

Housework as a Woman's Job? What Looks Like Gender Ideologies Could Also Be Stereotypes

Katrin Auspurg,^a Sabine Düval^{a,b}

a) LMU Munich; b) German Youth Institute (DJI)

Abstract: We question the validity of standard measures of gender ideology. When asked about “men” and “women” in general, respondents may imagine women (men) with lower (higher) labor market resources. Therefore, standard measures may conflate gender ideologies (injunctive norms) with stereotypical beliefs (descriptive norms). We test this hypothesis with an experiment in the German family panel *pairfam*: ~1,200 respondents rated the appropriate division of housework in ~3,700 hypothetical couples. By gradually adding information about labor market resources, we were able to override respondents' stereotypical beliefs. We find that with more information, even “traditional” respondents support egalitarian housework arrangements. The main difference between “traditional” and “egalitarian” respondents is not in their ideologies (as previously thought), but in their interpretation of vague items. This leads us to conclude that standard measures overestimate traditional gender ideologies. Our study also illustrates how varying the amount of information can help identify respondents' implicit beliefs.

Keywords: housework; gender ideologies; stereotypical beliefs; injunctive and descriptive norms; survey experiment; causal mediation analysis

Reproducibility Package: The data we used (*pairfam* data release 10.0) can be accessed here: <https://www.pairfam.de/en/data/data-access>. Our replication files (Stata do-files and data on response times not included in the *pairfam* release) are available on the following OSF platform: <https://osf.io/3fqw9> (Auspurg and Düval 2024).

Citation: Auspurg, Katrin, and Sabine Düval. 2024. “Housework as a Woman's Job?: What Looks Like Gender Ideologies Could Also Be Stereotypes.” *Sociological Science* 11: 789–814.

Received: September 21, 2023

Accepted: February 22, 2024

Published: September 3, 2024

Editor(s): Arnout van de Rijt, Maria Abascal

DOI: 10.15195/v11.a29

Copyright: © 2024 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

SIGNIFICANT gender inequalities in paid and unpaid work persist in Germany and other Western countries. Although women have caught up with men in terms of education, they still often work fewer hours in the labor market and earn less than men (e.g., for the United States: Killewald and Gough 2010; for Germany: Nitsche and Grunow 2016). At the same time, women still bear the main responsibility for unpaid work: according to official statistics for Germany, women are on average responsible for about two-thirds of the housework in heterosexual couples (German Federal Statistical Office 2015; for similar statistics on other countries: OECD 2017). This gendered division of labor is coupled with other inequalities, such as unequal access to pensions (Eurostat 2022). In addition, if housework requires a “second shift” after paid work outside the home, women's greater involvement in routine housework tasks may create an additional burden for them, limiting the time and energy they have available for more rewarding activities (Grunow, Begall, and Buchler 2018; Hochschild and Machung 1989).

Accordingly, there is a great deal of interest in the question of “why do women most of the housework?” One central explanation is traditional gender ideologies (which might also be called injunctive gender norms; that is, norms that tell

what individuals should do), which have been examined in hundreds of studies. According to these ideologies, gender per se determines who should do which part of the work (Bartley, Blanton, and Gilliard 2005; Davis and Greenstein 2009). In this study, however, we argue that standard survey item measures have likely overestimated the explanatory power of gender ideologies by conflating them with stereotypical beliefs about what is likely to be the case (what might be called descriptive norms: that is, norms that describe what individuals actually do; see, e.g., Cialdini, Kallgren, and Reno 1991 for definitions of different types of norms). Support for gendered arrangements may not reflect the assumed normative support for gendered housework (i.e., traditional gender ideologies), but rather support for egalitarian ideologies combined with interpretations of vague item questions in light of existing gender inequalities that respondents experience (but may not support): respondents may simply have thought it more appropriate for the partner with more time availability (fewer hours in the labor market) and/or lower contributions to the household income (lower earnings) to do more of the housework. In all traditional contexts with strong structural gender inequalities, this is typically the female partner. If our assumptions are correct, the standard items would confound the measurement of ideologies with that of stereotypical beliefs (i.e., they would conflate injunctive and descriptive norms). They would (1) misclassify some “egalitarian” respondents as “traditional” and thus (2) overestimate the prevalence of traditional gender ideologies.

To test these hypotheses empirically, one would like to control for the confounding factor of stereotypical beliefs. However, testing implicit beliefs is challenging. Contrasting contexts that are supposed to trigger different beliefs or asking respondents directly about their beliefs and controlling for them in multivariate models are likely to introduce confounder bias (Montgomery, Nyhan, and Torres 2018; VanderWeele 2015). Experiments allow for more internal validity, but typically do not allow testing of mediations in which a causal factor runs through another variable (all factors are manipulated simultaneously). Recently, however, an experimental design has been proposed that allows us to probe respondents' implicit beliefs under conditions of low information (Acharya, Blackwell, and Sen 2018; Imai, Tingley, and Yamamoto 2013). We use this design to unpack respondents' implicit beliefs when rating vague item questions to measure gender ideologies. In 2017/2018, about 1,200 respondents participated in an experiment embedded in the German family panel *pairfam*. Respondents' task was to indicate the appropriate division of housework in about 3,700 hypothetical couples. In the descriptions of these couples, we manipulated the (amount of) information about the partners' labor market characteristics. This allowed us to test exactly what we wanted to know: whether some respondents are misclassified as sharing “traditional” ideologies in the case of low information. In addition, our multifactorial experimental design allows us to estimate appropriate exchange rates between inputs to the relationship in the form of paid and unpaid work, conditioned on different family statuses (married or with children yes/no). Although we use this information only in extended analyses, this evidence may be of interest to scholars concerned with multidimensional gender ideologies, in which some inputs are assumed to be ideally equally distributed, whereas others are not.

With this experimental design, our study makes several contributions. First, we actually find support for the hypothesized misclassification and overestimation of gender ideologies. In our general population sample, respondents who are classified as being more or less supportive of traditional ideologies by the standard measures differ essentially only in their implicit beliefs about men's and women's labor market resources (i.e., the descriptive norms they use in rating items on ideal housework arrangements), but not in their support for the gendered spheres that define traditional gender ideologies.

Second, consistent with previous research on preferences (Auspurg, Iacovou, and Nicoletti 2017; Jacobs and Gerson 2016; Pedulla and Thébaud 2015), we find that almost all respondents support gender-neutral (equal) housework arrangements when gender inequalities in labor market hours and contributions to the household income are eliminated in an experimental design. At the same time, these results imply that only when women catch up with men in terms of both paid hours *and* earnings they will no longer be seen as more responsible for unpaid work. A key policy implication is therefore that reducing preexisting gender inequalities in paid work, rather than changing gender ideologies, is central to reducing gender inequalities in unpaid housework.

Third, although our study focuses on a specific issue, it may have broader lessons for empirical research. Different interpretations of vague items are likely to introduce not only measurement noise, but also bias. Endogeneity bias occurs when (1) vague items elicit context-specific interpretations and (2) contextual features that elicit these different interpretations are then included in analyses of the concepts they measure (as predictors, dependent, or control variables). The experimental design we present seems promising not only for unpacking heterogeneous interpretations of vague gender items, but also for other literatures that seek to explore stereotypes and beliefs as possible mediators of group-specific evaluations (e.g., in research on discrimination or trust in out-groups). We will offer some recommendations in the conclusions.

Background and State of Research

Gender Ideologies as an Explanation for Gender Inequalities

"Gender ideologies" (also called "attitudes about gender" or "gender role attitudes") encompass societal expectations of appropriate behavior attributed to men and women (see, e.g., Bartley et al. 2005; Davis and Greenstein 2009). By definition, gender ideologies measure the support for injunctive norms, that is, norms that prescribe what individuals ought to do. In classical conceptions, they range from traditional to egalitarian. Individuals on the *traditional* pole support gendered spheres in which childcare and domestic work should be done primarily by women, regardless of their income or career status (Davis and Wills 2014:810; Pedulla and Thébaud 2015). At the other pole are *nontraditional* or *egalitarian ideologies* that promote gender-neutral divisions of labor, most often in the form of equity norms (Gager 1998; Gager 2008; Thompson 1991). These view appropriate shares of housework and childcare as inversely proportional to other inputs into the relationship:

the higher the share of paid hours/earnings (or other efforts to produce household goods), the lower the share of housework/childcare a partner should do, regardless of gender.¹ More recent theories view gender ideologies as multidimensional (see, e.g., Grunow et al. 2018), as individuals may connect essentialist ideas about the distinct “natures” and responsibilities of women and men in some domains (such as the provision of childcare) with normative ideals about egalitarian, symmetrical responsibilities in others (such as the provision of income).² What all concepts still have in common, however, is the definition of the traditional pole: respondents are considered “traditional” if they assign at least certain responsibilities (such as housework) to women solely on the basis of their gender. This normative support for female housework (and male breadwinning) is often seen as a major explanation for existing gender inequalities in paid and unpaid work (see, e.g., Correll, Benard, and Paik 2007; Düval 2023; Ridgeway 2011).

Indeed, numerous studies have found that traditional ideologies are an important predictor of couples' division of labor (for reviews, see, e.g., Coltrane 2000; Davis and Greenstein 2009; Düval 2023). Traditional gender ideologies have also been found to explain a substantial part of cross-country variation in housework patterns and women's participation in the labor force (e.g., Hook 2006; Nordenmark 2004). They have also been shown to be an important source of variation in work arrangements between married and cohabiting couples, and between parents and nonparents (see, e.g., Baxter, Hewitt, and Haynes 2008; Gupta 1999). Some studies have used a design in which they attribute gender differences in housework or care that are found after accounting for the partners' different labor market resources (and sometimes some other inputs into the relationship) as evidence of traditional gender ideologies (e.g., Grunow, Schulz, and Blossfeld 2012; Kühhirt 2011). However, it is not possible to extract ideologies or norms from behavior (Bicchieri 2017). Residual gender differences could also be due to other unmeasured mechanisms. Therefore, much research has tried to identify traditional gender ideologies through direct measures based on item questions in surveys.

Why Standard Measurements May Conflate Ideologies and Beliefs

The standard approach to measuring gender ideologies in surveys (such as the International Social Survey Program, ISSP; the General Social Survey, GSS; or European Value Survey, EVS) is to ask respondents to rate several item statements that represent injunctive norms on the division of labor within couples, such as, “A man's job is to earn money; a woman's job is to take care of the home and family,” or, “Housework should be equally divided between men and women.” Other statements ask about the ideal division of childcare or policy issues, such as the ideal distributions of scarce jobs to men and women in times of economic recession (for an overview, see Walter 2018). Respondents generally rate these statements on a five-point scale from “strongly disagree” to “strongly agree.” According to their response patterns, they are then classified on a summated scale as being more or less supportive of traditional versus egalitarian ideologies: respondents who see women (men) as primarily responsible for typically female (male) work are classified as having traditional ideologies. All other respondents (who do not support gendered

responsibilities) are classified as egalitarian. Ideologies measured with such items have been used in hundreds of studies (for some reviews: Davis and Greenstein 2009; Walter 2018).

We argue that these measures likely overestimate traditional gender ideologies because some respondents with egalitarian ideologies are misclassified. The reason is that existing gender inequalities (which prime descriptive norms about what is the case) are likely to provide a stereotypical lens through which the abstract items are interpreted. The standard items only ask what “men” and “women” should do, without providing relevant background information, for example, on labor market resources (if at all, there is only information on the presence and age of children). We assume that, particularly in contexts with large structural gender inequalities, a significant proportion of respondents will classify persons described as women as second earners, working fewer hours and earning less (for justifications by theories of statistical discrimination, heuristics, and status beliefs, see, e.g., Correll and Benard 2006). If egalitarian ideologies in the form of equity norms are applied against the backdrop of such a stereotypical scenario, it may also be considered appropriate for women to take over most of the unpaid housework. In other words, we argue that instead of the intended gender ideologies, the standard item questions may measure stereotypical beliefs (i.e., descriptive norms).

The hypothesized causal structure is illustrated in Figure 1. For traditional gender ideologies (see the first panel on the left), the treatment effect of interest X is the direct effect of gender on the appropriate share of housework Y . In the absence of explicit information (as in the case of standard item measures), the gender information is likely to activate respondents' stereotypical beliefs: respondents are likely to imagine women as the partner with relatively fewer paid hours (i.e., more time available for housework) and earnings (i.e., lower contribution to the household income). This would also lead them to agree to a higher appropriate share of housework for the female partner if they only adhere to equity norms that consider paid work hours/income as gender-neutral determinants of the appropriate share of housework (see the indirect mediation effect by stereotypical beliefs M in the second panel in Figure 1). Unless this indirect effect is netted out, egalitarian respondents would be misclassified as traditional.

By overlooking this issue, much of the existing literature may have overestimated not only the prevalence but also the impact of traditional gender ideologies. This is because in particular respondents in contexts with high gender inequalities can be expected to have gendered stereotypes in mind and thus be misclassified. Therefore, the association of traditional gender ideologies with gender inequality across countries or over time observed in many studies may not be a causal effect. (Technically, this is due to endogeneity issues/a violation of exclusion restrictions; see Dafoe, Zhang, and Caughey 2018.) In sum, we assume that the standard measures without explicit information on labor market resources (1) confound traditional gender ideologies with stereotypes about “men” and “women,” thereby misclassifying some egalitarian respondents as traditional and (2) thus lead to an overall overestimation of the prevalence and impact of traditional gender ideologies.

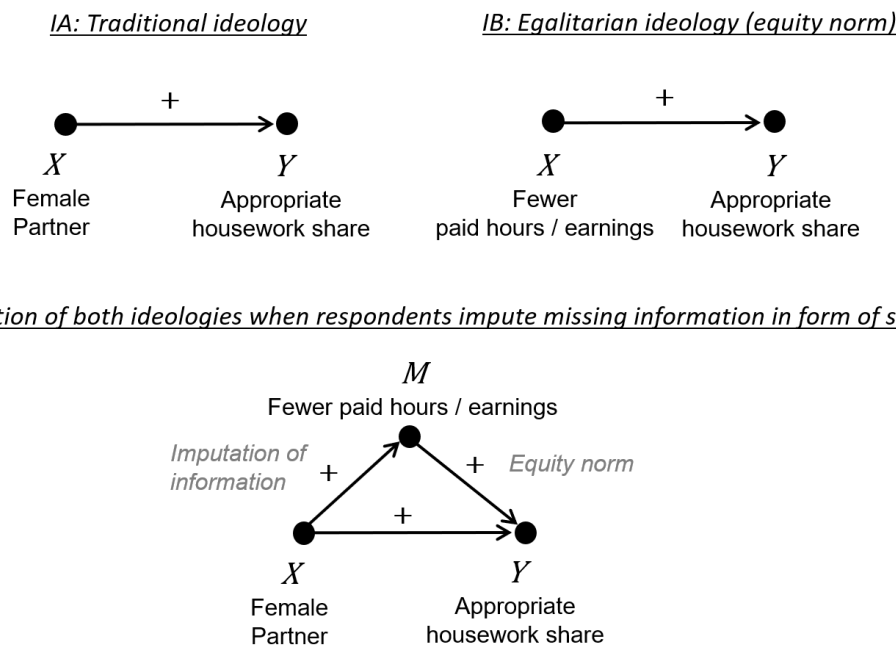


Figure 1: Overview on ideologies and assumed mediation by respondents' beliefs. *Notes:* Respondents are likely to impute missing information about the relative share of paid hours and earnings based on existing labor market inequalities (see the second panel). If not controlled for, this mediator leads to an overestimation of the direct effect of gender on the appropriate share of housework.

The Need for Experimental Manipulations

Similar issues have been discussed for other abstract measures, such as questions about “Black” and “White” Americans (in which respondents are likely to attribute different political ideologies; see Acharya et al. 2018; Dafoe et al. 2018). However, we are not aware of any studies that have focused on the widely used gender ideology measures discussed in this study (see our *Online Supplement, Part 4* for an overview on previous research).

A naïve approach to test the hypothesized mediation would be to ask respondents directly about their expectations of labor market characteristics when reading statements about men and women, and then to control for these beliefs in multi-variable regressions (for applications of such a “control by model” approach in discrimination research, see, e.g., Sterkens et al. 2022). However, the belief measured in this way is not based on an experimental manipulation. When including such measures in regression analyses, even experiments may suffer from confounder or collider bias (respondent characteristics may confound the association between beliefs and experimental outcome, see VanderWeele 2015; in the literature, which warns against “ruining” experimental designs by these methods, this bias is known as posttreatment bias, see, e.g., Montgomery et al. 2018).

We therefore enrich the literature by using a “control by design” approach. We use an experimental design that has been recently proposed in causal mediation

analysis literature precisely for the purpose of identifying respondent beliefs as an unmeasured mediator (Acharya et al. 2018; Imai et al. 2013). Standard experiments are not well suited to identifying these causal channels because they manipulate all treatments simultaneously and thus cannot test causal pathways in which a treatment effect X passes through a mediator M ($X \rightarrow M \rightarrow Y$). However, one can manipulate the treatment of interest (here gender) and information about the mediator (here assumed labor market characteristics) in separate arms of the experiment. Such designs allow, under certain assumptions (which will be evaluated later in the *Section on Extended Analyses*) to identify intermediate mechanisms (here respondents' implicit beliefs in situations without explicit information on labor market resources).

Methods and Data

Design of Our Experiment

All respondents were presented with three short vignettes about work-family scenarios in hypothetical heterosexual couples. Respondents' task was to rate the appropriateness of the vignette person's share of the couple's housework on an 11-point rating scale, ranging from -5 "Her/His housework share should be much smaller" over 0 "...is appropriate" to $+5$ "...should be much larger." Across the vignettes, we manipulated two factors central to our identification strategy:

- *Gender of the vignette person* (two levels). The vignettes varied in describing the share of housework done by the female or male partner. This is the treatment X to identify gender ideologies.
- *Amount of information* (three levels). There was either no information on labor market status ("low information" condition), information on both partners' labor market hours ("medium information"), or information on their labor market hours and relative contribution to the household income ("full information"). The provision of our gender-neutral information M' is expected to eliminate the mediation by implicit beliefs M that respondents use in the low information conditions (see more details later, *Subsection on Identification*).

Crossing these factors resulted in a 2×3 factorial design with six cells. Respondents were randomly assigned to one of these cells. In addition to this between-respondent variation, we also manipulated information within the vignettes presented to individual respondents:

- *Housework share* (five levels). The relative share of housework done by the partner described in the vignette varied from "30 percent (9 hours per week)" over "50 percent (15 hours per week)" up to "70 percent (21 hours per week)". We implemented an indirect rating task (instead of directly asking respondents for the appropriate housework share/hours), as this question format mirrors the standard items and is recommended for survey experiments to reduce respondent burden (Auspurg and Hinz 2015).

Low information	A <u>married</u> couple has <u>no child</u>. She is normally responsible for <u>70% (21 hours a week)</u> of the weekly housework (e.g., laundry, cooking, cleaning, repairs).										
Added for medium	She works <u>30 hours per week</u>, he works <u>40 hours per week</u>.										
and full Information	Her contribution to their monthly household income is <u>approximately half</u> of his.										
	How appropriate do you think <u>her share of the housework</u> is?										
	<u>Her share of the housework...</u>										
	should be much smaller					is appropriate					should be much higher
	-5	-4	-3	-2	-1	0	1	2	3	4	5
	o	o	o	o	o	o	o	o	o	o	o

Figure 2: Sample vignette for different information conditions. *Notes:* Sentences in gray were added only in the medium and full information conditions. Experimental manipulations are underlined. For a tabular overview on all experimental variations see *Supplement, Part 1.1*.

- *Paid hours and contribution to the household income* ($3 \times 3 \times 3$ levels). In the case of information provision, we independently manipulated the following levels: both partners could have 20, 30, or 40 labor market hours (which in combination results in a range of relative paid hours from “20 hours less” to “20 hours more”) and they could contribute “half as much,” “the same,” or “twice as much” as the partner to the household income. The independent manipulation allows us also to analyze the appropriate exchange rates between time and monetary inputs (this is possible with the full information condition that includes both dimensions; see our *Extended Analyses*).
- *Marital status and childcare* ($2 \times 3 \times 4$ levels). Some couples were explicitly described as married and some as having a child (aged two or eight years). Similar information on family status is often, but not always, provided in the standard item questions. In the case of children, we varied information about sharing childcare (“smaller share,” “same share,” or “larger share” than partner or no information). These manipulations allow us to explore in extended analyses whether childcare in particular is seen as a female responsibility.

We fully crossed these factors and used a *D*-efficient fraction of vignettes, which minimizes correlations between the experimental factors and their two- and three-way interactions (Auspurg and Hinz 2015). With this design, the vignettes were overall gender neutral: both partners performed, on average, 50 percent of the housework, and both partners had, on average, the same labor market status. All respondents were randomly assigned to three vignettes. A sample vignette text is shown in Figure 2.

Standard Measure of Gender Ideologies

We used a standard battery of items implemented by default in *pairfam* to measure gender ideologies. We focused on one item that asks about the appropriate division of housework, as this comes closest to our experimental setting: “Men should participate in housework to the same extent as women.” Respondents could answer on a five-point scale ranging from one “Strongly disagree” over three “Neither nor” to five “Strongly agree.” We classified respondents that answered this item with one or two as supporters of “traditional gender ideologies” (in our analysis sample: 3.2 percent, $N = 40$ respondents providing 120 valid vignette ratings), those that responded four or five as “egalitarian” (79.3 percent, $N = 990$ respondents providing 2,965 vignette ratings), and the remaining respondents (answer 3) are “neither nor” (17.4 percent, $N = 217$ respondents, providing 653 vignette ratings). In robustness analyses, we also use measures based on other items and an additive index.

To avoid possible “posttreatment bias” (interpretation of the items in light of the experimental manipulation), the standard item battery was administered prior to the experiment. Respondents first answered the standard item battery (which was asked shortly after the start of the survey), then answered questions on other topics (such as their family life, which took on average ~50 minutes), and finally responded to the experiment.

Participants and Survey Mode

Our experiment was carried out in the 10th wave of the German family panel *pairfam* conducted in 2017/2018 (Brüderl et al. 2019). Germany is a traditional welfare state with particularly strong gender inequalities: in 2017/2018, women’s hourly earnings were on average 21 percent lower than men’s (German Federal Statistical Office 2019). Although women have almost caught up with men in terms of general labor force participation, there are still large differences in the number of hours worked: only a minority of male employees (about one-fifth), but about half of female employees, work part-time (Schmitt and Auspurg 2022). This leads to strong gender inequalities in time available for housework and relative contributions to the household income. At the same time, Germany is a country with a relatively strong prevalence of traditional gender ideologies as measured by standard items (OECD 2017). Taken together, these aspects make Germany a prime example where one would expect the presumed overestimation of traditional gender ideologies.

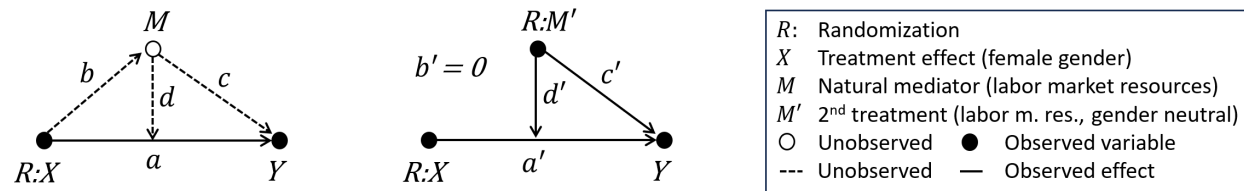
Pairfam is an annual panel study that was first launched in 2008/2009. It collects data from a nationwide, randomly selected sample of three birth cohorts of the German population: 1971 to 1973, 1981 to 1983, and 1991 to 1993 (for a detailed description: Huinink et al. 2011). Respondents are interviewed using computer-assisted personal interviews. Of the 4,750 respondents who participated in the 10th wave, a randomly selected subsample of 1,283 respondents was assigned to our experiment.³ The experiment was included in a self-completion module to reduce social desirability bias and to allow respondents to carefully read the vignettes themselves. After dropping missing values on the variables used (vignette ratings, standard measure of gender ideologies), the net sample size is 3,738 vignettes rated by 1,247 respondents.

Hypotheses and Identification Strategy

We have two main hypotheses⁴: we expect (1) that the provision of gender-neutral information about labor market resources will override respondents' stereotypical beliefs and thus lead to a decline in "traditional ideologies." Some respondents who support women doing more housework in low information conditions (i.e., who are classified as "traditional" based on standard items) should switch to supporting a more egalitarian division of housework once they have more explicit information about partners' labor market resources. We (2) hypothesize that the decline in support for traditional housework arrangements is mainly driven by respondents who are classified as "traditional" by the standard items. "Egalitarian" respondents are expected to be more immune to additional information for the following reasons: to disagree with housework as female's responsibility, they are likely to have already assumed mostly egalitarian labor market resources in the low information conditions (because they interpreted these questions as "all else being equal" between the partners, or because they themselves experience less traditional work arrangements). Or they may have thought that all couples should decide on their preferred division of housework without any norms being imposed on them (which would also count as a nontraditional, "egalitarian" stance; see Braun 2008). Either of these would imply that adding information would make little difference to egalitarian respondents.

Figure 3 shows a causal diagram for the identification strategy for the hypothesized mediation by implicit beliefs (for details, see Acharya et al. 2018; VanderWeele 2015). We expect our gender-neutral information M' to largely override respondents' stereotypical beliefs M . These implicit beliefs are called the "natural mediator" in the causal inference literature. Because our information M' is not correlated with gender, the mediation is eliminated (assuming that M' really overrides M), which allows us to identify the direct effect of gender (i.e., the effect net of gendered labor market characteristics respondents might input in the low information conditions). With this design, we can identify the following three empirical estimands:

1. *Controlled direct effect of gender* (a' in Figure 3). We can observe to what extent gender by itself (i.e., net of labor market resources) makes a difference: to what extent is there (still) an effect of gender when respondents' beliefs (the natural mediator M) are eliminated in the medium and full information conditions? In our view, only the direct gender effect measured in the full information condition is a valid measure of traditional gender ideologies.
2. *Causal effect of information/eliminated gender effect* ($a - a'$ in Figure 3). We assume that the provision of gender-neutral information eliminates evidence that women are seen as responsible for housework (i.e., traditional gender ideologies): the share of housework that is considered appropriate for the female partner should fall from $\gg 50$ percent to > 50 percent or even ~ 50 percent when gender-neutral information on labor market characteristics is added. We expect this to be particularly strong for respondents who are classified as adherents of traditional ideologies by the standard item questions. This can be analyzed by subgroup comparisons of respondents with more or less traditional ideologies (measured by the standard item questions, not



A: Low information condition

B: High information conditions

Figure 3: Blocking the natural mediation by gender-neutral information. *Notes:* In the “low information” condition (left panel), respondents are likely to impute missing information with implicit background beliefs M (labor market characteristics) that are correlated with X (gender). Despite the randomization, one can only estimate the total gender effect a (direct effect plus indirect effect in the form of the unmeasured mediation $b \cdot c$). In the conditions with more information (right panel), there is by design no causal relationship between X and M' (which is now gender-neutral) and thus no mediation. Here, the gender effect now measures the direct gender effect a' , which indicates support for traditional gender ideologies, whereas effects of labor market resources c' measure support for egalitarian ideologies (equity norms). In order to identify the indirect effect $b \cdot c$ (the assumed mediation, estimated by $a - a'$), it is necessary to assume that there is no moderation of X and M , which we denote by d and measure with d' .

shown in Figure 3 for better readability).⁵ The amount of the gender effect that is eliminated when the information is added is a measure of the indirect mediation effect ($b \cdot c$). This is an estimate of the extent to which standard items misclassify respondents as “traditional” due to missing information and thus overestimate traditional gender ideologies.

3. *Possible interactions between gender and labor market characteristics* (d and d' in Figure 3). These interaction effects may also explain why gender effects change when moving from the low to the medium and full information conditions. The assumption that they do not exist is therefore a necessary condition for correctly identifying the hypothesized mediation by the difference in gender effects between the low and high information conditions ($a - a'$; more details in the *Section on Extended Analyses*).

The ability to identify and rule out (3) as an alternative explanation to the hypothesized mediation by beliefs is a novelty of our multifactorial design, achieved by manipulating not only whether there is information (yes/no) about the hypothesized mediator in one experimental arm, but also manipulating the type of information provided. Another feature of our design is that we can estimate the amount of housework that the female partner is eventually expected to do more (by estimating cross-elasticities with housework hours, see Auspurg and Hinz 2015 and *Supplement, Part 1*). We can quantify the exchange rates that are considered appropriate, and thus also contribute to the question of what counts as egalitarian inputs into the relationship (more details in our *Section on Extended Analyses*).

All analyses are based on the standard estimation approach for multifactorial survey experiments (Auspurg and Hinz 2015; Hox, Kreft, and Hermkens 1991). We use linear regressions with random intercepts for respondents to predict vignette ratings. For these analyses we transform the vignette ratings so that our outcome measure, unless otherwise specified, is the rating of the female partner’s housework

share. Due to our experimental design, no control variables are needed in the main analysis. Descriptive overviews on our sample and more details on the experimental design, identification strategy, and robustness of the results are provided in *Supplement (Parts 1 – 3)*.

Results

Main Results

We start with the overall effect of adding information. Adding information about the partners' paid labor market hours (which are on average equal for men and women in our vignette world) reduces the share of housework considered appropriate for the female partner by 0.18 scale points on the 11-point rating scale, with this effect being at borderline significance ($p = 0.046$). The addition of information about the partners' relative contribution to the household income reduces it by a further 0.07 and thus 0.25 scale points compared to the low information condition ($p < 0.01$). This supports our first hypothesis that low information questions overestimate support for traditional housework arrangements (i.e., traditional ideologies). The effect size is small. However, we also expect changes primarily among respondents classified as being "traditional" on the standard item measure, who are a small group in our sample (three percent; providing $N = 120$ vignette ratings).

Our second hypothesis is that "traditional" respondents in particular will switch to supporting a (more) egalitarian division of housework once they receive more information. To test this, we split the respondents according to their ideology as measured by the standard item question. For ease of interpretation, the results are presented visually. Regression tables for the underlying model estimates are provided in *Supplement, Part 2*. Figure 4 shows predicted shares of housework considered appropriate for the female partner, estimated for the three different information conditions, along with their 95 percent confidence intervals. The middle line always symbolizes an equal division between the partners (50 percent), effects to the right (left) mean that the female partner should do a larger (smaller) share of the housework. The results are shown for three groups of respondents, which we classified based on their answers to the standard item measure.

As expected, in the low information condition, "traditional" respondents support a higher share of housework done by the female partner (>50 percent; see the first estimate in the left column). Thus, the item and vignette measures show a high degree of reliability. However, the responses to the vignettes provide valuable additional information: on average, the "traditional" respondents' ratings indicate that women are expected to do 10 hours more housework than men, or 67 percent of the housework per week (20 hours instead of 10 hours; for the technical estimation of these cross-elasticities, see *Supplement, Part 1*). This is close to the actual average share of housework done by women in Germany (about 60 percent, see German Federal Statistical Office 2015).

In line with our expectations, this measurement equivalence between vignette ratings and item questions holds only in the low information condition. Adding the information that the female partner has, on average, the same labor market hours

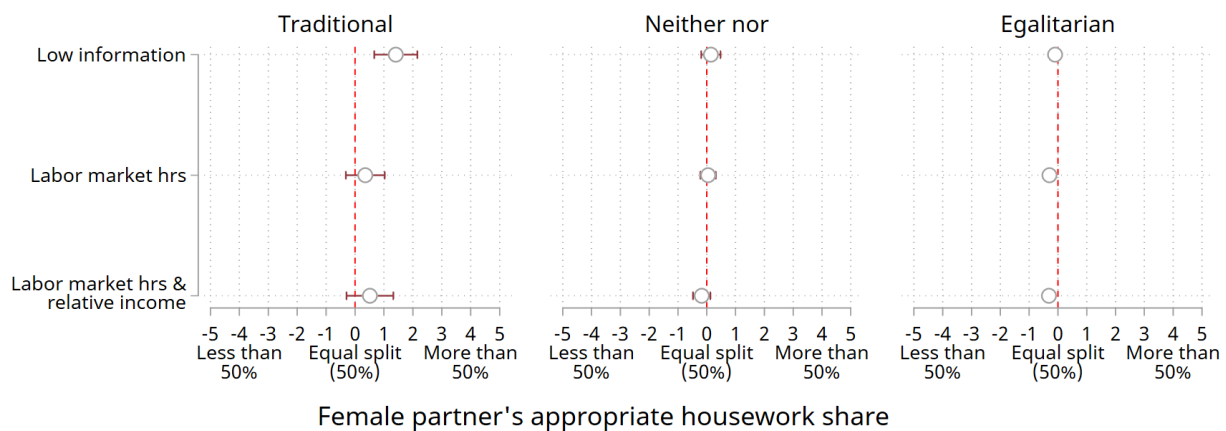


Figure 4: Female partners' appropriate housework share by amount of information on labor market characteristics and respondent group as classified by the item question on gender ideologies. *Notes:* This figure shows predictions of the appropriate share of housework done by the female partner as a function of the amount of information about labor market characteristics ("low information" with no information about labor market characteristics, information on "labor market hours," and information on "labor market hours and relative income"). Effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" respondents (last column). Number of vignette ratings for "traditional" respondents: 120; "neither nor": 653; "egalitarian": 2,965.

(and income) as the male partner eliminates the gender effect. (The reduction in the prediction of the appropriate housework share is substantial with an effect size of about one scale point: -1.056 , $p < 0.05$ for the medium; and -0.895 , $p = 0.111$ for the full information condition.)

For "egalitarian" respondents, the gender effect even turns slightly negative (see the right column). According to our estimates, "egalitarian" respondents find it appropriate for women with equal labor market hours and earnings to perform 1.7 hours less than their male partner, or about 47 percent of total housework per week.

In sum, additional information shifts the appropriate share of housework for women to smaller amounts, especially for "traditional" respondents.⁶ The most plausible explanation for this is that higher labor market hours and financial contributions entitle individuals to do less housework. Without explicit information, especially "traditional" respondents seem to assume that women have less of these labor market resources, as would be the standard case in Germany.

To gain a deeper insight into these implicit beliefs, we show in Figure 5 how appropriate housework shares vary by labor market hours. The first row again shows the appropriate female housework share in the "low information" condition, with the other rows now showing the appropriate shares conditional on the different relative labor market hours we indicated in the information conditions.

First, we see that labor market hours have a strong effect in the direction predicted by equity norms: the higher women's relative share of labor market hours, the lower their appropriate share of housework. Second, these gender-egalitarian norms are shared by all three groups of respondents. Third, for both "egalitarian"

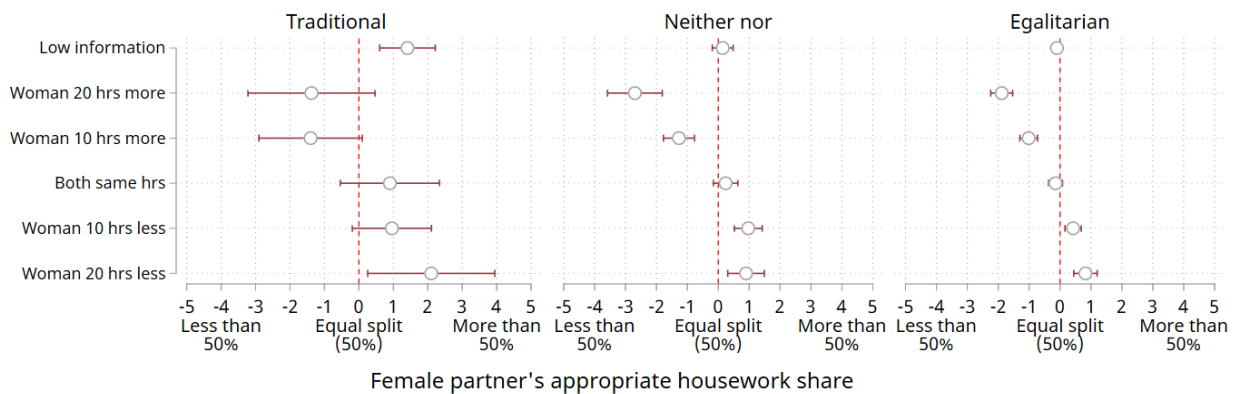


Figure 5: Female partners' appropriate housework share by information on the relative labor market hours and respondent group as classified by the item question on gender ideologies. *Notes:* This figure shows predictions of the appropriate share of housework done by the female partner as a function of different relative labor market hours (ranging from women having "20 hours less" to "20 hours more" than their partner). The effects are shown separately for "traditional" (first column), "neither nor" (second column), and "egalitarian" (last column) respondents. Number of vignette ratings for "traditional" respondents: 87; "neither nor": 442; "egalitarian": 1,931.

and "neither nor" respondents, the point estimate for the "low information" condition best matches with the estimate for the "both same hours" condition. If there are no interactions between gender and labor market status (which will be evaluated later), such equivalence is evidence that respondents assumed, on average, a similar labor market level M when information was missing; that is, they assumed that the female vignette partner had, on average, a *similar* labor market participation as her male partner. This is different for the "traditional" respondents: for this group, the estimate in the "low information" condition is on par with women having the same or a *lower* labor market participation (10 to 20 hours less than her partner).

Traditional respondents expect women to do slightly more housework even if they have the same labor market hours as their partner (see the prediction in the left panel for "both same hours"). Although this effect has a large confidence interval and does not reach statistical significance, this may indicate that these respondents assume that female vignette persons with equal labor market hours nevertheless make relatively lower contributions to the household income. An additional analysis found support for this assumption (see Figure A3 in *Supplement, Part 3.1*): traditional respondents indicated that women's share of housework should be higher only in the absence of information on relative income contributions. The point estimate for the appropriate housework share in the low information condition was between that found for women contributing between 33 percent and 50 percent of the household income. This, again, would reflect very well the standard case in Germany, where the unadjusted gender pay gap was 21 percent lower hourly wages for women in 2017/2018 (German Federal Statistical Office 2019), implying that even in dual-earner couples with roughly equal working hours, women would make lower contributions to the household income.

In sum, our results suggest that low information questions primarily measure different interpretations: the main difference between traditional and egalitarian respondents appears to be not in their support for traditional gender ideologies, but in the implicit beliefs they hold in the absence of information. When provided with the same, standardized information, both “traditional” and “egalitarian” respondents appear to support an egalitarian division of housework, rather than the traditional division indicative of traditional ideologies. If this interpretation is correct, then the literature based on the standard item measures has overstated the prevalence of traditional gender ideologies.

Extended Analyses: Are Female Labor Market Resources Devaluated?

We also test specific variants of traditional gender ideologies that assume not only a positive main effect of gender, but also a different strength of this effect depending on who is the primary breadwinner. Some theories suggest that in constellations with a female breadwinner, this violation of the traditional gender ideology is neutralized by women contributing a much *higher* share of housework (see, e.g., theories on “deviance neutralization” or “gender display,” which, however, have found so far only little empirical support; for an overview: England 2011). Technically, this means that a traditional ideology would also be revealed by a nonlinear interaction effect: the effects of labor market resources vary by gender. Once women’s relative labor market hours and earnings exceed a 50 percent share, a positive rather than a negative association with women’s appropriate share of housework is expected (and the reverse pattern of association for men). Another rationale could be a general devaluation of women’s labor market resources (in the sense that they count relatively less for women to buy themselves out of housework), in which case one would always observe a positive interaction effect between a female gender and labor market resources.

The absence of such an interaction is also important in our case to identify the mediation we are assuming (see the *Methods and Data* section, in particular the notes to Figure 3). By adding the gender-neutral information M' , both a possible mediation M (mediation by implicit beliefs) and the moderation $X \cdot M$ (gender-specific effects of the assumed labor market resources) are replaced by gender-neutral effects ($M' + X \cdot M'$). Thus, our gender-neutral information M' could in principle have eliminated a mediation, a moderation, or a mixture of both. Only if there is no moderation (d in Figure 3) is the eliminated gender effect evidence for the hypothesized mediation. The best strategy for testing whether the assumption of no moderation is plausible is to evaluate it for many possible levels of M' : if we find no interaction of gender X with many different levels of labor market characteristics M' , we are more confident that respondents’ implicit beliefs M do not also interact with gender.

Figure 6 shows how the effects of the two labor market characteristics vary by gender: in the first row, the effects of being the male or female partner on the appropriate housework share are shown by that partner’s share of labor market hours; the second row shows the effects by that partner’s relative contribution

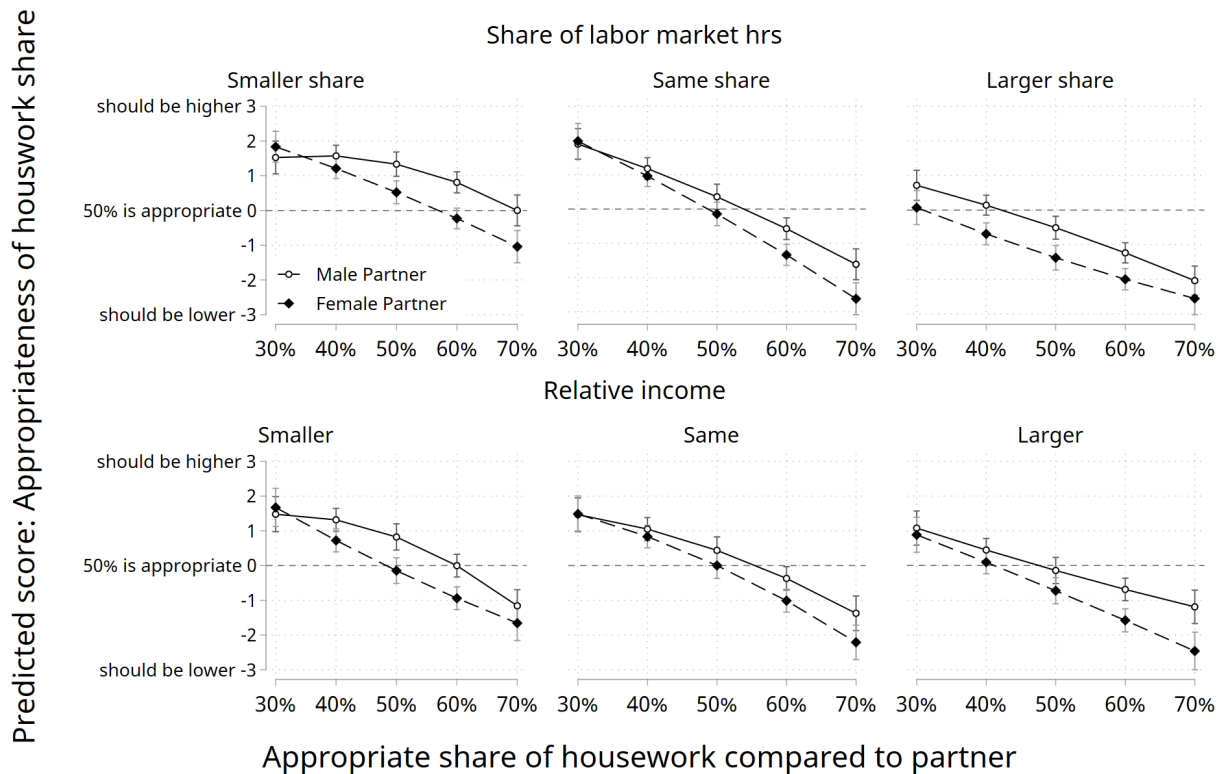


Figure 6: Appropriateness of the housework share by gender and relative labor market status. *Notes:* This figure shows the predicted appropriateness score of the male partner's vs. the female partner's housework share for different amounts of that partner's labor market hours (first row), and his/her contribution to the household income (second row). Predictions were estimated with separate regression models for labor market hours and relative income contribution. To show separate predictions for male and female partners, each regression model included interactions with the partner's gender. To adjust for possible nonlinear effects, squared terms of the housework hours were also included. Number of vignette ratings for labor market hours and for relative income: 1,278. Due to the small number of cases, we do not split the analyses by respondents' gender ideologies (as classified by the standard item questions).

to household income. The shares can be either smaller (first column), the same (second column), or larger (last column) than the respective partner's share. Only vignettes from the full information condition were included; men and women have on average the same characteristics on both labor market dimensions (hours and relative income contribution).

First, it can be seen that gender again has little effect on the ratings of appropriate shares of housework: in most constellations, the ratings for female and male vignette persons differ only slightly. In the few constellations with a statistically significant gender difference, it is always the *male* partner who is expected to do a slightly larger share of the housework. This is further evidence against traditional gender ideologies. Second, there is again broad support for equity norms: a housework share of around 50 percent is seen as appropriate in all constellations where both partners have roughly the same share of labor market hours or income (see the

second column of Figure 6). At the same time, partners with a smaller (larger) share of labor market hours or income contributions are expected to do more (less) of the housework. Third, the effects are throughout symmetric for men and women (compare the first and last columns of Figure 6). Thus, there is no evidence for gender-specific patterns in line with gender display or deviance neutralization theories. There is also no evidence of other nonlinear effects that would contradict the strong prevalence of egalitarian ideologies in the form of equity norms.

Although not the primary focus, our data also allow us to contribute to knowledge about what is considered an appropriate exchange rate between paid and unpaid contributions. In particular, time investments seem to matter: according to our estimates of cross-elasticities, having a one percentage point higher share of paid hours entitles a partner, on average, to a one percentage point lower share of housework (point estimate: -1.01 ; 95 percent CI: $-1.10, -0.92$).⁷ A one percent higher share of income contribution (recall that this was varied independently of hours worked and thus only indicates monetary and not time inputs) entitles, on average, to a 0.25 percentage point lower share of housework (point estimate: -0.25 ; 95 percent CI: $-0.34, -0.15$). Overall, and especially in the full information condition, women are entitled to do slightly less housework (two hours; see also the *Section on Main Results*). Given our design, which does not include information on labor market effort, we can only speculate whether respondents assumed higher effort on the part of women to achieve the same income contribution. Readers interested in more detail on these cross-elasticities (such as their variation by family background) or other measures of the relative “worth” of paid versus unpaid hours (which might also be quantified in terms of proportions of explained variance when rating couples’ total workload) are invited to use our replication materials (Auspurg and Düval 2024) to conduct their own analyses (see also Düval 2023 for reports on some additional findings).

Robustness Checks

For our research goal, it is particularly important that all covariates are balanced across male and female vignette partners and across the different information conditions. This was verified with randomization checks (see *Supplement, Part 1*).

Second, we tested whether information about the hypothetical couples’ marital status or childcare, which was included in some vignettes (see the *Methods and Data* section), would change the results. Our findings are again consistent throughout with equity norms, but not traditional gender ideologies: respondents agree that the appropriate share of housework should be greater if one partner does less childcare. Interestingly, this effect does not vary by gender either (see *Supplement, Part 3.1*). We also tested the robustness of our main findings by focusing only on married couples or couples with small children, in which a stronger activation of traditional ideologies might be expected (Grunow et al. 2012; Gupta 1999; Kühhirt 2011). Splitting the analyses by the family status of the hypothetical couples results in smaller numbers of cases and thus larger confidence intervals for the effects, but the main findings are still the same.

Third, in a split sample with other respondents, ratings of the appropriate share of the *total* workload (i.e., housework, paid employment, and childcare if applicable) were collected: is there evidence of ideologies that women should work longer shifts, and is this still (or less) true when they have the same labor market characteristics as men? Overall, the results are similar: doing more of the housework allows both men and women to reduce their total workload. Adding information on relative income contributions shifts the appropriate share of total workload for women to lower levels (especially so for “traditional” respondents). Similarly, the finding that there is no gendered valuation of labor market resources (technically: no interaction) is robust (see Figure A7 and A8 in *Supplement, Part 3.2*).

Finally, we also tested the robustness of the results with another item question (“Women should care more about their family than their career”) and an additive index based on three items measuring traditional gender ideologies (which allows for a larger group of respondents to be classified as “traditional,” see Figures A9–A12 in *Supplement, Part 3.3*). We also tested the robustness with other statistical models, when splitting the analyses for male and female respondents (to check for treatment heterogeneity), and when excluding the fastest respondents (*Supplement, Part 3.4*). None of these analyses changed our main conclusions.

Summary

Our combination of an experiment and causal mediation analysis allowed us to decompose the effect of gender on appropriate housework shares into two distinct mechanisms: support for female housework due to traditional gender ideologies, or support due to gendered beliefs about labor market inequalities combined with egalitarian gender ideologies. Our main findings, based on a random sample of German residents aged 24 to 47, can be summarized in three points:

1. In low information conditions, some respondents based their evaluations on stereotypical beliefs about men’s and women’s labor market resources. When this information was overridden by gender-neutral information, “traditional” individuals turned out to actually hold egalitarian ideologies. This confirms our main hypothesis that low information measures conflate ideologies (injunctive norms) with stereotypical beliefs (descriptive norms) and thereby overestimate the prevalence of traditional gender ideologies.
2. We found strong support for egalitarian ideologies in form of equity norms. When men’s and women’s labor market hours and contributions to household income were artificially equalized in our experiment, women were entitled to an even slightly smaller share of the housework. The appropriate share of housework depended strongly on time availability, and, to a lesser extent, on monetary contributions.
3. Our nuanced measurement with a multifactorial design revealed that, according to respondents’ ideologies, the exchange rates between labor market hours/income and housework should be roughly equal for both genders. Thus, consistent with previous research, there was no support for “deviance neutralization” or “gender display” theories.

Discussion

What else, if not traditional gender ideologies, could explain the gender inequalities in housework? It is worth noting that in our experiment, men's and women's labor market resources were artificially equalized. In the real world, (unobserved) differences in endowments could provide strong explanations. For example, in our respondent sample, women had on average about two-thirds of the net income, and about three-fourths of the labor market hours compared to men. Equity norms alone might explain the unequal share of housework hours in our respondent sample (wherein two-thirds of respondents reported that the women in their partnership did most of the housework). Based on the exchange rates revealed by our experiment, this higher female housework share would be normatively appropriate even if respondents only endorsed egalitarian ideologies (equity norms) and not traditional gender ideologies. Moreover, it is important to remember that gender is a social construct that is not limited to gender ideologies. Doing (or avoiding) housework may simply help individuals to signal their gender identity (Bittman et al. 2003; Thébaud, Kornrich, and Ruppanner 2021). However, these self-definitions are conceptually distinct from whether people subscribe to traditional gender ideologies or injunctive gender norms (Kroska 2000; Wood and Eagly 2015). Particularly in a liberal context with high levels of gender equality in the political, economic, and educational spheres, individuals may use gendered practices in other domains to affirm their gender identity (see the findings on the "gender equality paradox;" e.g., Stoet and Geary 2018).

Consistent with previous experimental studies (Auspurg et al. 2017; Jacobs and Gerson 2016; Pedulla and Thébaud 2015), we found that once important gender inequalities in labor market options were eliminated in hypothetical scenarios, individuals appear to prefer a largely gender-neutral division of unpaid labor. We extended this literature by showing that in such constellations, equal division of housework is not only the preferred option, but is also seen as a normative ideal, that is, individuals agree that the division of labor *should* then be equal. However, there is also a caveat to interpreting this as a consistently equal exchange (we are grateful to the reviewers for pointing this out): we only informed respondents about relative labor market hours and income contributions. Given the lower hourly wage for typical female jobs and the lower hourly wage even for women doing the same job as men (see, e.g., Schmitt and Auspurg 2022 for evidence on the German labor market), women typically have to work harder to achieve equal pay. If individuals see paid hours and the resulting income contributions as the main "currency" in exchange for housework, the seemingly "equal" exchange of paid and unpaid resources still requires more work effort on the part of most women.

In any case, our results suggest that preexisting structural inequalities in the labor market rather than gender ideologies are responsible for inequalities in the division of unpaid work. This suggests that, rather than trying to change ideologies, it would be much more effective to reduce labor market inequalities, for example, by removing part-time penalties (which seem to be particularly high for men; see Schmitt and Auspurg 2022) and investing in other measures that help to equalize labor force participation (see, e.g., Goldin 2014) or earnings for both genders. In

terms of practical implications, we also want to reiterate that we found no evidence for “deviance neutralization” or “gender display” theories. In contrast to previous research, we were able to rely on the strong internal validity of an experiment. We therefore agree with Paula England (2011) that research efforts might now be better spent on understanding other, more important levers for achieving gender equality.

What can be learned for measuring of gender ideologies? The main recommendation is to provide respondents with more information. When important contextual conditions are not specified in vague questions, (subgroup) differences found in the responses can mean both: different ideologies/attitudes, but also different interpretations (Braun 1998; Braun and Scott 2009). Our experiment with many different splits was not designed to measure individual scores on traditional gender norms. But one could also include more information in item questions. One has to consider the trade-off between question length and possible respondent exhaustion and satisficing (Dafoe et al. 2018). However, in our case, even a small amount of information on labor market characteristics significantly reduced the discrepancies in interpretation between respondent groups. Pilot studies can help strike the right balance between asking questions that are easy to answer but at the same time not too vague.

Limitations and Future Research

The internal validity of our results may be threatened by a social desirability bias. Methodological research suggests that respondents in factorial surveys show fewer socially desirable responses (Auspurg and Hinz 2015). In our experiment, we implemented a between-respondent variation in gender that “by design” hides sensitive topics (Walzenbach 2019). In addition, the vignettes were administered in a self-completion mode, which is known to minimize social desirability bias (Groves et al. 2004). Moreover, it can be argued that it is not at all plausible that a social desirability bias could have distorted the measurement of widely shared gender ideologies. Ideologies with such a low desirability that they cannot even be expressed in an anonymous interview situation certainly do not represent the common view. All of this adds to the confidence that our results do indeed capture widely shared ideologies.

It is worth noting, however, that our design only allowed us to examine the appropriate division of housework in different scenarios, but not gender ideologies or norms about the appropriate specialization in “male” versus “female” tasks (Schneider 2012). In an additional split, we asked respondents to rate the overall division of labor, not just housework. Again, no gender effects were found (see our *Robustness Checks*). In both cases, however, we asked respondents for their opinion on a division of labor that the described couples had already chosen (for whatever reason). Traditional gender ideologies may in particular guide who should generally specialize in which task. Future work could explore this particular aspect of gender ideologies.

We also did not examine ideologies about appropriate parenting (although we could test whether traditional gender ideologies apply specifically to parents, which was not the case). Typical items on these ideologies include, for example,

the question of whether a “preschool child is likely to suffer if the mother has a job instead of just focusing on the home” (Walter 2018). We assume that the problems of missing information also apply to these items: answering them requires assumptions about childcare during the mother’s working hours (e.g., are there childcare facilities available, could the father or grandparents also provide care). The assessment of these and other normative items may also depend on assumptions about whether two (full) earnings are needed for the family to be self-sustaining. Future research could explore these ideologies and implicit beliefs.⁸

The external validity of our results may be limited by the fact that only younger cohorts are included in the *pairfam* survey (the age range was 24 to 47 years, with a mean age of 35). This homogeneous age sample limited the variance in gender ideologies. Further studies could explore whether the age–attitude relationship found in previous research (older generations holding more traditional ideologies, see Perales, Lersch, and Baxter 2019) still holds when more informative survey questions are used.

It would also be helpful to replicate this experiment in other national contexts (such as more liberal welfare states like the United States). We could show a substantial effect of more information in the context of a traditional welfare state with strong gender inequalities (which is probably the most likely case for the hypothesized overestimation of traditional gender norms). We were also able to show that respondents’ implicit beliefs are consistent with actual gender inequalities in Germany. However, we could not test whether more standardized information would also reduce cross-country differences in gender ideologies. Cross-country research would be important to draw more firm conclusions about the sources of variable gender inequalities (and how to reduce them through policy measures).

Finally, we hope that the experimental design presented in this research will inspire further studies that go beyond the specific application of measuring gender ideologies. Standard experiments are helpful in identifying treatment effects (such as gender or ethnic identity), but not in providing insights into why they occur (i.e., the mediators). Simply controlling for respondents’ beliefs in multivariable analyses (“control by model”) introduces confounder bias (Montgomery et al. 2018). Varying the amount of information as an experimental factor instead (i.e., implementing a “control by design”) introduces the potential for mediation analyses that can rely on the high internal validity of experimental designs. In our view, this design has great potential for studying many other social mechanisms involving stereotypes and even unconscious beliefs (e.g., in discrimination research, or research on trust in cooperation partners, research on investment in education based on beliefs about their payoffs, or research on the acceptance of policies based on different beliefs about their effectiveness). Implicit stereotypes and beliefs are often considered important mediators, and experimental designs can help to identify them as explicit factors.

Notes

- ¹ Equality (everything is shared equally, i.e., 50 percent done by both partners) is a special case here. Some definitions of nontraditional ideologies also include the idea that there

should be no prescriptions at all: couples should decide only on their own idiosyncratic preferences (Braun 2008).

- 2 Ultimately, it is then an empirical question as to what exactly the nontraditional division of labor should look like: different appropriate exchange ratios between the various inputs to the partnership come into question here, and our research design will allow us to reveal some of them.
- 3 The other respondents participated in another experiment in which they were asked to rate the distribution of the total workload (paid plus unpaid work). We use these ratings for robustness checks.
- 4 Our hypotheses were not preregistered. Our data collection took place in 2017/2018, and the preparatory work was done in 2016/2017, when preregistration was less common in sociology. (The only kind of preregistration we have is that we presented these hypotheses in presentations invited by the *pairfam* board, which decides on proposals for questionnaire modules.) But we can at least provide some arguments that our hypotheses are not post hoc, because not only the effect on the specific subgroup of traditional respondents but also the overall effect reaches statistical significance. In addition, we provide a variety of robustness analyses.
- 5 We are interested in the total moderation of the effect of more information by gender ideologies because we are only interested in the overall level of possible misclassification of the different groups (we primarily expect “traditional” respondents to be misclassified, i.e., to react to more information). Therefore, we do not control for other possible differences between respondents.
- 6 The reduction in the gender effect due to the addition of information is much stronger for the “traditional” respondents (about one scale point), whereas the gender effect is reduced by only about a quarter of a scale point for the “neither nor” and “egalitarian” respondents. The difference between the groups is statistically significant at the 10 percent level when tested in a joint regression with interaction effects. There is probably not enough power (small number of “traditional” respondents) to reach a higher level of statistical significance.
- 7 Measured in hours, each hour of additional paid work hour entitles on average to 0.28 hours less housework (95 percent CI: $-0.24, -0.31$).
- 8 Moreover, experiments such as ours could be used to measure norms and ideologies in general and also related concepts such as gender identities (Akerlof and Kranton 2000; Kroska 2000). According to recent literature, gender ideologies should be understood as a multidimensional concept, in which different roles (e.g., as breadwinner, but also as parent) are combined into complex profiles that no longer fit into a one-dimensional continuum from traditional to egalitarian roles (Grunow et al. 2018; Perales et al. 2019). Multifactorial survey experiments appear to be particularly well suited to capturing such multifaceted concepts.

References

- Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. “Analyzing Causal Mechanisms in Survey Experiments.” *Political Analysis* 26(4):357–78. <https://doi.org/10.1017/pan.2018.19>
- Akerlof, George A. and Rachel E. Kranton. 2000. “Economics and Identity.” *The Quarterly Journal of Economics* 115(3):715–53. <https://doi.org/10.1162/003355300554881>

- 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. <https://osf.io/3fqw9>.
- Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks, CA: Sage. <https://doi.org/10.4135/9781483398075>
- Auspurg, Katrin, Maria Iacovou, and Cheti Nicoletti. 2017. "Housework Share between Partners: Experimental Evidence on Gender-Specific Preferences." *Social Science Research* 66:118–39. <https://doi.org/10.1016/j.ssresearch.2017.01.003>
- Bartley, Sharon J., Priscilla W. Blanton, and Jennifer L. Gilliard. 2005. "Husbands and Wives in Dual-Earner Marriages: Decision-Making, Gender Role Attitudes, Division of Household Labor, and Equity." *Marriage & Family Review* 37(4):69–94. https://doi.org/10.1300/J002v37n04_05
- 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. <https://doi.org/10.1111/j.1741-3737.2008.00479.x>
- Bicchieri, Cristina. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780190622046.001.0001>
- Bittman, Michael, Paula England, Liana Sayer, Nancy Folbre, and George Matheson. 2003. "When Does Gender Trump Money? Bargaining and Time in Household Work." *American Journal of Sociology* 109(1):186–214. <https://doi.org/10.1086/378341>
- Braun, Michael. 1998. "Gender Roles." Pp. 111–34 in *Comparative Politics. The Problem of Equivalence*, edited by J. W. van Deth. New York, NY: Routledge.
- Braun, Michael. 2008. "Using Egalitarian Items to Measure Men's and Women's Family Roles." *Sex Roles* 59(9):644–56. <https://doi.org/10.1007/s11199-008-9468-5>
- Braun, Michael and Jacqueline Scott. 2009. "Gender-Role Egalitarianism – Is the Trend Reversal Real?" *International Journal of Public Opinion Research* 21(3):362–67. <https://doi.org/10.1093/ijpor/edp032>
- Brüderl, Josef, Sonja Drobnič, Karsten Hank, Bernhard Nauck, Franz J. Neyer, Sabine Walper, Philipp Alt, Christiane Bozoyan, Petra Buhr, Christine Finn, Madison Garrett, Henriette Greischel, Nicolai Gröpler, Kristin Hajek, Michel Herzig, Bernadette Huyer-May, Rüdiger Lenke, Lara Minkus, Bettina Müller, Timo Peter, Claudia Schmiedeberg, Philipp Schütze, Nina Schumann, Carolin Thönnissen, Martin Wetzels, and Barbara Wilhelm. 2019. "The German Family Panel (Pairfam)." *GESIS Data Archive, Cologne*. ZA5678. *Data file Version 10.0.0*, doi: 10.4232/pairfam.5678.10.0.0.
- Cialdini, Robert B., Carl A. Kallgren, and Raymond R. Reno. 1991. "A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior." *Advances in Experimental Social Psychology* 24:201–34. [https://doi.org/10.1016/S0065-2601\(08\)60330-5](https://doi.org/10.1016/S0065-2601(08)60330-5)
- Coltrane, Scott. 2000. "Research on Household Labor: Modeling and Measuring the Social Embeddedness of Routine Family Work." *Journal of Marriage and Family* 62(4):1208–33. <https://doi.org/10.1111/j.1741-3737.2000.01208.x>
- Correll, Shelley J. and Stephen Benard. 2006. "Biased Estimators? Comparing Status and Statistical Theories of Gender Discrimination." *Advances in Group Processes* 23:89–116. [https://doi.org/10.1016/S0882-6145\(06\)23004-2](https://doi.org/10.1016/S0882-6145(06)23004-2)

- Correll, Shelley J., Stephen Benard, and In Paik. 2007. "Getting a Job: Is There a Motherhood Penalty?". *American Journal of Sociology* 112(5):1297–338. <https://doi.org/10.1086/511799>
- Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(4):399–416. <https://doi.org/10.1017/pan.2018.9>
- Davis, Shannon N. and Theodore N. Greenstein. 2009. "Gender Ideology: Components, Predictors, and Consequences." *Annual Review of Sociology* 35:87–105. <https://doi.org/10.1146/annurev-soc-070308-115920>
- Davis, Shannon N. and Jeremiah B. Wills. 2014. "Theoretical Explanations Amid Social Change: A Content Analysis of Housework Research (1975-2012)." *Journal of Family Issues* 35(6):808–24. <https://doi.org/10.1177/0192513X13513020>
- Düval, Sabine. 2023. "Why Should Women Take Care of the Housework? Experimental Evidence Concerning Gender-Specific Norms for the Division of Paid and Unpaid Work among Couples." PhD thesis, Department of Sociology, LMU Munich.
- England, Paula. 2011. "Missing the Big Picture and Making Much Ado about Almost Nothing: Recent Scholarship on Gender and Household Work." *Journal of Family Theory & Review* 3(1):23–26. <https://doi.org/10.1111/j.1756-2589.2010.00077.x>
- Eurostat. 2022. "Gender Pension Gap by Age Group - Eu-Silc Survey", https://www.destatis.de/EN/Press/2023/03/PE23_N015_12_63.html (last access: September 16, 2023).
- Gager, Constance T. 1998. "The Role of Valued Outcomes, Justifications, and Comparison Referents in Perceptions of Fairness among Dual-Earner Couples." *Journal of Family Issues* 19(5):622–48. <https://doi.org/10.1177/019251398019005007>
- Gager, Constance T. 2008. "What's Fair Is Fair? Role of Justice in Family Labor Allocation Decisions." *Marriage & Family Review* 44(4):511–45. <https://doi.org/10.1080/01494920802454116>
- German Federal Statistical Office. 2015. "Zeitverwendungserhebung. Aktivitäten in Stunden und Minuten für Ausgewählte Personengruppen. 2012/2013." <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Zeitverwendung/Publikationen/Downloads-Zeitverwendung/zeitverwendung-5639102139004.html> (Last Access: February 19, 2024).
- German Federal Statistical Office. 2019. "Gender Pay Gap Remained Unchanged at 21% in 2018." Press Release No. 098. https://www.destatis.de/EN/Press/2019/03/PE19_098_621.html (Last Access: February 19, 2024).
- Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104(4):1091–119. <https://doi.org/10.1257/aer.104.4.1091>
- Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*, edited by Robert M. Groves. Hoboken, NJ: J. Wiley.
- Grunow, Daniela, Katia Begall, and Sandra Buchler. 2018. "Gender Ideologies in Europe: A Multidimensional Framework." *Journal of Marriage and Family* 80(1):42–60. <https://doi.org/10.1111/jomf.12453>
- Grunow, Daniela, Florian Schulz, and Hans-Peter Blossfeld. 2012. "What Determines Change in the Division of Housework over the Course of Marriage?" *International Sociology* 27(3):289–307. <https://doi.org/10.1177/0268580911423056>

- Gupta, Sanjiv. 1999. "The Effects of Transitions in Marital Status on Men's Performance of Housework." *Journal of Marriage and Family* 61(3):700–11. <https://doi.org/10.2307/353571>
- Hochschild, Arlie Russell and Anne Machung. 1989. *The Second Shift*. New York, NY: Viking.
- Hook, Jennifer L. 2006. "Care in Context: Men's Unpaid Work in 20 Countries, 1965–2003." *American Sociological Review* 71(4):639–60. <https://doi.org/10.1177/000312240607100406>
- Hox, Joop J., Ita G. G. Kreft, and Piet L. J. Hermkens. 1991. "The Analysis of Factorial Surveys." *Sociological Methods & Research* 19(4):493–510. <https://doi.org/10.1177/0049124191019004003>
- Huinink, Johannes, Josef Brüderl, Bernhard Nauck, Sabine Walper, Laura Castiglioni, and Michael Feldhaus. 2011. "Panel Analysis of Intimate Relationships and Family Dynamics (Pairfam): Conceptual Framework and Design." *Zeitschrift für Familienforschung* 23(1):77–101. <https://doi.org/10.20377/jfr-235>
- Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society. Series A* 76(1):5–32. <https://doi.org/10.1111/j.1467-985X.2012.01032.x>
- 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. <https://doi.org/10.1177/0891243215597445>
- Killewald, Alexandra and Margaret Gough. 2010. "Money Isn't Everything: Wives' Earnings and Housework Time." *Social Science Research* 39(6):987–1003. <https://doi.org/10.1016/j.ssresearch.2010.08.005>
- Kroska, Amy. 2000. "Conceptualizing and Measuring Gender Ideology as an Identity." *Gender & Society* 14(3):368–94. <https://doi.org/10.1177/089124300014003002>
- Kühhirt, Michael. 2011. "Childbirth and the Long-Term Division of Labour within Couples: How Do Substitution, Bargaining Power, and Norms Affect Parents' Time Allocation in West Germany?" *European Sociological Review* 28(5):565–82. <https://doi.org/10.1093/esr/jcr026>
- Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62(3):760–75. <https://doi.org/10.1111/ajps.12357>
- Nitsche, Natalie and Daniela Grunow. 2016. "Housework over the Course of Relationships: Gender Ideology, Resources, and the Division of Housework from a Growth Curve Perspective." *Advances in Life Course Research* 29:80–94. <https://doi.org/10.1016/j.alcr.2016.02.001>
- Nordenmark, Mikael. 2004. "Does Gender Ideology Explain Differences between Countries Regarding the Involvement of Women and of Men in Paid and Unpaid Work?" *International Journal of Social Welfare* 13(3):233–43. <https://doi.org/10.1111/j.1369-6866.2004.00317.x>
- OECD. 2017. *The Pursuit of Gender Equality. An Uphill Battle*. Paris, France: OECD Publishing.
- 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. <https://doi.org/10.1177/0003122414564008>
- Perales, Francisco, Philipp M. Lersch, and Janeen Baxter. 2019. "Birth Cohort, Ageing and Gender Ideology: Lessons from British Panel Data." *Social Science Research* 79:85–100. <https://doi.org/10.1016/j.ssresearch.2018.11.003>

- Ridgeway, Cecilia L. 2011. *Framed by Gender. How Gender Inequality Persists in the Modern World*. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199755776.001.0001>
- Schmitt, Laila and Katrin Auspurg. 2022. "A Stall Only on the Surface? Working Hours and the Persistence of the Gender Wage Gap in Western Germany 1985–2014." *European Sociological Review* 38(5):754–69. <https://doi.org/10.1093/esr/jcac001>
- Schneider, Daniel. 2012. "Gender Deviance and Household Work: The Role of Occupation." *American Journal of Sociology* 117(4):1029–72. <https://doi.org/10.1086/662649>
- Sterkens, Philippe, Stijn Baert, Claudia Rooman, and Eva Deros. 2022. "Why Making Promotion after a Burnout is Like Boiling the Ocean." *European Sociological Review* 39(4):516–31. <https://doi.org/10.1093/esr/jcac055>
- Stoet, Gijsbert and David C. Geary. 2018. "The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education." *Psychological Science* 29(4):581–93. <https://doi.org/10.1177/0956797617741719>
- Thébaud, Sarah, Sabino Kornrich, and Leah Ruppanner. 2021. "Good Housekeeping, Great Expectations: Gender and Housework Norms." *Sociological Methods & Research* 50(3):1186–214. <https://doi.org/10.1177/0049124119852395>
- Thompson, Linda. 1991. "Family Work: Women's Sense of Fairness." *Journal of Family Issues* 12(2):181–96. <https://doi.org/10.1177/019251391012002003>
- VanderWeele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford, UK: Oxford University Press.
- 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. <https://doi.org/10.1007/s11135-017-0491-x>
- Walzenbach, Sandra. 2019. "Hiding Sensitive Topics by Design?: An Experiment on the Reduction of Social Desirability Bias in Factorial Surveys." *Survey Research Methods* 13(1):103–121.
- Wood, Wendy and Alice H. Eagly. 2015. "Two Traditions of Research on Gender Identity." *Sex Roles* 73(11):461–73. <https://doi.org/10.1007/s11199-015-0480-2>

Acknowledgments: We thank the participants of the Conference of the European Survey Research Association (ESRA) in 2019, the Annual Meeting of the American Sociological Association (ASA) in 2019, the *pairfam* User Conference in 2019, and the seminar on "Analytical Sociology: Theory and Empirical Applications" at the Venice International University in 2018 for helpful suggestions. We are also grateful for comments on an earlier version we received from Josef Brüderl. We used data from the German Family Panel *pairfam*, coordinated by Josef Brüderl, Sonja Drobniè, Karsten Hank, Johannes Huinink, Bernhard Nauck, Franz J. Neyer, and Sabine Walper. From 2004 to 2022, *pairfam* was funded as priority program and long-term project by the German Research Foundation (DFG). Sabine Düval worked on the manuscript and data analysis mainly during her PhD studies at the LMU Munich. Part of this work was done while Katrin Auspurg was a Visiting Fellow at the European University Institute (EUI) in Florence.

Katrin Auspurg: Department of Sociology, LMU Munich. E-mail: katrin.auspurg@lmu.de.

Sabine Düval: German Youth Institute (DJI). E-mail: dueval@dji.de.