ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 447

Reverse Gender Bias in Performance Ratings

Simon Dato Svenja Hippel Asri Özgümü

November 2026 www.econtribute.de

UNIVERSITÄT BONN UNIVERSITÄT ZU KÖLN

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866 – is gratefully acknowledged.

Reverse Gender Bias in Performance Ratings*

Simon Dato† Svenja Hippel‡ Asri Özgümüs§

November 13, 2026

In a pre-registered hiring experiment ($N = 2{,}680$) embedded in a real tutor-hiring process, we investigate how student evaluators rate two fixed teaching samples submitted by two tutor applicants, one female- and one male-authored. We randomize whether the tutor applicants' gender pronouns are visible, whether ratings determine a real hire, and whether student evaluators are rewarded for matching the ratings of a gender-blind expert committee. Because only the pronoun is edited, differences in the rating gap between the two samples across pronoun conditions identify the effect of observable gender without requiring equally good samples. Contrary to the usual expectation of bias against women, making the two applicants' gender observable induces a reverse bias: it shifts the female-minus-male rating gap toward the female applicant and lowers the probability that evaluators choose the male applicant in a hypothetical hiring decision. Neither removing hiring consequences for the applicant nor rewarding student evaluators for matching expert ratings significantly changes this bias. The reversal is not a uniform shift toward the female applicant. It is concentrated among evaluators who state that women explain Microeconomics better, whereas evaluators who see no gender difference show no significant shift. Stated beliefs predict the rating gap only when applicant gender is visible.

JEL classification: D90, J71.

Keywords: gender, performance rating, hiring decision, experiment.

*We would like to thank Dan Banko, Urs Fischbacher, Simon Gächter, Sebastian Goerg, Henning Hermes, Helmut Herwartz, Michael Kosfeld, Pauline Madiès, Daniel Müller, Muriel Niederle, Petra Nieken, Fabio Römeis, Antonio Rosato, Johannes Schrön, Philipp Struthmann, Stefan T. Trautmann, as well as the members of LakeLab Uni Konstanz, participants at ESA North America Tucson 2025, EBS Economics Group Research Workshop 2026, and ESA Europe Barcelona 2026 for helpful comments and suggestions. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2126/2 - 390838866. The experiment was pre-registered on OSF (pre-registration). Ethical approval was obtained from the Ethics committee of EBS University. The authors are listed in alphabetical order. All authors contributed equally to the manuscript.

†EBS Universität für Wirtschaft und Recht, EBS Business School, Rheingaustr. 1, D-65375 Oestrich-Winkel, E-mail address: simon.dato@ebs.edu.

‡University of Bonn, Center for Advanced Studies in Law and Economics, Adenauerallee 24-42, D-53113, Bonn, E-mail address: shippel@uni-bonn.de.

§University of Würzburg, Chair for Contract Theory and Information Economics, Sanderring 2, D-97070, Würzburg, E-mail address: asri.oezguemues@uni-wuerzburg.de.

1. Introduction

Performance evaluations play a central role in labor markets. Decisions about hiring, promotions, and compensation often depend on subjective assessments of individual performance. Because these evaluations frequently determine important economic outcomes, understanding whether and when they are systematically biased is crucial for ensuring fair and efficient personnel decisions.

A large literature documents gender differences in subjective evaluations, including in hiring and educational settings (e.g., Moss-Racusin et al., 2012; MacNell et al., 2015; Boring et al., 2016; Boring, 2017; Mengel et al., 2019; Özgümüs et al., 2020; Coffman et al., 2021; Barron et al., 2025; Bohren et al., 2025; Hochleitner et al., 2025). A recurring finding in this literature is discrimination against women. Some more recent contributions, however, suggest a less uniform picture: gender effects vary by setting, evaluator characteristics, and institutional environment, and in some contexts evaluations do not disadvantage women or may even favor them (e.g., Bohnet et al., 2016; Card et al., 2020; Binderkrantz et al., 2022; Binderkrantz and Bisgaard, 2024; Andersson et al., 2025; Johannesen and Muchardt, 2026). These findings shift attention from whether gender bias exists to understanding the direction of gender bias and the forces that shape it. When does making gender observable advantage rather than disadvantage female applicants, and what explains the direction of this effect?

We study these questions in a pre-registered hiring experiment embedded in an actual hiring process for an undergraduate teaching assistant in Microeconomics at a German university. Students evaluate teaching samples submitted by two real applicants, one female and one male. The experiment independently varies the observability of applicant gender and the institutional environment in which evaluations are made. To identify the effect of gender observability, we randomly display or hide gendered pronouns in the teaching materials while holding the samples themselves fixed. Comparing the female-minus-male rating difference across these pronoun conditions identifies how making gender observable changes the relative evaluation of the two samples, without requiring them to be of equal quality. Because all evaluators rate the same two teaching samples, the experiment identifies the effect of making gender observable for this particular pair of applicants and this population of German university students, not an average effect across applicant pairs or evaluator populations. Across institutional treatments, we vary whether evaluations determine the hiring decision and whether evaluators are financially incentivized to align their ratings with those of an independent, gender-balanced expert committee that is blind to applicants' gender. We also ask evaluators to make a direct, hypothetical hiring decision by giving them the option to choose the female applicant, choose the male applicant, or leave the hiring decision to a random draw. Finally, we elicit their beliefs about whether women or men are, in general, better at explaining concepts in Microeconomics tutorials.

Our main result is that making our two applicants' gender observable significantly shifts evaluations toward the female applicant. When gender is hidden, the female-minus-male rating gap weakly favors the male applicant. Revealing gender significantly shifts this gap such that

evaluations favor the female applicant when gender is observable. This shift is driven primarily by higher ratings for the female applicant, accompanied by a smaller decline in ratings for the male applicant. Thus, rather than disadvantaging the female applicant, gender observability generates a reverse gender bias in her favor. This favorable rating is provided by female as well as male evaluators: both shift their evaluations toward the female applicant by a similar amount. Moreover, the interaction between pronoun visibility and rater gender is insignificant, providing no evidence for female in-group favoritism.

We next ask whether the effect of making gender observable varies with the institutional environment in which evaluations are made. We vary (i) whether evaluations contribute to an actual hiring decision and (ii) whether evaluators are rewarded for matching the assessments of an independent, gender-balanced expert committee that is blind to applicants' gender. These institutional characteristics could moderate gender-dependent evaluations by changing the costs of acting on gender-related preferences or the weight evaluators place on gender-related beliefs about performance (e.g., Becker, 1957; Phelps, 1972; Arrow, 1973; Bartoš et al., 2016; Coffman et al., 2021; Stroube, 2022; Gioia and Immordino, 2024). We find, however, that the shift toward the female applicant is directionally consistent across all three institutional environments: making gender observable reverses the sign of the female-minus-male evaluation gap when ratings have no hiring consequences, when they contribute to an actual hiring decision, and when hiring consequences are combined with incentives to match gender-blind expert assessments. We find no significant evidence that hiring consequences or incentives to match expert assessments moderate this effect. Thus, we interpret these findings as evidence for the directional robustness of the reverse gender bias across the institutional environments we study, not as evidence that consequences or incentives have no effect in general.

The reversal extends to the direct hypothetical hiring choice. Making gender observable reduces the probability of selecting the male applicant by 6.9 percentage points, a decline of about 17 percent relative to the gender-hidden condition. Thus, the effect of gender observability carries over to the type of personnel decision that these evaluations are intended to inform. We also examine what shapes the response to gender observability and find an important role for evaluators' gender-related beliefs. The reversal is concentrated among the roughly one quarter of belief-stating evaluators who believe that women explain Microeconomics better: in this group, making gender observable strongly raises the female-minus-male rating gap and the probability of choosing the female applicant in the hypothetical hiring decision. Evaluators who see no gender difference, by contrast, show a small and statistically insignificant shift. Moreover, stated beliefs do not predict the female-minus-male rating gap when gender is hidden, but strongly predict it when gender is observable. This pattern is consistent with gender information allowing evaluators to act on beliefs they already hold.

Our findings contribute to the literature on gender bias in subjective evaluations in three ways. First, we provide causal evidence that making a female applicant's gender observable can improve rather than worsen her evaluation in an actual hiring process, illustrating that gender bias need not have a fixed direction. Second, we show that the effect of gender observability varies

systematically with evaluators' beliefs about gender differences in task-specific ability and is concentrated among evaluators whose beliefs favor women, suggesting that such beliefs may help explain heterogeneity in the direction of gender bias. Third, we show that reverse gender bias is robust to institutional changes. Taken together with the broader evidence, our findings caution against equating gender bias in subjective evaluations with discrimination against women. For personnel selection, such bias can matter regardless of its direction: it can alter hiring decisions even when observed performance is held fixed, particularly when the candidates' performance is similar.

The remainder of the paper proceeds as follows. Section 2 discusses the related literature. Section 3 describes the experimental design and derives testable hypotheses. Section 4 presents the main results and additional analyses. Section 5 discusses the findings and concludes.

2. Literature Review

Our paper relates to three strands of the literature. First, it connects to the literature on gender bias in teaching evaluations and hiring decisions. A large body of research examines whether evaluations are systematically affected by gender. Experimental and quasi-experimental studies document that female instructors often receive lower student evaluations despite similar teaching performance or learning outcomes in higher education (e.g., MacNell et al., 2015; Boring et al., 2016; Boring, 2017; Mengel et al., 2019; Özgümüs et al., 2020). Likewise, studies of hiring decisions report disadvantages for equally qualified female applicants (e.g., Moss-Racusin et al., 2012; Coffman et al., 2021).

More recent evidence, however, paints a more nuanced picture. Some studies find no overall gender bias but substantial heterogeneity across evaluators, for example, with male students tending to evaluate male instructors more favorably and female students tending to evaluate female instructors more favorably (Binderkrantz and Bisgaard, 2024). Other studies find little or no evidence of gender bias in different social and institutional contexts. For example, Binderkrantz et al. (2022) and Andersson et al. (2025) experimentally vary the perceived gender of instructors in educational settings and find no evidence of a disadvantage for female instructors in Denmark and Sweden. Binderkrantz et al. (2022) suggest that the relatively egalitarian gender norms in Scandinavian countries may be one possible explanation, as such norms could weaken traditional gender stereotypes about instructors. While they do not directly test this explanation, it illustrates one way in which the broader context may matter for gender-dependent evaluations. More broadly, these and related findings have contributed to a growing focus on when and how gender affects evaluations, rather than on the existence of gender discrimination per se. Beyond broader contextual factors, recent work highlights more specific mechanisms that may shape gender differences in evaluations, including differences in the information available to evaluators, such as through gendered recommendation letters, as well as inaccurate beliefs and stereotypes (Eberhardt et al., 2023; Hochleitner et al., 2025).

Against this background, our results provide causal evidence that making gender observable

does not necessarily disadvantage women. In our hiring setting for a Microeconomics teaching assistant, making gender observable leads to a more positive evaluation of the female applicant and reverses the female-minus-male evaluation gap. This finding contrasts with evidence of lower evaluations of female instructors but complements recent evidence that the direction and magnitude of gender bias can vary across contexts. By embedding the evaluation of standardized teaching materials in an actual hiring process, our experimental design combines causal identification of the effect of gender observability with a real-world personnel selection setting.

This design also addresses a central methodological challenge in the literature on gender bias in teaching evaluations: the trade-off between causal identification and external validity. In naturally occurring teaching evaluations, teaching performance cannot be held constant across instructors, making it difficult to isolate the causal effect of gender. Experimental studies overcome this problem by standardizing performance but often rely on hypothetical evaluations, deception, or laboratory environments in which participants may not perceive their decisions as consequential. We bridge this gap by conducting a hiring experiment embedded in a real tutor-hiring process, in which participants evaluate standardized teaching materials submitted by actual applicants for an actual undergraduate tutor position. Because we vary only whether applicant gender is observable while holding the two teaching samples fixed, the difference in the female-minus-male evaluation gap across conditions identifies the effect of gender observability. At the same time, our institutional treatments allow us to examine this effect when evaluations contribute to actual hiring decisions and when evaluators are incentivized to match gender-blind expert assessments. Our design therefore combines the internal validity of a controlled experiment with a real-world personnel selection setting and provides a framework for studying gender bias when performance is otherwise difficult to standardize.

Second, our paper relates to the literature on how features of the decision environment shape discriminatory behavior. A growing body of research examines whether discrimination responds to the stakes and institutional conditions under which decisions are made. For example, Bartoš et al. (2025) show that unconscious racial bias against Black tutors among politically conservative learners reduces learning and performance, but disappears when learners face higher monetary incentives for their own performance. Closest to our setting, Gioia and Immordino (2024) find that employers exhibit a gender-specific preference for female workers, which is significantly reduced and no longer statistically detectable in intentional hiring decisions when employers face stronger monetary incentives to select productive workers.[1] Previous studies further show that discriminatory behavior responds to the anonymity of candidates' identities (Goldin and Rouse, 2000), evaluation procedures (Bohnet et al., 2016), image concerns (Coffman et al., 2021), the visibility of discriminatory actions (Barron et al., 2025), and interventions designed to raise awareness of bias (Peterson et al., 2019; Genetin et al., 2022; Hoorens et al., 2021; Ayllón and

[1] A key difference between our experiment and that of Gioia and Immordino (2024) lies in the incentive structure. In their experiment, employers monetary payoffs directly depend on the productivity of the selected worker, creating a financial incentive to select more productive workers. In our setting, the hiring chair also has an interest in selecting a qualified tutor, but this incentive is not experimentally varied.

Zamora, 2025).

We contribute to this literature by studying whether making gender observable has a different effect when evaluators know that their assessments directly affect those they evaluate. This question connects to research showing that the purpose of an appraisal affects ratings (Jawahar and Williams, 1997). In contrast to Bartoš et al. (2025) and Gioia and Immordino (2024), where incentives increase the decision makers' own financial returns to better performance or more productive choices, respectively, our hiring consequences change what is at stake for the evaluated applicants. In teaching evaluations, evaluators may differ in whether and to what extent they perceive their assessments as consequential, making it difficult to determine how consequences shape gender-dependent evaluations. Our design fixes these consequences experimentally and makes them explicit by varying whether evaluations contribute to actual hiring decisions. In addition, we study whether evaluations respond to incentives to align with a gender-blind expert benchmark. Whereas hiring consequences increase the stakes of evaluations for applicants, benchmark incentives give evaluators a reason to align their assessments with an external standard of performance. Importantly, because the expert committee is blind to applicant gender, the benchmark provides a gender-neutral external assessment of performance rather than one that may itself reflect gender-dependent evaluations. We experimentally vary whether evaluators are rewarded for matching gender-blind expert assessments, allowing us to identify how hiring consequences and incentives to match expert assessments affect gender-dependent evaluations.

Our results show that neither of these features significantly moderates the effect of gender observability. We find no significant evidence that attaching consequences for the applicants to evaluations moderates the effect of gender observability, and incentives to match gender-blind expert assessments do not significantly moderate it either. Unlike Gioia and Immordino (2024), who find that stronger monetary incentives reduce the gender-specific preference for intentionally hiring female workers, we do not detect a significant attenuation of pro-female bias under incentives. However, whereas their incentives reward the selection of more productive workers, ours reward agreement with a gender-blind expert benchmark. The absence of significant moderation in our setting is therefore not inconsistent with their findings. The shift in favor of the female candidate is present in the same direction across all three institutional settings. We interpret these results as showing that the reversal persists across the institutional settings we examine, while recognizing that the absence of significant moderation does not imply that hiring consequences or incentives to match expert assessments have no effect on its magnitude.

Third, our paper relates to the literature on the mechanisms underlying discrimination. Classical economic theories distinguish between taste-based discrimination (Becker, 1957) and statistical discrimination (Arrow, 1973; Phelps, 1972). More recent work emphasizes inaccurate beliefs, stereotypes, and belief-based discrimination as potential sources of discriminatory behavior (Coffman et al., 2021; Fischbacher et al., 2024; Barron et al., 2025). Coffman et al. (2021) compare hiring decisions over workers labeled by gender with decisions over the same workers labeled by a non-stereotypical characteristic (birth month) and show that discrimination against women is largely driven by beliefs about average group differences rather than by

gender-specific considerations. Fischbacher et al. (2024) further demonstrate that productivity beliefs and gender stereotypes shape hiring evaluations, while Barron et al. (2025) show that belief-based discrimination depends on whether discriminatory actions are observable. Together, these studies highlight the role of evaluators' beliefs in shaping gender-dependent evaluations and how the influence of these beliefs depends on the decision environment.

We contribute to this emerging literature on belief-based discrimination by linking it to the literature on teaching evaluations. We elicit evaluators' beliefs about gender differences in teaching ability and examine whether these beliefs predict how evaluations respond to the observability of applicant gender. This allows us to assess whether the role of beliefs documented in hiring experiments extends to an educational setting involving real personnel decisions. We find that heterogeneity in the response to gender observability is systematically related to evaluators' beliefs about gender differences in teaching ability. These beliefs predict relative evaluations when applicant gender is observable, but not when it is hidden, and the shift toward the female applicant is concentrated among evaluators whose beliefs favor women.

3. Study Design

We first describe how we retrieved the teaching samples from the tutor applicants (3.1) and then present the design of the rating experiment (3.2). We next present our testable hypotheses (3.3) and experimental procedures (3.4). The full instructions for the rating experiment, including the teaching samples used in the experiment, are provided in Appendix D.

3.1. Teaching Samples

We collected teaching samples from applicants for an actual open position as a tutor for an introductory Microeconomics course at the University of Würzburg (Germany). The successful applicant would be hired to teach one or two tutorials per week during the following semester. In addition to standard application materials (CV, grade transcript), applicants were asked to submit a short, five-slide presentation on exercises similar to those covered in the tutorial. Applicants received a pre-structured presentation template containing the exercises and formal solutions but no explanations of the solution path. They were asked to add two to three sentences of explanation per slide within a predefined text box.

The presentation included a title slide with the name of the course. Applicants were also asked to fill in their full name and gender pronouns (he/him, she/her, they/them), both on the title slide and in the footer of each subsequent slide. For the experiment, applicants' names were always redacted, while the visibility of their gender pronouns was experimentally manipulated.

From the applicant pool we selected two teaching samples, one from a female and one from a male applicant, whose explanations were comparable in length and content to minimize differences between the materials. Because this was an actual teaching position, we restricted the selection to candidates whom the chair responsible for the tutorial considered acceptable to hire. This selection constitutes a researcher degree of freedom: we selected one female and

one male sample from a larger set, and our results may not generalize to every combination of female and male applicants. Importantly, identification does not require the two samples to be of equal quality because differencing across pronoun conditions removes any fixed quality difference between them.

3.2. Rating Experiment

In the rating experiment, we presented both applicants' teaching samples to student raters in randomized order. They rated the quality of each slide set separately. Participants were enrolled students from various German universities.[2] Regarding the quality of the explanations, i.e., the part of the slides completed by the applicants themselves, we elicited two measures: (1) a composite score, defined as the average of responses to five detailed evaluation statements, and (2) an overall assessment. All items were rated on a Likert scale from 1 (do not agree at all) to 10 (fully agree). For the composite score, evaluators rated the following five statements:

- The explanations of the solutions are easy to understand.
- The explanations are clearly structured.
- The explanations make the mathematical solution intuitively understandable.
- The explanations are precise.
- The slides are suitable for independent study (without attending the tutorial).

Additionally measuring the overall assessment allows us to compare raters' overall assessment with the composite score based on the five detailed items. For the overall assessment, raters indicated the extent to which they agreed with the statement:

- Overall, the explanations on the slides are very well done.

Between raters, we varied pronoun visibility and the order in which the two teaching samples were presented and rated. In the conditions with pronoun visibility, applicants' names were redacted, but their gender pronouns remained visible on the title slide and in the footer of each subsequent slide, allowing raters to infer whether the sample was submitted by the female or male applicant. In the conditions without pronoun visibility, both names and gender pronouns were redacted, preventing raters from inferring applicant gender from the teaching materials.

[2]Restricting participation to enrolled students ensured that participants were familiar with tutorials and the context of the hiring decision, even if they did not study an economics-related subject. Participants were invited through university participant databases. Nevertheless, 14 raters (0.5%) reported in the post-experimental questionnaire that they were not studying; excluding them leaves the pooled effect of pronoun visibility essentially unchanged (0.198 instead of 0.202 for the composite rating).

3.2.1. Treatments

In addition to the pronoun condition, each rater was randomly assigned to one of three institutional treatments, summarized in Table 1. These treatments vary along two dimensions: (1) whether ratings contribute to the actual hiring decision and (2) whether raters are monetarily rewarded for matching gender-blind expert assessments.

In the *Baseline*, ratings contributed to the hiring decision, with the hiring outcome determined by average ratings across raters. Raters were informed that the applicant with the highest average rating would be hired for the tutor position. They were also informed that it would be randomly determined whether the overall assessment or the composite score based on the detailed evaluation items would be used for this comparison. Because any individual rater had a low probability of being pivotal, the hiring consequence should be interpreted as a *perceived collective* consequence rather than as a large individual stake.

In *No Consequences*, raters were informed that their ratings would not affect the hiring decision.

The *Incentives* treatment extends the *Baseline* by adding a direct monetary incentive to match an independent, gender-blind expert assessment. As in the *Baseline*, ratings contributed to the hiring decision. In addition, raters were rewarded for matching the assessments of an independent expert committee that was blind to applicant gender.[3] The gender-balanced expert committee consisted of six members: one female and one male professor, one female and one male member of the non-tenured academic staff (PhD student or postdoc), and one female and one male former tutor for the same course. The committee did not observe applicant gender because its members received all pages of the teaching samples except the title slide, but of each page only a cut-out containing the exercises, formal solutions, and applicants' explanations. Raters in the *Incentives* treatment were informed about this procedure and shown an example slide on which the area submitted to the committee was framed in red.[4] For each rater, one rating item from one of the two teaching samples was randomly selected. The expert committee's average rating for this item was rounded to the nearest integer. Raters received an additional €1 if their rating matched the resulting expert assessment, on top of the fixed participation fee of €4.

3.2.2. Hypothetical Hiring Decision, Beliefs and Post-Experimental Questionnaire

After the main rating task, participants completed several non-incentivized follow-up measures. First, we elicited a hypothetical hiring decision. Participants were asked to imagine that they could decide which applicant would teach the tutorial in the upcoming semester. They could

[3] The expert assessments should not be interpreted as an objective measure of true teaching quality. Rather, they provide an external evaluation standard that is unaffected by information about applicant gender.

[4] As shown in Appendix D, the cut-outs excluded the title slide and slide footers, where applicants had provided their names and gender pronouns. This procedure ensured that the expert committee remained unaware that the experiment varied the observability of gender pronouns and allowed us to communicate the committee's information set to raters without a lengthy explanation that might have drawn attention to the gender manipulation.

Table 1: Experimental Treatments

	Treatment		
	No Consequences	Baseline	Incentives
Ratings contribute to hiring decision	No	Yes	Yes
Monetary reward for matching expert assessments	No	No	Yes

choose the applicant presented first, the applicant presented second, or indicate that the hiring decision should be determined by chance.

We also elicited participants' beliefs about gender differences in teaching ability in Microeconomics tutorials, independent of their own field of study. Participants indicated on a nine-point Likert scale which gender they believed provides better explanations in Microeconomics tutorials, with -4 indicating that men are much better, 4 indicating that women are much better, and 0 indicating no difference. An additional "cannot judge" response option was provided.[5]

The experiment concluded with a standard socio-demographic questionnaire eliciting age, gender, field of study, degree stage, and high-school math grade. Participants were also asked whether they had previously taken a Microeconomics course and whether they had prior experience as a tutor.

These additional measures allow us to complement the main rating outcomes with a hypothetical hiring decision. While the rating measures capture assessments of teaching quality, the hypothetical hiring decision requires participants to choose directly between the two applicants or leave the decision to chance. We also examine whether participants' hiring choices are consistent with their performance ratings. Finally, we use the belief measure and rater gender to examine heterogeneity in the effect of gender observability, assessing whether this effect varies systematically with raters' beliefs about gender differences in teaching ability and between male and female raters.

3.3. Testable Hypotheses

Denote the average rating of the five detailed questions of a rater i in treatment $t \in \{N, B, I\}$ (N = *No Consequences*, B = *Baseline*, I = *Incentives*) with pronoun visibility $p \in \{v, n\}$ (v = visible, n = not visible) of an applicant with gender $g \in \{f, m\}$ (f = female, m = male) by $r^t_{i,g,p}$ so that the female-minus-male gap of a rater i is:[6]

$$R^t_{i,p} = r^t_{i,f,p} - r^t_{i,m,p}.$$

[5] In analyses that use the belief as a continuous measure, we treat "cannot judge" responses as missing. In the analyses by belief category, raters who answered "cannot judge" form a separate category.

[6] We define R and O as the female-minus-male difference. This reverses the algebraic sign relative to our preregistration, which used the male-minus-female difference; the two are equivalent up to sign, and we adopt the female-minus-male convention so that a positive pronoun-visibility effect corresponds to a shift toward the female applicant.

Likewise, let the overall assessment of a rater i in treatment t with pronoun visibility p of an applicant with gender g be $o^t_{i,g,p}$ so that the female-minus-male gap of a rater i is:

$$O^t_{i,p} = o^t_{i,f,p} - o^t_{i,m,p}.$$

Each rater evaluates both female and male applicants but is randomly assigned to one treatment and one pronoun-visibility condition. Our hypotheses therefore concern the population means of these gaps, which we denote by $\mu^t_{R,p} \equiv \mathbb{E}[R^t_{i,p}]$ and $\mu^t_{O,p} \equiv \mathbb{E}[O^t_{i,p}]$.

Our first hypothesis states that pronoun visibility affects the average female-minus-male gap in *Baseline*, i.e., that a gender bias is present when evaluations have hiring consequences. This hypothesis builds on the existing literature and serves as a starting point for examining whether the effect of gender observability varies with the consequences and incentives attached to ratings:

Hypothesis 1. $\mu^B_{R,v} \neq \mu^B_{R,n}$ *and* $\mu^B_{O,v} \neq \mu^B_{O,n}$

To assess the causal effect of hiring consequences on gender-dependent evaluations and the presence of a gender bias, we compare the results of the *Baseline* treatment, in which performance ratings determine the hiring decision, to the results of the *No Consequences* treatment, in which the ratings are not used for the hiring decision. Hiring consequences may moderate the effect of gender visibility through two opposing channels. Both presume that raters care about selecting the better teacher and that this concern weighs more heavily when the rating determines an actual hire, for instance due to efficiency concerns (Charness and Rabin, 2002; Engelmann and Strobel, 2004) or meritocratic fairness views (Cappelen et al., 2007). First, hiring consequences may weaken the gender bias. If raters have an intrinsic preference for one gender, hiring consequences raise the importance of selecting the better teacher and thus make acting on this preference more costly. This potentially attenuates the gender bias (e.g., Becker, 1957; Coffman et al., 2021; Gioia and Immordino, 2024). Second, hiring consequences may strengthen the gender bias. If raters regard gender as informative about teaching performance, hiring consequences turn the rating into a forecast of future performance, and cues believed to be predictive become more important. This potentially amplifies the gender bias in the direction of these beliefs (e.g., Phelps, 1972; Arrow, 1973; Coffman et al., 2021; Stroube, 2022). Because these channels imply opposing directional predictions, our second hypothesis is two-sided and states that hiring consequences moderate the effect of gender visibility:

Hypothesis 2. $\mu^B_{R,v} - \mu^B_{R,n} \neq \mu^N_{R,v} - \mu^N_{R,n}$ *and* $\mu^B_{O,v} - \mu^B_{O,n} \neq \mu^N_{O,v} - \mu^N_{O,n}$

Additionally, to assess the causal effect of rater incentives, we compare the *Incentives* treatment with the *Baseline*. The *Incentives* treatment is identical to the *Baseline* except that raters are rewarded for matching assessments provided by a gender-balanced expert committee that is blind to applicant gender. These incentives may likewise moderate the effect of gender visibility through opposing channels. If the effect of gender visibility reflects gender-related preferences rather than beliefs about performance, rewarding agreement with gender-blind expert assessments makes deviations from these assessments individually costly and may therefore attenuate

the effect of gender visibility. Conversely, if raters regard gender as informative about underlying teaching performance, incentives to match expert assessments may increase their reliance on gender-related beliefs when performance is imperfectly observed, potentially amplifying the effect of gender visibility in the direction of these beliefs. Given these opposing predictions, our third hypothesis is also two-sided and states that rater incentives to match gender-blind expert assessments moderate the effect of gender visibility:

Hypothesis 3. $\mu^B_{R,v} - \mu^B_{R,n} \neq \mu^I_{R,v} - \mu^I_{R,n}$ *and* $\mu^B_{O,v} - \mu^B_{O,n} \neq \mu^I_{O,v} - \mu^I_{O,n}$

Hypothesis 1 concerns the existence and direction of a gender bias, while Hypotheses 2 and 3 concern whether it is moderated by the institutional environment. Importantly, we do not formulate directional predictions for this moderation because both hiring consequences and incentives to match gender-blind expert assessments can, in principle, either attenuate or amplify gender-dependent evaluations, depending on whether they primarily constrain gender-related preferences or increase reliance on gender-related beliefs about performance.

3.4. Procedures

The experiment was programmed in Qualtrics and data were collected in February and March 2026 at nine German universities. Eligibility for participation required enrollment at a German university at the time of the study, which was stated in all invitation emails and on the introductory page of the experiment. The experiment took a median of 11 minutes to complete and could be completed asynchronously.[7] To ensure attention and sufficient understanding of the instructions, the experiment included an attention check, failure of which resulted in exclusion from further participation, as well as comprehension questions that participants had to answer correctly before proceeding to the main experiment (both are reproduced in Appendix D).

We distributed survey links either to eligible members of the subject pools of experimental laboratories (BonnEconLab at the University of Bonn, Göttingen Laboratory of Behavioral Economics at the University of Göttingen, Cologne Laboratory for Economic Research (CLER) at the University of Cologne, Lakelab at the University of Konstanz, Mannheim Laboratory for Experimental Economics (mLab) at the University of Mannheim, Potsdam Laboratory for Economic Experiments (PLEx) at the University of Potsdam) or through e-learning platforms accompanying undergraduate courses at universities without experimental laboratories (University of Bayreuth, EBS Business School, University of Würzburg).

We determined the target sample size using an a priori power analysis based on two technical pilots conducted during previous tutor-hiring rounds ($N = 170$ in 2024 and $N = 395$ in 2025), recruiting participants both through e-learning platforms and through in-person classroom experiments at the Universities of Würzburg and Bonn as well as EBS Business School. The pilot data are not used in the present study other than to inform the power analysis.[8]

[7] To encourage timely completion, each invitation email specified a deadline.

[8] The pilot studies differed slightly in design and included only some of the treatments used in the present study, but both included a *No Consequences* treatment similar to the one used here.

Based on the observed effect of pronoun visibility on the female-minus-male difference in performance ratings in the *No Consequences* treatment, we specified an effect size of Cohen's $d = 0.184$. An a priori power analysis for a two-sided independent-groups t-test ($\alpha = 0.05$, power $= 0.80$) yielded a required sample size of 465 observations per cell for the comparison between the visible and invisible pronoun conditions within the *No Consequences* treatment. Applying this target to each of the three institutional treatments yielded a total target sample size of $465 \times 2 \times 3 = 2{,}790$ observations.

We stopped data collection after 2,939 participants had completed the study.[9] A total of 4,154 participants had started the survey. Before excluding incomplete submissions, we screened all responses for duplicate participation using their email addresses and international bank account numbers.[10] This procedure removed 76 duplicates (61 complete and 15 incomplete). Removing the remaining 1,200 incomplete submissions left 2,878 complete and distinct responses.[11] To exclude potential participants from the pilot studies, we further dropped 194 participants who reported in the post-experimental questionnaire that they had previously participated in a similar study.[12] Finally, we excluded two participants whose bank accounts rejected the payment and two who had inadvertently completed the survey after previously testing the experiment program, yielding a final sample of 2,680 raters.

Table 2: Participant flow

Stage	*N*
Started the survey	4,154
less duplicate appearances	−76
less incomplete submissions	−1,200
less prior participation in a similar study	−194
less rejected payment	−2
less program testers	−2
Final analysis sample	2,680

[9] Because recruitment took place simultaneously at multiple universities, the exact number of observations collected on a given day was difficult to predict. We stopped data collection on the morning after the final deadline specified in the outstanding invitation emails, once the target sample size had been reached.

[10] Following our pre-registration ("If duplicates occur, only the first response will be used"), we retained each rater's first participation and dropped all subsequent participations. We screened for duplicates *before* removing incomplete submissions, so that a rater who reached randomization, aborted, and subsequently re-entered under a different assignment could not remain in the sample under the later assignment.

[11] Of the 1,200 participants who did not complete the study (after removing duplicate re-entries), 93.2% (1,118) exited before reaching randomization and therefore never saw treatment material. Among the 2,960 raters who reached randomization, after removing duplicate re-entries, post-randomization non-completion was 2.77% and was statistically balanced across treatments and pronoun visibility conditions ($p > 0.18$). Including the exclusions described below, the final sample retains 90.5% of these raters.

[12] Because this question was asked after the rating task, the exclusion could in principle depend on assignment. The exclusion rate does not differ significantly across treatments (6.5% to 6.9%; $p = 0.94$) or across the six treatment-by-pronoun cells ($p = 0.50$); it is somewhat higher with visible than with invisible pronouns (7.6% versus 6.0%; $p = 0.09$). Retaining these raters leaves the pooled effect of pronoun visibility essentially unchanged.

Assignment to institutional treatment, pronoun condition, and presentation order was random. Randomization yielded six cells of comparable size (invisible/visible pronoun condition): *No Consequences* 456/435, *Baseline* 455/446, and *Incentives* 447/441. The average cell size was 447, compared with the pre-registered target of 465. The pronoun conditions (1,358 invisible, 1,322 visible) and the two presentation orders (1,342/1,338) were likewise balanced.

Participants were on average 24.9 years old; 60.0% identified as female, 37.1% as male, 1.3% as diverse (non-binary), and 1.7% preferred not to report their gender. Participants came from a range of fields of study, with 25.9% studying economics.

Table 10 in Appendix A reports means of predetermined covariates (age, rater gender, field of study, degree stage, exposure to Microeconomics, tutoring experience, and high-school math grade) by treatment and pronoun condition, together with joint tests of balance. The covariates are balanced across conditions, consistent with successful randomization.

4. Results

We begin by presenting descriptive statistics and testing for a gender bias within each treatment (4.1). We also highlight the economic implications of the observed rating gaps for applicant outcomes. Next, we turn to the question of whether hiring consequences and rater incentives moderate the gender bias (4.2), before examining the hypothetical hiring decision (4.3). Finally, we analyze whether heterogeneity in the effect of gender visibility is systematically related to evaluators' beliefs about gender differences in teaching ability (4.4).

We present results in the main text only for the composite rating R, and report the corresponding results for the overall assessment in Appendix C. The only exception is Table 3, which reports descriptive statistics for both measures. Results are qualitatively identical across the two measures and we treat a result as robust only if it is supported by both measures. In the main text, we report the pre-registered regression analyses and provide the corresponding non-parametric tests in footnotes. Because all raters evaluated the same two teaching samples, our estimates identify the effect of making gender observable for *this particular pair* of applicants, not an average effect across applicant pairs.

4.1. Gender Bias Within Treatments

Our key dependent variables are the female-minus-male gaps R_i and O_i as defined in Section 3.3. Under pronoun invisibility, these gaps reflect any underlying quality difference between the two teaching samples. Because the teaching samples are held fixed across pronoun conditions, comparing the female-minus-male gaps across the randomly assigned visible and invisible pronoun conditions differences out this fixed quality difference and identifies the causal effect of making gender observable. The five detailed items comprising the composite rating exhibit high internal consistency (Cronbach's $\alpha = 0.88$ and 0.92 for the two applicants), supporting their aggregation into a composite measure.

Table 3 reports average ratings by treatment and pronoun condition for both the composite

Table 3: Average ratings by treatment and pronoun condition

Treatment	Pronoun	Composite rating			Overall assessment		
		Female applicant	Male applicant	Diff. R (F−M)	Female applicant	Male applicant	Diff. O (F−M)
No Consequences	Invisible	7.12	7.18	-0.07	7.15	7.27	-0.12
	Visible	7.19	7.09	0.10	7.16	7.11	0.05
Baseline	Invisible	7.57	7.65	-0.08	7.59	7.67	-0.08
	Visible	7.72	7.55	0.17	7.80	7.56	0.24
Incentives	Invisible	7.33	7.51	-0.18	7.40	7.57	-0.18
	Visible	7.55	7.54	0.01	7.59	7.56	0.03

Notes: Cell means of the composite rating (average of the five detailed items) and of the overall assessment, on a scale from 1 to 10. The columns headed Diff. R and Diff. O are the female-minus-male rating differences, computed on unrounded means; a positive value means the female applicant is rated higher. Pronoun indicates whether the applicants' pronoun was visible.

rating and the overall assessment. The two measures are strongly correlated ($\rho = 0.84$; only 4.4% of raters rank the two applicants in opposite directions on the two measures), and all results reported below for the composite rating also hold for the overall assessment. The raw data reveal a consistent pattern: in each treatment and for both measures, the female-minus-male gap is (weakly) negative when gender is hidden and positive when gender is observable. Thus, making gender observable shifts ratings toward the female applicant in all three treatments and on both measures.

Figure 1 displays the gender bias separately for each treatment for the composite rating (the corresponding overall assessment results are shown in Figure 6 in Appendix C.1). In all three treatments, making gender observable shifts evaluations toward the female applicant.

To test whether the reverse gender bias is statistically significant (Hypothesis 1), we regress the female-minus-male rating gap on pronoun visibility separately within each treatment, controlling for presentation order. Table 4 reports these regressions for the composite rating, without and with demographic controls (the corresponding results for the overall assessment are reported in Table 15 in Appendix C.1).

Three observations emerge. First, the bias is *directionally present in every treatment*: the estimated effect of gender visibility is positive in all three treatments, with and without demographic controls. Second, the effect is statistically significant in the *Baseline*: when raters face no monetary incentives but know that their collective assessments determine the hiring decision, making gender observable increases the female-minus-male evaluation gap by 0.26 points on the composite measure.[13] Third, the statistical evidence is weaker in the other two treatments. In *Incentives*, the OLS estimate is marginally significant, while the corresponding non-parametric test rejects equality ($p = 0.027$, WRS). In *No Consequences*, the OLS estimate is not statistically significant, consistent with the non-parametric test ($p = 0.112$, WRS).

Because the pronoun manipulation is identical across treatments, with the same teaching

[13] A two-sided Wilcoxon rank-sum test (WRS) likewise rejects equality for the composite rating ($p < 0.01$).

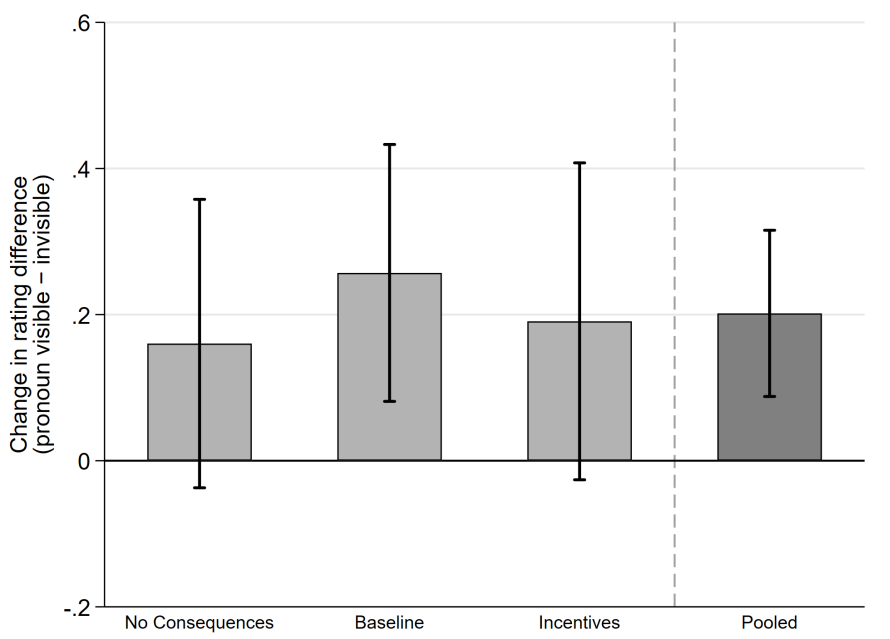


Figure 1: Reverse gender bias by treatment (composite rating): the change in the female-minus-male gap from the pronoun-invisible to the pronoun-visible condition, estimated by OLS controlling for presentation order (pooled bar: with treatment intercepts), i.e., the no-control estimates of Table 4. Error bars: 95% confidence intervals based on heteroskedasticity-robust standard errors.

samples and the same pronoun edit in each treatment, the three treatment-specific coefficients estimate the effect of the same manipulation in different institutional environments. To estimate the average effect of making gender observable across these environments, we pool the data and estimate a common pronoun-visibility coefficient while retaining treatment intercepts and controlling for presentation order. Table 4 shows that making gender observable has a highly significant positive effect on the female-minus-male rating gap.[14] Presentation order affects the female-minus-male rating gap, but does not significantly moderate the effect of pronoun visibility.[15]

The robustness of the reversal has a direct economic implication, which follows from Table 3 without further estimation. In all three treatments and on both rating measures, the male applicant receives the higher average rating when gender is hidden, whereas the female applicant

[14]The corresponding non-parametric test confirms this result ($p < 0.001$). Randomization was conducted separately within each of the nine universities, which balances the pronoun conditions within each university. Adding university fixed effects leaves the pooled estimate unchanged. Although Bonn accounts for the majority of observations, raters from Bonn do not respond significantly differently to pronoun visibility than raters from other universities, and the effect persists when Bonn is excluded. Tables 22 and 23 in Appendix C.5 report these specifications.

[15]Seeing the female sample first significantly decreases the female-minus-male rating gap. However, the pronoun visibility × presentation order interaction is small and statistically insignificant ($b = 0.11$, $p = 0.36$ for the composite rating R; $b = 0.10$, $p = 0.47$ for the overall assessment O; pooled specifications with treatment intercepts).

Table 4: Effect of pronoun visibility on the rating difference (composite rating)

	No Consequences		Baseline		Incentives		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pronoun visible	0.160	0.161	0.257***	0.245***	0.191*	0.197*	0.202***	0.204***
	(0.101)	(0.103)	(0.090)	(0.091)	(0.111)	(0.113)	(0.058)	(0.058)
Female first	-0.060	-0.038	-0.381***	-0.388***	-0.064	-0.060	-0.169***	-0.167***
	(0.101)	(0.101)	(0.090)	(0.092)	(0.110)	(0.110)	(0.058)	(0.058)
Treatment intercepts	No	No	No	No	No	No	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	891	891	901	901	888	888	2,680	2,680
R^2	0.003	0.030	0.028	0.045	0.004	0.022	0.009	0.020

Notes: OLS with heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male rating difference in the composite rating. Columns 1–6 estimate the visibility effect separately within each treatment; columns 7–8 report the average visibility effect across treatments while retaining treatment intercepts, which are estimated but not reported. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

receives the higher average rating when gender is observable. Recall that the actual hiring rule in the consequential treatments selects the applicant with the highest average rating, with the rating measure chosen at random. Under this rule, the reversal in the rating gap also reverses which applicant would be selected: in each treatment and on both rating measures, the male applicant would be selected when gender is hidden and the female applicant when gender is observable.

Three qualifications help put this finding in perspective. First, the underlying rating shift is modest in absolute terms (about 0.2 points) and the cell averages are close to one another. The hiring reversal reflects the discreteness of the hiring decision: a small shift in relative ratings can change the selected applicant when the initial comparison is close. Second, the reversal is identified for this particular pair of teaching samples and need not generalize to other applicant pairs. Third, the reversal of the *average-based* hiring decision does not imply that gender visibility also changes the choices of *individual* raters when asked to select one applicant directly. We examine this question in Section 4.3.

Result 1. *Making the applicants' gender observable induces a reverse gender bias: it shifts ratings toward the female applicant. The estimated effect is positive in all three treatments, statistically significant in the Baseline, and highly significant when pooled across treatments. In the No Consequences and Incentives treatments, the evidence is not robust across statistical tests.*

The gap shows that relative evaluations shift toward the female applicant, but not whether this shift reflects higher ratings for the female applicant, lower ratings for the male applicant, or both. We therefore estimate the effect of gender observability separately on each applicant's rating.

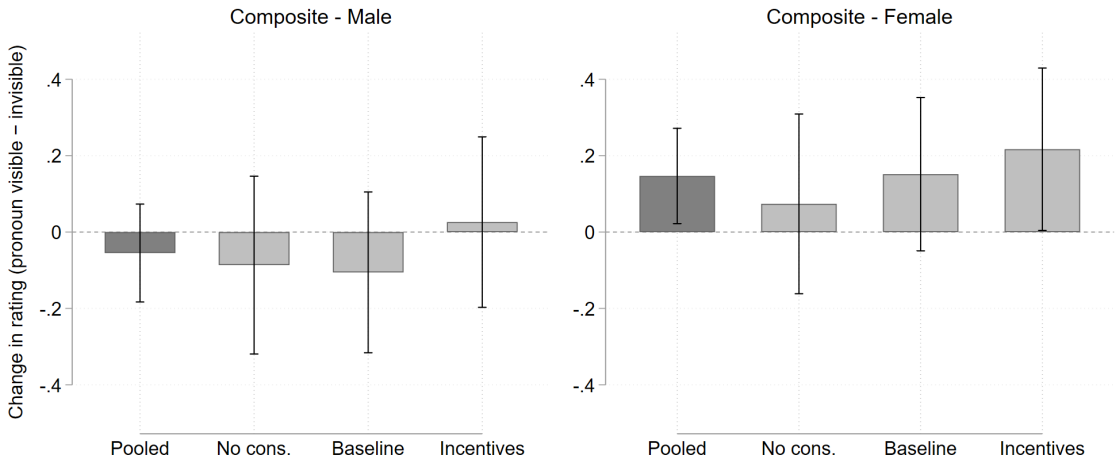


Figure 2: Effect of pronoun visibility on each applicant's composite rating, pooled and by treatment: OLS estimates controlling for presentation order (pooled: with treatment intercepts). Error bars: 95% confidence intervals based on heteroskedasticity-robust standard errors.

The decomposition shown in Figure 2 and formally analyzed in Table 5 (the corresponding overall-assessment results are reported in Figure 7 and Table 16 in Appendix C.1) indicates that the reversal operates mainly through higher ratings for the female applicant, alongside a smaller decline for the male applicant. Pooled across treatments, making gender observable increases the female applicant's composite rating by 0.147 points, while decreasing the male applicant's rating by 0.055 points. The former effect is statistically significant in the OLS specification, whereas the latter is not.[16]

We next examine whether the reverse bias differs by rater gender. As pre-registered, we restrict the analysis to raters who identify as female or male. One potential explanation is in-group favoritism: the reversal may be driven primarily by female evaluators favoring the female applicant. Table 6 reports the effect of pronoun visibility separately for female and male raters, as well as its interaction with rater gender (the corresponding results for the overall assessment are reported in Table 17 in Appendix C.1). Both female and male raters shift their evaluations toward the female applicant, and on the composite rating the effect of pronoun visibility is statistically significant for both groups.[17] Moreover, the interaction between pronoun visibility and rater gender is statistically insignificant, providing no evidence that female raters respond more strongly to gender observability than male raters.

[16]The non-parametric test provides weaker evidence for the increase in the female applicant's rating ($p = 0.068$, WRS), while the decline in the male applicant's rating remains statistically insignificant ($p = 0.196$, WRS). For the overall assessment, the pattern is similar, except that the non-parametric test also detects a significant decline in the male applicant's rating (Appendix C.1).

[17]On the overall assessment, the effect is significant for female raters and marginally significant for male raters ($p = 0.064$; Table 17).

Table 5: Effect of pronoun visibility on each applicant's rating (composite rating)

	Male applicant		Female applicant	
	(1)	(2)	(3)	(4)
Pronoun visible	-0.055	-0.042	0.147**	0.162***
	(0.065)	(0.063)	(0.064)	(0.062)
Female first	0.106	0.113*	-0.062	-0.054
	(0.065)	(0.063)	(0.064)	(0.062)
Demographic controls	No	Yes	No	Yes
Observations	2,680	2,680	2,680	2,680
R^2	0.015	0.093	0.017	0.075

Notes: OLS with heteroskedasticity-robust standard errors in parentheses, pooled across treatments with treatment intercepts (omitted treatment: Baseline). The dependent variable is the composite rating given to the male or to the female applicant. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: Pronoun visibility and rater gender: pooled interaction and by rater gender (composite rating)

	Pooled interaction		Female raters		Male raters	
	(1)	(2)	(3)	(4)	(5)	(6)
Pronoun visible	0.203***	0.212***	0.204***	0.210***	0.189**	0.184**
	(0.077)	(0.077)	(0.077)	(0.077)	(0.090)	(0.091)
Male rater	-0.048	-0.038				
	(0.083)	(0.085)				
Pronoun visible × Male rater	-0.015	-0.029				
	(0.118)	(0.119)				
Demographic controls	No	Yes	No	Yes	No	Yes
Observations	2,601	2,601	1,608	1,608	993	993

Notes: OLS, heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male rating difference in the composite rating. Columns 1 and 2 pool both rater genders and interact pronoun visibility with a male-rater indicator. Columns 3 to 6 estimate the visibility effect separately within each rater gender. All columns include treatment intercepts and the presentation-order control, which are estimated but not reported. All columns are restricted to raters who identify as female or male, excluding those who selected diverse or prefer not to say. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.2. The Impact of Hiring Consequences and Rater Incentives

We now examine whether hiring consequences and rater incentives moderate the effect of gender visibility. First, to estimate the effect of hiring consequences, we compare the *No Consequences* and *Baseline* treatments and test whether the effect of pronoun visibility differs between them. Second, to estimate the effect of rater incentives, we compare the *Baseline* and *Incentives* treatments and test whether the effect of pronoun visibility differs between these treatments. Table 7 reports both sets of regressions for the composite rating (the corresponding results for the overall assessment are reported in Table 18 in Appendix C.2).

Table 7: Interaction of pronoun visibility with hiring consequences (H2) and rater incentives (H3) (composite rating)

	H2		H3	
	(1)	(2)	(3)	(4)
Pronoun visible	0.254***	0.251***	0.254***	0.246***
	(0.090)	(0.090)	(0.090)	(0.090)
No Consequences	0.024	0.021		
	(0.098)	(0.098)		
Incentives			-0.088	-0.089
			(0.100)	(0.100)
Visible × No cons.	-0.095	-0.083		
	(0.135)	(0.137)		
Visible × Incent.			-0.064	-0.050
			(0.142)	(0.143)
Female first	-0.221***	-0.215***	-0.223***	-0.225***
	(0.067)	(0.068)	(0.071)	(0.072)
Constant	0.023	-0.136	0.024	-0.067
	(0.070)	(0.212)	(0.071)	(0.208)
Demographic controls	No	Yes	No	Yes
Observations	1,792	1,792	1,789	1,789
R^2	0.011	0.023	0.013	0.026

Notes: OLS with heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male rating difference in the composite rating. The omitted treatment is Baseline, so the interaction is the difference in the pronoun-visibility effect between treatments: consequences (H2) or incentives (H3). Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We find no statistically significant evidence that either hiring consequences or rater incentives moderate the reverse bias. For hiring consequences, the interaction with *Pronoun Visible* is negative and statistically insignificant. If anything, the negative estimate suggests a somewhat smaller effect of gender visibility when hiring consequences are removed. Regarding rater incentives, the corresponding interaction is also statistically insignificant.[18]

[18]This result does not imply that incentives fail to improve rating accuracy, since our design does not observe true teaching quality.

Result 2. *We find no significant evidence that either hiring consequences or rater incentives moderate the effect of gender observability; we are, however, not powered to detect moderate treatment effects.*

How informative are these insignificant interactions? Our design is not powered to detect moderate treatment differences. The power analysis was designed to detect the visibility effect itself, not differences in it, and even in *Baseline* the realized visibility effect is only about as large as the effect size assumed there. Consequently, even a treatment that eliminated the reverse gender bias entirely would not reliably produce a significant interaction.[19] We therefore interpret the treatment comparisons as evidence for the *robustness* rather than the invariance of the reverse gender bias. The same pronoun-visibility manipulation shifts evaluations toward the female applicant in all three institutional environments, and the raw female-minus-male gap reverses sign in each treatment. However, the interaction estimates are not sufficiently precise to rule out meaningful differences in the magnitude of the effect across treatments.

Table 8: Effect of treatment on the level of ratings (composite rating)

	Average		Male appl.		Female appl.	
	(1)	(2)	(3)	(4)	(5)	(6)
No Consequences	-0.476***	-0.474***	-0.464***	-0.463***	-0.487***	-0.485***
	(0.072)	(0.069)	(0.080)	(0.077)	(0.079)	(0.077)
Incentives	-0.138**	-0.146**	-0.078	-0.090	-0.198***	-0.202***
	(0.067)	(0.066)	(0.078)	(0.076)	(0.074)	(0.073)
Pronoun visible	0.046	0.060	-0.055	-0.042	0.147**	0.162***
	(0.058)	(0.055)	(0.065)	(0.063)	(0.064)	(0.062)
Female first	0.022	0.029	0.106	0.113*	-0.062	-0.054
	(0.058)	(0.056)	(0.065)	(0.063)	(0.064)	(0.062)
Demographic controls	No	Yes	No	Yes	No	Yes
Observations	2,680	2,680	2,680	2,680	2,680	2,680
R^2	0.018	0.100	0.015	0.093	0.017	0.075

Notes: OLS with heteroskedasticity-robust standard errors in parentheses. The dependent variable is the level of the composite rating, not the female-minus-male difference: in columns 1–2 the rater's average rating across the two applicants (one observation per rater), in columns 3–6 the rating given to the male and to the female applicant respectively. The omitted treatment is Baseline. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade; non-response categories of gender and tutoring experience enter as separate indicators, so all raters are retained in the controlled columns. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The insignificant interactions raise the concern that the treatment manipulations may have been too weak to affect behavior. Table 8 therefore examines their effects on the *level* of the composite rating, measured as each rater's average rating across the two applicants (the

[19] Measured in the same way, as the visibility effect divided by the standard deviation of the female-minus-male gap, it amounts to $d = 0.13$ pooled across treatments on the composite rating (0.11 in *No Consequences*, 0.19 in *Baseline*, and 0.12 in *Incentives*), compared with the assumed $d = 0.184$.

corresponding results for the overall assessment are reported in Table 19 in Appendix C.2). Relative to the *Baseline*, average ratings are 0.476 points lower in *No Consequences* ($p < 0.01$) and 0.138 points lower in *Incentives* ($p < 0.05$). Thus, the treatment manipulations affect evaluation behavior even though they do not significantly moderate the effect of gender visibility. Ratings are highest when they contribute to an actual hiring decision without incentives to match gender-blind expert assessments, while such incentives partially attenuate this increase. This pattern is consistent with evidence that the purpose of an evaluation can affect rating levels (Jawahar and Williams, 1997).

Pronoun visibility produces a different pattern. It has virtually no effect on the average rating across applicants (0.046, statistically insignificant), while increasing the female applicant's rating by 0.147 points and reducing the male applicant's rating by 0.055 points, although the latter effect is statistically insignificant (columns 3 and 5 of Table 8). Thus, the institutional treatments shift the overall level of evaluations, whereas gender visibility shifts evaluations *between* the two applicants.

4.3. The Hypothetical Hiring Decision

In contrast to the rating measures, the hypothetical hiring decision requires participants to choose directly between the two applicants or defer to a random draw.[20] As shown in Figure 3, the choice data closely mirror the rating results. When gender is hidden, 35.8% of raters choose the female applicant, 41.0% choose the male applicant, and 23.2% defer to a random draw. When gender is visible, these shares are 40.5%, 34.1%, and 25.3%, respectively. Pooling across treatments, pronoun visibility significantly shifts the choice distribution ($\chi^2(2) = 13.73$, $p = 0.001$). A multinomial logit shows that making gender observable reduces the probability of choosing the male applicant by 6.9 percentage points and increases the probability of choosing the female applicant by 4.8 percentage points; both effects are statistically significant ($p < 0.01$ and $p = 0.010$, respectively). Relative to the 41.0% probability of choosing the male applicant when gender is hidden, the former (unrounded: 6.94 percentage points) corresponds to a 16.9% decline. As shown in Panel (c), the direction of these effects is the same in all three treatments.

The pre-registered consistency check further shows that hiring choices closely align with the rating measures. Among the 2,409 raters who strictly rank one applicant above the other on the composite measure, 74.0% choose the higher-rated applicant, 20.0% defer to a random draw, and only 6.0% choose the lower-rated applicant. Among raters who make a direct choice between the applicants, 92.5% select the applicant they rated more highly. The corresponding consistency is even stronger for the overall assessment (Appendix C.3).

Result 3. *The hypothetical hiring decision mirrors the rating results: making gender observable shifts choices toward the female applicant, with the same directional pattern in all three treatments. Individual hiring choices are also closely aligned with the rating measures.*

[20] This measure gives each participant a direct choice between the applicants, rather than making an individual rating one component of an aggregate hiring decision. Its limitation is that the choice is hypothetical.

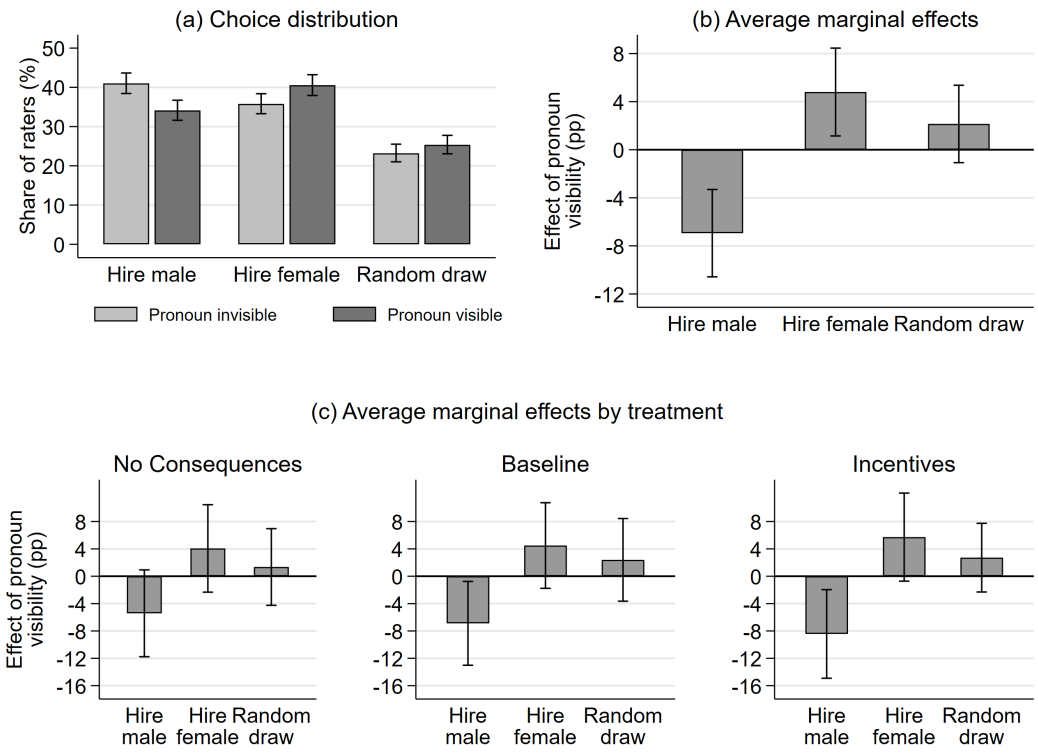


Figure 3: The hypothetical hiring decision. Panel (a) shows the distribution of the hiring choice in each pronoun condition (95% confidence intervals of the shares). Panel (b) shows the average marginal effect of making gender observable on the probability of each choice, in percentage points, from a multinomial logit pooled across the three treatments (95% CIs based on robust standard errors). Panel (c) shows the same average marginal effects estimated separately by treatment (multinomial logit, 95% CIs).

4.4. Mechanisms: Gender-Related Beliefs and Other Channels

What helps explain the reverse gender bias? One potential channel is belief-based: raters may hold beliefs about gender differences in teaching ability and condition their evaluations on these beliefs when applicant gender is observable. Other channels need not operate through perceived differences in teaching ability, including taste- or identity-based preferences, corrective-fairness motives, experimenter demand, or social desirability (Coffman et al., 2021). Our elicited belief about which gender provides better explanations in Microeconomics tutorials allows us to examine whether heterogeneity in the effect of gender visibility is systematically related to raters' gender-related beliefs. We present four pieces of evidence on the role of beliefs and then turn to experimenter demand.

First, many raters perceive no gender difference in explaining Microeconomics (43.4% of all raters); 15.8% believe that women explain better, 8.1% that men do, and 32.6% answer "cannot judge." Stated beliefs are only weakly related to observable rater characteristics (Table 11 in

Appendix B; $R^2 \leq 0.074$). Their strongest correlate is the rater's own gender: male raters' beliefs are 0.53 points less favorable to women, and 9.7% of male raters compared with 19.7% of female raters believe that women explain better (Figure 5 in Appendix B). Answering "cannot judge" mainly reflects not having taken Microeconomics (-16 percentage points for raters who have) and is unrelated to rater gender. Importantly, we find no evidence that the pronoun manipulation affects stated beliefs: pronoun visibility predicts neither the belief ($p = 0.62$) nor the probability of answering "cannot judge" ($p = 0.63$), and neither do treatment or presentation order (Table 11).[21] This is reassuring for interpreting the belief measure as capturing heterogeneity that is not affected by the pronoun manipulation. However, because beliefs were elicited after the rating task and without incentives, we cannot treat them as strictly predetermined.

Second, beliefs matter only when gender is observable. With invisible pronouns, stated beliefs do not predict the female-minus-male rating gap (slope -0.015, $p = 0.72$), whereas they strongly predict it when gender is observable (slope 0.163, $p < 0.001$; Table 12 in Appendix B). The same holds for the hiring choice: with invisible pronouns, the distribution of choices does not differ across belief categories ($p = 0.39$), whereas it does when gender is observable ($p < 0.001$). This placebo comparison speaks against the concern that the belief reflects a general preference for one of the two teaching samples. It also provides a consistency check on the concern that stated beliefs were formed after the fact to justify the ratings: rationalization that systematically moved stated beliefs toward the favored gender when gender is visible would shift their distribution, yet the distribution of beliefs does not differ between pronoun conditions ($\chi^2(8) = 3.44$, $p = 0.90$, across the nine belief values).[22] These checks are descriptive. Because beliefs were elicited after the rating task, they cannot rule out rationalization that leaves the overall distribution unchanged, and the comparisons by belief below should be read as contrasts conditional on beliefs stated after the task rather than as effects for predetermined belief types.

Third, the reversal is concentrated among raters who believe that women explain better. Table 9 and Figure 4 estimate the effect of gender visibility separately within five belief categories (the corresponding results for the overall assessment are reported in Table 20 and Figure 8 in Appendix C.4). The effect differs significantly across categories ($p = 0.012$). Among raters who believe that women explain better, making gender observable raises the female-minus-male gap by 0.64 points (beliefs +1 and +2, $p < 0.001$) and 0.72 points (beliefs +3 and +4, $p = 0.029$). We detect no difference between these two effects ($p = 0.83$), although the estimate for strong beliefs is imprecise. Among raters who see no gender difference, the effect is small and statistically insignificant (0.08), and among raters who believe that men explain better it is

[21] 1,805 of the 2,680 raters state a belief. Missingness does not differ significantly across treatments ($\chi^2 = 2.72$, $p = 0.257$) or pronoun conditions ($\chi^2 = 0.37$, $p = 0.543$). The distribution over belief categories differs somewhat across treatments ($\chi^2(10) = 19.63$, $p = 0.033$): a pro-female belief is 6.3 percentage points more common in *No Consequences* than in *Baseline* ($p = 0.010$). Because treatment was assigned before the belief was elicited, this difference reflects either chance or a small effect of the treatment on stated beliefs; it does not affect the comparisons across pronoun conditions reported below, which are randomized within each treatment.

[22] Pronoun visibility also does not change the share of raters who state a gender difference or the mean absolute belief.

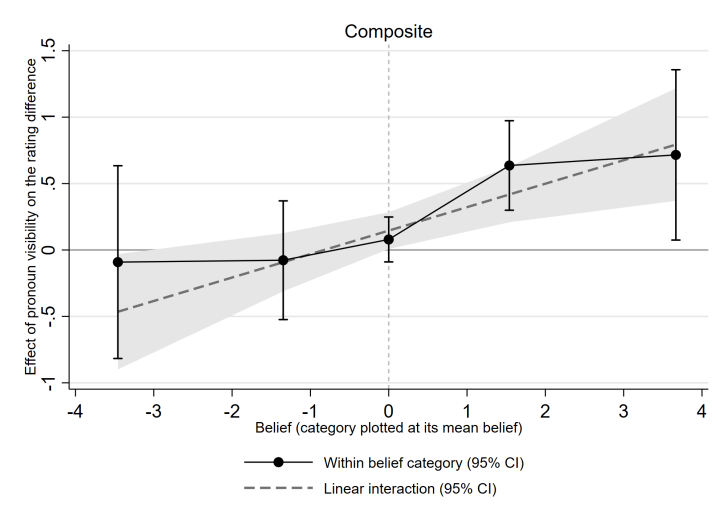


Figure 4: Effect of pronoun visibility by belief category (composite rating): the visibility effect estimated within each belief category, plotted at the category's mean belief (dots, 95% CI; Table 9, column 2), against the prediction of the linear visibility × belief interaction (dashed line, 95% CI band; column 1).

close to zero (−0.09 and −0.08) and insignificant as well. Raters who answered "cannot judge" shift by 0.21 points ($p = 0.043$), close to the sample average. Although raters who believe that women explain better make up only 23% of belief-stating raters, they account for roughly half of the pooled visibility effect.[23]

The same pattern appears in the hypothetical hiring decision (Table 13 in Appendix B). Among raters who believe that women explain better, making gender observable increases the probability of choosing the female applicant by 27 (beliefs +1 and +2) and 19 percentage points (beliefs +3 and +4), and reduces the probability of choosing the male applicant by 17 to 18 percentage points. Among raters with a zero belief, the choice distribution does not respond to gender visibility.

Fourth, the belief heterogeneity is not a proxy for rater gender. Because female raters more often hold pro-female beliefs, belief and rater gender could capture the same heterogeneity. When pronoun visibility is interacted with both, the belief interaction remains significant (0.160, $p = 0.007$), whereas no additional moderation by rater gender is detected (−0.084, $p = 0.56$; Table 14 in Appendix B), although this estimate is imprecise (95% CI $[-0.36, 0.19]$). Moreover, among raters with a pro-female belief, both female and male raters shift toward the female applicant (0.63, $p = 0.001$, $N = 316$, and 0.51, $p = 0.042$, $N = 96$), and the two shifts do not differ significantly (difference −0.06, $p = 0.85$).[24] Belief-related heterogeneity thus remains

[23] Weighting each category's effect by its share of the full sample, the raters with a pro-female belief contribute about 0.10 of the pooled effect of 0.20 on the composite rating, the "cannot judge" raters about 0.07, and the raters with a zero belief about 0.03.

[24] The three-way interaction of pronoun visibility, a pro-female belief, and a male-rater indicator is also insignificant (0.07, $p = 0.85$; overall assessment: 0.32, $p = 0.47$). For the overall assessment, the shifts among pro-female-belief raters are 0.80 ($p < 0.001$) for female and 0.84 ($p = 0.014$) for male raters.

Table 9: Effect of pronoun visibility by belief category (composite rating)

	(1) Linear	(2) Categories	(3) + Cannot judge
Pronoun visible	0.146**		
	(0.070)		
Pronoun visible × Belief	0.177***		
	(0.058)		
Visible: belief −4 and −3		-0.091	-0.095
		(0.370)	(0.369)
Visible: belief −2 and −1		-0.077	-0.077
		(0.228)	(0.228)
Visible: belief 0		0.079	0.080
		(0.086)	(0.086)
Visible: belief +1 and +2		0.636***	0.633***
		(0.172)	(0.172)
Visible: belief +3 and +4		0.716**	0.714**
		(0.327)	(0.326)
Visible: cannot judge			0.206**
			(0.102)
Observations	1,805	1,805	2,680
R^2	0.019	0.021	0.017

Notes: OLS, heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male difference in the composite rating. All columns include treatment intercepts and the presentation-order control. Belief indicates which gender explains Microeconomics better in general (−4 = males much better to 4 = females much better, 0 = no difference). Column 1 interacts pronoun visibility linearly with the belief; its *Pronoun visible* row is the visibility effect at belief 0 implied by the linear specification. Column 2 replaces the linear term by five belief categories; each *Visible* row is the visibility effect estimated within that category. Columns 1 and 2 use the raters who state a belief; column 3 adds the raters who answered "cannot judge" as a sixth category. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

detectable after accounting for rater gender, consistent with the absence of a significant rater-gender interaction in Section 4.1.

As raters may have inferred the purpose of the study and adjusted their ratings to what they believed we expected, the results could be explained by experimenter demand. Two indirect checks speak against this. First, gender becomes a salient contrast between the applicants only once both samples have been seen (Hsee, 1996; Bohnet et al., 2016), so experimenter demand should produce a larger shift when the female sample is rated second. Second, accommodating presumed expectations is costless for the applicants only when ratings have no hiring consequences, so the shift should be smaller when ratings are consequential. Neither pattern appears. The effect of gender visibility is, if anything, larger when the female sample is rated first (0.26 versus 0.15 points) and when ratings are consequential (0.22 versus 0.16 points), although neither difference is statistically significant ($p = 0.36$ and $p = 0.61$; the same pattern holds for the overall assessment). These checks cannot rule out experimenter demand, but they provide no

indication that it drives the reverse gender bias.

Taken together, the results show that the reverse gender bias is strongly related to raters' stated beliefs about gender differences in teaching ability. These beliefs predict relative evaluations and hiring choices only when applicant gender is observable, and the shift toward the female applicant is concentrated among the raters who believe that women explain better.

Result 4. *The reverse gender bias varies systematically with raters' stated beliefs about gender differences in teaching ability: these beliefs predict relative evaluations when applicant gender is observable, but not when it is hidden. The shift toward the female applicant, in both ratings and hiring choices, is concentrated among raters who believe that women explain better; raters who see no gender difference show no significant shift.*

5. Conclusion

Contrary to much of the earlier evidence on discrimination against women, we find that making applicants' gender observable shifts subjective performance evaluations toward the female applicant. This shift is directionally consistent across all three institutional environments we study, extends to direct hypothetical hiring choices, and operates mainly through higher ratings for the female applicant, alongside a smaller decline in ratings for the male applicant. We find no significant evidence that either hiring consequences or incentives to match gender-blind expert assessments moderate this effect.

Our belief results provide insight into the structure of this reversal. The shift toward the female applicant varies systematically with raters' stated beliefs about gender differences in teaching ability: these beliefs predict relative evaluations when applicant gender is observable, but not when it is hidden. The reversal is concentrated among the raters who believe that women explain better, whereas raters who see no gender difference show no significant shift. This pattern is consistent with raters conditioning their evaluations on gender-related beliefs when gender information is available (Coffman et al., 2021). The reversal also does not appear to be driven by female raters alone. Both female and male raters shift their evaluations toward the female applicant, we find no significant interaction between pronoun visibility and rater gender, and the heterogeneity related to beliefs remains detectable after accounting for rater gender.

These results do not identify a particular mechanism. Because beliefs were elicited after the evaluation task, they cannot be treated as predetermined. Moreover, the association between stated beliefs and evaluations could reflect beliefs about teaching ability or related gender attitudes and preferences. For example, raters who favor women in their stated beliefs may also be more inclined to correct for perceived disadvantages faced by women or to respond in a socially desirable manner. Such channels could generate the observed relationship between stated beliefs and evaluations without differences in perceived teaching ability being the underlying mechanism. We therefore interpret the belief results as showing that the response to gender information is systematically related to raters' gender-related beliefs and attitudes, rather than

as identifying why this relationship arises.

Several features of our setting may shape this response. First, raters assess two relatively similar teaching samples one after the other, making applicant gender a salient dimension along which the samples differ once gender becomes observable. Preferences and judgments can differ between joint and separate evaluation (Hsee, 1996; Hsee et al., 1999; Bohnet et al., 2016), and the effect we identify may therefore differ when candidates are evaluated in isolation. Second, our evaluations are embedded in a hiring context rather than conventional end-of-course teaching evaluations. Evaluators may consequently consider not only perceived teaching performance but also the distributive consequences of their assessments, including concerns about fairness or representation. Finally, our study was conducted among university students in a setting and period in which attitudes toward gender and representation may differ from those in populations examined in earlier studies. These features provide potential explanations for why our findings differ from evidence of disadvantages for women in other settings, but our design does not allow us to distinguish among them.

Our findings also have several limitations. The reversal is estimated for a single pair of teaching samples. Although the effect of gender observability is causally identified for this pair, its magnitude and direction across other applicant pairs remain to be established. Similarly, we reveal gender through pronouns stated on otherwise anonymous slides; whether gender cues such as names, photographs, voices, or personal interaction have the same effect is not tested. Moreover, our treatment-moderation estimates are not sufficiently precise to rule out meaningful differences in the magnitude of the effect across institutional environments. We therefore conclude that the reversal persists across the institutional environments we study, while recognizing that its magnitude may differ across them. Finally, our analysis concerns exclusively the contrast between female and male applicants and raters and does not permit conclusions about non-binary applicants or raters. Future research could address these limitations by studying multiple applicant pairs and populations, varying the strength of institutional incentives, eliciting incentivized beliefs before the evaluation task, and extending the design beyond binary gender comparisons.

Gender-dependent evaluations can affect hiring decisions regardless of which gender they favor. Whether such shifts improve or worsen selection cannot be determined from our experiment because we do not observe true teaching quality. More generally, our findings caution against equating gender bias in subjective evaluations with discrimination against women. The direction of gender bias can vary across settings, and our results suggest that raters' beliefs may help explain this variation. Identifying when and why gender information advantages one group rather than another, and establishing how these effects generalize across applicants and populations, remains an important task for future research.

References

Andersson, O., M. Backman, N. Bengtsson, and P. Engström (2025). Are economics students biased against female teachers? Evidence from a randomized, double-blind natural field experiment. *Journal of Economic Behavior & Organization 238*, 107153.

Arrow, K. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*, pp. 3–33. Princeton University Press.

Ayllón, S. and C. Zamora (2025). Gender bias in student evaluations of teaching: Do debiasing campaigns work? *IZA Discussion Paper No. 17632*.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2025). Explicit and implicit belief-based gender discrimination: A hiring experiment. *Management Science 71*(2), 1600–1622.

Bartoš, V., M. Bauer, J. Chytilová, and F. Matějka (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review 106*(6), 1437–1475.

Bartoš, V., U. Glogowsky, and J. Rincke (2025). The color of knowledge: Impacts of tutor race on learning and performance. *CESifo Working Paper No. 12352*.

Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Binderkrantz, A. S. and M. Bisgaard (2024). A gender affinity effect: the role of gender in teaching evaluations at a Danish university. *Higher Education 87*(3), 591–610.

Binderkrantz, A. S., M. Bisgaard, and B. Lassesen (2022). Contradicting findings of gender bias in teaching evaluations: Evidence from two experiments in Denmark. *Assessment & Evaluation in Higher Education 47*(8), 1345–1357.

Bohnet, I., A. Van Geen, and M. Bazerman (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science 62*(5), 1225–1234.

Bohren, J. A., P. Hull, and A. Imas (2025). Systemic discrimination: Theory and measurement. *Quarterly Journal of Economics 140*(3), 1743–1799.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics 145*, 27–41.

Boring, A., K. Ottoboni, and P. B. Stark (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review 97*(3), 818–827.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020). Are referees and editors in economics gender neutral? *Quarterly Journal of Economics 135*(1), 269–327.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*(3), 817–869.

Coffman, K. B., C. L. Exley, and M. Niederle (2021). The role of beliefs in driving gender discrimination. *Management Science 67*(6), 3551–3569.

Eberhardt, M., G. Facchini, and V. Rueda (2023). Gender differences in reference letters: Evidence from the economics job market. *Economic Journal 133*(655), 2676–2708.

Engelmann, D. and M. Strobel (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review 94*(4), 857–869.

Fischbacher, U., D. Kübler, and R. Stüber (2024). Betting on diversity – Occupational segregation and gender stereotypes. *Management Science 70*(8), 5502–5516.

Genetin, B., J. Chen, V. Kogan, and A. Kalish (2022). Mitigating implicit bias in student evaluations: A randomized intervention. *Applied Economic Perspectives and Policy 44*(1), 110–128.

Gioia, F. and G. Immordino (2024). Preference-based hiring decisions and incentives. *Journal of the Economic Science Association 10*(1), 1–14.

Goldin, C. and C. Rouse (2000). Orchestrating impartiality: The impact of blind auditions on female musicians. *American Economic Review 90*(4), 715–741.

Hochleitner, A., F. Tufano, G. Facchini, V. Rueda, and M. Eberhardt (2025). How tinted are your glasses? Gender views, beliefs and recommendations in hiring. *NHH Dept. of Economics Discussion Paper* (07).

Hoorens, V., G. Dekkers, and E. Deschrijver (2021). Gender bias in student evaluations of teaching: Students' self-affirmation reduces the bias by lowering evaluations of male professors. *Sex Roles 84*(1), 34–48.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes 67*(3), 247–257.

Hsee, C. K., G. F. Loewenstein, S. Blount, and M. H. Bazerman (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin 125*(5), 576–590.

Jawahar, I. and C. R. Williams (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology 50*(4), 905–925.

Johannesen, N. and S. Muchardt (2026). Is the bar higher for female scholars? Evidence from career steps in economics. *Journal of Public Economics 257*, 105611.

MacNell, L., A. Driscoll, and A. N. Hunt (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education 40*(4), 291–303.

Mengel, F., J. Sauermann, and U. Zölitz (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association 17*(2), 535–566.

Moss-Racusin, C. A., J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences 109*(41), 16474–16479.

Özgümüs, A., H. Rau, S. Trautmann, and C. König-Kersting (2020). Gender bias in the evaluation of teaching materials. *Frontiers in Psychology 11*, 1074.

Peterson, D. A., L. A. Biederman, D. Andersen, T. M. Ditonto, and K. Roe (2019). Mitigating gender bias in student evaluations of teaching. *PLOS One 14*(5), e0216241.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–661.

Stroube, B. K. (2022). Economic consequences and the motive to discriminate. *Administrative Science Quarterly 67*(1), 207–236.

A. Randomization Balance

Table 10 reports means of the predetermined covariates by treatment and by pronoun condition, with per-covariate balance tests and an omnibus joint-orthogonality test. Consistent with successful randomization, no covariate differs detectably across treatments or pronoun conditions and the joint tests do not reject balance.

Table 10: Covariate balance across treatments and pronoun conditions

	Treatment				Pronoun		
Covariate	No cons.	Baseline	Incent.	p	Invisible	Visible	p
Age (years)	25.10	24.93	24.62	0.115	24.86	24.91	0.807
Male (share of female and male raters)	0.38	0.39	0.37	0.772	0.37	0.39	0.228
Field of study (shares)				0.758			0.924
Economics	0.25	0.25	0.28		0.27	0.25	
Social sciences	0.08	0.09	0.07		0.08	0.08	
Psychology	0.04	0.04	0.06		0.04	0.05	
Natural sciences	0.18	0.18	0.18		0.18	0.18	
Computer science	0.04	0.04	0.03		0.04	0.04	
Medicine / health care	0.06	0.07	0.06		0.06	0.06	
Humanities	0.14	0.15	0.13		0.15	0.13	
Law	0.12	0.13	0.12		0.13	0.12	
Other field	0.07	0.06	0.06		0.06	0.07	
Not a student	0.01	0.00	0.00		0.00	0.01	
Degree stage (shares)				0.979			0.760
Bachelor	0.44	0.46	0.47		0.45	0.46	
Master	0.32	0.30	0.31		0.31	0.31	
State examination	0.16	0.16	0.16		0.17	0.16	
Doctorate	0.05	0.05	0.04		0.05	0.05	
Other	0.03	0.02	0.02		0.02	0.03	
Took a Microeconomics course (share)	0.37	0.37	0.38	0.868	0.39	0.36	0.146
Prior tutoring experience (share)	0.21	0.20	0.19	0.517	0.20	0.19	0.555
HS mathematics grade (1=best, 6=worst)	2.13	2.13	2.13	0.996	2.15	2.11	0.379
Joint orthogonality (χ^2)		–		0.980		–	0.941

Notes: Cell entries are covariate means (for shares, the proportion equal to one). Per-covariate p-values are F-tests from a regression of the covariate on the assignment indicator (OLS, heteroskedasticity-robust standard errors; the two-level pronoun contrast is equivalent to a two-sample t-test). For field of study and degree stage, the table reports the share of raters in each category; their p-values are Pearson χ^2 tests of the category distribution across treatments or pronoun conditions. The final row is an omnibus orthogonality test: the joint significance (χ^2) of all covariates, including field-of-study and degree-stage indicators, in a model predicting assignment – a multinomial logit for the three treatments and a logit for the two pronoun conditions. Missing covariate values are handled pairwise in the per-covariate rows and listwise in the omnibus test; in particular, the male share is computed among the 2,601 raters who report female or male gender, and tutoring experience excludes the raters who preferred not to answer.

B. Additional Belief Analyses

Table 11 relates the stated belief to predetermined rater characteristics and to the randomized conditions. Column 1 uses the belief itself and column 2 an indicator for answering "cannot judge." Figure 5 shows the distribution over the belief categories separately for female and male raters.

Table 11: Correlates of the belief

	(1) Belief	(2) Cannot judge
Male rater	-0.534***	0.027
	(0.064)	(0.019)
Age	-0.002	-0.002
	(0.006)	(0.002)
Took Microeconomics	0.077	-0.161***
	(0.081)	(0.023)
Tutoring experience	0.137*	-0.012
	(0.080)	(0.024)
Math grade (1 = excellent)	0.000	0.016*
	(0.035)	(0.009)
Pronoun visible	0.031	0.009
	(0.063)	(0.018)
No Consequences	0.058	-0.022
	(0.075)	(0.022)
Incentives	0.030	-0.030
	(0.075)	(0.022)
Female first	-0.030	-0.016
	(0.064)	(0.018)
Observations	1,741	2,568
R^2	0.074	0.055

Notes: OLS (column 1) and linear probability model (column 2), heteroskedasticity-robust standard errors in parentheses. Dependent variables: the belief about which gender explains Microeconomics better in general (-4 = males much better to 4 = females much better, 0 = no difference; column 1); an indicator for answering "cannot judge" (column 2). Column 1 uses the raters who state a belief, column 2 all raters; raters who did not report female or male gender (79) or who preferred not to report their tutoring experience are not included. Both columns include field-of-study, degree-stage, and university fixed effects (not reported). Math grade is the high-school grade (1 = excellent to 6 = unsatisfactory). Treatment coefficients are relative to the Baseline treatment. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 12 reports the linear relation between the stated belief and the rating gap, pooled with a visibility × belief interaction (column 1) and separately within each pronoun condition (columns 2 and 3).

Table 13 reports the effect of pronoun visibility on the hypothetical hiring choice within each belief category.

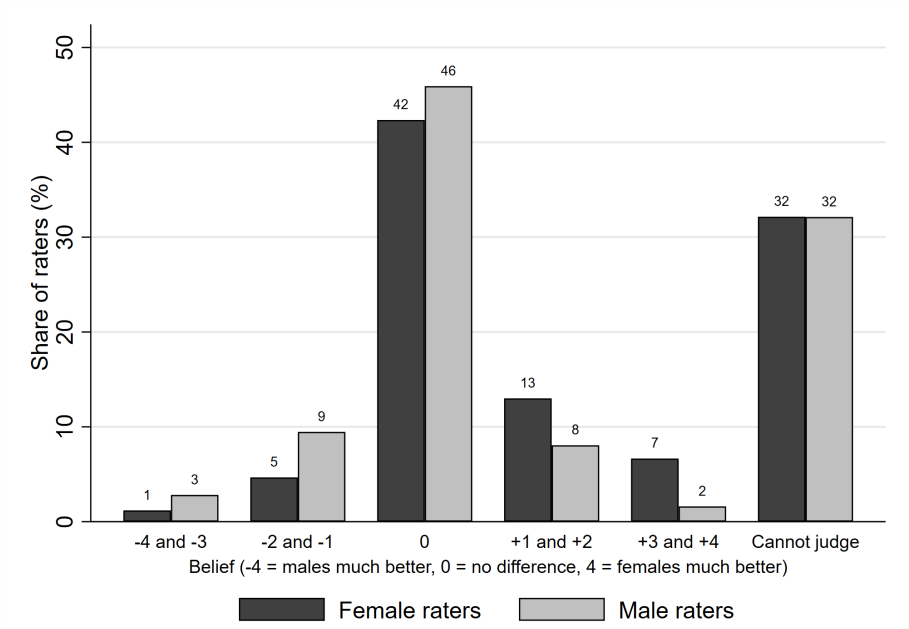


Figure 5: Distribution of stated beliefs by rater gender: share of female and male raters in each belief category (in percent). Raters who did not report female or male gender are not included.

Table 12: Rating difference on belief, by pronoun condition (composite)

	(1) Pooled	(2) Pronoun invisible	(3) Pronoun visible
Pronoun visible	0.146**		
	(0.070)		
Belief	-0.015	-0.015	0.163***
	(0.041)	(0.041)	(0.041)
Pronoun visible × Belief	0.177***		
	(0.058)		
Female first	-0.158**	-0.191*	-0.124
	(0.070)	(0.102)	(0.098)
Constant	0.023	0.088	0.097
	(0.080)	(0.105)	(0.093)
Observations	1,805	922	883
R^2	0.019	0.007	0.025

Notes: OLS, robust SE. Dependent variable: the female-minus-male rating difference in the composite rating. Column 1 pools both conditions and interacts pronoun visibility with belief; columns 2 and 3 report the belief slope separately within each condition. Belief indicates which gender explains Microeconomics better in general (-4 = males much better to 4 = females much better, 0 = no difference). Treatment fixed effects included. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 13: Effect of pronoun visibility on the hiring choice, by belief category

	(1) Female applicant	(2) Male applicant	(3) Random draw
Visible: belief −4 and −3	0.033	-0.072	0.039
	(0.127)	(0.145)	(0.129)
Visible: belief −2 and −1	-0.176**	0.053	0.123**
	(0.073)	(0.075)	(0.060)
Visible: belief 0	0.023	-0.044	0.021
	(0.028)	(0.029)	(0.025)
Visible: belief +1 and +2	0.271***	-0.184***	-0.087**
	(0.055)	(0.054)	(0.043)
Visible: belief +3 and +4	0.187**	-0.169**	-0.017
	(0.086)	(0.082)	(0.073)
Visible: cannot judge	0.027	-0.070**	0.044
	(0.033)	(0.032)	(0.030)
Observations	2,680	2,680	2,680

Notes: Linear probability models, heteroskedasticity-robust standard errors in parentheses. Dependent variable: an indicator for choosing the female applicant (column 1), the male applicant (column 2), or the random draw (column 3) in the hypothetical hiring decision. Each row is the effect of pronoun visibility on that probability within the belief category; the three effects in a row sum to zero. Belief indicates which gender explains Microeconomics better in general (−4 = males much better to 4 = females much better, 0 = no difference); raters who answered "cannot judge" form a separate category. All columns include category intercepts, treatment intercepts, and the presentation-order control. Exploratory analysis, not pre-registered. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 14 interacts pronoun visibility with both the belief and rater gender in the same specification, for both rating measures.

Table 14: Belief or rater gender: interaction of pronoun visibility with both

	(1) Composite rating	(2) Overall assessment
Visible × Belief	0.160***	0.256***
	(0.059)	(0.070)
Visible × Male rater	-0.084	-0.051
	(0.142)	(0.167)
Pronoun visible	0.175*	0.177
	(0.096)	(0.110)
Belief	-0.018	-0.073
	(0.042)	(0.049)
Male rater	-0.051	-0.044
	(0.104)	(0.121)
Observations	1,765	1,765
R^2	0.019	0.024

Notes: OLS, heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male difference in the composite rating (column 1) and the overall assessment (column 2). Both columns interact pronoun visibility with the belief and with a male-rater indicator in the same specification and include treatment intercepts and the presentation-order control. Sample: raters who state a belief and report female or male gender. Belief indicates which gender explains Microeconomics better in general (-4 = males much better to 4 = females much better, 0 = no difference). Exploratory analysis, not pre-registered. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C. Overall-Assessment Results

The main text presents every estimated result for the composite rating R. This appendix reproduces the same regression tables and figures for the overall assessment O. The estimates are qualitatively identical to their composite-rating counterparts throughout, so we present them here without repeating the discussion. Each figure and table below names its composite-rating counterpart in the main text or in Appendix B. (The descriptive statistics are already reported for both measures in Table 3.)

C.1. Gender Bias Within Treatments

Figure 6 and Table 15 are the overall-assessment counterparts of Figure 1 and Table 4: the reverse bias is positive in every treatment and individually significant in *Baseline* (the visibility effect there is 0.33 points; $p < 0.01$, WRS).

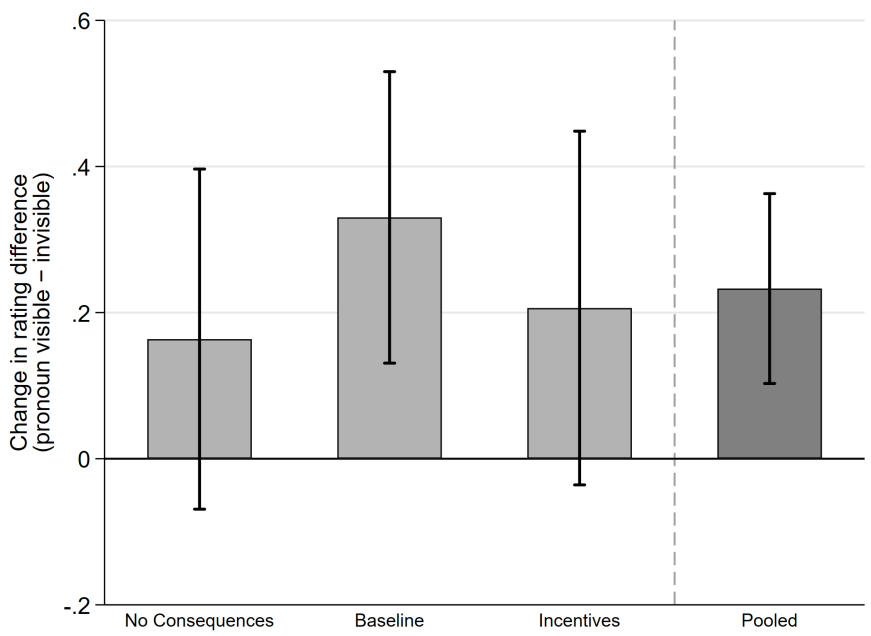


Figure 6: Reverse gender bias by treatment (overall assessment): the change in the female-minus-male gap from the pronoun-invisible to the pronoun-visible condition, estimated by OLS controlling for presentation order (pooled bar: with treatment intercepts), i.e., the no-control estimates of Table 15. Error bars: 95% confidence intervals based on heteroskedasticity-robust standard errors. Overall-assessment counterpart of Figure 1.

Table 15: Effect of pronoun visibility on the rating difference (overall assessment)

	No Consequences		Baseline		Incentives		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pronoun visible	0.164	0.187	0.330***	0.335***	0.206*	0.211*	0.233***	0.237***
	(0.119)	(0.122)	(0.102)	(0.104)	(0.123)	(0.125)	(0.066)	(0.066)
Female first	-0.255**	-0.239**	-0.418***	-0.394***	-0.076	-0.084	-0.250***	-0.240***
	(0.118)	(0.120)	(0.102)	(0.103)	(0.123)	(0.123)	(0.066)	(0.066)
Treatment intercepts	No	No	No	No	No	No	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	891	891	901	901	888	888	2,680	2,680
R^2	0.007	0.030	0.029	0.044	0.004	0.031	0.011	0.021

Notes: OLS with heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male rating difference in the overall assessment. Columns 1–6 estimate the visibility effect separately within each treatment; columns 7–8 report the average visibility effect across treatments while retaining treatment intercepts, which are estimated but not reported. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 7 and Table 16 decompose the pooled effect into the female premium and the male penalty on the overall assessment (counterparts of Figure 2 and Table 5). The reversal is again driven by a female premium, but one difference worth noting is that on the overall assessment the non-parametric test detects a significant male decline ($p = 0.043$, WRS), whereas on the composite it does not.

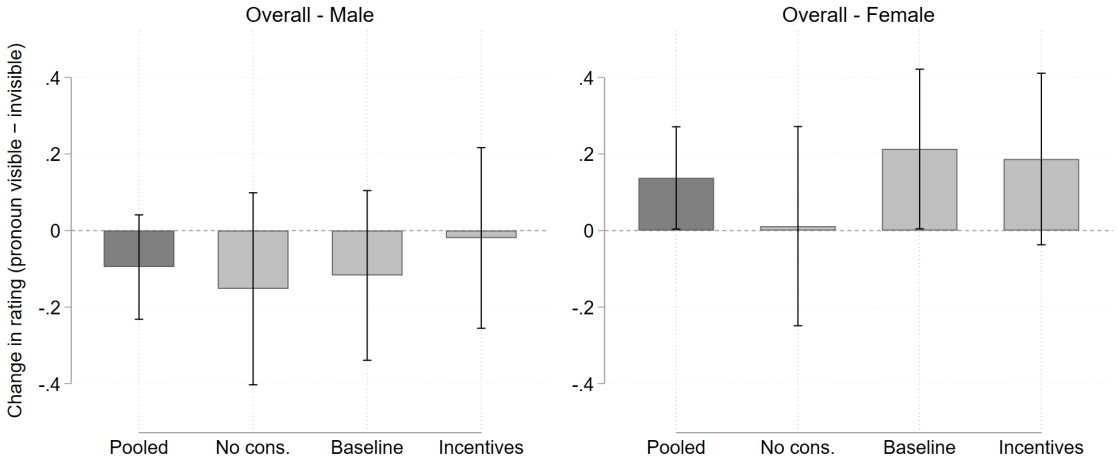


Figure 7: Effect of pronoun visibility on each applicant's overall assessment, pooled and by treatment: OLS estimates controlling for presentation order (pooled: with treatment intercepts). Error bars: 95% confidence intervals based on heteroskedasticity-robust standard errors. Overall-assessment counterpart of Figure 2.

Table 16: Effect of pronoun visibility on each applicant's rating (overall assessment)

	Male applicant		Female applicant	
	(1)	(2)	(3)	(4)
Pronoun visible	-0.095	-0.083	0.137**	0.154**
	(0.070)	(0.068)	(0.068)	(0.067)
Female first	0.135*	0.137**	-0.114*	-0.103
	(0.070)	(0.068)	(0.068)	(0.067)
Demographic controls	No	Yes	No	Yes
Observations	2,680	2,680	2,680	2,680
R^2	0.013	0.068	0.018	0.064

Notes: OLS with heteroskedasticity-robust standard errors in parentheses, pooled across treatments with treatment intercepts (omitted treatment: Baseline). The dependent variable is the overall assessment given to the male or to the female applicant. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 17 is the overall-assessment counterpart of Table 6: both rater genders shift toward the female applicant and the rater-gender interaction is insignificant throughout.

Table 17: Pronoun visibility and rater gender: pooled interaction and by rater gender (overall assessment)

	Pooled interaction		Female raters		Male raters	
	(1)	(2)	(3)	(4)	(5)	(6)
Pronoun visible	0.244***	0.254***	0.244***	0.255***	0.192*	0.185*
	(0.087)	(0.087)	(0.087)	(0.087)	(0.104)	(0.105)
Male rater	0.004	0.013				
	(0.094)	(0.096)				
Pronoun visible × Male rater	-0.052	-0.065				
	(0.135)	(0.136)				
Demographic controls	No	Yes	No	Yes	No	Yes
Observations	2,601	2,601	1,608	1,608	993	993

Notes: OLS, heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male rating difference in the overall assessment. Columns 1 and 2 pool both rater genders and interact pronoun visibility with a male-rater indicator. Columns 3 to 6 estimate the visibility effect separately within each rater gender. All columns include treatment intercepts and the presentation-order control, which are estimated but not reported. All columns are restricted to raters who identify as female or male, excluding those who selected diverse or prefer not to say. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.2. The Impact of Hiring Consequences and Rater Incentives

Table 18 is the overall-assessment counterpart of Table 7: neither the interaction of hiring consequences (-0.166) nor the interaction of rater incentives (-0.121) with *Pronoun visible* is significant.

Table 19 is the overall-assessment counterpart of Table 8 and confirms the same pattern: the treatments shift the level of ratings substantially (-0.480 and -0.127 relative to *Baseline*) while pronoun visibility does not, raising only the female applicant's rating.

C.3. The Hypothetical Hiring Decision

The pre-registered consistency check reported in the main text on the composite ranking is even sharper on the overall assessment: among the 1,866 raters who strictly rank one applicant above the other on the overall assessment, 82.2% choose the higher-rated applicant, 14.1% defer to the random draw, and only 3.7% choose the lower-rated applicant.

C.4. Mechanisms: Gender-Related Beliefs and Other Channels

Table 20 and Figure 8 are the overall-assessment counterparts of Table 9 and Figure 4. The effect of gender visibility again differs significantly across belief categories ($p < 0.001$) and is concentrated among raters who believe that women explain better (0.81 and 0.95 points), while it is small and insignificant among raters with a zero belief (0.11, $p = 0.27$). Table 21 is the overall-assessment counterpart of Table 12. As with the composite, the belief predicts the

Table 18: Interaction of pronoun visibility with hiring consequences (H2) and rater incentives (H3) (overall assessment)

	H2		H3	
	(1)	(2)	(3)	(4)
Pronoun visible	0.329***	0.330***	0.328***	0.329***
	(0.102)	(0.103)	(0.102)	(0.103)
No Consequences	-0.026	-0.024		
	(0.111)	(0.112)		
Incentives			-0.088	-0.080
			(0.115)	(0.115)
Visible × No cons.	-0.166	-0.147		
	(0.156)	(0.159)		
Visible × Incent.			-0.121	-0.116
			(0.160)	(0.161)
Female first	-0.337***	-0.319***	-0.248***	-0.244***
	(0.078)	(0.078)	(0.080)	(0.080)
Constant	0.079	-0.068	0.036	-0.121
	(0.083)	(0.250)	(0.084)	(0.247)
Demographic controls	No	Yes	No	Yes
Observations	1,792	1,792	1,789	1,789
R^2	0.017	0.029	0.014	0.029

Notes: OLS with heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male rating difference in the overall assessment. The omitted treatment is Baseline, so the interaction is the difference in the pronoun-visibility effect between treatments: consequences (H2) or incentives (H3). Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 19: Effect of treatment on the level of ratings (overall assessment)

	Average		Male appl.		Female appl.	
	(1)	(2)	(3)	(4)	(5)	(6)
No Consequences	-0.480***	-0.478***	-0.426***	-0.428***	-0.534***	-0.529***
	(0.076)	(0.074)	(0.085)	(0.084)	(0.085)	(0.084)
Incentives	-0.127*	-0.136**	-0.053	-0.069	-0.202***	-0.203***
	(0.070)	(0.068)	(0.082)	(0.081)	(0.078)	(0.077)
Pronoun visible	0.021	0.036	-0.095	-0.083	0.137**	0.154**
	(0.060)	(0.059)	(0.070)	(0.068)	(0.068)	(0.067)
Female first	0.011	0.017	0.135*	0.137**	-0.114*	-0.103
	(0.060)	(0.059)	(0.070)	(0.068)	(0.068)	(0.067)
Demographic controls	No	Yes	No	Yes	No	Yes
Observations	2,680	2,680	2,680	2,680	2,680	2,680
R^2	0.017	0.079	0.013	0.068	0.018	0.064

Notes: OLS with heteroskedasticity-robust standard errors in parentheses. The dependent variable is the level of the overall assessment, not the female-minus-male difference: in columns 1–2 the rater's average rating across the two applicants (one observation per rater), in columns 3–6 the rating given to the male and to the female applicant respectively. The omitted treatment is Baseline. Controls (predetermined): age, rater gender, field of study, degree stage, exposure to microeconomics, tutoring experience, and high-school math grade; non-response categories of gender and tutoring experience enter as separate indicators, so all raters are retained in the controlled columns. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

gap only when gender is observable, and the visibility×belief interaction is positive and highly significant.

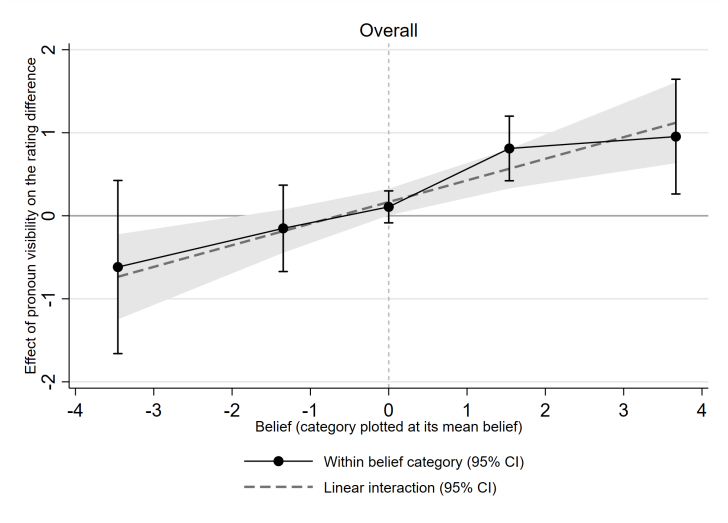


Figure 8: Effect of pronoun visibility by belief category (overall assessment). Overall-assessment counterpart of Figure 4; the dots are the column-2 estimates of Table 20, the dashed line the column-1 prediction.

Table 20: Effect of pronoun visibility by belief category (overall assessment)

	(1) Linear	(2) Categories	(3) + Cannot judge
Pronoun visible	0.166^{**}		
	(0.082)		
Pronoun visible × Belief	0.260^{***}		
	(0.068)		
Visible: belief −4 and −3		-0.618	-0.615
		(0.532)	(0.531)
Visible: belief −2 and −1		-0.151	-0.151
		(0.265)	(0.265)
Visible: belief 0		0.107	0.106
		(0.098)	(0.098)
Visible: belief +1 and +2		0.811^{***}	0.813^{***}
		(0.198)	(0.198)
Visible: belief +3 and +4		0.954^{***}	0.955^{***}
		(0.352)	(0.352)
Visible: cannot judge			0.210^{*}
			(0.115)
Observations	1,805	1,805	2,680
R^2	0.024	0.028	0.022

Notes: OLS, heteroskedasticity-robust standard errors in parentheses. Dependent variable: the female-minus-male difference in the overall assessment. Otherwise as Table 9. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 21: Rating difference on belief, by pronoun condition (overall)

	(1) Pooled	(2) Pronoun invisible	(3) Pronoun visible
Pronoun visible	0.166**		
	(0.082)		
Belief	-0.064	-0.064	0.198***
	(0.046)	(0.046)	(0.049)
Pronoun visible × Belief	0.260***		
	(0.068)		
Female first	-0.256***	-0.296**	-0.217*
	(0.081)	(0.115)	(0.113)
Constant	-0.007	0.093	0.050
	(0.093)	(0.119)	(0.116)
Observations	1,805	922	883
R^2	0.024	0.010	0.032

Notes: Overall-assessment counterpart of Table 12 in Appendix B. OLS, robust SE. Dependent variable: the female-minus-male rating difference in the overall assessment. Column 1 pools both conditions and interacts pronoun visibility with belief; columns 2 and 3 report the belief slope separately within each condition. Belief indicates which gender explains Microeconomics better in general (-4 = males much better to 4 = females much better, 0 = no difference). Treatment fixed effects included. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.5. Robustness

Table 22: Robustness: university fixed effects, excluding Bonn, and Bonn heterogeneity (composite rating)

	(1) University FE	(2) Drop Bonn	(3) Bonn interaction
Pronoun visible	0.202***	0.175**	0.176**
	(0.058)	(0.082)	(0.082)
Female first	-0.169***	-0.171**	-0.169***
	(0.058)	(0.082)	(0.058)
Bonn			-0.047
			(0.083)
Visible × Bonn			0.049
			(0.116)
Observations	2,680	1,232	2,680
R^2	0.009	0.012	0.009

Notes: OLS, robust SE. Dependent variable: the female-minus-male rating difference in the composite rating. Treatment fixed effects included. Column 1 adds university fixed effects; column 2 excludes Bonn; column 3 interacts pronoun visibility with a Bonn indicator, so the pronoun coefficient is the effect outside Bonn and the interaction is the Bonn differential. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 23: Robustness: university fixed effects, excluding Bonn, and Bonn heterogeneity (overall assessment)

	(1) University FE	(2) Drop Bonn	(3) Bonn interaction
Pronoun visible	0.233***	0.161*	0.160*
	(0.066)	(0.093)	(0.093)
Female first	-0.250***	-0.276***	-0.250***
	(0.066)	(0.093)	(0.066)
Bonn			-0.065
			(0.094)
Visible × Bonn			0.135
			(0.132)
Observations	2,680	1,232	2,680
R^2	0.012	0.013	0.012

Notes: OLS, robust SE. Dependent variable: the female-minus-male rating difference in the overall assessment. Treatment fixed effects included. Column 1 adds university fixed effects; column 2 excludes Bonn; column 3 interacts pronoun visibility with a Bonn indicator, so the pronoun coefficient is the effect outside Bonn and the interaction is the Bonn differential. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

D. Instructions

The following instructions are a translation from German. The original instructions are available from the authors upon request.

Welcome to this study!

Please read the following information about the study and how we use your data carefully.

1. Purpose of the Study
   The goal of this study is to collect evaluations of teaching materials. Your participation in this study will help researchers gain insights into human decision-making.

2. Voluntary Participation
   Participation is voluntary. You may withdraw your consent at any time up to 3 months after participating in the study by contacting [blinded for review]. You do not need to provide a reason for this decision.

3. Procedure
   The study takes about 10 to 15 minutes. At the beginning, you will need to complete a few attention tests in order to proceed. You will receive additional instructions for each decision and answer a few comprehension questions. The study does not involve any deception. The study poses no risks beyond those of normal daily life. Aside from your compensation, there are no known benefits to you from participating in this study.

4. Payment
   You must complete the entire study to receive payment. All participants will receive a fixed payment of 4 euros. In addition, under certain randomly assigned conditions, you may receive a bonus of 1 euro. Do not switch to another browser tab or window during the study, as this may invalidate the study and result in you not receiving compensation.

5. Data and Confidentiality
   As part of this study, we collect your email address, as well as your name, address, and bank account information to process your payment. This information is required to pay you via bank transfer. Your personal data will be processed solely for the purpose of making the payment and will not be used for any other purposes. The research data collected during the experiment will be analyzed in an anonymized form. Results will be processed and published exclusively in an anonymized form, so that no conclusions can be drawn about your identity. For your security, your data will be transmitted to us in encrypted form.

6. Research Group
   The research group consists of [blinded for review].

Do you agree to participate in the study under the conditions described in items 1, 2, 3, 4, 5, and 6?

○ I agree

○ I DO NOT agree

Your email address

The version reproduced here is the one used for invitations through university participant databases; for invitations via e-learning platforms, only the eligibility text on this screen differed.

Please note that you may only participate in this study if you are enrolled at a German university. Furthermore, you must be registered in a university's participant database and have received an invitation email for this study through that database.

In addition, you may not participate in this study more than once.

Please enter the email address you used to register in the participant database in the field below.

A payment can only be guaranteed if you enter the correct email address here.

Email address: _____

Your bank account information

All payments made as part of this study will be processed via bank transfer.

For legal reasons, in order to transfer your compensation for this experiment, we will need your name and address in addition to your account's IBAN (which must be from the SEPA region) and your bank's BIC.

If you do not wish to provide us with this information, you will unfortunately not be able to participate in this experiment.

IBAN: _____

BIC (only for foreign accounts): _____

First Name: _____

Last Name: _____

Street and house number: _____

City: _____

ZIP code: _____

○ I confirm that, during the current calendar year, I will receive less than 1,500 euros in income from a university outside an employment relationship. Should this limit be exceeded during the calendar year, contrary to expectations, I agree to report this to the appropriate office at my university. I am aware that any income earned here may be subject to reporting requirements in connection with tax returns, scholarships, BaföG, and similar matters. I will take it upon myself to ensure that this is properly verified.

In studies like ours, there are sometimes participants who do not read the text questions carefully and simply click their way through the study quickly. This affects the results of research studies.

Please indicate whether you agree or disagree with each of the following four statements. If you fail a screening, you will not be able to participate.

	Disagree	**Agree**
No human on Earth has hands.	○	○
No U.S. citizen has ever driven a car.	○	○
There are dogs on Earth.	○	○
Not a single person was born in the year 2000.	○	○

Introduction

We recommend that you maximize your browser window before you begin.

This questionnaire consists of two parts. In each part, you will have the opportunity to evaluate a set of slides from the Microeconomics course.

Each set of slides was submitted by an applicant as a teaching demonstration as part of the current application process for open teaching assistant positions in Microeconomics I at the

Faculty of [blinded for review]. These are teaching demonstrations on the topic of consumer decisions.

The applicants received a PowerPoint template from the chair in which the problems and their respective solutions were already included. Their task was to add explanations to the solutions.

You evaluate a teaching demonstration by answering a question about your overall impression of the explanations, as well as additional detailed questions concerning individual aspects of the teaching demonstration (e.g., clarity, structure).

The Applicant

Displayed if: Pronoun Visibility = No

The applicant has included their first and last names as well as their pronouns ("she/her" or "he/him") on the title page of the presentation and in the footer of every page.

We have redacted the first and last names as well as the pronoun.

Displayed if: Pronoun Visibility = Yes

The applicant has included their first and last names as well as their pronouns ("she/her" or "he/him") on the title page of the presentation and in the footer of each page.

We have redacted the first and last names.

Selection Decision

Displayed in NO CONSEQUENCES

Please note that the selection of tutors is made independently of your evaluations.

Your evaluations therefore have no consequences whatsoever for the applicant.

Displayed in BASELINE AND INCENTIVES

Please note that we use your ratings to select a tutor.

Ultimately, the participants' ratings for both applicants are compared. A random draw will determine whether the overall impression or the average of the detailed questions is used for the comparison. The applicant with the higher average rating will be selected to offer a tutorial once or twice a week during the upcoming summer semester of 2026, in conjunction with the Microeconomics I lecture. In the event of a tie, the decision will be made at random.

Accordingly, each of your evaluations can have a direct impact on the candidate whose set of slides you are evaluating.

Displayed in INCENTIVES

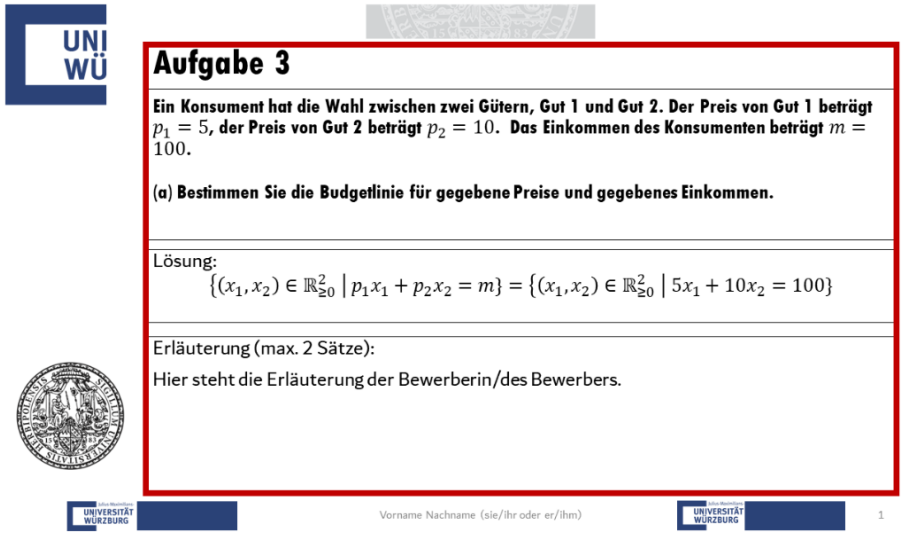
Expert Committee

In addition, the applicant's set of slides will be submitted to an independent expert committee for evaluation.

This independent expert committee consists of six expert representatives from various groups, selected on an equal basis (one female and one male professor, one female and one male member of the academic staff, and one female and one male former tutor).

Only the content presented by the applicant, and not any other information contained on the slides, is submitted to the expert committee for evaluation.

If the sample slide below were part of an actual set of slides submitted by an applicant, the expert committee would only be presented with the area framed in red for evaluation.



Displayed in INCENTIVES

Your Bonus

The expert committee's ratings do not influence the decision among applicants. However, they are relevant for calculating your bonus. The bonus you receive depends on how close your rating is to the expert committee's average rating.

To determine this, one of the two sections and one of the questions you answered are selected at random. The expert committee's average rating for that question is rounded to the nearest whole number (2.49 is rounded down to 2; 8.50 is rounded up to 9; etc.).

If your rating matches the expert committee's rounded rating, you will receive a bonus of 1 euro in addition to the fixed payment of 4 euros.

Comprehension Questions

Each question was shown on a separate screen, with a button ("Show/Hide Instructions") that displayed the corresponding instructions again. Participants could only proceed after selecting the correct answer, marked here with an asterisk (*).

Displayed if Pronoun Visibility = 0 (correct: first option) and Pronoun Visibility = 1 (correct: second option)

Which statement regarding the applicant's personal information is correct?

- ○ We have redacted the first and last names as well as the pronoun.
- ○ We have redacted only the first and last names.

Displayed in all treatments (correct: first option in NO CONSEQUENCES, second option in BASELINE and INCENTIVES)

Which statement regarding the consequences of your evaluations for the selection decision is correct?

- ○ The selection is independent of my evaluations, so they have no consequences whatsoever for the applicant.
- ○ My evaluations are used for the selection decision, so they can have a direct impact on the applicant.

Displayed in INCENTIVES

Which statement regarding the expert committee is correct?

- ○ Only the applicant's explanations, but not personal details or other identifying information, are submitted to the expert committee for evaluation.*
- ○ The applicant's explanations and personal details are submitted to the expert committee for evaluation.

Displayed in INCENTIVES

Which statement regarding your bonus is correct?

- ○ I cannot receive a bonus.
- ○ I receive a bonus of 1 euro if my rating for a randomly selected question matches the expert committee's rounded average rating for that question.*
- ○ I receive a bonus of 1 euro for sure.

Your Evaluation

Please take your time to review the candidate's set of slides first. Afterward, answer a series of questions to evaluate the teaching demonstration.

Please take your time to review all the slides.

Displayed if Pronoun Visibility = 0 AND Female First = 1

Mikroökonomik I

Konsumentenentscheidungen

[redacted]

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze):

Die maximale Menge von Gut 1 x_1^{max} wird berechnet, indem das gesamte verfügbare Einkommen m nur für dieses Gut x_1 ausgegeben wird. In diesem Fall teilt man das Einkommen m= 100 durch den Preis p_1 = 5, was einen maximalen Konsum von 20 Einheiten ergibt.

[redacted]

2

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze):

Die Budgetlinie beschreibt alle Güterbündel, bei denen das Einkommen m durch die Gesamtausgaben (p_1*x_1 + p_2*x_2) vollständig ausgeschöpft wird. Durch die Umstellung nach x_2 wird eine Steigung von - ½ deutlich, welche die Opportunitätskosten von Gut 1 widerspiegelt, da für jede zusätzliche Einheit dieses Gutes auf eine halbe Einheit von Gut 2 verzichtet werden muss.

[redacted]

3

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze):

Die Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$ wird zunächst mit dem konstanten Nutzenniveau $\overline{U}$ gleichgesetzt und anschließend nach x_2 aufgelöst. Die daraus resultierende Funktion beschreibt alle Güterbündel, die dem Konsumenten denselben Nutzen stiften. In diesem Fall stellen sich die Indifferenzkurven als Gerade mit einer Steigung von -½ dar, deren vertikaler Achsenabschnitt durch das Nutzenniveau ($\overline{U}/4$) bestimmt wird.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze):

Die Grenzrate der Substitution (GRS) entspricht dem Absolutwert der Steigung und gibt das individuelle Tauschverhältnis an, zu dem ein Konsument bereit ist, Gut 2 durch Gut 1 zu ersetzen. Dies ergibt sich aus dem Quotienten der ersten Ableitungen nach x_1 und x_2, welcher hier konstant bei ½ liegt. Da die Indifferenzkurve in diesem Fall eine Gerade mit konstanter Steigung ist, ist die GRS unabhängig vom betrachteten Güterbündel.

Displayed if Pronoun Visibility = 1 AND Female First = 1

Mikroökonomik I

Konsumentenentscheidungen

[redacted]
(sie/ihr)

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze):

Die maximale Menge von Gut 1 x_1^{max} wird berechnet, indem das gesamte verfügbare Einkommen m nur für dieses Gut x_1 ausgegeben wird. In diesem Fall teilt man das Einkommen m= 100 durch den Preis p_1 = 5, was einen maximalen Konsum von 20 Einheiten ergibt.

[redacted] (sie/ihr) 2

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze):

Die Budgetlinie beschreibt alle Güterbündel, bei denen das Einkommen m durch die Gesamtausgaben (p_1*x_1 + p_2*x_2) vollständig ausgeschöpft wird. Durch die Umstellung nach x_2 wird eine Steigung von - ½ deutlich, welche die Opportunitätskosten von Gut 1 widerspiegelt, da für jede zusätzliche Einheit dieses Gutes auf eine halbe Einheit von Gut 2 verzichtet werden muss.

[redacted] (sie/ihr) 3

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze):

Die Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$ wird zunächst mit dem konstanten Nutzenniveau $\overline{U}$ gleichgesetzt und anschließend nach x_2 aufgelöst. Die daraus resultierende Funktion beschreibt alle Güterbündel, die dem Konsumenten denselben Nutzen stiften. In diesem Fall stellen sich die Indifferenzkurven als Gerade mit einer Steigung von -½ dar, deren vertikaler Achsenabschnitt durch das Nutzenniveau ($\overline{U}/4$) bestimmt wird.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze):

Die Grenzrate der Substitution (GRS) entspricht dem Absolutwert der Steigung und gibt das individuelle Tauschverhältnis an, zu dem ein Konsument bereit ist, Gut 2 durch Gut 1 zu ersetzen. Dies ergibt sich aus dem Quotienten der ersten Ableitungen nach x_1 und x_2, welcher hier konstant bei ½ liegt. Da die Indifferenzkurve in diesem Fall eine Gerade mit konstanter Steigung ist, ist die GRS unabhängig vom betrachteten Güterbündel.

Displayed if Pronoun Visibility = 0 AND Female First = 0

Mikroökonomik I

Konsumentenentscheidungen

[redacted]

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze): Um die maximale Menge von Gut 1 zu bestimmen, geht man davon aus, dass der Konsument sein gesamtes Einkommen für Gut 1 ausgibt. Daher wird das Einkommen des Konsumenten von 100 durch den Preis von Gut 1 (p_1 = 5) dividiert.

[redacted] 2

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze): Die Budgetlinie ist gegeben durch $p_1x_1 + p_2x_2 = m$; setzt man p_1 = 5, p_2 = 10 und m = 100 ein, ergibt sich $5x_1 + 10x_2$ = 100, welche zur Bestimmung der Steigung nach x_2 aufgelöst wird.

Die Steigung ist der Koeffizient vor x_1 und entspricht $-\frac{p_1}{p_2} = -\frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet (Opportunitätskosten).

[redacted] 3

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze): Um die Indifferenzkurve für ein Nutzenniveau $\overline{U}$ zu bestimmen, setzt man $u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U}$. Danach löst man die Gleichung nach auf x_2 auf. Das Ergebnis zeigt, wie viel von Gut 2 der Konsument konsumieren muss, um bei einem festen Nutzenwert $\overline{U}$ mit unterschiedlicher Menge von Gut 1 indifferent zu bleiben.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze): Die Grenzrate der Substitution im Güterbündel (x_1, x_2) erhält man, indem man die partielle Ableitung der Nutzenfunktion nach x_1 durch die partielle Ableitung nach x_2 dividiert.

Hier ergibt sich $GRS(x_1, x_2) = \frac{2}{4} = \frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet.

Displayed if Pronoun Visibility = 1 AND Female First = 0

Mikroökonomik I

Konsumentenentscheidungen

[redacted]

(er/ihm)

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze): Um die maximale Menge von Gut 1 zu bestimmen, geht man davon aus, dass der Konsument sein gesamtes Einkommen für Gut 1 ausgibt. Daher wird das Einkommen des Konsumenten von 100 durch den Preis von Gut 1 (p_1 = 5) dividiert.

[redacted] (er/ihm) 2

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze): Die Budgetlinie ist gegeben durch $p_1x_1 + p_2x_2 = m$; setzt man p_1 = 5, p_2 = 10 und m = 100 ein, ergibt sich $5x_1 + 10x_2$ = 100, welche zur Bestimmung der Steigung nach x_2 aufgelöst wird.

Die Steigung ist der Koeffizient vor x_1 und entspricht $-\frac{p_1}{p_2} = -\frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet (Opportunitätskosten).

[redacted] (er/ihm) 3

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze): Um die Indifferenzkurve für ein Nutzenniveau $\overline{U}$ zu bestimmen, setzt man $u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U}$. Danach löst man die Gleichung nach auf x_2 auf. Das Ergebnis zeigt, wie viel von Gut 2 der Konsument konsumieren muss, um bei einem festen Nutzenwert $\overline{U}$ mit unterschiedlicher Menge von Gut 1 indifferent zu bleiben.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze): Die Grenzrate der Substitution im Güterbündel (x_1, x_2) erhält man, indem man die partielle Ableitung der Nutzenfunktion nach x_1 durch die partielle Ableitung nach x_2 dividiert.

Hier ergibt sich $GRS(x_1, x_2) = \frac{2}{4} = \frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet.

To view the slide set again, please click the button: Show/Hide Slide Set

Please answer the following detailed questions:

	Do not agree at all									Fully agree
	1	2	3	4	5	6	7	8	9	10
The explanations for the solutions are easy to understand.	○	○	○	○	○	○	○	○	○	○
The explanations are clearly structured.	○	○	○	○	○	○	○	○	○	○
The explanations make the mathematical solution intuitively understandable.	○	○	○	○	○	○	○	○	○	○
The explanations are precise.	○	○	○	○	○	○	○	○	○	○
The slides are suitable for independent study (without attending the tutorial).	○	○	○	○	○	○	○	○	○	○

Please answer the following question regarding your overall impression:

	Do not agree at all									Fully agree
	1	2	3	4	5	6	7	8	9	10
Overall, the explanations on the slides are very well done.	○	○	○	○	○	○	○	○	○	○

Introduction to Part 2

Displayed in NO CONSEQUENCES

Part 2 of the questionnaire now begins. In Part 2, you will have the opportunity to evaluate another set of slides from the same application process.

Please note that the selection of tutors is again independent of your evaluations.

Your evaluations therefore have no consequences whatsoever for the applicant.

Displayed in BASELINE

Now Part 2 of the questionnaire begins. In Part 2, you will have the opportunity to evaluate another set of slides from the same application process.

Please note that your evaluations will again be used by us to select the most suitable tutors.

Accordingly, just as in Part 1, each of your evaluations may have a direct impact on the applicant whose set of slides you are evaluating.

Displayed in INCENTIVES

Now Part 2 of the questionnaire begins. In Part 2, you have the opportunity to evaluate another set of slides from the same application process.

Please note that we will again use your evaluations to select the most suitable applicant.

Please take your time to review all the slides.

Displayed if Pronoun Visibility = 0 AND Female First = 1

Mikroökonomik I

Konsumentenentscheidungen

[redacted]

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze): Um die maximale Menge von Gut 1 zu bestimmen, geht man davon aus, dass der Konsument sein gesamtes Einkommen für Gut 1 ausgibt. Daher wird das Einkommen des Konsumenten von 100 durch den Preis von Gut 1 (p_1 = 5) dividiert.

[redacted] 2

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze): Die Budgetlinie ist gegeben durch $p_1x_1 + p_2x_2 = m$; setzt man p_1 = 5, p_2 = 10 und m = 100 ein, ergibt sich $5x_1 + 10x_2$ = 100, welche zur Bestimmung der Steigung nach x_2 aufgelöst wird.

Die Steigung ist der Koeffizient vor x_1 und entspricht $-\frac{p_1}{p_2} = -\frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet (Opportunitätskosten).

[redacted] 3

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze): Um die Indifferenzkurve für ein Nutzenniveau $\overline{U}$ zu bestimmen, setzt man $u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U}$. Danach löst man die Gleichung nach auf x_2 auf. Das Ergebnis zeigt, wie viel von Gut 2 der Konsument konsumieren muss, um bei einem festen Nutzenwert $\overline{U}$ mit unterschiedlicher Menge von Gut 1 indifferent zu bleiben.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze): Die Grenzrate der Substitution im Güterbündel (x_1, x_2) erhält man, indem man die partielle Ableitung der Nutzenfunktion nach x_1 durch die partielle Ableitung nach x_2 dividiert.

Hier ergibt sich $GRS(x_1, x_2) = \frac{2}{4} = \frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet.

Displayed if Pronoun Visibility = 1 AND Female First = 1

Mikroökonomik I

Konsumentenentscheidungen

[redacted]

(er/ihm)

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze): Um die maximale Menge von Gut 1 zu bestimmen, geht man davon aus, dass der Konsument sein gesamtes Einkommen für Gut 1 ausgibt. Daher wird das Einkommen des Konsumenten von 100 durch den Preis von Gut 1 (p_1 = 5) dividiert.

[redacted] (er/ihm) 2

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze): Die Budgetlinie ist gegeben durch $p_1x_1 + p_2x_2 = m$; setzt man p_1 = 5, p_2 = 10 und m = 100 ein, ergibt sich $5x_1 + 10x_2$ = 100, welche zur Bestimmung der Steigung nach x_2 aufgelöst wird.

Die Steigung ist der Koeffizient vor x_1 und entspricht $-\frac{p_1}{p_2} = -\frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet (Opportunitätskosten).

[redacted] (er/ihm) 3

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze): Um die Indifferenzkurve für ein Nutzenniveau $\overline{U}$ zu bestimmen, setzt man $u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U}$. Danach löst man die Gleichung nach auf x_2 auf. Das Ergebnis zeigt, wie viel von Gut 2 der Konsument konsumieren muss, um bei einem festen Nutzenwert $\overline{U}$ mit unterschiedlicher Menge von Gut 1 indifferent zu bleiben.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze): Die Grenzrate der Substitution im Güterbündel (x_1, x_2) erhält man, indem man die partielle Ableitung der Nutzenfunktion nach x_1 durch die partielle Ableitung nach x_2 dividiert.

Hier ergibt sich $GRS(x_1, x_2) = \frac{2}{4} = \frac{1}{2}$, was bedeutet, dass der Konsument für eine zusätzliche Einheit von Gut 1 auf eine halbe Einheit von Gut 2 verzichtet.

Displayed if Pronoun Visibility = 0 AND Female First = 0

Mikroökonomik I

Konsumentenentscheidungen

[redacted]

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze):

Die maximale Menge von Gut 1 x_1^{max} wird berechnet, indem das gesamte verfügbare Einkommen m nur für dieses Gut x_1 ausgegeben wird. In diesem Fall teilt man das Einkommen m= 100 durch den Preis p_1 = 5, was einen maximalen Konsum von 20 Einheiten ergibt.

[redacted]

2

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze):

Die Budgetlinie beschreibt alle Güterbündel, bei denen das Einkommen m durch die Gesamtausgaben (p_1*x_1 + p_2*x_2) vollständig ausgeschöpft wird. Durch die Umstellung nach x_2 wird eine Steigung von - ½ deutlich, welche die Opportunitätskosten von Gut 1 widerspiegelt, da für jede zusätzliche Einheit dieses Gutes auf eine halbe Einheit von Gut 2 verzichtet werden muss.

[redacted]

3

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze):

Die Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$ wird zunächst mit dem konstanten Nutzenniveau $\overline{U}$ gleichgesetzt und anschließend nach x_2 aufgelöst. Die daraus resultierende Funktion beschreibt alle Güterbündel, die dem Konsumenten denselben Nutzen stiften. In diesem Fall stellen sich die Indifferenzkurven als Gerade mit einer Steigung von -½ dar, deren vertikaler Achsenabschnitt durch das Nutzenniveau ($\overline{U}/4$) bestimmt wird.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze):

Die Grenzrate der Substitution (GRS) entspricht dem Absolutwert der Steigung und gibt das individuelle Tauschverhältnis an, zu dem ein Konsument bereit ist, Gut 2 durch Gut 1 zu ersetzen. Dies ergibt sich aus dem Quotienten der ersten Ableitungen nach x_1 und x_2, welcher hier konstant bei ½ liegt. Da die Indifferenzkurve in diesem Fall eine Gerade mit konstanter Steigung ist, ist die GRS unabhängig vom betrachteten Güterbündel.

Displayed if Pronoun Visibility = 1 AND Female First = 0

Mikroökonomik I

Konsumentenentscheidungen

[redacted]
(sie/ihr)

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(a) Wie hoch ist die maximale Menge von Gut 1, die der Konsument bei gegebenen Preisen und gegebenem Einkommen konsumieren kann?

Lösung:

$$x_1^{max} = \frac{m}{p_1} = \frac{100}{5} = 20$$

Erläuterung (max. 2 Sätze):

Die maximale Menge von Gut 1 x_1^{max} wird berechnet, indem das gesamte verfügbare Einkommen m nur für dieses Gut x_1 ausgegeben wird. In diesem Fall teilt man das Einkommen m= 100 durch den Preis p_1 = 5, was einen maximalen Konsum von 20 Einheiten ergibt.

Aufgabe 1

Ein Konsument hat die Wahl zwischen zwei Gütern, Gut 1 und Gut 2. Der Preis von Gut 1 beträgt $p_1 = 5$, der Preis von Gut 2 beträgt $p_2 = 10$. Das Einkommen des Konsumenten beträgt $m = 100$.

(b) Bestimmen Sie die Budgetlinie. Wie hoch ist die Steigung der Budgetlinie? Wie ist diese ökonomisch zu interpretieren?

Lösung:

$$5x_1 + 10x_2 = 100 \Leftrightarrow x_2 = 10 - \frac{1}{2}x_1$$

Erläuterung (max. 2 Sätze):

Die Budgetlinie beschreibt alle Güterbündel, bei denen das Einkommen m durch die Gesamtausgaben (p_1*x_1 + p_2*x_2) vollständig ausgeschöpft wird. Durch die Umstellung nach x_2 wird eine Steigung von - ½ deutlich, welche die Opportunitätskosten von Gut 1 widerspiegelt, da für jede zusätzliche Einheit dieses Gutes auf eine halbe Einheit von Gut 2 verzichtet werden muss.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(a) Bestimmen Sie die Funktionsvorschrift der Indifferenzkurve, die Nutzenniveau $\overline{U}$ repräsentiert.

Lösung:

$$u(x_1, x_2) = 2x_1 + 4x_2 = \overline{U} \quad \Rightarrow \quad x_2 = \frac{\overline{U}}{4} - \frac{1}{2}x_1$$

Erläuterung (max. 3 Sätze):

Die Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$ wird zunächst mit dem konstanten Nutzenniveau $\overline{U}$ gleichgesetzt und anschließend nach x_2 aufgelöst. Die daraus resultierende Funktion beschreibt alle Güterbündel, die dem Konsumenten denselben Nutzen stiften. In diesem Fall stellen sich die Indifferenzkurven als Gerade mit einer Steigung von -½ dar, deren vertikaler Achsenabschnitt durch das Nutzenniveau ($\overline{U}/4$) bestimmt wird.

Aufgabe 2

Die Präferenzen eines Konsumenten über Güterbündel (x_1, x_2) werden repräsentiert von der Nutzenfunktion $u(x_1, x_2) = 2x_1 + 4x_2$.

(b) Berechnen Sie die Grenzrate der Substitution im Güterbündel (x_1, x_2).

Lösung:

$$GRS(x_1, x_2) = \frac{\frac{\partial u(x_1, x_2)}{\partial x_1}}{\frac{\partial u(x_1, x_2)}{\partial x_2}} = \frac{2}{4} = \frac{1}{2}$$

Erläuterung (max. 3 Sätze):

Die Grenzrate der Substitution (GRS) entspricht dem Absolutwert der Steigung und gibt das individuelle Tauschverhältnis an, zu dem ein Konsument bereit ist, Gut 2 durch Gut 1 zu ersetzen. Dies ergibt sich aus dem Quotienten der ersten Ableitungen nach x_1 und x_2, welcher hier konstant bei ½ liegt. Da die Indifferenzkurve in diesem Fall eine Gerade mit konstanter Steigung ist, ist die GRS unabhängig vom betrachteten Güterbündel.

To view the slide set again, please click the button: Show/Hide Slide Set

Please answer the following detailed questions:

	Do not agree at all								Fully agree
	1	2	3	4	5	6	7	8	9	10
The explanations for the solutions are easy to understand.	○	○	○	○	○	○	○	○	○	○
The explanations are clearly structured.	○	○	○	○	○	○	○	○	○	○
The explanations make the mathematical solution intuitively understandable.	○	○	○	○	○	○	○	○	○	○
The explanations are precise.	○	○	○	○	○	○	○	○	○	○
The slides are suitable for independent study (without attending the tutorial).	○	○	○	○	○	○	○	○	○	○

Please answer the following question regarding your overall impression:

	Do not agree at all								Fully agree
	1	2	3	4	5	6	7	8	9	10
Overall, the explanations on the slides are very well done.	○	○	○	○	○	○	○	○	○	○

Hypothetical Selection Decision

Imagine you could decide which applicant would be allowed to give the tutorial next semester.

Who would you choose?

○ Applicant from Part 1

○ Applicant from Part 2

○ The selection should be made by random draw.

General Assessment

Regardless of your field of study, we would like to ask you for a general assessment. This is based on your personal perception; there are no right or wrong answers.

On average: Who explains concepts better in microeconomics tutorials?

- ○ 1 Men are significantly better
- ○ 2
- ○ 3
- ○ 4
- ○ 5 No difference
- ○ 6
- ○ 7
- ○ 8
- ○ 9 Women are significantly better
- ○ I can't say

Note: In the analysis, responses are recoded to -4 *(men are significantly better) to* 4 *(women are significantly better) by subtracting* 5*.*

Personal Information

How old are you (in years)? _____

What is your gender?

- ○ Male
- ○ Female
- ○ Diverse
- ○ I prefer not to specify

In which academic field are you primarily studying?

○ Economics

○ Social Sciences

○ Psychology

○ Natural Sciences

○ Computer Science

○ Medicine/Health

○ Humanities

○ Law

○ Other

○ I am not a student

Have you already taken a microeconomics course?

○ Yes

○ No

What stage of your studies are you currently in?

○ Bachelor's

○ Master's

○ State Examination

○ Doctorate

○ Other

Do you already have experience as a tutor?

○ Yes

○ No

○ I prefer not to say

What grade did you receive in mathematics on your high school diploma?

○ Very good

○ Good

○ Satisfactory

○ Sufficient

○ Poor

○ Failing

Have you ever participated in the same or a very similar study?

○ Yes

○ No

Displayed if "Yes" was selected for question "Have you ever participated in the same or a very similar study?"

In what year did you participate in this or a very similar study?

○ 2026

○ 2025

○ 2024

○ 2023 or earlier