

Replication: Relative income within the household, gender norms, and well-being

Jules Tucher

2025-12-16

```
library(tidyverse)
library(haven)
library(rdrobust)
library(fixest)
library(knitr)
library(kableExtra)
library(pwrss)
library(rdhte)
library(ggpubr)

# set local directory and open analysis datafile
root_dir <- getwd()
data_raw <- read_dta(file.path(root_dir, "Replication Files Plos", "tables.dta"))
```

Introduction and Background

The article for this replication, “Relative income within the household, gender norms, and well-being,” is a 2024 study that uses regression discontinuity design (RDD) on longitudinal observation data to examine the effects of women out-earning their husbands on overall life satisfaction and health outcomes (Gihleb, et al. 2024). It explores the differences in effect between East and West Germany, where social norms about women’s participation in the workforce differs.

In this project, I will replicate the main figures from the original paper, as well as performing the two RD analyses. The replication package includes a Stata dataset and two Stata program files to replicate the main findings in the paper. The raw data is not included, but instead a clean analysis file that includes some of the constructs from the analysis, while others need to be created.

The remainder of this document is structured as follows. The introduction concludes with the research hypothesis. The methods section includes a description of the data, the quantitative measures used in the analysis and their construction in R, and the RD model proposed by the original authors. The results section includes a visual exploration of the running variable, an implementation of the base RD model in R, plots for the relationships between key outcomes and the running variable near the threshold for men and women, and results from the base and region-specific RD models. Finally, the conclusion contains a discussion of the results and how they differ from the original paper and the assumptions and limitations of the methods employed.

Research Hypothesis

Although the research team does not explicitly state their hypotheses in the published article, the null hypothesis is that relative income has no effect on the satisfaction and mental and physical health of men and women in Germany. The confirmatory hypothesis of the study is that women earning more than their husbands causes people to be less satisfied with their lives, with a theoretical framing based on cultural and gender norms that expect men to be the breadwinners. Exploratory hypotheses investigate the moderating effect of region on this causal relationship, based on an understanding of the differing cultural and gender norms in East and West Germany. Because it became more normalized for women to enter the work force during the Cold War in East Germany, cultural values shifted toward more equal work expectations of men and women.

Methods

Discription of the Data

The replication dataset provided by the research team is at the couple-level, with values for the male and female members of the heterosexual couple. The unit of analysis is the couple. The population is married adults in Germany. This observational data was collected from the German Socioeconomic Panel (GSOEP), a longitudinal socioeconomic survey. Note that this data includes multiple observations for the same couple, as the same sample completed the survey year after year. The authors account for this by modeling outcomes adjusted for fixed effects.

The methods employ a regression discontinuity design, comparing outcomes between couples where women earn just a bit more than their husbands. Therefore, the explanatory or running variable is the difference in income between the two members of the couple. The outcomes of interest include satisfaction with life, health, and work, as well as physical and mental health ratings. Finally, the research team was interested in regional differences between East and West Germany, specifically as associated with different cultural and gender norms in the two regions. Other covariates include age, wage, number of children, and college attainment of each member of the couple, as measured by the GSOEP.

The couple-level dataset provided in the replication package has already handled matching spouses and some covariate cleaning. However, construction of the running variable, standardization of outcomes (including accounting for fixed effects), and cleaning of the remaining covariates is required for replication. Operationalization strategies from the original paper will be followed for this procedure. Obtaining the original raw data from the GSOEP is out of scope for this project, as the data is publicly available but requires an application process to obtain.

Quantitative Measures

The data provided by the research team is a Stata dataset. Once loaded into R, remove labels from labeled numeric variables before converting all variables with value labels to factors. Then, construct a dummy variable for college education. Finally, filter the dataset by the constraints described in the original paper: husband between 18 and 65, couple-level (remove duplicate female rows), and no missing covariates. Other covariates include region (East/West), number of children, age, and education.

```
# convert categorical variables with value labels to dummy variables
data_sample <- data_raw %>%
  # remove value label for numeric variables
  mutate(
    across(
      starts_with(c("partner", "pglabgro", "plc0013_h", "plc0016",
```

```

        "ijob1", "overallllifesat", "worksat", "m11125")),
      zap_labels
    )
  ) %>%
  # convert value labels to factors
  mutate(
    across(
      where(haven::is.labelled),
      haven::as_factor
    )
  ) %>%
  # create dummy variables
  mutate(college = (edu_isced == "College or More") * 1,
         college_female = (edu_isced_female == "College or More") * 1
  ) %>%
  # filter according to missing covariates and age limitations
  # age constraints for males
  filter(age > 17, age < 65) %>%
  # keep partnered males, since female data already added in columns and dataset at partner level
  filter(partner == 1) %>%
  filter(female == 0) %>%
  # filter if missing education of either partner
  filter(!is.na(college), !is.na(college_female)) %>%
  # filter if missing number of children for either partner
  filter(!is.na(nchildren), !is.na(nchildren_female))

```

Wage variables for both members of the couple are provided in the data set. To construct the running variable, subtract the husband wage from the wife's wage. Also create a recoded version that is scaled to 1,000s of Euros and an integer version that will be used for error clustering. Finally, create a treatment indicator for if the running variable is above or below 0. Filter out any observations with missing running variable.

```

# create constructs
data_rv <- data_sample %>%
  mutate(
    ## Running variable / treatment
    # female-male wage difference (running variable)
    diff_wage_OR = ifelse(!is.na(wage) & !is.na(wage_female), wage_female - wage, NA_real_),
    # scaled
    recode_diff_wage_OR = diff_wage_OR/1000,
    # integer (for error clustering)
    recode_diff_wage_OR_int = round(diff_wage_OR/1000, 0),
    # treatment variable based on RV
    treatment = (diff_wage_OR>0) * 1,
  ) %>%
  # filter if missing running variable
  filter(!is.na(recode_diff_wage_OR))

```

Next, consider the five outcome variables. The three satisfaction variables (life, work, and health) are constructed based on survey questions on an 11-point Likert scale. No other construction details are provided, but exploration of the provided data indicates that variables are averages on that same scale. The other two outcome variables in the main analysis, mental and physical health, are pulled from the 12-Item Short Form Health Survey, a subsurvey of the GSOEP administered every 2 years. The 12 questions are asked on

the 0 to 100 scale, and converted to averages and standardized for this analysis. The final consideration in the construction of the analysis dataset is to account for fixed-effects of time-varying outcomes, since the data set contains multiple records per couple (across survey years). To account for the fixed effects of these time-varying outcomes, regress each outcome to a constant with couple-level fixed-effects. Then, extract residuals from the model and replace the outcome with the residual.

```
# declare outcome variables and running var
running_var = "recode_diff_wage_OR"
outcomes = c("overalllifecat", "worksat", "m11125", "std_mcs", "std_pcs",
              "overalllifecat_female", "worksat_female", "m11125_female",
              "std_mcs_female", "std_pcs_female")

# index the data so we can lookup later
analysis_data <- data_rv %>%
  # standardize health and mental-health outcomes
  mutate(
    across(
      c('mcs', 'pcs', 'mcs_female', 'pcs_female'),
      scale,
      .names='std_{.col}')
    ) %>%
  rowid_to_column()

## recalculate outcomes to account for individual fixed effects
for (outcome in outcomes) {
  # fitting regression on participant ID
  fe_model <- feols(as.formula(paste(outcome, "~ 1 | pid")), data = analysis_data)

  # get the rows used in the exact sample
  model_rows <- fixest_data(fe_model, sample = "estimation")$rowid

  # create residual variable for each outcome, storing NA when not in the model
  analysis_data[[paste0(outcome, "_res")]] <- NA
  analysis_data[model_rows, paste0(outcome, "_res")] <- resid(fe_model)

  # confirm rows not included in model are those with NA resid
  stopifnot(analysis_data[is.na(analysis_data[[paste0(outcome, "_res")]]), "rowid"] ==
            analysis_data[!(analysis_data$rowid %in% model_rows), "rowid"])
}
```

Estimation Procedure

This study implements an RDD analysis to estimate the local average treatment effect (LATE) using OLS linear regression. The model is specified as:

$$Y_{iht} = \beta_1 I[E_{iht}^{Wife} \geq E_{iht}^{Husband}] + \beta_2 (E_{iht}^{Wife} - E_{iht}^{Husband}) + \beta_3 (E_{iht}^{Wife} - E_{iht}^{Husband}) * I[E_{iht}^{Wife} \geq E_{iht}^{Husband}] + \epsilon_{iht} + \eta_i$$

Here, Y is the metric of well-being or health, $E_{iht}^{Wife} - E_{iht}^{Husband}$ is the running variable constructed for the treatment, β_1 is the coefficient of interest, η_i represents the fixed effects at the individual level, and ϵ_{iht} is the couple-level error term. Controls are omitted from this equation but present in the model.

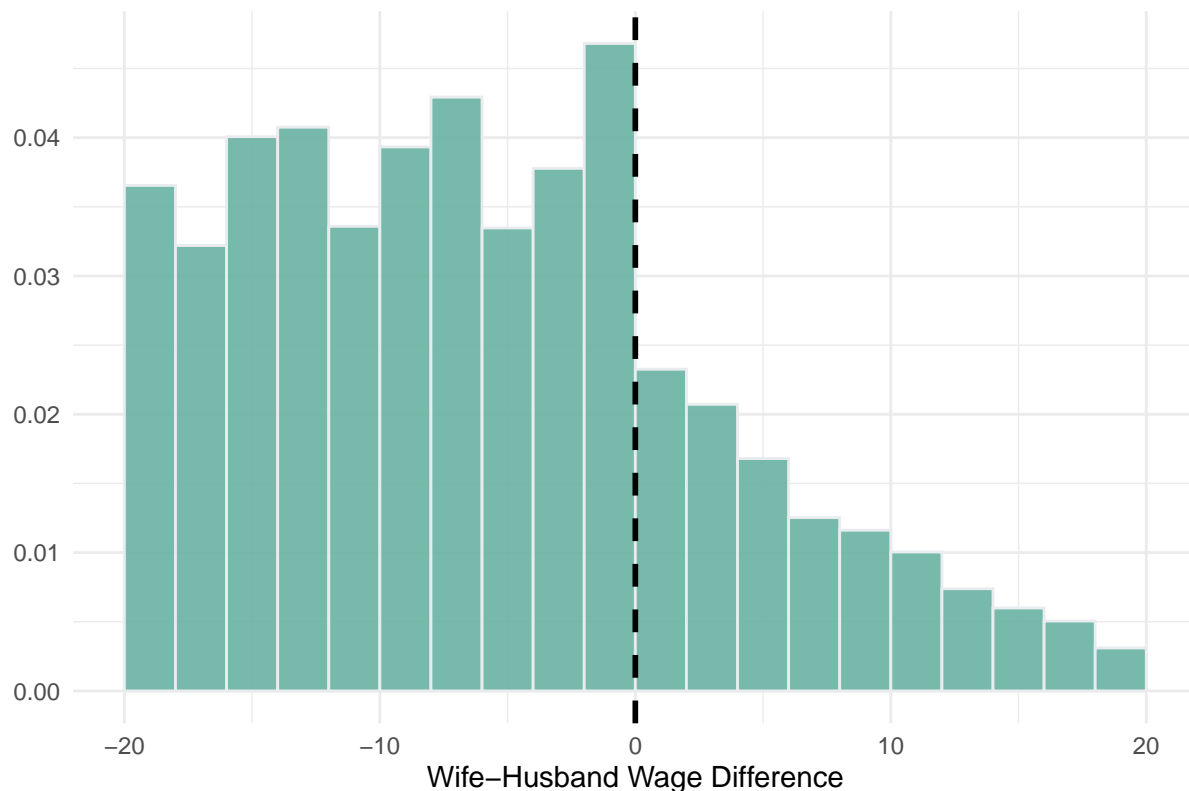
Results

Running Variable: Wife-Husband Wage Difference

```
# restrict to within 20k of threshold
threshold_data <- analysis_data %>%
  filter(abs(recode_diff_wage_OR) < 20)

threshold_data %>%
  ggplot(aes(x=recode_diff_wage_OR)) +
    geom_histogram(aes(y = after_stat(density)), boundary = 0, binwidth=2,
                  color="#e9ecef", fill="#69b3a2", alpha=0.9) +
    theme_minimal() +
    theme(
      plot.title = element_text(size=15)
    ) +
    geom_vline(aes(xintercept = 0), color = "black", linetype = "dashed", linewidth = 1) +
    labs(
      x = "Wife-Husband Wage Difference",
      y = "",
      title = paste0("Fig 1. Running variable density near threshold"),
      fill = ""
    )
)
```

Fig 1. Running variable density near threshold



The above density plot (Fig. 1) depicts the distribution of the running variable (wife-husband wage difference in €1,000) for those observation whose running variable is within €20,000 of the threshold. As seen in the

original paper, there is a sharp decline in the number of couples immediately after 0, indicating a discontinuity at the cutoff.

Implementing the model in R

To implement the base regression discontinuity model in R, we employ the `rdrobust` package, which provides estimation and plotting tools. A model is estimated for each of the 10 outcomes (5 variables for each member of the couple) by providing the outcome variable (adjusted for individual fixed effects), the running variable, the RD threshold (in this case, 0), and data for covariate controls (gender-dependent age, wage, number of children and dummy for college degree). No age variable for the female member of the couple was included in the replication data set, so the presumably male age covariate is substituted in the female outcome models. Two other parameters allow specification of the bandwidth size and clustering. Optimal bandwidth is determined for each outcome using mean square error (MSE) selection criteria and robust inference following Gihleb et al. Standard errors are clustered by relative income bins (€1,000 intervals). Finally, the `rdplot` function is employed to plot the relationship between running variable and outcome within the bandwidth.

```
# store covariate data (consistent across outcomes)
male_covs = c("age", "wage", "nchildren", "college")

# store female covs
female_covs = c("age", "wage_female", "nchildren_female", "college_female")

# set threshold to 0 and pull out running variable
threshold <- 0.00001

results <- list()
plots <- list()
for (outcome in outcomes) {

  # set covariate names depending on male/female
  covs <- if (grepl("female", outcome)) female_covs else male_covs

  # rdrobust: outcome ~ running | cutoff, clustered SE
  rd <- rdrobust(y = analysis_data[[paste0(outcome, "_res")]],
                x = analysis_data[[running_var]],
                c = threshold,
                covs = analysis_data[, covs, drop=FALSE], # controls
                bwselect = "mserd", # use MSE selection criteria
                cluster = analysis_data$recode_diff_wage_OR_int # integer version of RV
                )

  # get sample used in RD so we can report stats (non-missing outcome and within benchmark)
  rd_sample <- !is.na(analysis_data[[paste0(outcome, "_res")]]) & # outcome not missing
  # check if treated
  ifelse(analysis_data[[running_var]] <= threshold,
        # if not treated, use left bandwidth
        abs(analysis_data[[running_var]] - threshold) <= rd$bws["h", "left"],
        # if treated, use right bandwidth
        abs(analysis_data[[running_var]] - threshold) <= rd$bws["h", "right"])
  rd_effective_data <- analysis_data[rd_sample,]

  # store results for later table
```

```

results[[outcome]] <- list(
  # pull robust effect estimate, error, and pval
  rd_effect = rd$coef[3],
  rd_error = rd$se[3],
  rd_pval = rd$pv[3],
  n_effective = nrow(rd_effective_data),
  y_mean = mean(rd_effective_data[[outcome]]),
  y_sd = sd(rd_effective_data[[outcome]])
)

# create RD plots (for lifesat and health)
if (grepl("overalllifesat", outcome) | grepl("m11125", outcome)) {

  plots[[outcome]] <- rdplot(
    y = rd_effective_data[[paste0(outcome, "_res")]],
    x = rd_effective_data$recode_diff_wage_OR,
    c = threshold,
    covs = rd_effective_data[, covs, drop=FALSE],
    binselect = "es",
    h = rd$bws["h",],
    p = 1,
    ci = 95
  )$rdplot
}
}

```

Replicating Figure 2: Gender norms, well-being and health, men.

Following Fig. 2 in the original publication, the below figure depicts the relationship between relative income levels and satisfaction with life and health among men in heterosexual couples. Standard errors are clustered by relative income bins (€1,000 bins) and specifications include fixed effects. Bandwidth and binning selection were conducted with MSE-optimization and the IMSE-optimal evenly spaced method, respectively. Bars depict 95% confidence intervals.

```

ggarrange(
  plots[["overalllifesat"]]+
    ggtitle("Satisfaction with life") +
    xlab("Scaled relative income near threshold") +
    ylab("Outcome with individual FE"),
  plots[["m11125"]]+
    ggtitle("Satisfaction with health") +
    xlab("Scaled relative income near threshold") +
    ylab(""),
  ncol = 2, nrow = 1) %>%
  annotate_figure(bottom = text_grob("Fig 2. Gender norms, well-being and health, men.))

```

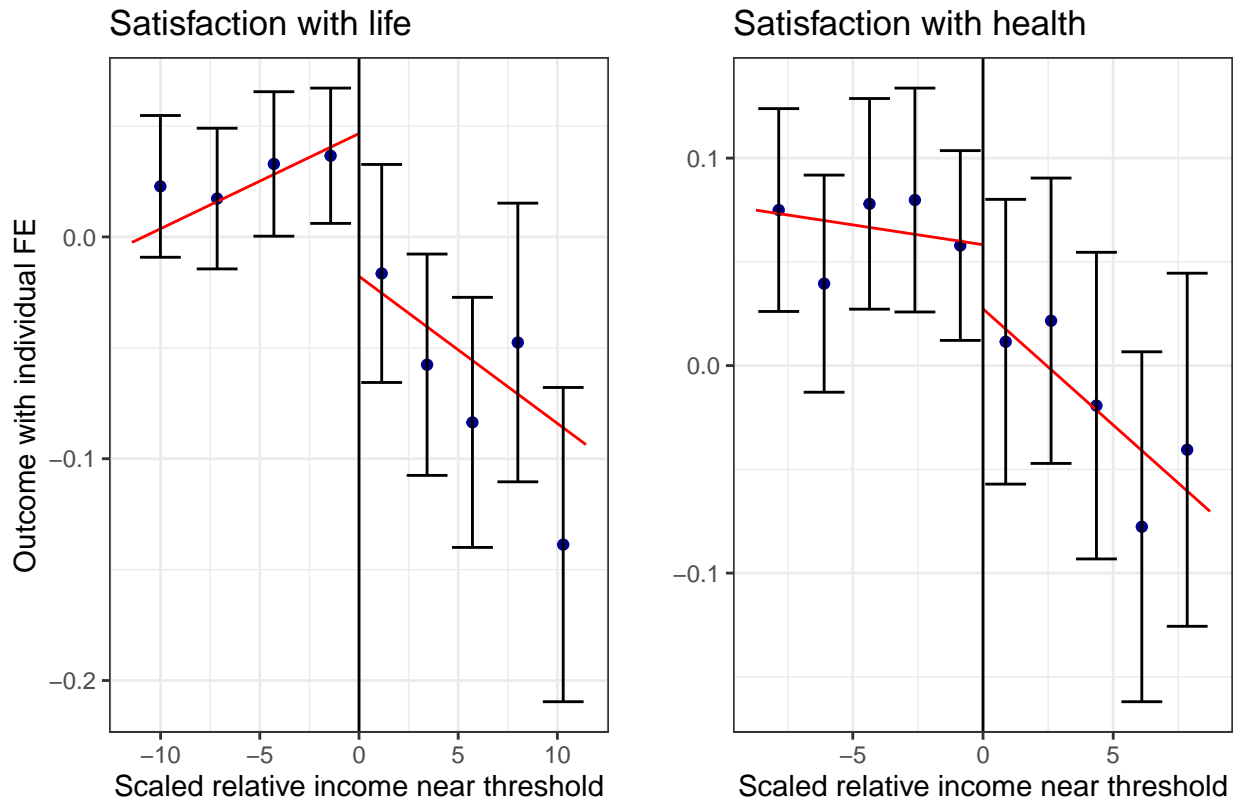


Fig 2. Gender norms, well-being and health, men.

In couples in which wives out-earn their husbands, life and work satisfaction among men appears significantly lower.

Replicating Figure 3: Gender norms, well-being and health, women.

Following Fig. 3 in the original publication, this figure depicts the relationship between relative income levels and satisfaction with life and health among women in heterosexual couples. Standard errors are clustered by relative income bins (€1,000 bins) and specifications include fixed effects. Bandwidth and binning selection were conducted with MSE-optimization and the IMSE-optimal evenly spaced method, respectively. Bars depict 95% confidence intervals.

```
ggarrange(
  plots[["overalllifefemale"]]+
    ggtitle("Satisfaction with life") +
    xlab("Scaled relative income near threshold") +
    ylab("Outcome with individual FE"),
  plots[["m11125_female"]]+
    ggtitle("Satisfaction with health") +
    xlab("Scaled relative income near threshold") +
    ylab("") +
    scale_y_continuous(
      breaks = c(-0.1, -0.05, 0, 0.05, 0.1),
      limits = c(-0.1, 0.1)),
  ncol = 2, nrow = 1) %>%
  annotate_figure(bottom = text_grob("Fig 2. Gender norms, well-being and health, women."))
```

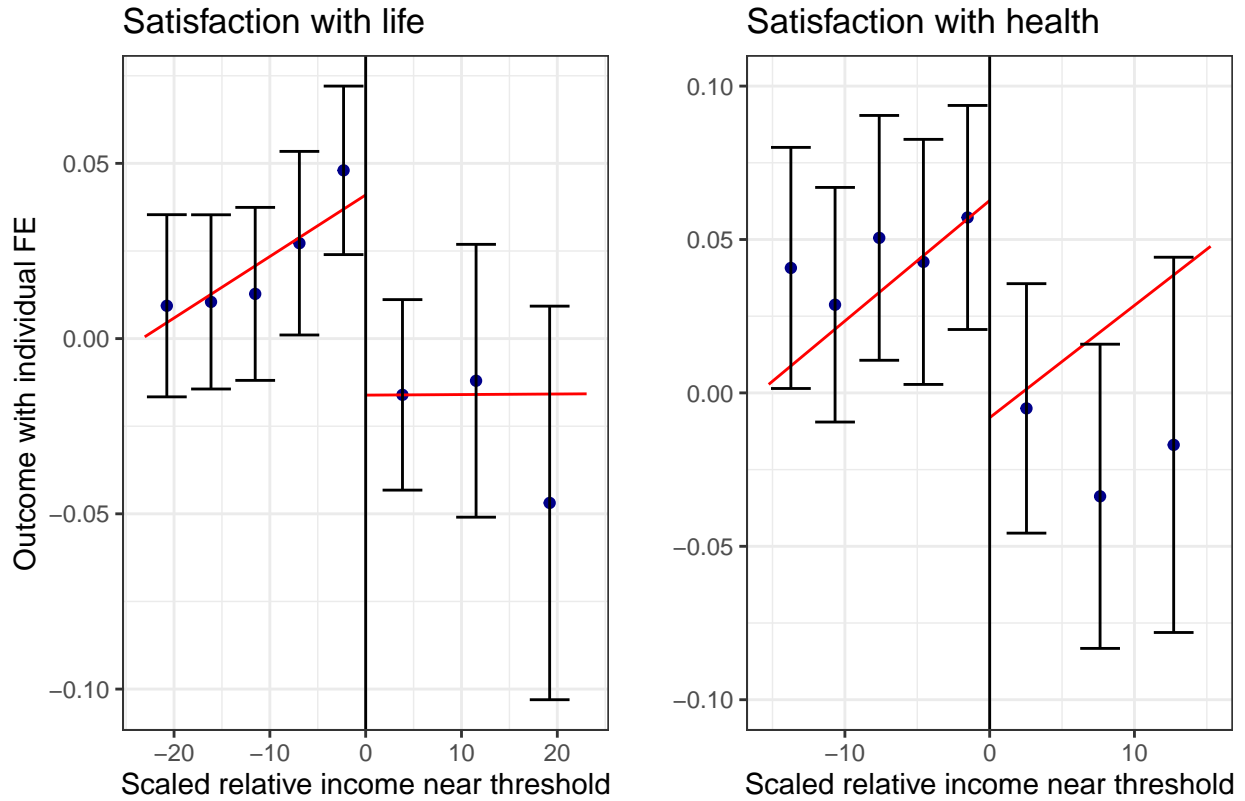



Fig 2. Gender norms, well-being and health, women.

Women's satisfaction with life and health appears lower when they just out-earn their husbands.

Replicating Table 1: RD estimates: Gender norms, well-being, and health—Including individual fixed effects

In the below table, which replicates Table 1, standard errors are reported in parentheses and are clustered by relative income bins (in €1,000 intervals). As demonstrated in the code, bandwidths were selected using MSE optimization.

```
# convert results to a table that we like
results %>%
  # combine results for all models
  lapply(as.data.frame) %>%
  bind_rows() %>%
  mutate(outcome = outcomes) %>%
  # add gender column, p-value sig stars, and combine values for final RD expression
  separate(outcome, into = c("outcome", "gender"), fill = "right", sep = "_female") %>%
  mutate(pstar = ifelse(rd_pval < 0.01, "***",
                        ifelse(rd_pval < 0.05, "**",
                              ifelse(rd_pval < 0.1, "*", "")))) %>%
  mutate(gender = ifelse(is.na(gender), "male", "female"),
         rd_effect = paste0(round(rd_effect, 3), pstar),
         rd_error = paste0("(", round(rd_error, 3), ")"),
         n_effective = paste(format(n_effective, big.mark=",")))
  ) %>%
  # format estimates
```

```

mutate(
  across(
    all_of(c("y_mean", "y_sd")),
    ~paste(round(.x, 3))
  )
) %>%
# reshape so columns are each model
select(-all_of(c("rd_pval", "pstar"))) %>%
pivot_longer(cols = -c("outcome", "gender")) %>%
pivot_wider(id_cols = c("gender", "name"), names_from = "outcome") %>%
# order rows and label
mutate(name = factor(name, levels = c("rd_effect", "rd_error", "n_effective", "y_mean", "y_sd")),
       gender = factor(gender, levels = c("male", "female"))) %>%
arrange(gender, name) %>%
mutate(name = rep(c("RDD wife earns more", "", "Observations (effective)",
                  "Mean of dep. var.", "SD of dep. var."), 2)) %>%
select(-gender) %>%
# final kable
kable(
  col.names = c("", "Satisfaction with life", "Satisfaction with work",
                "Satisfaction with health", "Mental health", "Physical health"),
  align = c("l", rep('c', 5)),
  escape = FALSE,
  caption = "RD estimates: Gender norms, well-being, and health-Including individual FE."
) %>%
kableExtra::add_header_above(
  header = c(" " = 1, "Dependent variable" = 5)
) %>%
kableExtra::group_rows("Panel A: Male", 1, 5) %>%
kableExtra::group_rows("Panel B: Female", 6, 10) %>%
kableExtra::landscape()

```

Table 1: RD estimates: Gender norms, well-being, and health—Including individual FE.

	Dependent variable				
	Satisfaction with life	Satisfaction with work	Satisfaction with health	Mental health	Physical health
Panel A: Male					
RDD wife earns more	-0.057*** (0.014)	-0.043*** (0.011)	-0.033*** (0.013)	0.013 (0.01)	-0.043*** (0.011)
Observations (effective)	28,185	20,708	22,401	6,801	7,552
Mean of dep. var.	7.089	6.886	6.781	-0.031	-0.096
SD of dep. var.	1.69	2.095	2.084	1.019	1.034
Panel B: Female					
RDD wife earns more	-0.071*** (0.01)	-0.007 (0.018)	-0.084*** (0.013)	-0.084*** (0.021)	0.031** (0.013)
Observations (effective)	49,267	43,628	35,822	6,895	10,169
Mean of dep. var.	7.245	7.08	6.819	-0.041	-0.058
SD of dep. var.	1.648	2	2.045	1.014	1

Table 1 shows the estimated RD effects among men and women for each of the five main outcomes. Among men, satisfaction declines when they are out-earned by their wives. An effect of -0.06 standard deviations was found for overall life satisfaction, with smaller but still significant effects found for work and health satisfaction. No effect was found for mental health, but a similar negative effect of -0.04 SD was found for physical health. Among women, negative effects were found for overall and health satisfaction, while no effect was found on work satisfaction. A negative effect of -0.08 SD was found on mental health, but a mild increase in physical health was found (0.03 SD).

Replicating Table 2: RD estimates: Gender norms, well-being, and health—West and East Germany—Including individual fixed effects

To implement the region-specific model, use the `rdhte` package which complements `rdrobust` by adding functionality that handles heterogeneous treatment effects. In this case, we are interested in heterogeneous treatment effects between West and East Germany. Only a few adjustments need to be made from the initial base model, including the addition of a factor covariate representing region of residence.

```
# create analysis dataset for region-specific model
analysis_data <- analysis_data %>%
  mutate(
    # region * treatment
    west_treatment = west * treatment
  )

results_west <- list()
for (outcome in outcomes) {

  # set gender-dependent covariates
  covs <- if (grepl("female", outcome)) female_covs else male_covs

  # rdhte: estimation for heterogeneous treatment effects in RD designs
  rd_hte <- rdhte(y = analysis_data[[paste0(outcome, "_res")]],
    x = analysis_data[[running_var]],
    c = threshold,
    covs.hte = factor(analysis_data$west),
    covs.eff = as.matrix(analysis_data[, covs, drop=FALSE]),
    bwselect = "mserd", # use MSE selection criteria
    cluster = analysis_data$recode_diff_wage_OR_int # integer version of RV
  )

  # get sample used in RD so we can report stats (non-missing outcome and within benchmark)
  rd_sample <- !is.na(analysis_data[[paste0(outcome, "_res")]]) &
    ifelse(analysis_data[[running_var]] <= threshold,
      abs(analysis_data[[running_var]] - threshold) <= rd_hte$h[1,1], # left bandwidth
      abs(analysis_data[[running_var]] - threshold) <= rd_hte$h[1,2]) # right bandwidth
  rd_effective_west <- analysis_data[rd_sample & (analysis_data$west == 1),]
  rd_effective_east <- analysis_data[rd_sample & (analysis_data$west == 0),]

  # store results for later table
  results_west[[outcome]] <- list(
    # results for west
    rd_effect_west = rd_hte$coef[2], # pull effect for west = 1
    rd_error_west = rd_hte$se[2], # pull se for west
    rd_pval_west = rd_hte$pvalue[2], # pull p-value for west
  )
}
```

```

n_effective_west = nrow(rd_effective_west),
y_mean_west = mean(rd_effective_west[[outcome]]),
y_sd_west = sd(rd_effective_west[[outcome]]),
# results for east
rd_effect_east= rd_hte$coef[1], # pull effect for west = 0
rd_error_east = rd_hte$se[1], # pull se for east
rd_pval_east = rd_hte$pv[1], # pull p-value for east
n_effective_east = nrow(rd_effective_east),
y_mean_east = mean(rd_effective_east[[outcome]]),
y_sd_east = sd(rd_effective_east[[outcome]])
)
}

```

In the below table, which replicates Table 2 in the original paper, standard errors are reported in parentheses and are clustered by relative income bins (in €1,000 intervals). As demonstrated in the code, bandwidths were selected using MSE optimization.

```

# convert results to a table that we like
results_west %>%
  # combine results for all models
  lapply(as.data.frame) %>%
  bind_rows() %>%
  mutate(outcome = outcomes) %>%
  # reshape region
  pivot_longer(!outcome) %>%
  separate(name, into = c("estimate_pre", "estimate", "region"), sep="_") %>%
  mutate(estimate = paste(estimate_pre, estimate, sep="_")) %>%
  select(-estimate_pre) %>%
  pivot_wider(id_cols = c("outcome", "region"), names_from="estimate") %>%
  # add gender column, p-value sig stars, and combine values for final RD expression
  separate(outcome, into = c("outcome", "gender"), fill = "right", sep = "_female") %>%
  mutate(pstar = ifelse(rd_pval < 0.01, "***",
                        ifelse(rd_pval < 0.05, "**",
                              ifelse(rd_pval < 0.1, "*", "")))) %>%
  mutate(gender = ifelse(is.na(gender), "male", "female"),
         rd_effect = paste0(round(rd_effect, 3), pstar),
         rd_error = paste0("(", round(rd_error, 3), ")"),
         n_effective = paste(format(n_effective, big.mark=","))
  ) %>%
  # format estimates
  mutate(
    across(
      all_of(c("y_mean", "y_sd")),
      ~paste(round(.x, 3))
    )
  ) %>%
  # reshape so columns are each model
  select(-all_of(c("rd_pval", "pstar"))) %>%
  pivot_longer(cols = -c("outcome", "gender", "region")) %>%
  pivot_wider(id_cols = c("gender", "name", "region"), names_from = "outcome") %>%
  # order rows and label
  mutate(name = factor(name, levels = c("rd_effect", "rd_error", "n_effective", "y_mean", "y_sd")),
         gender = factor(gender, levels = c("male", "female")),

```

```

    region = factor(region, levels = c("west", "east")) %>%
  arrange(region, gender, name) %>%
  mutate(name = rep(c("RDD wife earns more", "", "Observations (effective)",
    "Mean of dep. var.", "SD of dep. var."), 4)) %>%
  select(-gender, -region) %>%
  # final kable
  kable(
    col.names = c("", "Satisfaction with life", "Satisfaction with work",
      "Satisfaction with health", "Mental health", "Physical health"),
    align = c("l", rep('c', 5)),
    escape = FALSE,
    caption = "RD estimates: Gender norms, well-being, and health-West and East Germany-Including indiv.
  ) %>%
  kableExtra::add_header_above(
    header = c(" " = 1, "Dependent variable" = 5)
  ) %>%
  kableExtra::group_rows("Panel A: Male (West)", 1, 5) %>%
  kableExtra::group_rows("Panel B: Female (West)", 6, 10) %>%
  kableExtra::group_rows("Panel C: Male (East)", 11, 15) %>%
  kableExtra::group_rows("Panel D: Female (East)", 16, 20) %>%
  kableExtra::landscape()

```

Table 2: RD estimates: Gender norms, well-being, and health—West and East Germany—Including individual FE.

	Dependent variable				
	Satisfaction with life	Satisfaction with work	Satisfaction with health	Mental health	Physical health
Panel A: Male (West)					
RDD wife earns more	-0.05*	-0.031	-0.044	0.003	-0.014
	(0.042)	(0.06)	(0.038)	(0.074)	(0.062)
Observations (effective)	13,252	10,954	21,753	3,940	2,804
Mean of dep. var.	7.295	7.017	6.863	-0.014	-0.074
SD of dep. var.	1.689	2.107	2.131	1.035	1.085
Panel B: Female (West)					
RDD wife earns more	-0.078*	-0.017	-0.099*	-0.062	0.035
	(0.047)	(0.046)	(0.058)	(0.055)	(0.067)
Observations (effective)	21,934	21,567	14,229	3,846	3,963
Mean of dep. var.	7.391	7.151	6.853	-0.043	-0.04
SD of dep. var.	1.643	2.011	2.112	1.016	1.024
Panel C: Male (East)					
RDD wife earns more	-0.098**	-0.093	-0.034	0.041	-0.132*
	(0.07)	(0.153)	(0.117)	(0.132)	(0.117)
Observations (effective)	7,075	5,799	10,483	2,310	1,690
Mean of dep. var.	6.648	6.613	6.623	-0.074	-0.133
SD of dep. var.	1.644	2.051	1.934	0.992	0.984
Panel D: Female (East)					
RDD wife earns more	-0.078	-0.005	-0.085*	-0.086	0.012
	(0.096)	(0.129)	(0.09)	(0.132)	(0.121)
Observations (effective)	10,551	10,036	7,584	2,254	2,311
Mean of dep. var.	6.789	6.92	6.665	-0.059	-0.074
SD of dep. var.	1.641	1.967	1.932	1.014	0.975

Table 2 explore the heterogeneity of the effect of region, specifically between East and West Germany. The difference between this model and the base model is that this estimation allows for heterogenous effects in the two different regions. Therefore, 4 panels are included in this table to depict effects for men and women in East and West Germany. A negative effect of double the magnitude is found for overall life satisfaction among men in East Germany compared to West Germany. While out-earning their husbands has a negative effect on overall life satisfaction among women in West Germany, no evidence for effect is found among women in East Germany. No support for negative effect on any other outcome other than health satisfaction, which is negative among women in both West and East Germany, indicating that the result in the overall sample is consistent across region.

Table 3: Adjusting p-values for multiple correlated outcomes

The following table depicts RD effect estimates with original and adjusted significance stars (* < 0.1, ** < 0.05, *** < 0.01).

```
# convert results to a table that we like
results_west %>%
  # combine results for all models
  lapply(as.data.frame) %>%
  bind_rows() %>%
  mutate(outcome = outcomes) %>%
  # reshape region
  pivot_longer(!outcome) %>%
  separate(name, into = c("estimate_pre", "estimate", "region"), sep="_") %>%
  mutate(estimate = paste(estimate_pre, estimate, sep="_")) %>%
  select(-estimate_pre) %>%
  pivot_wider(id_cols = c("outcome", "region"), names_from="estimate") %>%
  bind_rows(
    results %>%
    # combine results for all models
    lapply(as.data.frame) %>%
    bind_rows() %>%
    mutate(outcome = outcomes,
           region = "both")
  ) %>%
  # add gender column
  separate(outcome, into = c("outcome", "gender"), fill = "right", sep = "_female") %>%
  mutate(gender = ifelse(is.na(gender), "male", "female")) %>%
  group_by(region) %>%
  mutate(holm_pval = p.adjust(rd_pval, method="holm")) %>%
  ungroup() %>%
  mutate(pstar = ifelse(rd_pval < 0.01, "***",
                       ifelse(rd_pval < 0.05, "**",
                              ifelse(rd_pval < 0.1, "*", ""))),
         pstar_adj = ifelse(holm_pval < 0.01, "***",
                            ifelse(holm_pval < 0.05, "**",
                                   ifelse(holm_pval < 0.1, "*", "")))) %>%
  mutate(rd_effect_orig = paste0(round(rd_effect, 3), pstar),
         rd_effect_adj = paste0(round(rd_effect, 3), pstar_adj)) %>%
  select(outcome, gender, region, rd_effect_orig, rd_effect_adj) %>%
  pivot_wider(id_cols = c("outcome", "gender"), names_from = "region",
              values_from = c("rd_effect_orig", "rd_effect_adj"),
              names_glue = "{region}_{.value}") %>%
```



```

select(outcome, starts_with("both"), starts_with("west"), starts_with("east")) %>%
mutate(outcome = rep(c("Satisfaction with life", "Satisfaction with work",
                      "Satisfaction with health", "Mental health", "Physical health"), 2)) %>%
# final kable
kable(
  col.names = c("", rep(c("RD effect", "RD effect (adj. p-value)"), 3)),
  align = c("l", rep('c', 5)),
  escape = FALSE,
  caption = "Adjusted p-value comparison for RD estimates."
) %>%
kableExtra::add_header_above(
  header = c("Region:" = 1, "Germany (all)" = 2, "West Germany" = 2, "East Germany" = 2)
) %>%
kableExtra::group_rows("Panel A: Male", 1, 5) %>%
kableExtra::group_rows("Panel B: Female", 6, 10) %>%
kableExtra::landscape()

```

Table 3: Adjusted p-value comparison for RD estimates.

Region:	Germany (all)		West Germany		East Germany	
	RD effect	RD effect (adj. p-value)	RD effect	RD effect (adj. p-value)	RD effect	RD effect (adj. p-value)
Panel A: Male						
Satisfaction with life	-0.057***	-0.057***	-0.05*	-0.05	-0.098**	-0.098
Satisfaction with work	-0.043***	-0.043***	-0.031	-0.031	-0.093	-0.093
Satisfaction with health	-0.033***	-0.033**	-0.044	-0.044	-0.034	-0.034
Mental health	0.013	0.013	0.003	0.003	0.041	0.041
Physical health	-0.043***	-0.043***	-0.014	-0.014	-0.132*	-0.132
Panel B: Female						
Satisfaction with life	-0.071***	-0.071***	-0.078*	-0.078	-0.078	-0.078
Satisfaction with work	-0.007	-0.007	-0.017	-0.017	-0.005	-0.005
Satisfaction with health	-0.084***	-0.084***	-0.099*	-0.099	-0.085*	-0.085
Mental health	-0.084***	-0.084***	-0.062	-0.062	-0.086	-0.086
Physical health	0.031**	0.031*	0.035	0.035	0.012	0.012

Table 3 depicts the same RD estimates as Tables 1 and 2 in the first column of each region, but in the second column, provides adjusted p-values according to step-down Bonferroni algorithm. Results from the full sample remain significant, but no region-specific effects are significant.

Discussion and Conclusion

In this replication, two adjustments were made to the presentation of results to correct for misleading or omitted results from the original paper. First, the sample size, mean, and SD included in the tables are that of the sample as a whole, not of those included in the regression sample due to being within the bandwidth. This is not consistent with the principle of RD design, which reduces the sample to just those observations near the threshold of the running variable. In this project’s presentation of the results table, the sample statistics (size, mean, and SD for each outcome) are restricted to just those observations near the threshold for each outcome, as determined by each outcome’s calculated bandwidth. Second, only the effects for West Germany were included for the region-specific analysis. By omitting the results for East Germany, the comparison is only between West Germany and the entire sample, rather than between regions. Including effect size estimates for both regions in one table allows for clearer comparisons across regions. Aside from these two differences, results in this replication are presented to exactly mirror those in the original paper for best comparison.

In the implementation of the model in R, a couple error and warning messages indicate assumption violations that were not included in the original paper. First, the regression discontinuity model implementation in R, as per the `rdrobust` package, throws a warning: “mass points detected in the running variable.” This indicates a pattern observed in Fig. 1: discontinuity at the cutoff of the running variable. This pattern violates an assumption of RDD that assumes local continuity at the cutoff point of the running variable. This indicates that treatment (above or below earning difference threshold) is not “as-good-as randomized,” which can effect sampling variability and bandwidth selection. Next, for the region-specific model, the `rdhte` package does not allow controlling for the same variable that is factoring into heterogeneous treatment effects. The original paper they “add to Eq (1) an interaction term between a dummy variable that identifies individuals in couples where the wife out-earns the husband with an indicator variable for residing in West Germany. Furthermore, we control for a dummy variable for individuals residing in West Germany” (Gihleb, et al., 2024, 5). However, this implementation is not possible in R because of the redundancies in including a variable as a control and as a grouping variable for heterogeneous effects. This redundancy may be technically possible in a Stata implementation, but it introduces confounders that may bias results.

Results in this replication projects closely mirror the original paper for the complete sample. Effect sizes are small but significant and indicate that on average, wives out-earning their husbands has a negative effect on life satisfaction in Germany. In Table 1, magnitude and direction of coefficients is very similar across all gender-outcome combinations except for physical health among women. By including sample mean and SD in Table 1, we can see that the sample mean health is slightly higher than the standardized health across the full sample, which may explain this result. Table 2 Panels A and B closely reflect the original Table 2, with two exceptions. The original research found no effect on life satisfaction among men in West Germany. However, this replication in R produced an effect of comparable magnitude between men and women in East and West Germany. Furthermore, common negative effect was found among women in West Germany for life and health satisfaction, but the replication found no effect on work satisfaction. Omitting redundant controls when replicating the heterogeneous effects model may account for these differences.

As previously mentioned, the data contains multiple observations for each couple. This is not suggested as a limitation in the original paper, as the authors consider it handled by fixed-effects residualization. While this could be an effective strategy for covariates, it is not applied to covariates, only to outcome variables. Since units in the analysis are not independent, clustering at the unit level would be prudent for more conservative confidence intervals (out of scope for this project due to computational constraints).

Additionally, modeling multiple, correlated outcomes in the same analysis (10 in the main article, with additional outcomes in the appendices), opens the door for “p-hacking,” where false positive significance is detected. To account for this, I adjusted p-values using a step-down Bonferroni approach to control for the

family-wise error rate. This adjustments accounts for multiple, correlated p-values. Table 3 table shows that estimate significance is robust to adjustment for the full sample, but evidence for region-specific effect disappears under adjustment. This is particularly interesting due to the exploratory hypothesis of this paper, which provides the main motivation for the study. There is not evidence for cultural shifts in acceptance of women in the workforce in East Germany reducing the negative effect of women out-earning their husbands on satisfaction and health measures.

Are these methods sufficient to draw causal claims? The use of RDD in this research is appealing because it captures the cultural effect of a woman technically out-earning her husband. The model does not account for duration, treating women who out-earn their husbands for only one year the same as women who out-earn their husbands for an entire career. Additionally, the paper does not complete causal identification because it does not suggest a robust causal mechanism between relative income and life satisfaction. In some ways, RDD does not seem well-suited for this research question because the magnitude and duration of relative income likely contributes to feelings of life satisfaction. While income is considered as a covariate, its direct effects are not examined. This design is insufficient to answer the research question of how relative spousal income affects individuals' well-being and health at the regional level. The original paper claims that effects are stronger for women in West Germany, but those results were unable to be replicated and any evidence disappeared when considering adjustments for multiple, correlated outcomes. The use of these methods seem insufficient to draw the causal claims made by the original paper upon replication.

Citations

Rania Gihleb, Osea Giuntella, & Stella, L. (2024). Relative income within the household, gender norms, and well-being. *PLoS ONE*, 19(10), e0306180–e0306180. <https://doi.org/10.1371/journal.pone.0306180>