sociological science

Supplement to:

Auspurg, Katrin, and Sabine Düval. 2024. “House-work as a Woman’s Job?: What Looks Like Gender Ideologies Could Also Be Stereotypes.” Sociological Science 11: 789-814.

## Supplementary Information for

# What Looks Like Gender Ideologies Could Be Stereotypes? Using an Experimental Design to Dig Deeper into Normative Attitudes and Beliefs Underlying Couples' Labor Division

Katrin Auspurg[a] and Sabine Düval[a,b]

[a] LMU Munich, Department of Sociology; Konradstr. 6, 80801 Munich, Germany
[b] German Youth Institute (DJI), Nockherstr. 2, 81541 Munich, Germany
Contact: katrin.auspurg@lmu.de, dueval@dji.de

## Content

# 1. Materials and Methods

## 1.1. Design of our Experiment

In our multifactorial survey experiment, we asked respondents to rate couples described by 7 dimensions, all of which are known to influence couples' work-sharing decisions (see, e.g., Baxter, Hewitt and Haynes 2008, Davis and Greenstein 2009): (1) the couple's marital status; (2) the presence and age of children in the household; (3) the division of childcare between partners; (4) the share of housework performed by both partners; (5) the male partner's labor market hours; (6) the female partner's labor market hours; and (7) the relative contribution of both partners to the household income. Between 2 and 5 categories (levels) were defined for each of the 7 dimensions (see Table A1 for all dimensions and levels).

**Table A1.** Vignette Dimensions and Levels

| **Dimensions** | **Levels** 1 | 2 | 3 | 4 | 5 | **Information Condition** |
|---|---|---|---|---|---|---|
| 1 Marital status | Unmarried | Married | — | — | — | Low |
| 2 Presence and age of children | No children | One 2-yrs old child | One 8-yrs old child | — | — | Low |
| 3 Share of childcare (relative to partner) | Larger share | Smaller share | Same share | No information | — | Low |
| 4 Share of housework | 70% (21 hrs. per week) | 60% (18 hrs. per week) | 50% (15 hrs. per week) | 40% (12 hrs. per week) | 30% (9 hrs. per week) | Low |
| 5 Labor market hours per week (man) | 40 h | 30 h | 20 h | No information | — | Medium |
| 6 Labor market hours per week (woman) | 40 h | 30 h | 20 h | No information | — | Medium |
| 7 Income contribution (relative to partner) | Twice as much | Half as much | Same | No information | — | High |

In addition, we varied the amount of information presented to respondents: (A) in the "low information" condition, we presented only information on family status, childcare, and the couple's division of housework (i.e., dimensions 1-4). The remaining dimensions were blanked out. This question format is very similar to the standard item questions on traditional gender ideologies. (B) In the "medium information" condition, we additionally presented information on both partners' labor market hours (i.e., dimensions 5 and 6), and (C) in the "full information" condition we presented all 7 dimensions, including information on both partners' relative contribution to the household income of (i.e., dimension 7). This between-respondent split allowed us to examine whether respondents agreed with the traditional housework arrangement only when they lacked information about the couple's labor market resources.

In a second between-respondent split, we implemented a variation of the rating task: About half of the respondents had to rate the housework share of a female vignette person, while the other half had to rate the housework share of a male vignette person. We deliberately used a between-respondent split to avoid social desirability bias and also to avoid confusion when respondents had to switch between rating men's and women's housework. Respondents were randomly assigned to both between-respondent splits (i.e., information condition and evaluation task). Figure A1 shows an example vignette, including all 3 information conditions, when a couple with a child was described (an example vignette for a childless couple can be found in the main text).

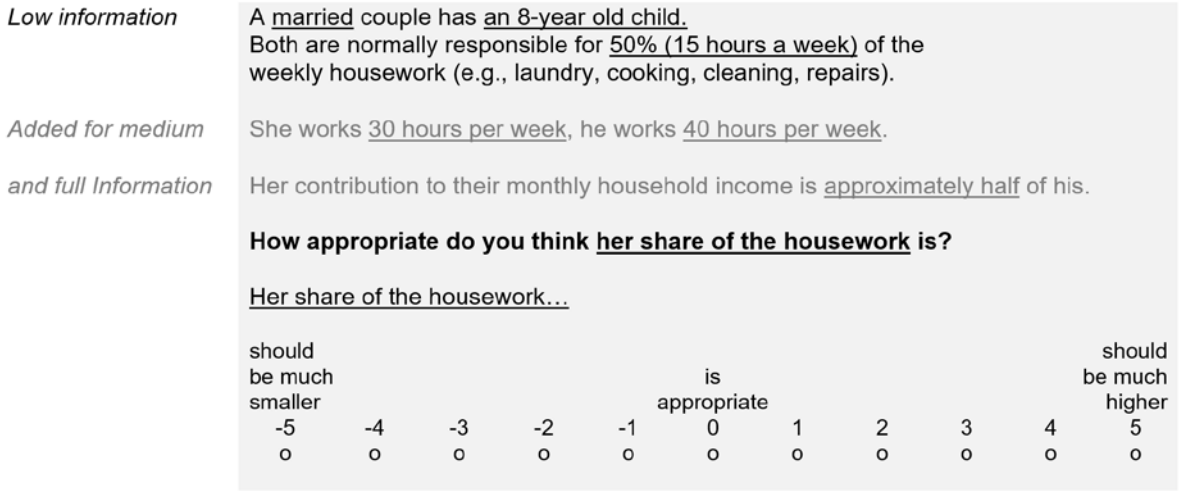


**Figure A1.** Sample Vignette for Different Information Conditions (Couple with Child)

*Notes*. This figure shows a sample vignette for the low, medium, and full information conditions (the sentences in gray were added only in these conditions). The experimental manipulations are underlined.

The full set of all possible scenarios (the vignette universe) includes all 7,680 possible combinations of dimension levels. To select our fraction of 750 vignettes, we used a *D*-efficient sampling technique that minimizes correlations between dimensions while maximizing the variance of each of the dimensions (for details, see Atzmüller and Steiner 2010, Auspurg and Hinz 2015). This procedure ensured that all main effects of the vignette dimensions and all two- and three-way interactions between dimensions were not confounded. Put differently, there is no gender inequality in our vignette sample, so in our fictional vignette sample, gender does not affect men's and women's labor market characteristics. In the realized sample, the correlation of gender with the income share was $r = .002$ ($p = .941$) and with labor market hours was $r = -.034$ ($p = .085$). Table A2 shows that all correlations within our vignette sample are very small ($r < 0.04$) and not statistically significant ($p > .08$). In addition, correlations between vignette dimensions and relevant respondent characteristics (e.g., age, gender, education) are also very small ($r < 0.06$). This shows that randomization was successful.

We divided the sample of 750 vignettes into 250 different questionnaire versions, each containing 3 vignettes. The different questionnaire versions were randomly assigned to the respondents, with the order of the 3 vignettes randomized for each respondent to neutralize possible effects of the vignette order (Auspurg and Jäckle 2017). Our experiment was conducted in a self-completion mode, in which the interviewer hands the CAPI laptop to the respondent. Self-completion is the recommended mode for multifactorial survey experiments, first because the vignettes may be better understood by respondents if they read them directly than if they are read by an interviewer, and second because this mode reduces possible social desirability bias (Auspurg and Hinz 2015). A series of pretests conducted prior to implementation in *pairfam* indicated that respondents coped well with the questions and the level of complexity. For more information on the vignette module, see Düval and Auspurg (2020).

**Table A2.** Correlation Matrix of Vignette Dimensions and Respondent Characteristics

| Pearson correlation (*p*-value) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Vignette Dimensions** | Married | Child present | Share of childcare | Housework hours | Labor market hrs. | Share of income | Gender | Info cond. |
| Married | 1.000 | | | | | | | |
| Child present | 0.007 | 1.000 | | | | | | |
| | (0.675) | | | | | | | |
| Share of childc.[a] | 0.007 | -0.008 | 1.000 | | | | | |
| | (0.760) | (0.742) | | | | | | |
| Housework hrs. | -0.002 | 0.011 | 0.007 | 1.000 | | | | |
| | (0.911) | (0.516) | (0.778) | | | | | |
| Labor m. hrs. | 0.007 | 0.011 | -0.019 | 0.005 | 1.000 | | | |
| | (0.714) | (0.593) | (0.504) | (0.789) | | | | |
| Share of income | -0.004 | -0.011 | 0.007 | 0.024 | 0.006 | 1.000 | | |
| | (0.892) | (0.706) | (0.852) | (0.388) | (0.836) | | | |
| Gender | 0.027 | 0.001 | -0.005 | 0.002 | -0.034 | 0.002 | 1.000 | |
| | (0.096) | (0.954) | (0.840) | (0.904) | (0.085) | (0.941) | | |
| Info condition[a] | -0.001 | 0.004 | 0.004 | 0.002 | -0.005 | -[c] | 0.016 | 1.000 |
| | (0.960) | (0.801) | (0.846) | (0.887) | (0.788) | - | (0.335) | |
| **Respondent characteristics** | | | | | | | | |
| Gender | 0.012 | 0.013 | -0.014 | 0.018 | -0.017 | 0.012 | 0.045* | 0.041* |
| | (0.465) | (0.419) | (0.541) | (0.279) | (0.389) | (0.661) | (0.006) | (0.013) |
| Age | -0.022 | 0.005 | 0.000 | -0.021 | -0.041* | 0.011 | -0.033* | 0.004 |
| | (0.184) | (0.781) | (0.994) | (0.202) | (0.041) | (0.691) | (0.046) | (0.814) |
| University degree | 0.019 | -0.018 | -0.036 | 0.017 | -0.032 | -0.018 | -0.045* | 0.017 |
| | (0.234) | (0.267) | (0.120) | (0.310) | (0.109) | (0.530) | (0.006) | (0.310) |
| In relationship | -0.013 | -0.016 | -0.012 | 0.001 | -0.021 | -0.030 | -0.017 | 0.050* |
| | (0.442) | (0.343) | (0.602) | (0.975) | (0.291) | (0.292) | (0.292) | (0.002) |
| Child in household | -0.007 | 0.005 | 0.007 | -0.004 | -0.023 | -0.003 | -0.022 | -0.012 |
| | (0.652) | (0.774) | (0.767) | (0.807) | (0.252) | (0.903) | (0.185) | (0.472) |

*Notes.* [a] Categorical variable from low to high share resp. information.
[b] Not calculated, since income is only available for one information condition (the full information condition).
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided tests).

## 1.2. Identification Strategy

### *Effects of Information and Measurement of Gender Ideologies*

We use linear regressions to predict the rating of the female vignette person's housework, denoted by $Y^F$. [1] Recall that this rating was measured on an 11-point rating scale, ranging from -5 "Her housework share should be much smaller" over 0 "Her housework share is appropriate" to +5 "Her housework share should be much larger."

We are mainly interested in three estimands (see the main paper). First, does gender still affect the appropriate share of housework when there is rich information on both partners' labor force participation and relative income? This is the *controlled direct effect of gender*.

Second, we are interested in whether the gender effect is smaller when information is provided (and stereotypes about male and female labor force characteristics should be disrupted) than when this information is not provided. This is called the *eliminated gender effect*: the gender effect that is

[1] Evaluations of male vignette persons' share of housework were brought to this metric by a simple transformation (multiplication by -1).

eliminated by adding information. We assume that this eliminated gender effect is mainly present for traditional respondents.[2]

Third, in further analyses we will also examine whether this elimination is due to the fact that adding information blocks the *indirect gender effect* we hypothesize, i.e., the mediation through respondents' inferences in the case of low information. The eliminated effect is a correct measure of this mediation only if the effects of respondents' inferences on labor market characteristics do not interact with gender. Such an interaction could exist, for example, if respondents devalue the exchange value of female labor market characteristics for buying out of housework. To test whether the assumption of no interaction is plausible, we will check whether the labor market characteristics indicated in the vignettes interact with gender (more details below; for a technical discussion, see Figure 3 in the main text).

In our basic regression, we use only two dummies (0/1) indicating the medium and full information conditions as predictors (denoted as $I_m$ respectively $I_h$). To adjust for the clustering of evaluations within respondents (3 vignette evaluations per respondent), we use regressions with random intercepts. Eq. 1 shows the regression equation, where $\alpha_j$ is the random intercept, $i$ is an indicator for the vignettes, and $j$ an indicator for the respondents.

$$Y_{ij}^{F} = \alpha_j + \beta_{Im} I_{m\_ij} + \beta_{Ih} I_{h\_ij} + \varepsilon_{ij} \quad ; \qquad i = 1, 2, 3 \text{ vignettes}; j = 1, \ldots N_{resp.} \quad \text{(Eq.1)}$$

Due to random assignment, the shares of housework described in the vignettes are equal for men and women (50% on average). Also, all labor market variables added in the medium and full information conditions are gender-neutral, with both men and women having on average 50% of the labor market hours and providing 50% of the household income. Thus, in our vignette scenarios, egalitarian ideologies (equity norms) are evident by respondents who say that an equal division of housework (50% done by women) is appropriate. Technically, this means that $\bar{Y}^F$ predicted by Eq.1 is zero (meaning "appropriate;" see Eq. 2a). In the case of the adherence to traditional gender ideologies, women are expected to do (much) more than 50% of the housework. This corresponds to a significant positive value of $\bar{Y}^F$ (c.f. Eq. 2b; with higher values indicating stronger adherence to traditional gender ideologies).

| | | |
|---|---|---|
| Equity norm: | $\bar{Y}^F = 0$ | (Eq.2a) |
| Trad. gender ideol.: | $\bar{Y}^F > 0$ | (Eq.2b) |

Our main interest is in the prevalence of traditional gender ideologies under different information conditions. The total gender effect corresponding to the measure of traditional ideologies by the standard item questions is indicated by the regression intercept (that is predicting $\bar{Y}^F$ when there is no information on labor market characteristics; both $I_m = 0$ and $I_h = 0$). We expect this total effect to be positive for traditional respondents (Eq. 3a). The effects eliminated by adding information we expect in our first hypothesis are estimated by the regression coefficients of the two dummy variables indicating the information conditions $I_m$ and $I_h$. We expect them to be negative (women's appropriate housework share decreases when information is added), with the effect of adding full information being larger in size than the effect of only adding medium information (Eq. 3b). Finally, the controlled direct gender effect, which measures gender ideologies in the full information condition is revealed by the sum of the intercept and the effect of this information condition (Eq. 3c). Here, we only expect this effect to be smaller than the

[2] This is tested by (i) splitting the analyses by respondents' gender ideologies (measured by the standard item) and (ii) pooled analyses with an interaction effect between the amount of information and gender ideologies (to see if subgroup differences are statistically significant).

total gender effect measured by the intercept (which indicates the gender effect in the case of the "low information" condition, i.e., when the assumed mediation by respondents' inferences is not "under control"). If this controlled direct gender effect were zero, we would conclude that respondents do not support traditional gender ideologies but only equity norms; if it were greater than zero, respondents would still support traditional gender ideologies.

| | | |
|---|---|---|
| Total gender effect (traditional ideology in low info cond.): | $\alpha > 0$ | (Eq. 3a) |
| Eliminated gender effects (Δgender effects due to adding info): | $\beta_{Ih} < \beta_{Im} < 0$ | (Eq. 3b) |
| Controlled direct gender effect (trad. ideol. in high info cond.): | $\alpha + \beta_{Ih} < \alpha$ | (Eq. 3c) |

The assumed differences across subgroups expected by our second hypothesis (stronger effect of adding information for "traditional" respondents) are tested by adding respondents' gender ideologies and interaction effects with the amount of information ($I_m$ and $I_h$) as further predictors to the regression equations (not shown).

Adding the hours of housework indicated in the vignettes $H$ as another predictor allows us to quantify the size of the gender effect in the metric of housework hours. This effect size can be estimated by the ratio of the two coefficients: $-\alpha/\beta_H$ gives the amount of housework hours that the female partner is expected to do more (or less). These cross-elasticities are estimated using the Stata ado `wtp` (Auspurg and Hinz 2015, Hole 2007). The logic is to calculate the amount by which the share of housework indicated in a vignette would have to be increased (or decreased) to bring the rating down to zero. For details and formulas, see Auspurg and Hinz (2015).

In extended regressions, we add the exact values of labor market hours and earnings to the basic regression (using "low information" as one possible value). In this way, we can see which value of these variables (which we refer to as $M'$ in the main tex) has a similar effect to the "low information" condition. The most plausible explanation for a similar effect would be that, on average, respondents in this condition imputed a similar value (i.e., the value is on par with respondents' implicit beliefs, which we denote by $M$). In line with our theoretical argumentation, we expect the effect for "low information" to be similar to the effects of female partners having lower earnings and labor market hours compared to male partners. However, this interpretation requires the assumption that these variables do not interact with gender, which we discuss in the next section.

***Separation of Mediator and Moderator Effects***

The finding of a similar effect of a labor market status level to the "low information" condition is only an indication, not a proof, that respondents made similar inferences. This is because the mediation we assume is only one possible reason for the changes we observe when information is added. Another reason for the eliminated gender effect is an interaction between the labor market characteristics $M$ that respondents assume in the low information condition and gender; this moderation is also changed by adding the information $M'$ (to $M' \cdot X$; see the *Extended Analyses* in the main text). To interpret the eliminated effect as an indirect gender effect (mediation), one must assume that this interaction does not exist.

Our design also includes rich variations in the levels of $M'$. This variation allows us to empirically test whether there are any interactions between these levels and being female. If not, this would make it more certain that the reason for a possible elimination of the gender effect by adding information is exactly the mediation we assume ($X \rightarrow M \rightarrow Y$) and not an interaction ($M \cdot X$).[3] To our knowledge, this

[3] One can only estimate interactions with observed values $M' \cdot X$; which might not cover $M \cdot X$ in case our vignette values do not include respondents' assumptions $M$ and the interactions hinge on these specific values. However,

extension is unique to our design: Multifactorial survey experiments allow for rich variation, allowing for a deep exploration of possible ideologies and stereotypical beliefs.

It should be noted, however, that we only measure average treatment effects (ATEs). If respondents differ in their beliefs (there is treatment heterogeneity), heterogeneous changes in the gender effect that go in different directions when information is added could cancel each other out (Green, Ha and Bullock 2009, Imai, Tingley and Yamamoto 2013). In this case, we would underestimate the mediation (elimination of the gender effect by adding information). However, our robustness analyses only showed the variation across respondents' gender ideologies we assume. This variation does not affect the elimination of gender effects we expect for "traditional" respondents. Treatment effects were found to be very homogeneous across other respondent groups (e.g., defined by age or gender). Moreover, in our study, gender effects were completely eliminated (no longer statistically significant) when the hypothesized mediation was blocked, which also makes an underestimation of this mediation unlikely.

# 1. Descriptive Statistics and Main Result

## 1.1. Descriptive Statistics

Table A3 presents descriptive statistics about our analysis sample. A total of 36 respondents (2.8%) had to be dropped, either due to missing values on the item question on gender ideologies (1 respondent) or because they did not provide any valid vignette rating (35 respondents / 2.7%). 55% of the sample is female. The average age of the respondents is 35 years, and 73% of our respondents are in a relationship.

The bottom rows of Table A3 show descriptive statistics for the vignette ratings. There was little item nonresponse in our analyses sample: Only 2.8% of the vignettes were not answered (for detailed statistics: Düval and Auspurg 2020).

**Table A3.** Descriptive Statistics

| | *N* | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Respondent female | 1,247 | 0 | 1 | 0.549 | 0.498 |
| Respondent age | 1,247 | 24 | 47 | 34.98 | 8.429 |
| Respondent university degree | 1,247 | 0 | 1 | 0.281 | 0.450 |
| Respondent in relationship | 1,243 | 0 | 1 | 0.734 | 0.442 |
| Resp. ≥ 1 child in household | 1,247 | 0 | 1 | 0.468 | 0.499 |
| Item question | 1,247 | 1 | 5 | 4.262 | 0.874 |
| Vignette ratings: Total | 3,738 | -5 | 5 | 0.036 | 2.095 |
| Vignette ratings: Female partner | 1,847 | -5 | 5 | -0.128 | 2.086 |
| Vignette ratings: Male partner | 1,891 | -5 | 5 | 0.197 | 2.093 |

*Notes*. Each respondent was asked to evaluate 3 vignettes.

The left side of Figure A2 shows the distribution of the standard item question on gender ideologies, which is the main focus of our analyses: "Men should participate in housework to the same extent as women." Respondents could answer on a 5-point scale ranging from 1 "Strongly disagree" to 5 "Strongly agree." Respondents who answered 1 or 2 are classified as "traditional," those that answered 4 or 5 as "egalitarian," while the others (answer 3) are "neither nor". With a mean of 4.3 (SD: 0.9) on the 5-point rating scale, most respondents agreed with the statement that men should participate in housework to the same extent as women (see Table A3). Almost 80% of the respondents (= 990

due to the rich variation we use, we can estimate also non-linear interactions, which makes overlooking interaction effects unlikely.

respondents) were classified as having "egalitarian" gender ideologies. Only 3.2% (= 40 respondents) classify as strictly "traditional," while the remaining 17.4% (= 217 respondents) are "neither nor."

The right side of Figure A2 shows the distribution of the responses to the multifactorial survey experiment. Recall that, on average, male and female vignette persons perform the same amount of housework (average share of 50%) and have exactly the same labor market characteristics (labor market hours and relative income contribution). Therefore, a mean rating of zero would indicate that male and female vignette persons should do the same amount of housework. With a mean rating of 0.20 (SD: 2.1) for male vignette persons and of -0.13 (SD: 2.1) for female vignette persons, we see hardly any difference by gender. This is a first indication of strong support for equity norms.

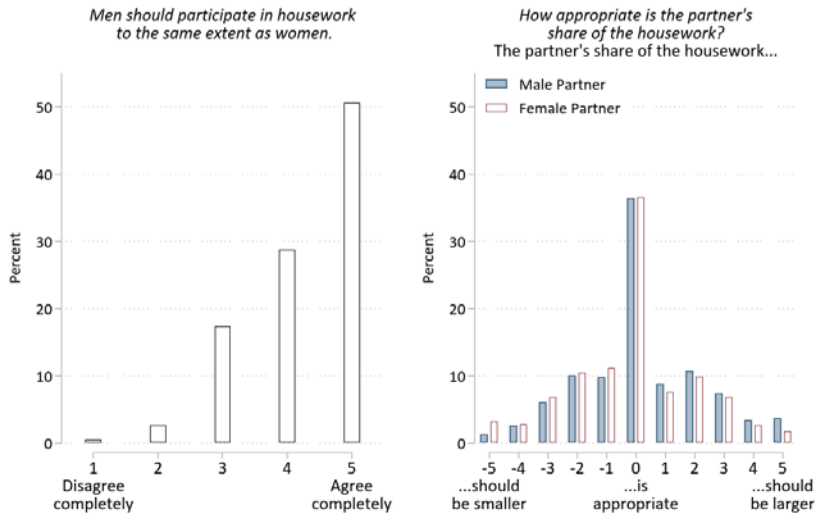


**Figure A2.** Distribution of Responses: Item Question and Factorial Survey Experiment

*Notes.* This figure shows the distribution of the responses to the item question on gender ideologies (left) and the multifactorial survey experiment (right). Number of respondents for the item question: 1,247; number of vignette ratings: 3,738, consisting of 1,847 ratings of female and 1,891 ratings of male vignette persons.

### 1.2. Detailed Regression Results

In the following, we present the regression results underlying the figures in the main text. All analyses can be reproduced with our replication package (Auspurg and Düval 2014).

**Table A4.** Regression Estimates for Results Presented in Figure 4
(Coefficients and in Parentheses *t*-Values)

| | (1)<br>Traditional resp. | (2)<br>Neither nor resp. | (3)<br>Egalitarian resp. |
|---|---|---|---|
| Information condition (ref.: low information) | | | |
| Labor market hours | -1.056* | -0.0982 | -0.195 |
| | (-2.063) | (-0.449) | (-1.919) |
| Labor market hours + relative income | -0.895 | -0.319 | -0.214* |
| | (-1.594) | (-1.382) | (-2.149) |
| Intercept | 1.410*** | 0.144 | -0.0988 |
| | (3.709) | (0.841) | (-1.389) |
| Number of observations | 120 | 653 | 2,965 |
| Number of individuals | 40 | 223 | 996 |
| Wald-test $\chi^2$ | 4.67 | 2.10 | 5.57 |
| Wald test *p*-value | 0.097 | 0.350 | 0.062 |
| Rho | 0.062 | 0.150 | 0.069 |

*Notes.* For the Wald test, the joint null hypothesis is that both dummies for the information conditions are equal to zero. The *p*-value of this test is shown in the row below its Chi² ($\chi^2$) value. Rho is the fraction of the variance in the unobserved component that is explained by the random intercept.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided tests).

**Table A5.** Regression Estimates for Results Presented in Figure 5
(Coefficients and in Parentheses *t*-Values)

| | (1)<br>Traditional resp. | (2)<br>Neither nor resp. | (3)<br>Egalitarian resp. |
|---|---|---|---|
| Labor market hours (ref.: low information) | | | |
| Woman 20 hours more | -2.787** | -2.841*** | -1.788*** |
| | (-2.713) | (-5.868) | (-9.198) |
| Woman 10 hours more | -2.813** | -1.415*** | -0.915*** |
| | (-3.235) | (-4.594) | (-5.693) |
| Both same hours | -0.510 | 0.0961 | -0.0467 |
| | (-0.606) | (0.363) | (-0.348) |
| Woman 10 hours less | -0.452 | 0.828** | 0.523*** |
| | (-0.632) | (2.887) | (3.472) |
| Woman 20 hours less | 0.689 | 0.755* | 0.921*** |
| | (0.671) | (2.170) | (4.484) |
| Intercept | 1.410*** | 0.143 | -0.0986 |
| | (3.424) | (0.833) | (-1.442) |
| Number of observations | 87 | 442 | 1,931 |
| Number of individuals | 29 | 152 | 650 |
| Wald-test $\chi^2$ | 18.77 | 94.47 | 173.43 |
| Wald-test *p*-value | 0.002 | 0.000 | 0.000 |
| Rho | 0.133 | 0.238 | 0.087 |

*Notes.* For the Wald test, the joint null hypothesis is that both dummies for the information conditions are equal to zero. The *p*-value of this test is shown in the row below its Chi² ($\chi^2$) value. Rho is the fraction of the variance in the unobserved component that is explained by the random intercept.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided tests)

**Table A6.** Regression Estimates for Results Presented in Figure 6
(Coefficients and in Parentheses *t*-Values)

| | (1)<br>Labor market hrs. | (2)<br>Relative income |
|---|---|---|
| Main effects: vignette factors | | |
| VP is woman (ref.: man) | 1.127 | 0.808 |
| | (0.486) | (0.331) |
| VP's share of housework | -3.356 | 2.726 |
| | (-0.494) | (0.374) |
| VP's share of housework squared | -5.418 | -9.827 |
| | (-0.803) | (-1.354) |
| VP's share of labor market hours (ref.: same hours) | | |
| less labor market hours | -3.713 | |
| | (-1.629) | |
| more labor market hours | -1.406 | |
| | (-0.646) | |
| VP's relative income (ref.: same income) | | |
| less income | | -1.554 |
| | | (-0.643) |
| more income | | 1.700 |
| | | (0.705) |
| Interactions: VP woman X | | |
| VP's share of housework | -3.783 | -2.840 |
| | (-0.388) | (-0.276) |
| VP's housework squared | 1.052 | 0.705 |
| | (0.109) | (0.069) |
| Interactions VP's share of housework X | | |
| less labor market hours | 13.79 | |
| | (1.428) | |
| more labor market hours | 0.270 | |
| | (0.029) | |
| less income | | 7.239 |
| | | (0.706) |
| more income | | -10.56 |
| | | (-1.029) |
| Interactions: VP's share of housework squared X | | |
| less labor market hours | -8.826 | |
| | (-0.924) | |
| more labor market hours | 1.618 | |
| | (0.177) | |
| less income | | -6.727 |
| | | (-0.662) |
| more income | | 11.99 |
| | | (1.175) |
| Interactions: VP woman X | | |
| less labor market hours | 2.535 | |
| | (0.790) | |
| more labor market hours | -0.361 | |
| | (-0.113) | |
| less income | | 4.197 |
| | | (1.207) |

| | | |
|---|---|---|
| more income | | -0.975 |
| | | (-0.283) |
| Interactions: VP woman X VP's share of housework X | | |
| less labor market hours | -10.73 | |
| | (-0.795) | |
| more labor market hours | -3.080 | |
| | (-0.228) | |
| less income | | -19.32 |
| | | (-1.320) |
| more income | | 3.871 |
| | | (0.265) |
| Interactions: VP woman X VP's share of housework sq. X | | |
| less labor market hours | 10.09 | |
| | (0.755) | |
| more labor market hours | 6.134 | |
| | (0.460) | |
| less income | | 19.73 |
| | | (1.363) |
| more income | | -4.432 |
| | | (-0.305) |
| Intercept | 3.391* | 1.533 |
| | (2.127) | (0.896) |
| Number of observations | 1,278 | 1,278 |
| Number of individuals | 428 | 428 |
| Wald-test $\chi^2$ | 882 | 536 |
| Wald-test p-value | 0.000 | 0.000 |
| Rho | 0.121 | 0.080 |

*Notes.* VP is the abbreviation for vignette person. For the Wald test, the joint null-hypothesis is that both dummies for the information conditions are zero. The *p*-value of this test is show in the line below its Chi² ($\chi^2$) value. Rho is the fraction of the variance of the unobserved component explained by the random intercept.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided tests)

# 2. Extended Analyses and Robustness Checks

### 2.1. Effects of Relative Share of Income and Childcare, Different Modeling Strategy

We conducted extensive robustness checks to test the stability of our results and also to examine some extensions, such as the effects of childcare. As in the main text, we present the results visually (predicted values or regression coefficients with 95% confidence intervals). The exact regression results underlying these estimates can be reproduced using our replication packages (Stata do-files; Auspurg and Düval 2024).

First, we perform additional analyses to test our assumption that respondents expect female vignette persons with the same labor market hours to make relatively smaller contributions to the household income. The results are presented in Figure A3. Note that for the cases shown here, full information on the labor market hours was also provided (vignettes with information on income contributions always included information on the labor market hours). Here, in the full information condition, even "traditional" respondents no longer show a statistically significant gender effect (i.e., these respondents also rated the appropriate share of housework to be equal for male and female vignette persons) when the vignettes portrayed both partners as having equal labor market resources (having at least on average equal contribution to the household income and equal labor market hours).

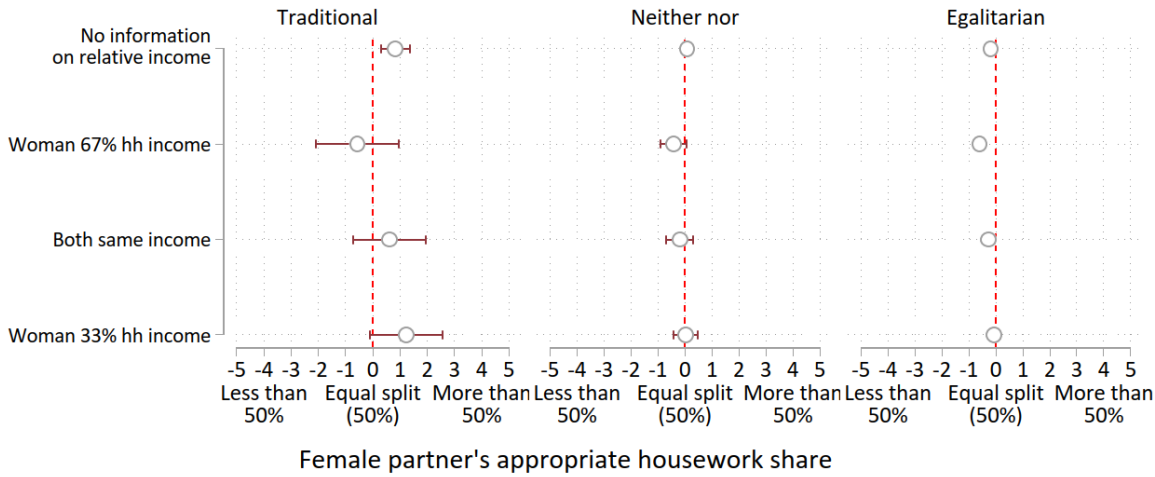


**Figure A3.** Female Partners' Appropriate Housework Share by Information on the Relative Income and Respondent Group as Classified by the Item Question on Gender Ideologies

*Notes.* This figure shows predictions of the appropriate housework share done by the female partner depending on the amount of information on the partners' relative income. "hh income" stands for household income. The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" (last column) respondents. Number of vignette evaluations for "traditional" respondents: 120; for "neither nor" respondents: 653; for "egalitarian" respondents: 2,965.

Furthermore, and even more interestingly, we can conclude from the results in Figure A3 that "traditional" respondents in the low information condition again assume traditional arrangements in which the female partner makes a lower contribution to the household income. Note also that only this traditional income arrangement leads "traditional" respondents to support a higher share of housework for the female partner. Although the difference is quite large, it is not statistically significant ($p = 0.074$). This is probably due to the relatively small number of "traditional" respondents. For the "egalitarian" respondents, we can again conclude that these respondents assumed egalitarian arrangements; the prediction for the low information condition is closest to the predictions for vignettes in which the male and female vignette partners were described as making an equal contribution to the household income. As might be expected, the "neither nor" group falls between the egalitarian and traditional respondents. For them, the low information condition is mostly on par with the condition where the woman contributes half as much income as the man. However, none of the effects are statistically significant. Across all respondent groups, the effects of income contributions are consistent with equity norms; but the effect sizes are much smaller than those found for labor market hours. (This can be seen as evidence for time availability instead of bargaining theories.)

In addition, we tested whether information about the hypothetical couples' childcare, which was included in some splits (see the *Method Section*), modified our results. Figure A4 shows the predictions of the female partner's appropriate housework share by the amount of childcare she provides. Overall, the results do not vary much by childcare, but are again (mostly) in line with equity norms. For "traditional" respondents, however, the results are somewhat inconclusive, probably due to the small number of cases: There are only 18 ratings of vignettes with information on childcare by respondents classified as "traditional" by the standard item question.

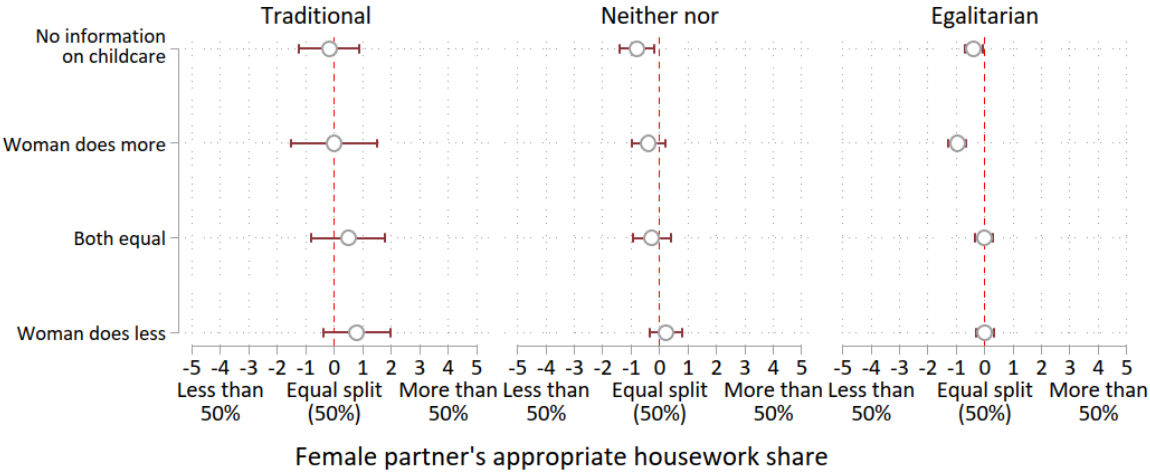


**Figure A4.** Female Partners' Appropriate Housework Share by Information on Childcare and Respondent Group as Classified by the Item Question on Gender Ideologies

*Notes.* This figure shows predictions of the appropriate housework share done by the female partner depending on the information on the division of childcare. The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" respondents (last column). Number of vignette evaluations for "traditional" respondents: 18; for "neither nor" respondents: 141; for "egalitarian" respondents: 694.

We also tested whether there are non-linear, gender-specific effects for childcare (see Figure A5). Again, a housework share of around 50% is considered appropriate if both partners have roughly the same share of childcare. At the same time, partners with a smaller (larger) share of childcare are expected to do more (less) of the housework. Again, the predictions are symmetric for male and female partners: Both genders should reduce their time spent on housework when they have a larger share of childcare. To our knowledge, this has not been tested before.

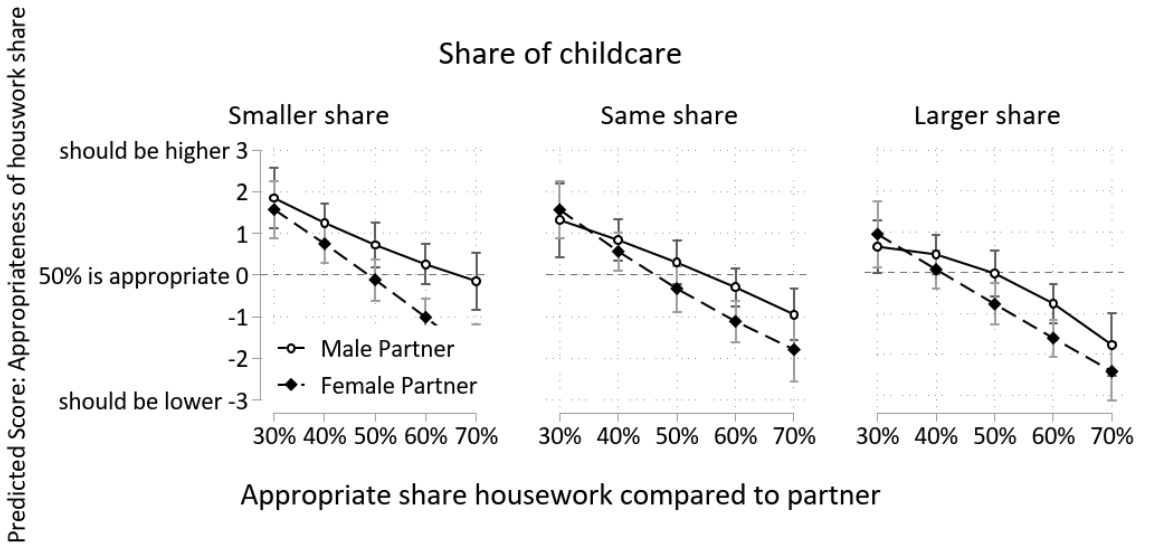


**Figure A5.** Appropriateness of the Housework Share by the Partner's Gender and Information on Childcare

*Notes.* This figure shows the predicted appropriateness score of the male partner's versus the female partner's housework share under different conditions: When doing a smaller (panel on the left), the same (middle panel), or larger share of childcare (panel on the right) than the partner. To estimate separate predictions for male and female partners, each regression model included interactions with the partner's gender. Number of vignette evaluations: 853. Due to the small number of cases, we do not split the analyses by respondents' gender ideologies (as classified by the traditional item questions).

Finally, we tested the robustness of our results by switching to a different modeling strategy, where we regressed the appropriateness of the vignette person's share of housework on all vignette dimensions together and tested for significant differences between male and female partners in pooled analyses with interaction terms. The results are presented in Figure A6. Again, our results are consistent with equity norms: Working more labor market hours or making a relatively higher income contribution allows for a lower share of the housework. At the same time, being responsible for a smaller share of the labor market work leads to a larger share of appropriate housework. Furthermore, we find no statistically significant difference between male and female vignette persons. This means that respondents do not differ in their evaluations for men and women. This is further evidence against traditional gender ideologies.

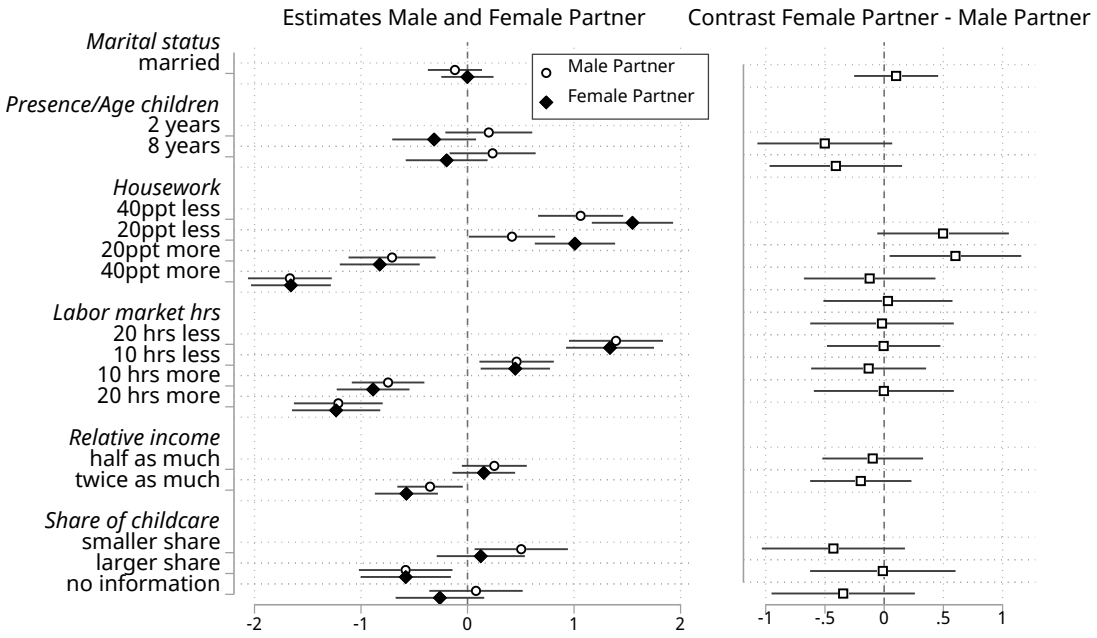


**Figure A6.** Impact of Vignette Dimensions by Gender Vignette Person

*Notes.* This figure shows regression coefficients for all vignette dimensions on the appropriateness of the vignette person's share of housework together with 95% confidence intervals. The coefficients are displayed separately by gender of the vignette partner. The left panel regressions were estimated separately for male and female vignette persons. The right panel shows the difference in effect sizes across both genders (estimated in pooled analyses with interaction effects). This figure includes only vignettes of the full information condition. Number of vignette evaluations: 1,278, consisting of 627 ratings of female and 651 ratings of male vignette persons.

### 2.2. Evaluations of the Total Workload

We also examined the results for another rating task. In a split sample with different respondents, ratings were collected on the appropriate share of a partner's total workload (i.e., housework, paid employment, and childcare if applicable). Is there evidence for traditional gender ideologies that women should work longer shifts, and is this still (or less) true when they have equal labor market characteristics than men? The experimental design for total workload was similar to that for housework with one major difference: We did not implement a low information condition in this additional experiment, i.e., respondents always received information about the labor market hours of both partners.

Overall, the results are comparable to the main results presented. Figure A7 supports the finding that adding information on relative income shifts the ratings further to the left, implying that women should do less of the total workload. This shift is observable especially for "traditional" respondents.

The main results are also supported by Figure A8. To control for non-linear effects, the appropriateness of the total workload was estimated by including interaction terms (see Figure 6 in the main text). It is clear from all the figures that doing more of the housework allows both men and women to reduce their overall workload. There are level differences by the amounts of labor market hours, relative income, and childcare; however, there is no evidence for non-linear effects, especially not in cases where women are responsible for larger shares of couples' labor market hours or household income. This is also evidence against the gender display thesis, which would have expected a U-shaped curve (higher share of female housework in situations where the couple's labor market constellation deviates from traditional gender ideologies).

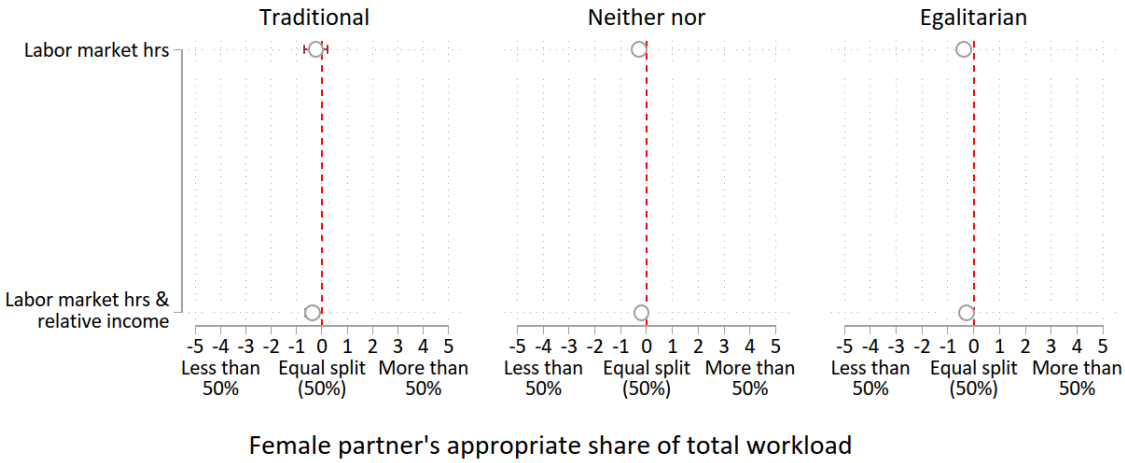


**Figure A7.** Female Partners' Appropriate Share of the Total Workload by Information on Labor Market Characteristics and Respondent Group as Classified by the Item Question on Gender Ideologies

*Notes.* This figure shows predictions of the appropriate share of the total workload (i.e., housework, paid employment, and childcare if applicable) done by the female partner depending on the amount of information on labor market characteristics. The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" respondents (last column). Number of vignette evaluations for "traditional" respondents: 374; for "neither nor" respondents: 1,630; for "egalitarian" respondents: 7,958.

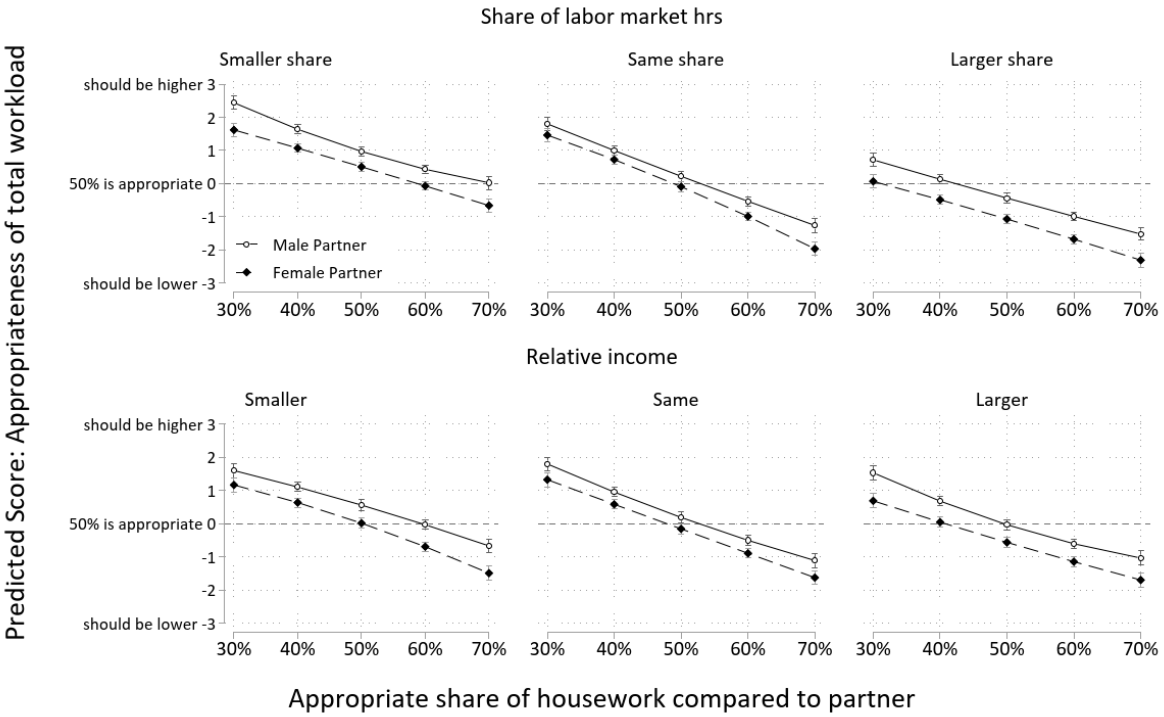


**Figure A8.** Appropriateness of the Share of the Total Workload by Gender and Relative Labor Market Status

*Notes.* This figure shows the predicted appropriateness score of the male versus female partner's share of total workload (i.e., housework, paid employment, and childcare if applicable) for a different amount of this partner's labor market hours (first row), and his/her share of income (second row). Predications were estimated with separate regression models for labor market hours and relative income contribution. To estimate separate predictions for male and female partners, each regression model included interactions with the partner's gender. To adjust for possible non-linear effects, also squared terms of the housework hours were included Number of vignette evaluations for labor market hours and for relative income: 6,930.

### 2.3. Other (Broader) Classifications of Traditional Respondents

In the main analyses, respondents were grouped into "traditional," "egalitarian," and "neither nor" in terms of their gender ideologies. This was done using the item question closest to the experimental design, namely: "Men should participate in housework to the same extent as women." Based on this item question, only 3% of respondents were classified as "traditional." To ensure that the results were not biased by the classification itself, two additional classifications were made.

First, respondents were classified according to their agreement with a second item question: "Women should care more about their family than their career." Based on this item, 16% of respondents were classified as "traditional" and 41% as "neither nor." The results were similar to the main analysis: "Traditional" respondents expect women to do more of the housework when no information on labor market hours and/or relative income is presented. Adding information again shifts the responses to the left (see Figure A9). Figure A10 again shows broad support for equity norms across all three respondent groups. Women who work more hours in the labor market than their partners are expected to reduce their share of housework, supporting the rejection of the gender display thesis and the finding that equity norms trump traditional gender ideologies.

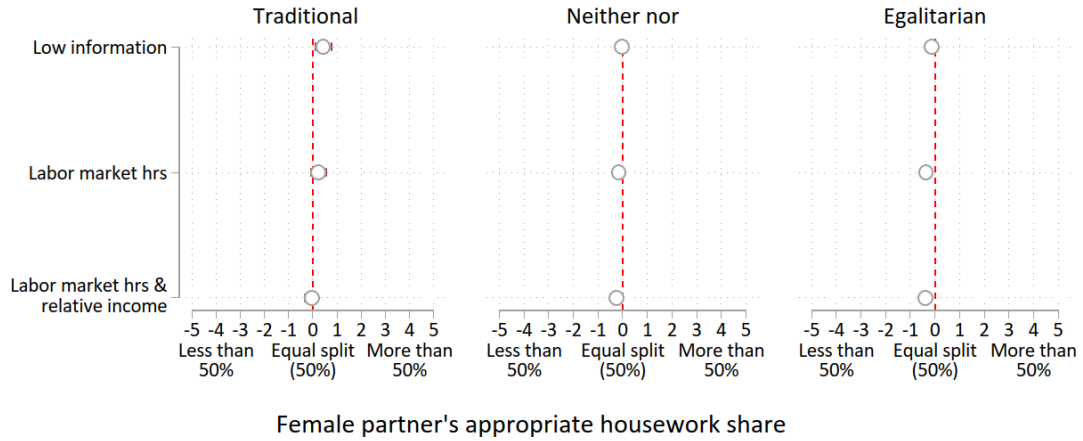


**Figure A9.** Female Partners' Appropriate Housework Share by Amount of Information on Labor Market Characteristics, Grouping Respondents in "Traditional" to "Egalitarian" based on an Alternative Item

*Notes.* This figure shows predictions of the appropriate housework share done by the female partner dependent on the amount of information on labor market characteristics. The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" (last column) respondents. Respondents are now grouped based on an alternative item question not probing on housework, but women's general responsibility for family work versus having a labor market career. Number of vignette evaluations for "traditional" respondents: 588; for "neither nor" respondents: 1,545; for "egalitarian" respondents: 1,599.

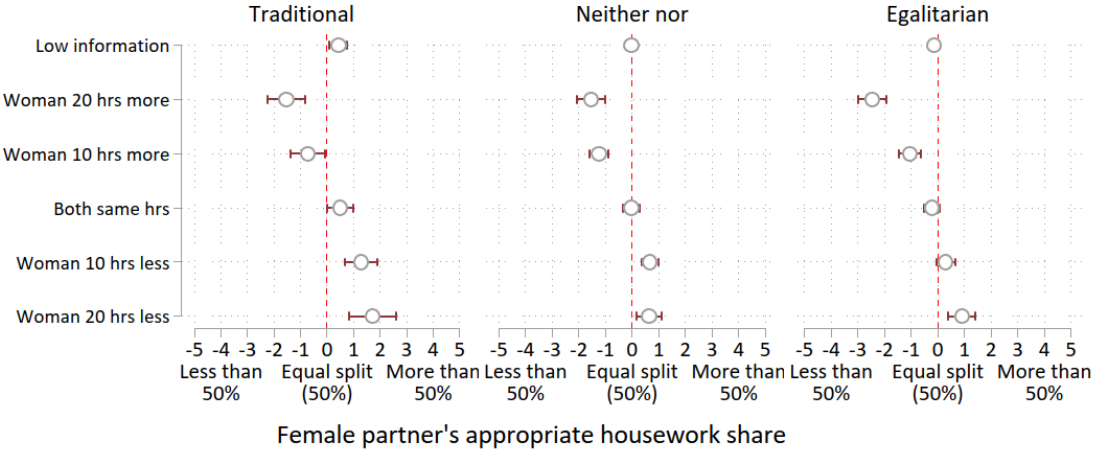


**Figure A10.** Female Partners' Appropriate Housework Share by Information on the Relative Labor Market Hours, Grouping Respondents in "Traditional" to "Egalitarian" based on an Alternative Item

*Notes.* This figure shows predictions of the appropriate housework share done by the female partner dependent on different relative labor market hours. The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" respondents (last column). Respondents are now grouped based on an alternative item question not probing on housework, but women's general responsibility for family work versus having a labor market career. Number of vignette evaluations for "traditional" respondents: 352; for "neither nor" respondents: 1,007; for "egalitarian" respondents: 1,095.

Second, an index of gender ideologies was generated using 3 items from *pairfam*'s standard gender ideologies measure. With a Cronbach's alpha of only .55, the reliability of this index is relatively low, the main argument for not including it in the main analyses. Based on a tercile split, respondents were again divided into the 3 groups. Even with this broader definition of "traditional," the conclusions are comparable to the main results (see Figures A11 and A12).

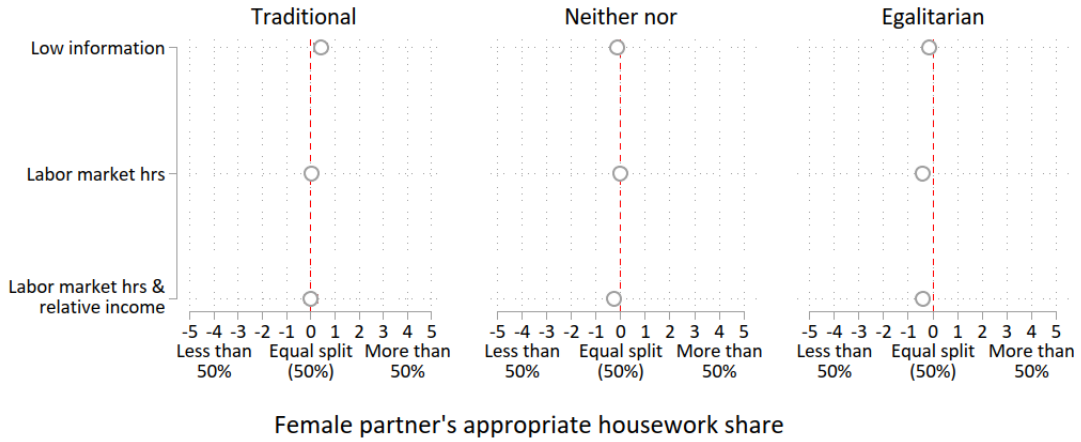


**Figure A11.** Female Partners' Appropriate Housework Share by Amount of Information, Grouping Respondents in "Traditional" to "Egalitarian" based on an Additive Index

*Notes.* This figure shows predictions of the appropriate housework share done by the female partner dependent on the amount of information on labor market characteristics. The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" respondents (last column). Respondents are now grouped based on an additive index summarizing 3 item questions. Number of vignette evaluations for "traditional" respondents: 867; for "neither nor" respondents: 1,160; for "egalitarian" respondents: 1,714.

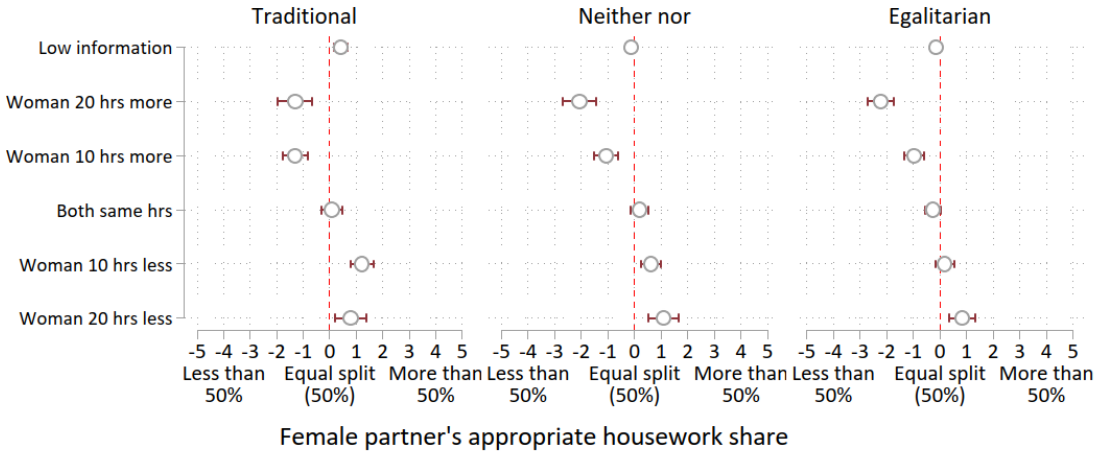


**Figure A12.** Female Partners' Appropriate Housework Share by Information on Labor Market Characteristics, Grouping Respondents in "Traditional" to "Egalitarian" based on an Additive Index

*Notes.* This figure shows predictions of the appropriate housework share done by the female partner dependent on different relative labor market hours. The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" (last column) respondents. Respondents are now grouped based on an additive index summarizing three item questions. Number of vignette evaluations for "traditional" respondents: 596; for "neither nor" respondents: 729; for "egalitarian" respondents: 1,135.

### 2.4. Checks With Other Sample Restrictions and Regression Models

It is well known from survey research that respondents who complete questionnaires the fastest are susceptible to method effects. For example, they may not take enough time to read the vignettes properly and therefore only remember the dimension they read last when answering related questions (e.g., Düval and Hinz 2020). Because this behavior would bias the results, the sample was restricted based on a

suggestion by Sauer et al. (2011)[4] to ensure that the results based on the original sample restrictions were not biased. The results were again comparable to the main results.

As a further check, the analyses were repeated separately for male and female respondents. Again, our main findings were found to be robust. An interesting side result of these analyses was that in the vignette ratings, male respondents did not appear to be more traditional in their gender ideologies than female respondents (Düval 2022). This is in contrast to the rating of the standard item questions, where we and many others have found this gender difference. This observation may indicate that especially male respondents impute gender-specific background characteristics when evaluating the item questions. Future research could test this hypothesis with larger numbers of respondents that allow for more reliable gender comparisons (our small number of "traditional" respondents only allows for initial, exploratory analyses; see Düval 2022 for a more detailed discussion).

Finally, a model specification check was conducted by switching from ordinary least squares (OLS) regressions with random intercepts to OLS regressions with cluster-robust standard errors to account for the hierarchical data structure. The results remained unchanged. These analyses can also be reproduced using our replication files.

## 3. Previous Research with Experimental Designs

There have been several studies on the measurement limitations of the abstract item questions on gender ideologies (e.g., Behr et al. 2013, Braun 1998, Braun 2008, Braun and Scott 2009, Constantin and Voicu 2014, Grunow, Begall and Buchler 2018, Walter 2018). However, we are not aware of any study that explicitly focused on the conflation of gender ideologies (injunctive norms) and stereotypical beliefs (descriptive norms) in combination with equity norms that we are interested in. As explained in the main text, ideally one would use an experimental design to study the effects of implicit beliefs and also to truly disentangle gender from the effects of different labor market resources. In this *Supplement*, we therefore provide information on some previous, related research on the division of labor in couples that used survey experiments.

First, Pedulla and Thébaud (2015) and Auspurg, Iacovou and Nicoletti (2017) focused on preferences for more or less gendered work arrangements. Interestingly, these studies, based on experimental rather than observational data, found that individuals have a strong preference for gender-blind, egalitarian arrangements. For the U.S., Pedulla and Thébaud (2015) found that respondents, regardless of their gender, preferred an egalitarian division of paid and unpaid work. Auspurg, Iacovou and Nicoletti (2017) used a factorial survey experiment with a British respondent sample to disentangle gender from other explanations for the preferred division of labor. Their main finding was that it is not gender that influences preferences for labor division. Instead, respondents based their preferences on gender egalitarian distribution rules. However, these preferences were tied to options that made them easily achievable, e.g., by removing (institutional) constraints such as a lack of childcare facilities or earnings inequalities that might make dual-earner arrangements difficult in reality. Jacobs and Gerson (2016) assigned respondents in the U.S. (Time Sharing Experiments in The Social Sciences, TESS) to three different scenarios (vignettes) with different family constellations and asked them whether they thought the described single or married mothers or fathers should continue to work full time, stay at home, or cut back to part time. Support of both mother's and father's employment increased substantially when the individuals were described as being satisfied with their jobs, or when respondents were made to believe that the family depended on their income. There was also a moderate gender gap in the expected

[4] The slowest 1% of respondents were excluded, as were all respondents who were two standard deviations below the average response time for the vignettes. In total, this affected only 23 respondents.

direction (respondents were overall more supportive of fathers' employment than mothers' employment), but this effect was small compared to the effects of the economic variables.

Second, experimental studies on (just) earnings differences between men and women (Auspurg, Hinz and Sauer 2017) and mothers and childless women (Correll, Benard and Paik 2007) found support for a gendered status value or gendered descriptive norms. Women and/or mothers were perceived as less competent and less committed to work in the labor market (Correll, Benard and Paik 2007). Auspurg, Hinz and Sauer (2017) found that experienced inequalities in men's and women's earnings influence perceptions and evaluations of justice. Gendered status beliefs are internalized by men and women and therefore earnings differentials are perceived as fair (Auspurg, Hinz and Sauer 2017). Of course, since the results only refer to descriptive beliefs about gender differences in income or in the labor market, they cannot be extrapolated to injunctive or descriptive norms about the division of labor in the household. However, they do shed light on the fact that gender status beliefs and descriptive gender norms have an impact on what is considered fair or appropriate.

Third, Carriero and Todesco (2017) and Schulz (2021) used factorial survey experiments to explore the fairness norms that men and women use to judge the division of paid and unpaid work in couples. In different regions of northwestern Italy, Carriero and Todesco (2017) explored the perceived fairness of different housework constellations of hypothetical couples. The fair amount of housework was found to depend on the working hours and gender; but in the opposite direction than expected from traditional gender ideologies. Also according to the authors, this result is likely an artefact caused by too little gradation regarding the share of housework, where the respondents could only choose between three extreme categories (woman does 20%, 50%, or 80% of the housework). Using a sample of German residents, Schulz (2021) compared men's and women's attitudes toward the division of housework. He found that labor market characteristics influence what men and women perceive as fair. Since he found that men should do more of non-routine housework tasks such as repairs or paperwork, gender seems to also matter. Schulz (2021) did not vary the amount of information presented to the respondents. He can only measure whether there is a direct effect of gender after controlling for labor market characteristics. However, since relevant factors (e.g., information on the division of childcare within the couple) were not included in his experimental design, it is unclear whether a remaining gender difference indicates support for traditional gender ideologies or whether the respondents simply assumed, for example, that women do a greater share of childcare and therefore consider it fair if she has to do less of the housework.

Finally, in a seminal study, Thébaud, Kornrich and Ruppanner (2021) provided insights into the different standards of housework to which men and women are held accountable. They used a factorial survey experiment in which U.S respondents were shown pictures of relatively clean or messy rooms of male and female occupants. The authors found an effect of the gender of the room occupant in that women occupying a relatively clean room were judged to be less clean and less moral than men. In messy rooms, no gender difference was found. They explain these results by the fact that women are exposed to higher standards and negative stereotypes regardless of the situation. In contrast, for men only messiness activates negative stereotypes. Particularly interesting in the context to our study is that Thébaud, Kornrich and Ruppanner (2021) found gender differences in the expected responsibility for housework. If they work full time, women are expected to do more housework. This is true both when they live alone and when they are parents or live with a spouse. However, the authors did not clearly separate the work and family conditions when asking about housework responsibilities.

This is where our study comes in. In our factorial survey experiment we not only experimentally vary the family constellations (e.g., marital status and presence and age of children), but also a wide range of work arrangements (e.g., both partner's shares of housework, childcare, and labor market hours and relative income contributions). This allows us to examine whether (women's) expected responsibility for housework depends on family conditions, work conditions, or interactions of these

spheres. In addition, by varying the amount of information, we can disentangle evidence for gender ideologies (injunctive norms) and stereotypical beliefs (descriptive norms).

**References**


Atzmüller, Christiane and Peter Steiner. 2010. "Experimental Vignette Studies in Survey Research." *Methodology: European Journal of Research Methods for The Behavioral and Social Sciences* 6:128-38.

Auspurg, Katrin and Sabine Düval. 2024. "Replication Files for 'What Looks Like Gender Ideologies Could Be Stereotypes? Using an Experimental Design to Dig Deeper into Normative Attitudes and Beliefs Underlying Couples' Labor Division.'" *OSF*. April 5. osf.io/3fqw9.

Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks, CA: Sage.

Auspurg, Katrin, Thomas Hinz and Carsten Sauer. 2017. "Why Should Women Get Less? Evidence on the Gender Pay Gap from Multifactorial Survey Experiments." *American Sociological Review* 82(1):179-210.

Auspurg, Katrin, Maria Iacovou and Cheti Nicoletti. 2017. "Housework Share between Partners: Experimental Evidence on Gender-Specific Preferences." *Social Science Research* 66:118-39.

Baxter, Janeen, Belinda Hewitt and Michele Haynes. 2008. "Life Course Transitions and Housework: Marriage, Parenthood, and Time on Housework." *Journal of Marriage and Family* 70(2):259-72.

Behr, Dorothée, Michael Braun, Lars Kaczmirek and Wolfgang Bandilla. 2013. "Testing the Validity of Gender Ideology Items by Implementing Probing Questions in Web Surveys." *Field Methods* 25(2):124-41.

Braun, Michael. 1998. "Gender Roles." Pp. 111-34 in *Comparative Politics. The Problem of Equivalence*, edited by J. W. van Deth. New York: Routledge.

Braun, Michael. 2008. "Using Egalitarian Items to Measure Men's and Women's Family Roles." *Sex Roles* 59(9):644-56.

Braun, Michael and Jacqueline Scott. 2009. "Gender-Role Egalitarianism – Is the Trend Reversal Real?". *International Journal of Public Opinion Research* 21:362-67.

Carriero, Renzo and Lorenzo Todesco. 2017. "The Interplay between Equity and Gender Ideology in Perceived Housework Fairness: Evidence from an Experimental Vignette Design." *Sociological Inquiry* 87(4):561-85.

Constantin, Andreea and Malina Voicu. 2014. "Attitudes Towards Gender Roles in Cross-Cultural Surveys: Content Validity and Cross-Cultural Measurement Invariance." *Social Indicators Research* 123.

Correll, Shelley, Stephen Benard and In Paik. 2007. "Getting a Job: Is There a Motherhood Penalty?". *American Journal of Sociology* 112(5):1297-338.

Davis, Shannon N and Theodore N. Greenstein. 2009. "Gender Ideology: Components, Predictors, and Consequences." *Annual Review of Sociology* 35(1):87-105.

Düval, Sabine. 2022. "Do Men and Women Really Have Different Gender Role Attitudes? Experimental Insight on Gender-Specific Attitudes toward Paid and Unpaid Work in Germany." *Social Science Research* 112:102804.

Düval, Sabine and Katrin Auspurg. 2020. "The Factorial Survey Experiment on "Distribution of Work in Partnerships" in the German Family Panel (Pairfam)." *Pairfam Technical Paper # 14.*

Düval, Sabine and Thomas Hinz. 2020. "Different Order, Different Results? The Effects of Dimension Order in Factorial Survey Experiments." *Field Methods* 32(1):23-37.

Green, Donald, Shang Ha and John Bullock. 2009. "Enough Already About 'Black Box' Experiments: Studying Mediation Is More Difficult Than Most Scholars Suppose." *The Annals of the American Academy of Political and Social Science* 628.

Grunow, Daniela, Katia Begall and Sandra Buchler. 2018. "Gender Ideologies in Europe: A Multidimensional Framework." *Journal of Marriage and Family* 80(1):42-60.

Hole, Arne. 2007. "Wtp: Stata Module to Estimate Confidence Intervals for Willingness to Pay Measures." Statistical Software Components: Boston College, Department of Economics.

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 176(1):5-32.
Jacobs, Jerry A. and Kathleen Gerson. 2016. "Unpacking Americans' Views of the Employment of Mothers and Fathers Using National Vignette Survey Data." *Gender & Society* 30(3):413-41.
Pedulla, David S. and Sarah Thébaud. 2015. "Can We Finish the Revolution? Gender, Work-Family Ideals, and Institutional Constraint." *American Sociological Review* 80(1):116-39.
Sauer, Carsten, Katrin Auspurg, Thomas Hinz and Stefan Liebig. 2011. "The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency." *Survey Research Methods* 5(3):89-102.
Schulz, Florian. 2021. "Attitudes Towards Sharing Housework in Couple Context: An Empirical, Factorial Survey Approach." *Journal of Family Research* 33(1): 148–183.
Thébaud, Sarah, Sabino Kornrich and Leah Ruppanner. 2021. "Good Housekeeping, Great Expectations: Gender and Housework Norms." *Sociological Methods & Research* 50(3):1186–214.
Walter, Jessica Gabriele. 2018. "The Adequacy of Measures of Gender Roles Attitudes: A Review of Current Measures in Omnibus Surveys." *Quality & Quantity* 52(2):829-48.