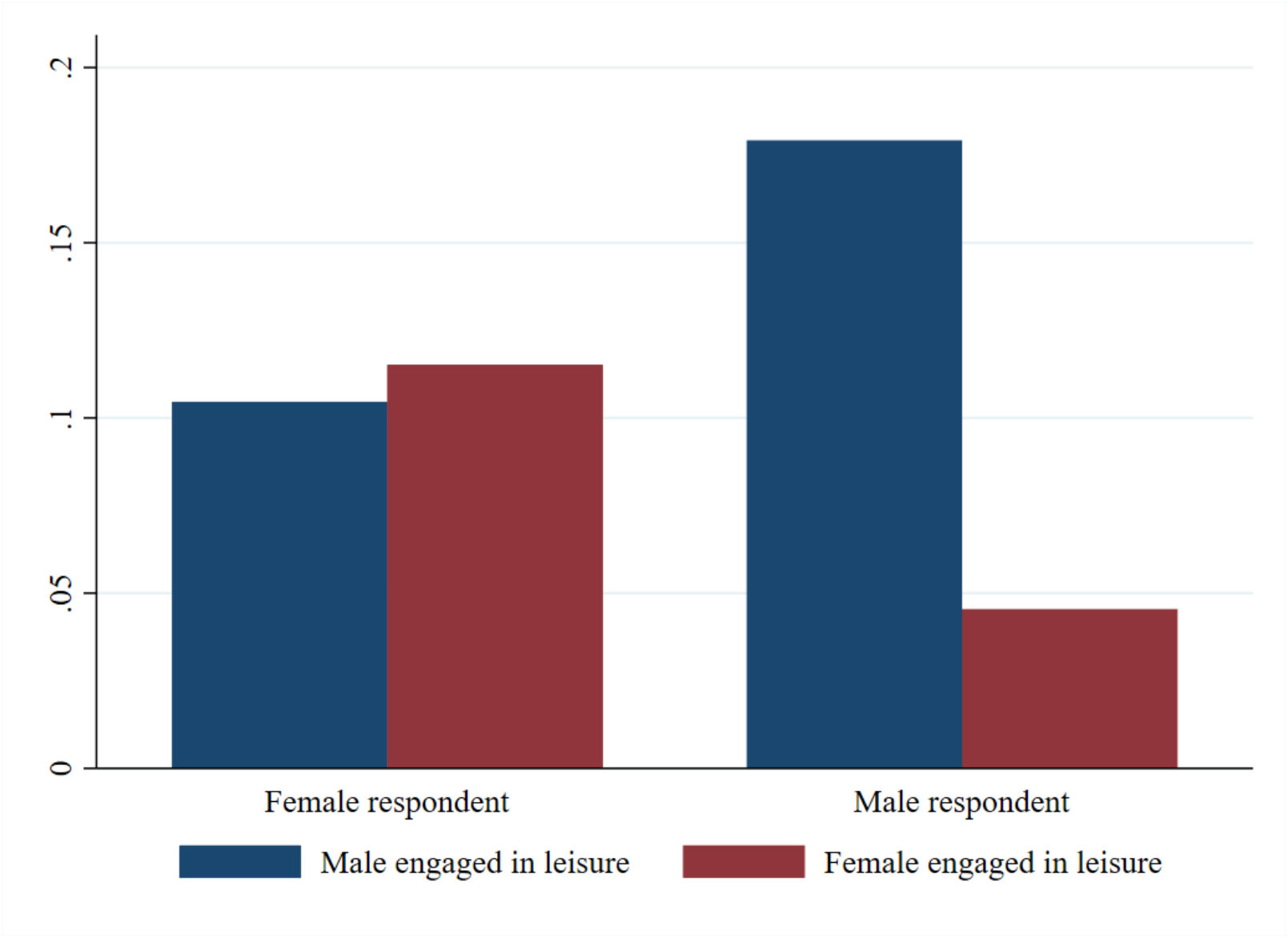
Figure: ATOC cost premium (vs DR) by number of endline rounds



Assumptions: ATOC surveys use an in-person baseline and also collect time use outcomes by phone at baseline and endline. Sample size held constant (1000).

Separate sample: Self- vs. spouse-reports

Engaged in leisure (ATOC method) - new paper: "Sharing is caring: Redistributing unpaid domestic and care work improves well-being for fathers and mothers" (Gupta et al., 2026)



References I

- Agness, D. et al. (Jan. 2025). "Valuing the Time of the Self-Employed". In: *The Review of Economic Studies* 92.6, pp. 3471–3503.
- Arthi, V. et al. (2018). "Not your average job: Measuring farm labor in Tanzania". In: *Journal of Development Economics* 130, pp. 160–172.
- Bick, A., N. Fuchs-Schündeln, and D. Lagakos (2018). "How Do Hours Worked Vary with Income? Cross-Country Evidence and Implications". In: *American Economic Review* 108.1, 170–99.
- Borjas, G. and D. Hamermesh (2024). "The Mismeasurement of Work Time: Implications for Wage Discrimination and Inequality". In: NBER No. 32025.
- Bound, J., C. Brown, and N. Mathiowetz (2001). *Measurement Error in Survey Data*. Vol. 5. Elsevier Masson SAS, pp. 3705–3843.

References II

- Breza, E. and S. Kaur (2025). "Labor Markets in Developing Countries". In: *Annual Review of Economics* 17.Volume 17, 2025, pp. 747–776.
- Csikszentmihalyi, M. (2014). *Flow and the Foundations of Positive Psychology*.
- Dabalen, A. et al. (2016). *Mobile Phone Panel Surveys in Developing Countries: A Practical Guide for Microdata Collection*. Tech. rep. World Bank Group: Directions in Development.
- Dillon, A. et al. (2012). "Explaining variation in child labor statistics". In: *Journal of Development Economics* 98.1. Symposium on Measurement and Survey Design, pp. 136–147.
- FAO and The World Bank (2018). *Food data collection in household consumption and expenditure surveys*. Vol. March, pp. 1–87.

References III

- Foster, A. D. and M. R. Rosenzweig (2022). "Are There Too Many Farms in the World? Labor Market Transaction Costs, Machine Capacities, and Optimal Farm Size". In: *Journal of Political Economy* 130.3, pp. 636–680.
- Gaddis, I. et al. (2021). "Measuring farm labor: survey experimental evidence from Ghana". In: *World Bank Economic Review* 35.3, pp. 604–34.
- Gollin, D., D. Lagakos, and M. E. Waugh (2014). "The agricultural productivity gap". In: *Quarterly Journal of Economics* 129.2, pp. 939–993.
- Gollin, D. and C. Udry (2021). "Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture". In: *Journal of Political Economy* 129.1, pp. 1–80.
- Gottlieb, C. et al. (2026). *The Gender Division of Work across Countries*. Tech. rep.

References IV

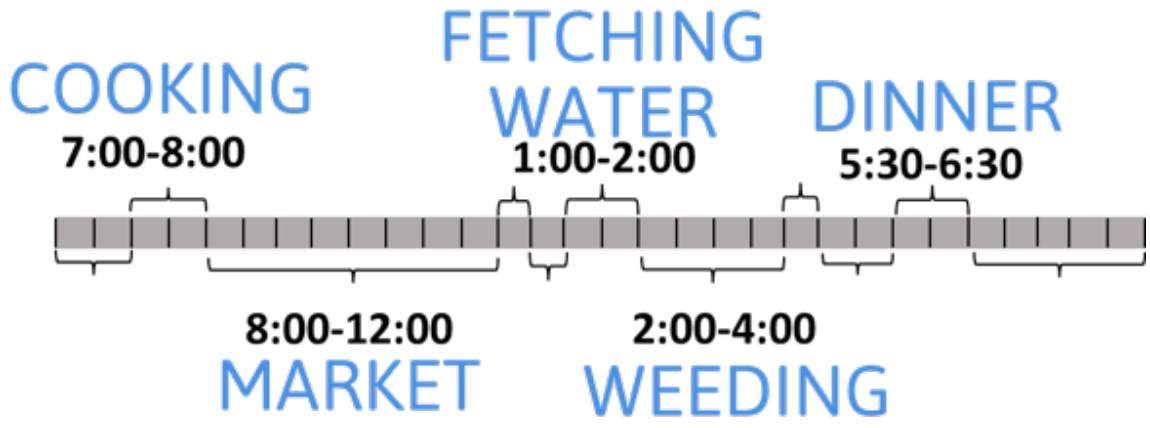
- Kahneman, D. et al. (2004). "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method". In: *Science (Washington, D.C.)* 306.5702, pp. 1776–1780.
- Krueger, A. B. and D. A. Schkade (2008). "The reliability of subjective well-being measures". In: *Journal of Public Economics* 92.8-9, pp. 1833–1845.

Appendix

Day Reconstruction Method—"Gold Standard"

▸ Kahneman, Krueger, et al. (2004) ▸ Neuroscience: Sensing Time

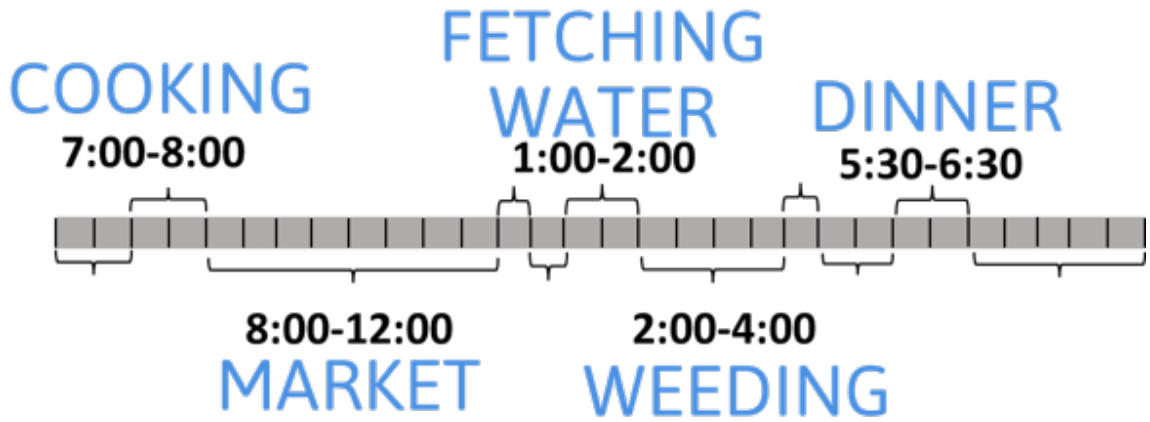
▸ Back to ATOC



Day Reconstruction Method—"Gold Standard"

▸ Kahneman, Krueger, et al. (2004) ▸ Neuroscience: Sensing Time

▸ Back to ATOC



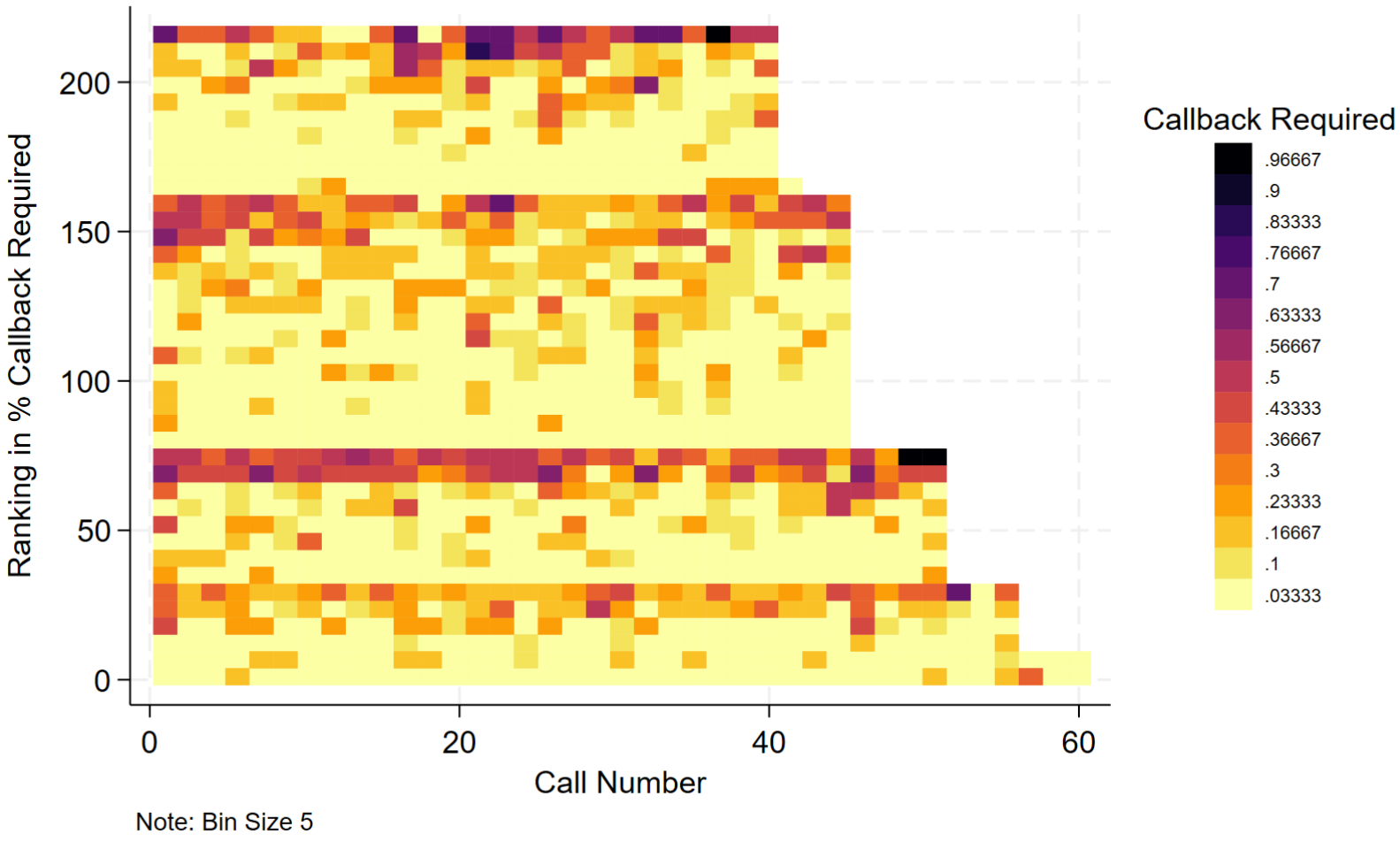
Challenges: ▸ Exacerbated in Low-Income Settings

- Long, tedious interview $\rightarrow$ limited info. on simultaneous activities, companions, quality of interactions
- Seasonality/informality (high within-person σ^2, irregular LFP)
- Innumeracy, (time) (il-)literacy, neuroscience of sensing time

(Borjas and Hamermesh, 2024; Gollin and Udry, 2021; Gaddis et al., 2021; Arthi et al., 2018; Dillon et al., 2012; Kahneman et al., 2004; Bound, Brown, and Mathiowetz, 2001)

Individual-level compliance

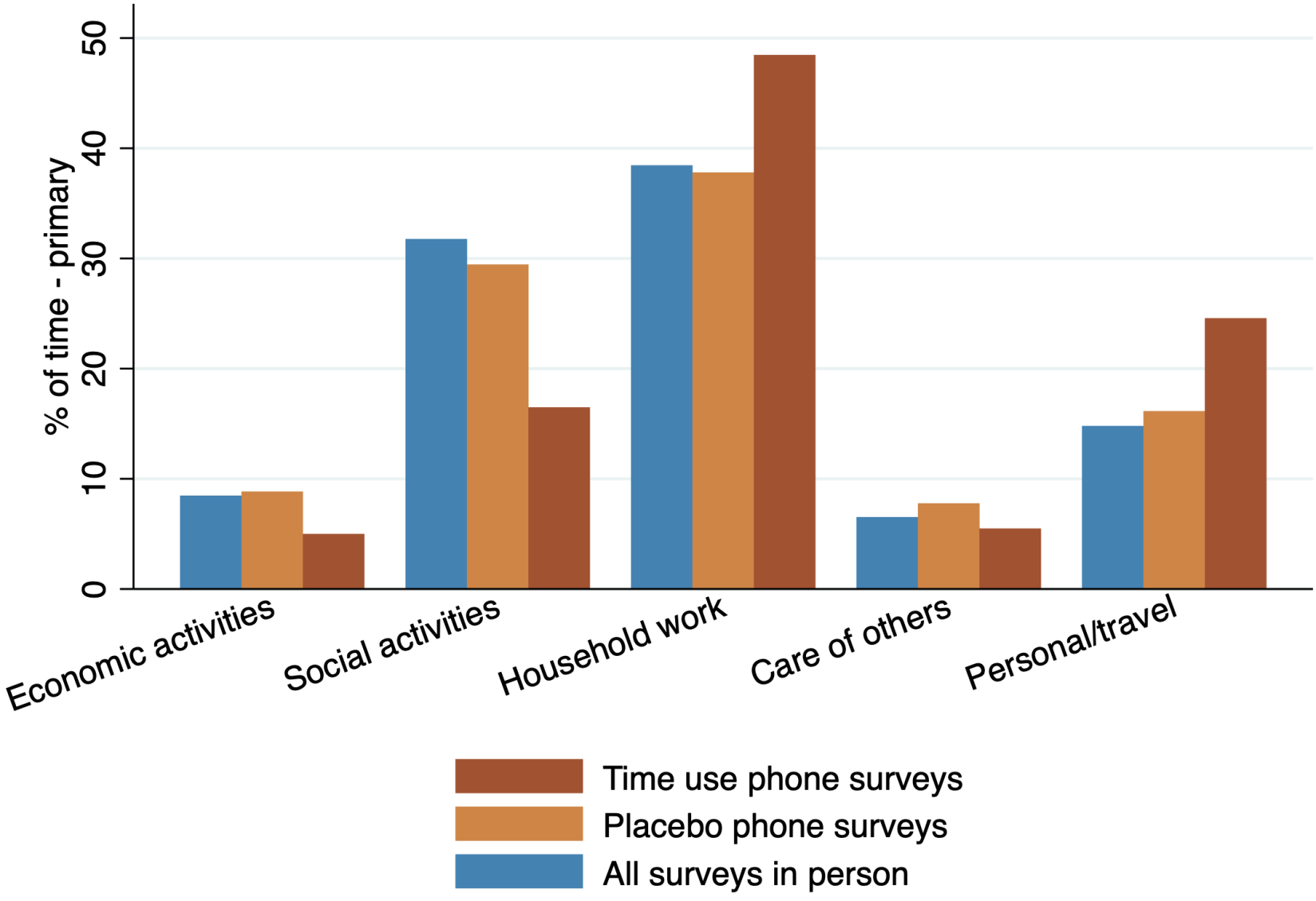
Back



Note: Bin Size 5

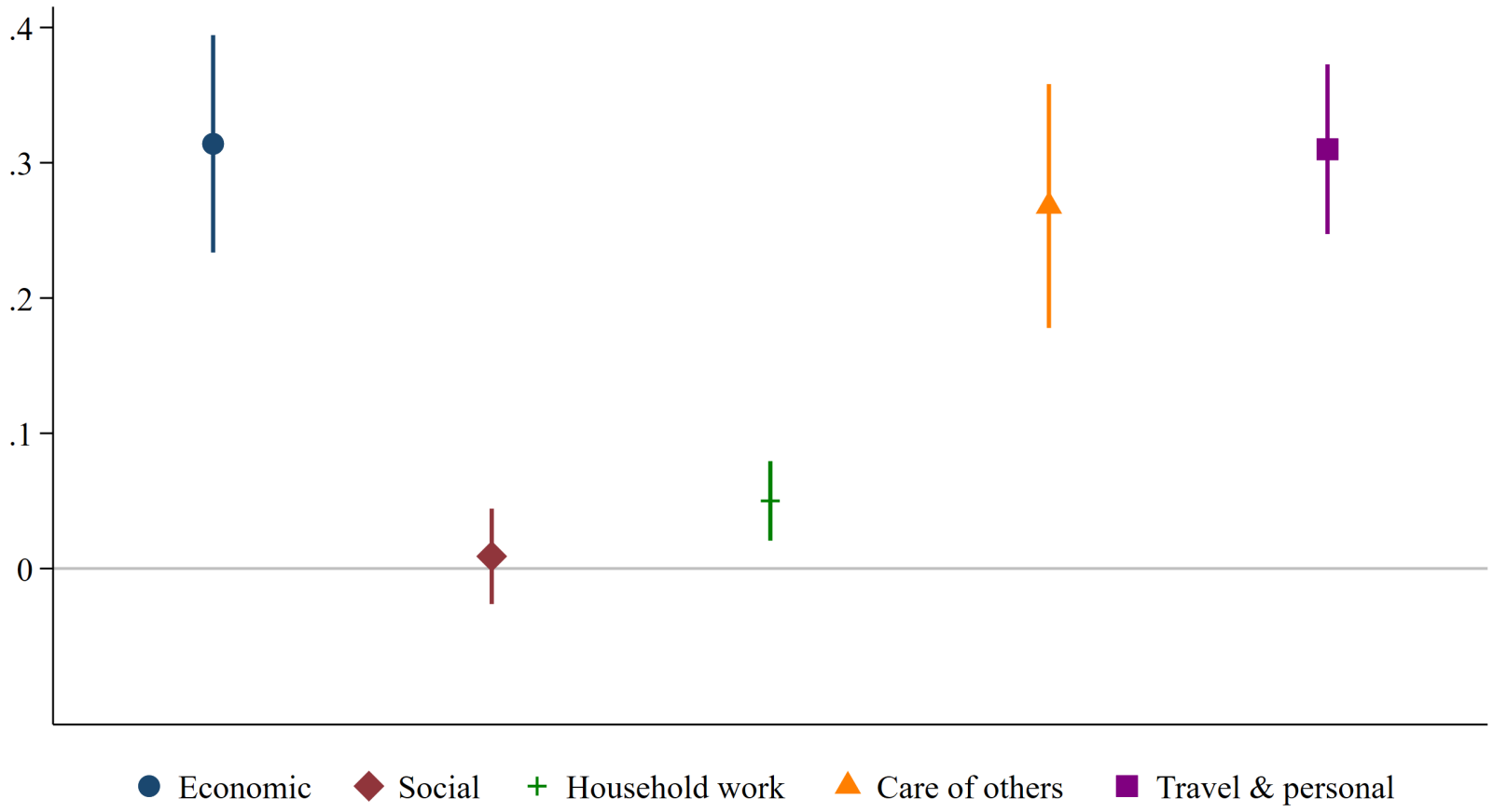
Average activity shares (DR vs. ATOC)

Back



Decomposition results: participation effect (extensive)

> 0 if ATOC participation higher ▸ Decomposition equations ▸ Back



Bars show 90 and 95% confidence intervals based on standard errors estimated for extensive and intensive margin of activity *a* jointly using the delta method.

Decomposition results: conditional duration (intensive)

> 0 if ATOC participation higher ▸ Decomposition equations

Bars show 90 and 95% confidence intervals based on standard errors estimated for extensive and intensive margin of activity *a* jointly using the delta method.

Decomposing average differences

Back

Define for activity a and survey method s:

μ_a^s = participation rate = % of respondents participating

σ_a^s = conditional duration = E [activity share | participated]

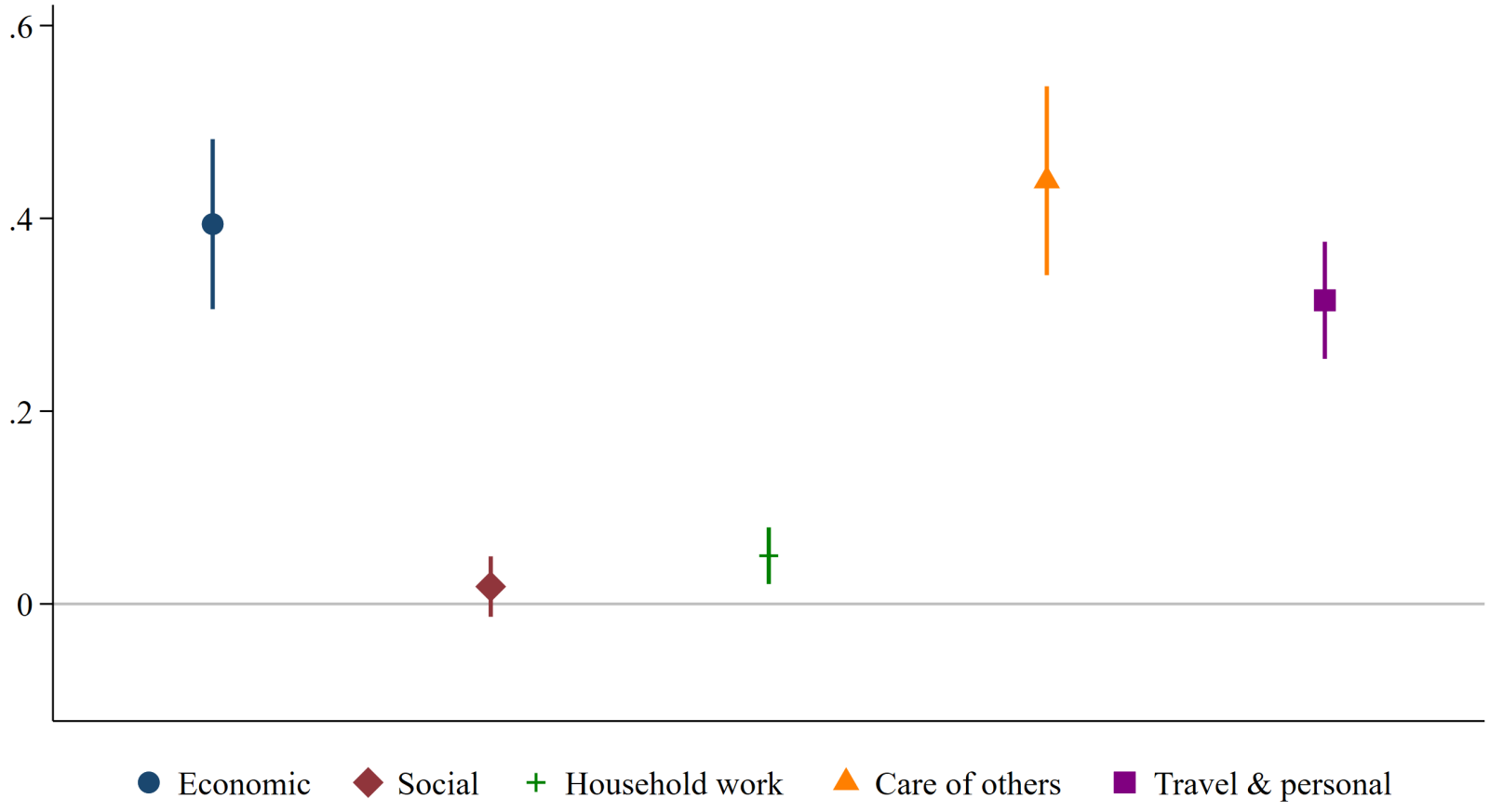
Decompose the average difference (LATE):

$$\begin{aligned}\beta_a &= \sigma_a^{\text{DR}} \cdot \left(\mu_a^{\text{ATOC}} - \mu_a^{\text{DR}}\right) + \left(\sigma_a^{\text{ATOC}} - \sigma_a^{\text{Dr}}\right) \cdot \mu_a^{\text{ATOC}} \\ &= \sigma_a^{\text{DR}} \cdot \Delta\mu_a + \Delta\sigma_a \cdot \mu_a^{\text{ATOC}}\end{aligned} \tag{1}$$

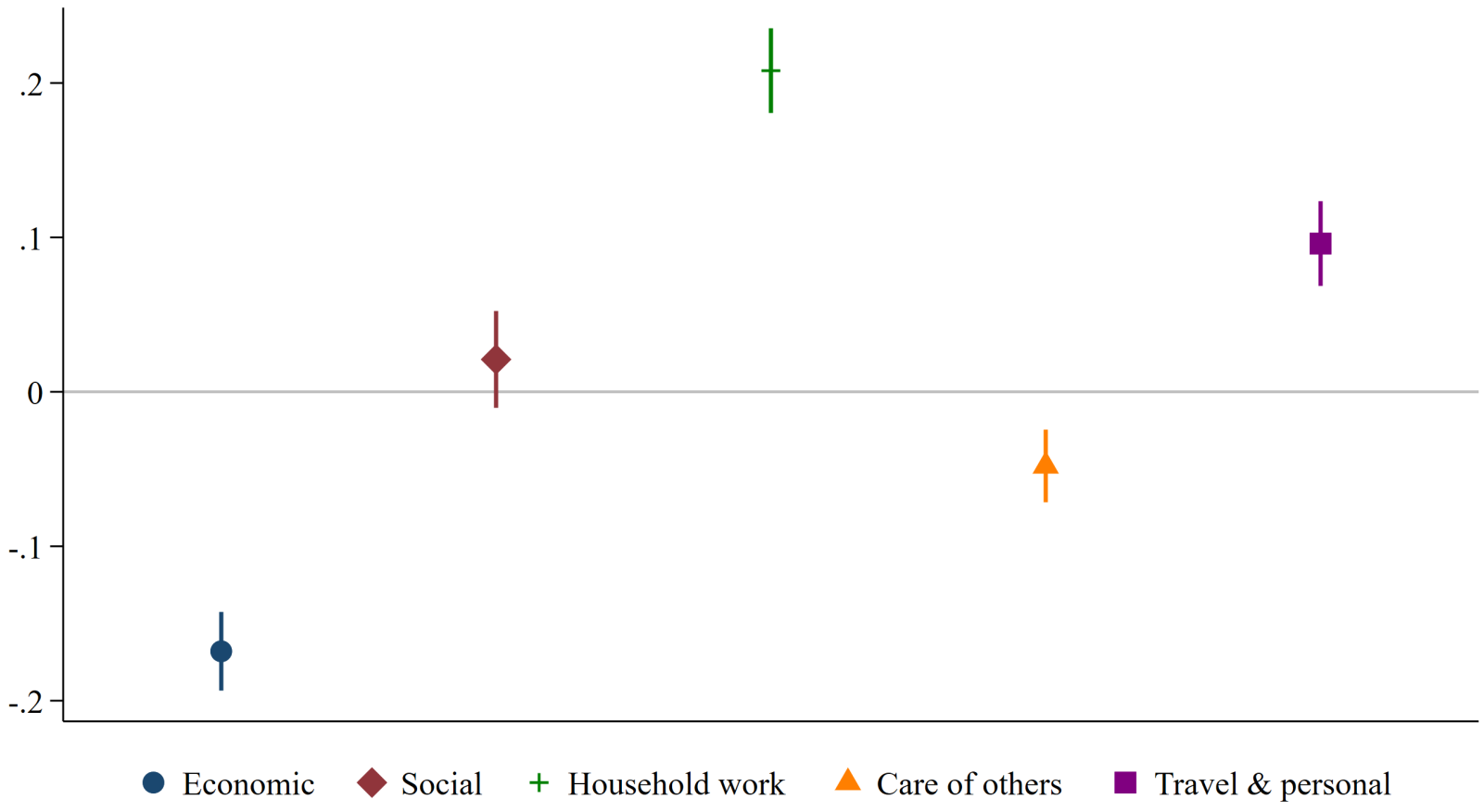
Joint estimation identifies effects and standard errors for:

- $\Delta\mu_a$ = a participation effect (extensive margin)
- $\Delta\sigma_a$ = a conditional duration effect (intensive margin)

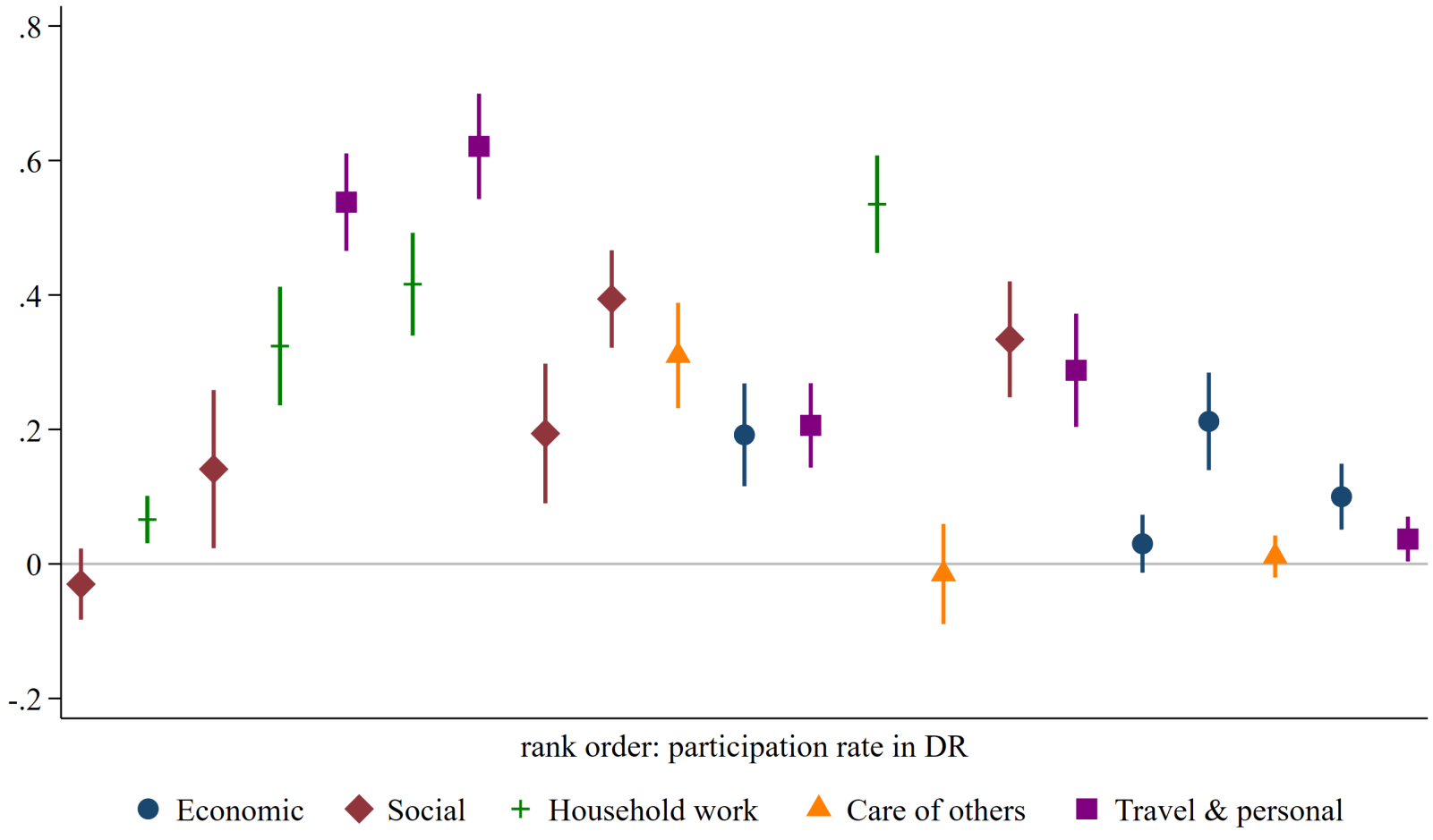
Decomposing average differences: Extensive margin (BR version)



Decomposing average differences: Intensive margin (BR version)

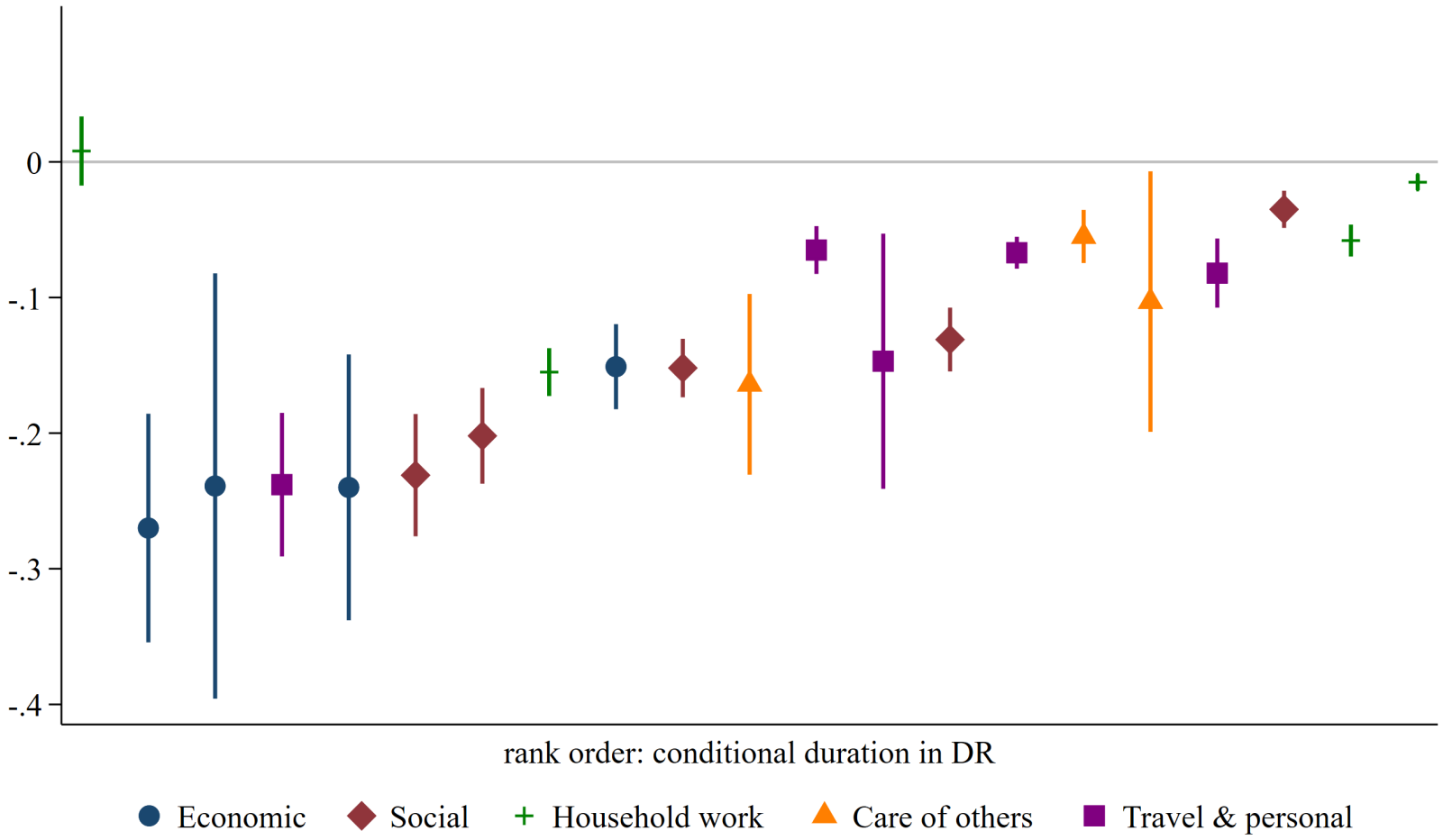


Participation effects vary across subcategories



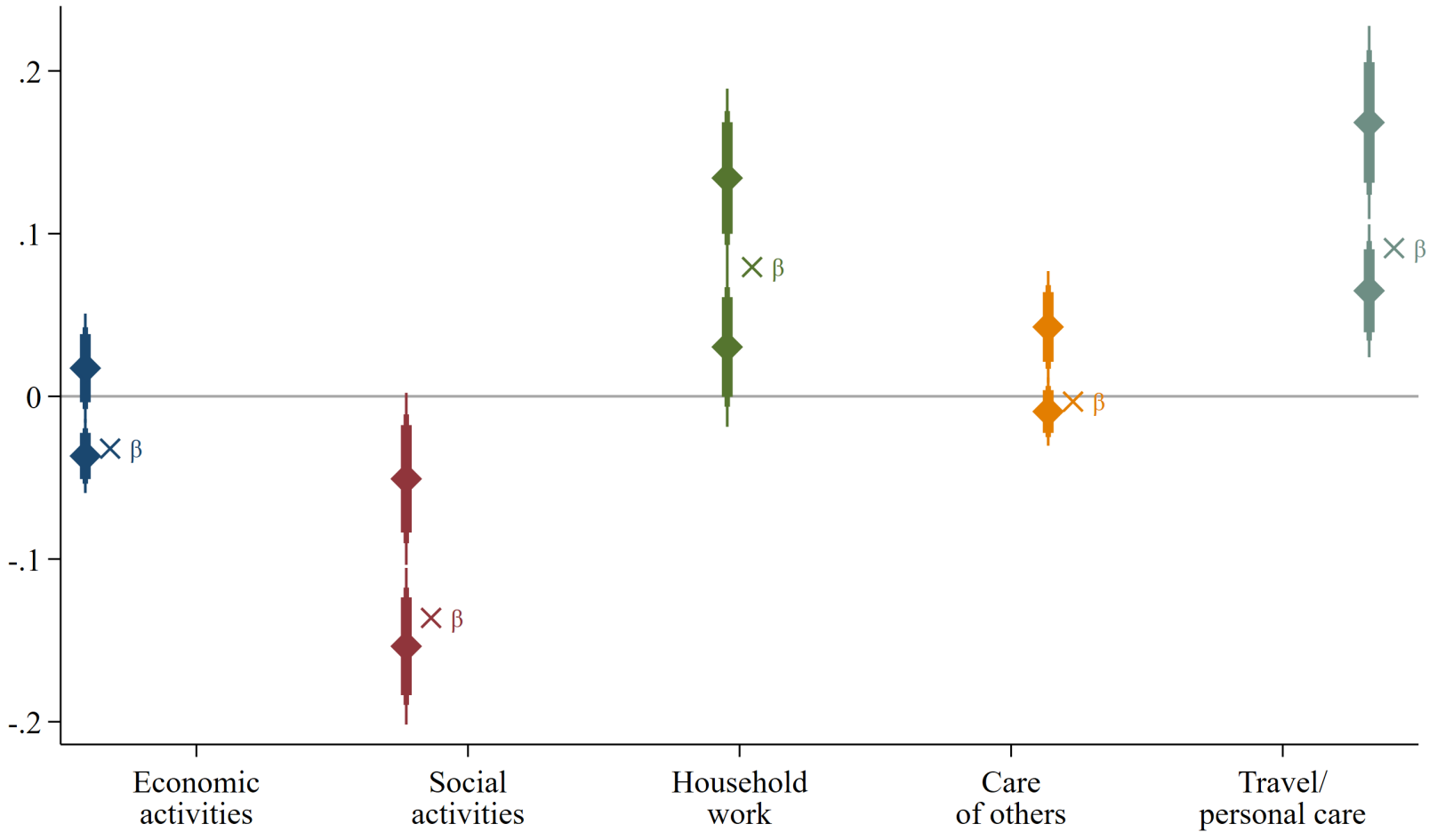
Bars show 90 and 95% confidence intervals based on standard errors estimated for extensive and intensive margin of activity *a* jointly using the delta method.

Conditional duration effects are correlated with reported duration



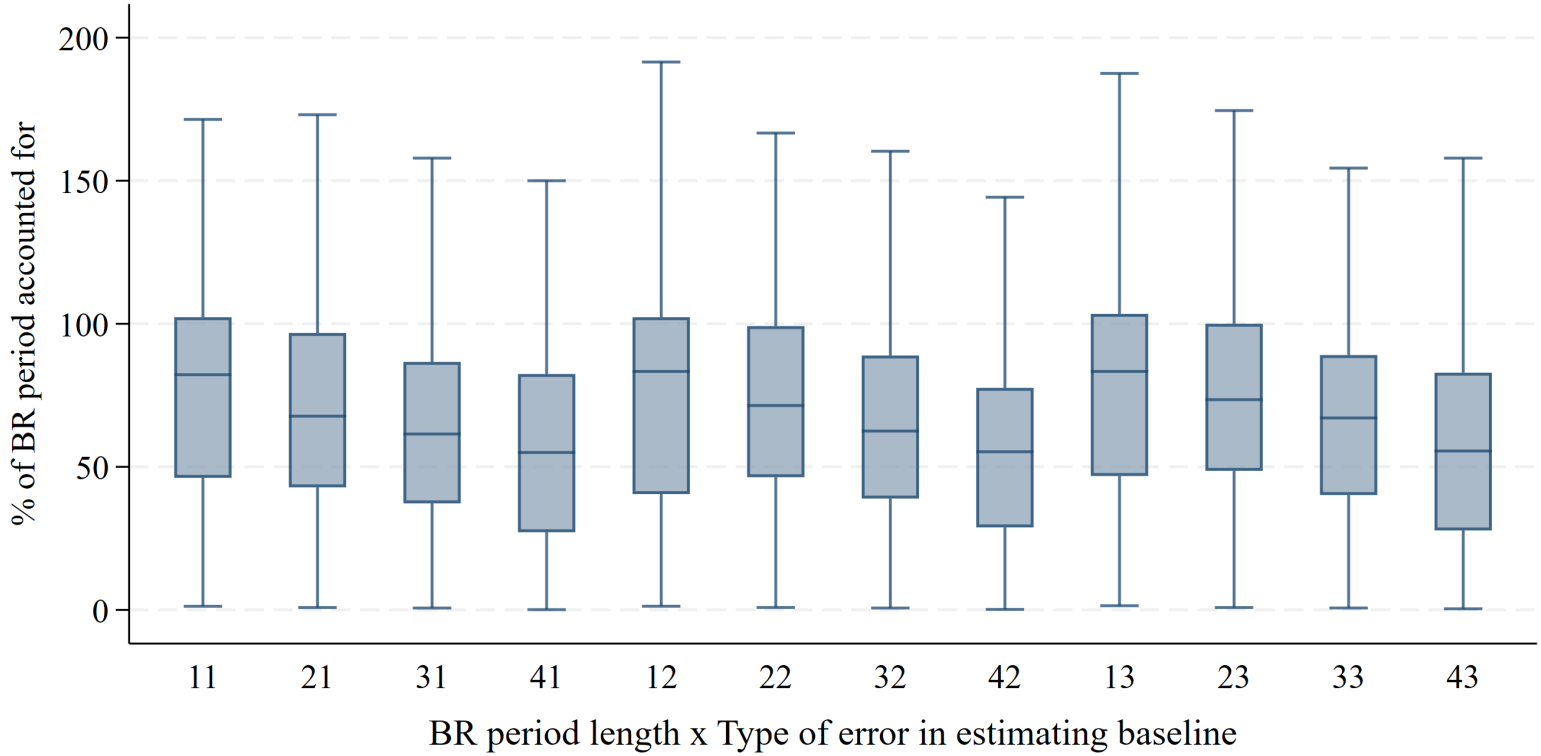
Bars show 90 and 95% confidence intervals based on standard errors estimated for extensive and intensive margin of activity *a* jointly using the delta method.

Bounding possible selection effects



Recall error mechanisms: Duration of bounded recall period matters but *not* respondent's baseline interview error

Back



X-axis labels (AB):
A denotes BRPL quartile (1 = shortest, 4 = longest);
B denotes baseline estimation error (1 = under, 2 = close, 3 = over)
Top and bottom of boxes are 75th and 25th percentiles; horizontal line within box is the median value for respondents in each decile; whiskers extend past the box by 1.5*IQR; omits outside values beyond this