Int J Semiot Law
https://doi.org/10.1007/s11196-023-10055-6

Is Epistemic Status Gender-Biased? Gender As a Predictor of Testimonial Reliability Assessments in Violent Crimes

Klaudyna Horniczak[1] · **Andrzej Porębski**[1] · **Izabela Skoczeń**[2]

Accepted: 20 September 2023

Abstract

It is rather uncontroversial that gender should have no influence on treating others as equal epistemic agents. However, is this view reflected in practice? This paper aims to test whether the gender of the testifier and the accused of assault is related to the perception of a testimony's reliability and the guilt of the potential perpetrator. Two experiments were conducted: the subjects (n = 361, 47% females, 53% males) assessed the reliability of the testifier in four scenarios of assault accusation, in which the only difference was the gender of the people presented. During the study, we have observed dependencies of gender and ascription of reliability, but only marginal differences in guilt attribution. The results of our research may constitute an argument for the existence of different epistemic status endowed on people depending on their gender and existing gender stereotypes. Our results suggest that gender bias may be situated at a deeper level than the linguistically triggered representation.

Keywords Gender bias · Assault · Reliability · Epistemic agency · Gendered language · Experimental jurisprudence

✉ Klaudyna Horniczak
klaudyna.horniczak@doctoral.uj.edu.pl

Andrzej Porębski
and.porebski@uj.edu.pl

Izabela Skoczeń
izabela.skoczen@uj.edu.pl

[1] Doctoral School in the Social Sciences, Jagiellonian University, Cracow, Poland

[2] Chair of Legal Theory and Jagiellonian Centre for Law, Language and Philosophy, Jagiellonian University, Cracow, Poland

Published online: 17 January 2024

1 Introduction

1.1 Gender Bias and Epistemic Beliefs

Is gender a factor that influences our perception of the epistemic status of another person? This question may be understood in at least a twofold sense. Firstly, ascribing reliability to an assertion made by a person may be influenced by whether the speaker is a man or a woman. Secondly, it might be the case that men and women differently perceive the credibility of other people's claims. Those two understandings overlap, and so it may also be true that women believe in assertions made by men and by women differently; the same may be true about men's beliefs. In this paper, we want to establish whether there are not only gender differences, but whether there is a gender bias present in the perception of reliability of speakers, and whether gender is an important factor in ascribing epistemic status to others.

At the beginning, some terminological remarks are in order. Two concepts, namely: gender difference and gender bias should be distinguished for the purpose of the study (for exhaustive account of biases see e.g., Kahneman & Tversky [28]; Kahneman [27]). Both occur when, in given circumstances, gender is the only varying factor, and still, the assessment of reliability changes. However, not in every situation this could be problematic or wrong. The gender of participants may carry additional information, for instance about a frequency assessment of the event described. Consequently, sometimes the testifier will be (not irrationally), regarded as less reliable, because their testimony will concern an event that is highly unlikely given the described gender arrangement. For example, a man, who testifies that he has been assaulted by five women may be considered less reliable than a woman who claims that she has been assaulted by five men—only because the former situation is deemed to be extremely infrequent. If you hear about an extremely rare event, it is a natural reaction to ask for more evidence than if you hear about an event that happens every day constantly. Thus, gender difference will occur, when the assessment varies across gender in a rational way, due to an actual incidence of phenomena. On the other hand, gender bias will occur in situations, when the assessment of reliability will change, with gender being the only altered variable, due to irrational and prejudiced beliefs about men or women. Gender bias then may be influenced by social gender norms and roles, as well as gender stereotypes. Moreover, in experimental circumstances the existence of gender bias will always imply gender differences because we cannot assume occurrence of the former, while no empirical difference can be observed.

The existence of a gender bias in endowing epistemic status to speakers is exceptionally relevant in cases of testimonies regarding violent crimes as well as sexual crimes. Especially, if in two cases the circumstances are similar, while the only differentiating factor is gender, and the assertion of reliability varies, it is plausible to assume that some external factors influence the perception of one as an epistemic agent. A disbelief in an utterance made by a woman and a belief in the same utterance made by a man in the context of violent and sexual crimes

calls for taking into consideration gender stereotypes that prevail in a society. These factors are particularly weighty in gender-based crimes, where the proportion of female victims is significantly higher than male victims, e.g., home abuse or sexual violence. The prevalence of female victims in this type of crimes may result in the perception of both gender-based crimes and its victims through the lens of gender stereotypes.

The discussion on the perceived reliability of perpetrators and victims and its relation to gender has been increasingly vocal since the 2018 accusation of Supreme Court candidate Brett Kavanaugh of sexual offense against Christine Blasey Ford [13]. In support of Ford's, there appeared an internet movement of "#BelieveWomen" and many women shared their experience of being victims of male sexual violence. Amidst the public discussion, the problem of disbelief towards testimonies provided by women who claimed they have been victims of such offenses was raised. Ford's case involved many factors that may be regarded as influencing the perceived reliability of both Ford and Kavanaugh, such as political views, socioeconomic hierarchy, etc. This case thus should serve only as an illustration to the problem described, as our goal is not to study reliability of people in power. We aim to investigate the existence of gender bias as such, which may occur in everyday human interactions, without the presence of other obscuring variables. As Kimberly Kessler Ferzan [17] points, one of the underlying demands of the "#BelieveWomen" movement is to treat women respectfully as epistemic agents, and to treat their testimonies regarding violence caused by men with equal respect and belief, just like testimonies of victims of any other crimes, regardless the gender. It should also be pointed out that some of the voices point to potential problems that may create "#BelieveWomen"-like movements, for example by attributing undue credibility to testimony based on gender, thereby reversing the gender bias [6]. The researched problem may also overlap with a possible misbelief towards victims of crimes as such and secondary victimization. For this reason, the goal of our research is to isolate the gender factor in testimonies and reliability ascriptions.

1.2 Reliability and Epistemic Justice

The presented research investigates, if (1) gender is a factor that differentiates the status of a person as an epistemic agent, and (2) if gender is related to the perception of reliability of a given testimony. Ideally, gender should have no influence on treating others as equal epistemic agents. This means that other relevant circumstances held fixed, similar testimonies should be given the same amount of trust. Seeing someone as an equal epistemic agent means attributing this person with at least an ordinary level of rationality and reliability. By the "at least ordinary level of rationality" we understand a level that we wish to ascribe to a person, when reasons to doubt their rationality have not (or not yet) been obtained by us. In other words, it is an attitude we usually have towards others in everyday encounters. It would be very inconvenient to assume and check if every person we meet may be irrational. It is advantageous and efficient to approach others as if they were at least ordinarily rational. Similarly, the "at least ordinary level of reliability" should be understood

as a level we may ascribe to an utterance made by a person we have no *prima facie* reasons to distrust. It is thus the level of reliability we have in everyday encounters with others, for example while inquiring about the time, or the road in a foreign city. Most of the knowledge we possess is not 'knowledge by acquaintance' but 'knowledge by description', to use Bertrand Russell's words [40]. To operate in the social world successfully, we often need to believe descriptions of things and situations we did not acquaint with directly. Therefore, if we receive someone's testimony, we treat it as epistemically equal and we can believe it to be true, as long as there are no significant reasons to do otherwise. As Kimberly Kessler Ferzan [17] indicates, we rely on testimonies of others all the time—doing otherwise would make everyday life impossible. Yet, as she claims, regarding violence against women, the tendency is to disbelieve female victims solely based on their gender. Ferzan [17] claims that in some cases, women are suffering from epistemic injustice, while there is no rational explanation for their epistemically inferior status. Such phenomenon would be an instance of hermeneutical injustice, described by Miranda Fricker as a product of a hermeneutical gap that occurs when members of marginalized groups are rendered as lesser epistemic or moral agents [19]. Specifically, regarding women as less credible in cases of violent crimes based solely on their gender would add up to testimonial injustice, a subtype of hermeneutical injustice occurring in testimonial exchanges, when stereotypes and prejudice lead to credibility deficit ascribed to members of a social group [18, 19]. Especially in cases of violence against women, since there is no reason to treat them as less reliable, there may be good reasons to acknowledge their epistemic superiority. For women are statistically more often victims of sexual crimes, their knowledge in this matter may exceed that of men [17]. Their expert knowledge may be an effect of standpoint epistemology [17]. Thus, it might be rational to perceive testimonies of female victims of violent crimes as more reliable than other testimonies. Of course, the question if in different cases there are any rational grounds to differentiate the reliability of men and women calls for separate research. Our goal here is not to determine the reason why such varying approaches in varying contexts to testimonies of men and women may appear. We rather want to investigate experimentally, if there are regularities in reliability ascriptions correlated with gender in the exact same context in the first place.

A practical dimension of the aforementioned discussion is for instance the so-called "he said-she said" type of cases [20]. These are mainly cases of sexual assault accusations, where the only evidence available are testimonies of both the victim and the assaulter. In cases of this kind, the presence of gender bias towards the epistemic status of women would be a highly harmful occurrence that could interfere with the fairness and impartialness of adjudicating a criminal case in court — it could unjustly favor the perspective of a male assaulter and result in insufficient protection of female victims. Georgi Gardiner claims that the presence of this kind of gender bias could be understood as a case of epistemic irrationality, meaning that the decision-makers would take into consideration factors that are irrelevant to the case, such as gender prejudice and beliefs about men and women that are based on a cultural understanding of gender roles, but not necessarily connected to the case adjudicated [21]. The emergence of gender bias in the findings of the presented research could provide an argument in favor of Gardiner's thesis. It also should

be noted that the presence of gender bias is frequently reported in situations that include a relation between an assaulter and a victim of violent crimes. Hence, it can be examined if it is more present in situations where a victim is female and an assaulter is male, but also the other way around — to investigate a possible gender bias towards male victims of female aggression. Comparison of those two variants may show either a bias towards one of the genders or towards victims versus perpetrators (or perhaps even both). Additionally, it should be investigated if in situations of male violence towards men and female violence towards women, there may be observed a tendency to ascribe a lower level of reliability towards victims. If this is the case, then it would suggest the presence of a victim bias, not a gender bias per se. Our experimental research is focused on examining all four of those variations of the victim-perpetrator relation.

Our envisioned research design permits to bring an additional insight into the potential structure of the gender bias in the "he said-she said" type of cases. Namely, the potential bias could occur at one of two levels. First, it could be that the bias is situated at a purely semiotic level. This would mean that the aforementioned epistemic irrationality suggested by Gardiner stems from linguistic representations encoded in the words we use. These semantic entities could be direct triggers of stereotypical mental representations leading to the discussed epistemic injustice [7, 36, 39]. If this is the case, then a vignette experiment employing rigid, binary gender categories such as "man" or "woman", should incite participants to answer any values-related question on the vignette in a biased manner, as it is the vignette language that triggers the bias. Moreover, it can be expected that the bias induced by the language should be systematic, with a consistent direction in situations where the concept occurs. By contrast, if the answers on different questions do not all yield gender effects, the linguistic hypothesis is more questionable and the gender bias is located at a deeper, mental level. This could indicate that the shared conceptions of "a man" or "a woman" are understood on a level of social identities [19]. As a result, this possibility would emphasize the need to fight entire stereotypes rather than the language itself, because gender-neutral terms would not be enough to challenge the gender bias present in society [31, 32].

There could be a relation of gender as well as beliefs about reliability and epistemic status. As it has been empirically researched, people expect different behavior of others in accordance with their gender, they adjust their behavior themselves to fit gender norms, and they expect others to adjust their behavior to gender norms [11, 12]. Research conducted by Stepnick and Orcutt [43] indicates that in the courts of law, it is female judges and female attorneys who are the most observant of the gender bias behavior towards women, even though both male and female judges and attorneys engage in such behaviour. When it comes to testimonies given by female victims of sexual assault, certain speech styles have been reported to affect victim's reliability. For instance, the use of uptalk, a manner of speaking characterized by risen intonation at the end of a declarative sentence, has been reported to negatively affect the perceived reliability of female victims of sexual assault. It did not have such effect on the reliability of male perpetrators, as well as it did not influence the perceived reliability of testimonies in a situation that did not involve a gender related crime, that is, in a medical malpractice trial [30]. In addition, the reliability of a

female victim of sexual assault would also drop when she was heard after the testimony of a man [30]. A powerful speech style on the other hand, which was reported to increase reliability of a speaker in court, has been reported to have an opposite effect, when employed by a victim of sexual assault [25]. Moreover, eye contact was a factor enhancing the credibility of experts, but only when they were male, and had no significant effect on credibility, when the experts were female [33]. Research also indicates a possible gender bias in specific relation to adjudicating criminal activity. For instance, men in the US tend to receive 63 percent higher sentences than women in cases where both the crimes and their relevant circumstances are similar; data also shows that women are more likely to avoid conviction and incarceration [42]. Interestingly enough, not only gender, but also masculine or feminine appearance may influence perception of reliability of victims and perpetrators of physical assault. As Wasarhaley et al. [45] indicated in the research on lesbian intimate partner violence, masculine-looking partners have been perceived as more reliable than a feminine-looking partners, when the perpetrator of violence was masculine. Moreover, male participants tended to have more sympathy towards the masculine-looking victims than towards feminine. The adopted gender roles, not the gender per se, have been observed to vary the perception of the victims and perpetrators. The relation of gender and ascribing reliability has not been exhaustively studied yet. Even though the aforementioned research partly covers this matter, the majority of experiments was conducted in relation to criminal court proceedings and studied prosecutors' and judges' decisions. Our aim is a wider approach, investigating gender bias in ascribing reliability to victims and perpetrators in the general population.

There have been numerous, insightful studies investigating the reliability of a witness' testimony. For instance, Ask and Granhag [3] find that a witness who disconfirmed a focal hypothesis was perceived as less reliable and credible; Dent and Stephenson [9] find that children witnesses are more susceptible to suggestibility and therefore are perceived as less reliable; Castelli et al. [8] argue that the interviewing style can impact the credibility of witnesses, especially children. By contrast, there have been less studies focusing on the victim's testimony. Among them, see, for instance, Voogt et al. [44] for an overview of measurements of a child-victim's credibility; van Doorn and Koster [10] for a systematic review of victim's emotionality and credibility, suggesting that emotionality can impact victim's credibility. Our study is part of the strand of research on victim's testimony. It goes one step further by investigating the additional impact of gender on testimony. Namely, we investigate simultaneously whether a putative victim's gender can impact perceptions of this victim's reliability as well as whether the gender of the suspect can impact guilt attributions. Our study employs the rigorous method of hypothesis testing through massive online surveys, which guarantees robust statistical effects, a balanced sample, preregistered hypotheses as well as rigorous attention and comprehension checks.

Last clarificatory remark is in order. What the experiments we propose in this paper can detect, formally speaking, are gender differences, because the formulation of the study does not provide a direct proof that it is a gender bias, rather than a rationally justified gender difference. As was mentioned in the beginning, rational and unprejudiced differences in the assessment of the reliability of assertions should

not be considered gender bias. On the other hand, should the differences occur on a population level, not in particular assessors, we are inclined to argue that it is a gender bias. In the discussion we will argue that gender differences, if found in the course of the study, will constitute a systemic gender bias, when they occur at a population, rather than at a particular level.

2 Methods

2.1 Study Design

In the course of the study, two experiments have been conducted.[1] The research's aim was to observe, whether gender bias, as defined above, occurs in identical scenarios, which differ only in the gender of the protagonists. Namely, whether reliability ascriptions to testifiers and suspects respectively will differ if everything is held fixed in the experimental scenarios, except for the gender of the protagonists.

The main questions and hypotheses of the research were as follows:

Q1: Are people gender-biased in ascribing reliability to testifiers when adjudicating violent crimes?

Q2a: Does the gender of participants impact the ascriptions of the testifier's reliability in violent crimes?

Q2b: Does the gender of the testifier or suspect impact testifier's reliability assessments pertaining to violent crimes?

H1 When everything else is held fixed in the scenario, except for the gender of testifier and suspect, there are differences in the assessments of the testifier's reliability.

H2 Participants' gender affects their assessments of the testifier's reliability.[2]

In the experiments conducted, participants had to pass an attention check and confirm being a native speaker of English. Participants were presented with one variation of the following scenario:

> An ordinarily looking (A) came to the police department to report that [he/she] has just been assaulted by [his/her] (B) boss after a company dinner. Although the police officer could smell a bit of alcohol from [his/her] breath, [he/she] answered all questions clearly nevertheless. The boss was a wealthy [man/woman], who never had any criminal record. There have been rumors about

[1] All data, preregistrations and appendix can be found in the OSF repository under the link: https://osf.io/wbu8c/?view_only=9bb307575ab54200a3b491b6fa399d2b. We preregistered two separate experiments but analyzed the data jointly. Link to preregistration of conditions MT and FT: https://aspredicted.org/blind.php?x=9k4zq9; link to preregistration of conditions BF and BM: https://aspredicted.org/7KJ_9XF.

[2] We also asked participants to assess the Guilt of the perpetrators, but found only negligible differences, and thus present the analysis of the answers on the Guilt question in the Appendix.

[him/her] mistreating [his/her] [male/female] workers, but no accusations have ever been proven.

Four variations of the scenario were randomly assigned to participants:

1. A = "woman", B = "female", rumors about mistreating female workers. Condition name: "both females" (**BF**).
2. A = "man", B = "male", rumors about mistreating male workers. Condition name: "both males" (**BM**).
3. A = "woman", B = "male", rumors about mistreating female workers. However the scenario does not suggest that the person reporting the crime actually is a victim, for the sake of clarity and conciseness we will use the name "female testifier" (**FT**) for this condition.
4. A = "man", B = "female", rumors about mistreating male workers. Condition name: "male testifier" (**MT**).

As can be seen, in each case rumors suggest that the boss could mistreat workers whose gender is the same as A's gender.

After reading the scenario, participants were asked to adjudicate the level of reliability of A:

"On a scale from 1 to 10, where 1 is completely unreliable and 10 is completely reliable, how reliable is the testimony of the [woman/man] in your opinion?"

2.2 Data Structure and Data Analysis

Collected data included dependent variable (highlighted in capital letters in the text), and four independent variables (for easier distinction, demographic variables are marked with a capital letter and the experiment variable with a lower case letter):

- *RELIABILITY (*ordinal variable approximated by interval scale)—participant's assessment of testimony's reliability: 1–10 Likert scale.
- *Condition* (factor)—the scenario which was presented to the participant in the questionnaire: BF *vs* BM *vs* FT *vs* MT.
- *Gender* (binary variable)—gender of the participant: 1—female *vs* 0—male (other values omitted, see: Sect. 3.3. of the present paper).
- *Age* (continuous variable)—age of participant.
- *Any legal expertise* (binary variable)—whether the participant has any legal knowledge, i.e., declares being a law student, legal academic, a judge or other lawyer: 1—having any legal expertise *vs* 0—lack of legal expertise.

Turning to the issue of data analysis, let's start with comments on the Likert scale. It is an ordinal scale, but it is indicated that it can be approximated as an interval scale, especially when it has many points [26, 34]. As the scale used in the study is a 10-point scale, it is not a misuse to treat it as an interval scale, which is why parametric statistical tests are used in the paper. However, as the ordinality of the Likert scale can manifest itself particularly strongly in the face

of controversial issues, and because of the desire for robustness of the analysis, non-parametric methods were additionally used in many issues.

Eventually, the study uses parametric methods (such as ANOVA), non-parametric methods (such as Kruskal–Wallis test or Dunn's test), and elements of visual analysis. When *p*-value is in-text reported, it refers to *t*-test, but in each case, the results of non-parametric statistical tests are also provided in the tables with the results of the post-hoc tests. Absolute value of Cohen's $|d|$ is reported in each *t*-test and η^2 are reported in each ANOVA. No arbitrary level of significance was adopted due to the many disadvantages and lack of a strong justification for this approach [22, 46]. Instead, according to test power calculations, any $p\text{-value} \leq 0.10$ was considered worthy of interest, while lower *p*-values were treated as leading to higher significance compared to higher *p*-values (hence, it was not a 0–1 decision-making rule). In the absence of a solid rationale for introducing a correction for multiple comparisons in current study [37], the *p*-values of the statistical tests without such correction were intentionally treated as reference values. However, as many researchers report *p*-values with corrections for multiple comparisons, for those interested and for comparability of the analysis with possible future similar studies, values for the Tukey HSD tests are additionally included in the results tables. The conclusions of the analysis would not change significantly if the *p*-values for the Tukey test were taken as reference values. We did not control the information about the legal expertise strictly enough to include variable *Any legal expertise* in the main analysis.

The statistical analysis was performed using programming language R ver. 4.2.2 [35] with packages: *tidyverse* set (including, especially, *ggplot2*, *readr*, *dplyr*, *broom*) [47], *afex* [41], *rstatix* [29], and IDE RStudio [38], JASP ver. 0.17.2, and Jamovi ver. 2.3.21.

2.3 The Sample Characteristics

Based on a predictive power analysis performed with the G-Power [15] software, we had estimated that a sample size of 400 participants, the 100 per condition and the 50 for condition-gender group, will be sufficient to achieve power for a two-tailed *t*-test at the level of 80 percent for the moderate size effects ($|d| = 0.50$) for the gender groups inside the condition ($\alpha = 0.10$, equal groups, precise $n = 102$) and small-moderate ($|d| = 0.35$) for the condition groups ($\alpha = 0.10$, equal groups, precise $n = 204$).

400 participants were recruited using the Amazon Mechanical Turk online platform. We filtered out participants who (i) were not native speakers of the English language; (ii) who failed an attention check; (iii) took less than 30 s to complete the entire questionnaire. The group of non-binary gender included only 2 participants, which was not enough to include it in the statistical analysis, so these participants were omitted. Finally, 361 observations have been left (192 males and 169 females). Distribution of the observations across conditions and gender, and post-hoc computed power of the Student's *t*-test are presented in Table 1.

Table 1 Contingency table of condition and *Gender*

Condition		Gender		Total	Power of the *t*-test (for $\|d\|=0.5$, $\alpha=.10$)
		Male	Female		
BF	Count	55	38	93	76%
	% [within row]	59%	41%	100%	
BM	Count	58	35	93	75%
	%	62%	38%	100%	
FT	Count	39	51	90	75%
	%	43%	57%	100%	
MT	Count	40	45	85	74%
	%	47%	53%	100%	
Total	Count	192	169	361	
	%	53%	47%	100%	

Table 2 Contingency table of *condition* and *Any legal expertise*

Condition		Any legal expertise		Total
		0	1	
BF	Count	62	31	93
	% [within row]	67%	33%	100%
BM	Count	62	31	93
	%	67%	33%	100%
FT	Count	78	12	90
	%	87%	13%	100%
MT	Count	79	6	85
	%	93%	7%	100%
Total	Count	281	80	361
	%	78%	22%	100%

The mean age in the sample was 39.4 years (SD = 11.1, range: 20–80) and did not significantly vary across the groups (F-test of the linear model predicting the age using *condition* and *Gender* variables: $F(4, 356) = 0.33$, $p = 0.858$).

A high number of participants declared having any legal expertise in the demographic questionnaire at the end of the survey, i.e. lawyers, judges, legal academics and law students (80 from 361, which is ca. 22 percent). Thus, the proportion of people who declared to have legal expertise in the sample is higher than in the population [2]. Precise number of lawyers per group is displayed in Table 2.

Table 3 Descriptive Statistics of *RELIABILITY*

RELIABILITY	Mode	Mean	SD	IQR	Skewness	Min	Max	1st quartile	Median	3rd quartile
	8	7.25	1.89	3	−0.87	1	10	6	8	9

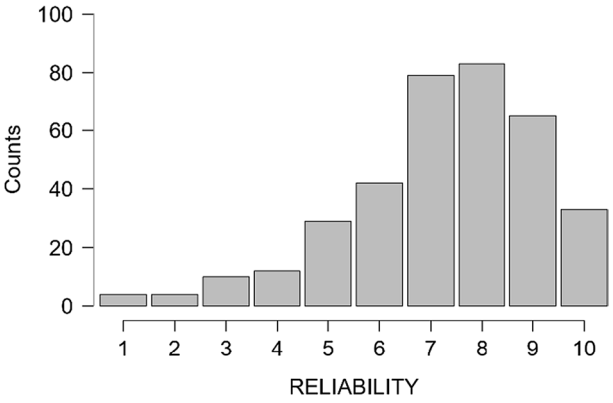

Fig. 1 Histogram of RELIABILITY

3 Results

3.1 Descriptive Statistics

The descriptive statistics for the dependent variable *RELIABILITY* are displayed in Table 3. In turn, the histograms for the values are indicated in Fig. 1. Comparison of the dispersion measures of *RELIABILITY* (SD = 1.89; IQR = 3) and its central parameters (mean = 7.25, mode = 8), allows one to conclude that its level of variability is fully sufficient for the statistical analysis (M/SD = 0.26). The skewness of the *RELIABLITY* distribution is around − 0.87, which means the distribution has slight negative skew.

The Spearman's correlation coefficient between *Age* and *RELIABILITY* was very low ($\rho = 0.03$, $p = 0.551$). In view of this, *Age* was not included in further analysis.

3.2 Statistical Analysis

An multi-way ANOVA including *condition* and *Gender* (note that *Gender* = 1 refers to female participants and *Gender* = 0 refers to male participants, as indicated in Sect. 3.2) as well as their interaction as independent variables was performed to assess the differences in *RELIABILITY* scores between experimental groups. The results are presented in Table 4. A sum of squares of the third type was used in each ANOVA.

Table 4 Multi-way ANOVA for *RELIABILITY*, including *Gender*, *condition* and their interaction

Cases	Sum of Sq	*Df*	Mean Sq	*F*	*p*		η^2
Gender	23.86	1	23.86	7.13	0.008	**	0.019
Condition	59.92	3	19.97	5.96	< 0.001	***	0.047
*Gender * condition*	18.17	3	6.06	1.81	0.145		0.014
Residuals	1182.08	353	3.35				

$^{**}p \leq 0.01$; $^{***}p < 0.001$

In the case of the *RELIABILITY* analysis, both *Gender* ($F(1, 353) = 7.13$, $p = 0.008$, $\eta^2 = 0.019$) and *condition* ($F(3, 353) = 5.96$, $p < 0.001$, $\eta^2 = 0.047$) were significant, see Table 4. There was too weak a basis to infer the significance of the interaction term ($F(3, 353) = 1.81$, $p = 0.145$, $\eta^2 = 0.014$), but again the p-value is relatively small and therefore a deeper analysis is recommended. To make the analysis more robust, as mentioned in Sect. 3.2., we decided to perform also non-parametric tests. Kruskal–Wallis rank sum test (which is known as non-parametric equivalent for the one-way ANOVA) for differences across *condition* variable validated aforementioned results ($\chi^2(3) = 15.60$, $p = 0.001$).

Post-hoc tests, which are displayed in Table 5, provide two highly significant relationships. Firstly, females tend to assess *RELIABILITY* as higher on average as compared to males ($t = -2.67$, $p_t = 0.008$, $|d| = 0.29$; 'p_t' refers to p-value of Student's t-test). Secondly, mean *RELIABILITY* ascription in the MT condition is lower than any other condition and this effect seems to have a medium strength (BF *vs* MT: $t = 2.85$, $p_t = 0.005$, $|d| = 0.43$; BM *vs* MT: $t = 4.10$, $p_t < 0.001$, $|d| = 0.62$; FT *vs* MT: $t = 2.80$, $p_t = 0.005$, $|d| = 0.43$). Other differences are not significant ($p_t > 0.188$, $p_D > 0.257$) The results of the non-parametric post-hoc tests (Table 5, right panel B) fully overlap with the conclusions from the parametric tests.

As visual inspection shows, on the one hand, for some conditions (FT, BF) there are noticeable differences in average answers of female and male participants, while for other conditions the difference is not apparent (MT, BM), see Fig. 2. On the other hand, when grouped by gender, the data reveals one group significantly different from the others, different for each gender (higher BM score for males, lower MT score for females), see Fig. 3. These observations make it worthwhile to undertake a further analysis: a comparison between genders, when grouping by condition (i.e., in each condition, separately), and between conditions, when grouping by gender (i.e. for females and for males, separately). The t-tests performed assumed equality of variances, as tests of equality of variances in three of the four cases did not give grounds to reject this assumption ($p = 0.053$ for BM; $p > 0.212$ for other conditions; for detailed results, see Table A1 in the online appendix).

Both Student's t and Mann–Whitney-Wilcoxon tests (presented in Table 6) allow for the conclusion that the assessments of *RELIABILITY* by female participants were significantly higher than for male participants in FT condition ($t(88) = -2.95$, $p_t = 0.004$; $W = 638$, $p_W = 0.003$; $|d| = 0.62$) and likely significantly higher in BF condition ($t(91) = -1.83$, $p_t = 0.071$; $W = 782$, $p_W = 0.036$;

Table 5 Post-hoc comparisons for *RELIABILITY*, by *Gender* and *condition*

Group 1	Group 2	A. Student's *t*- and Tukey's HSD tests						B. Dunn's tests				Flag
		Mean Diff	SE	*t*	|*d*|	p_t	p_{tukey}	*Z*	W_i	W_j	p_D	
Male	Female	−0.52	0.20	−2.67	0.29	0.008	0.008	−2.31	169.28	194.31	0.021	*
BF	BM	−0.35	0.27	−1.28	0.19	0.200	0.574	−1.13	184.91	201.96	0.258	
	FT	0.01	0.27	0.04	0.01	0.969	~1	−0.34	184.91	190.04	0.735	
	MT	0.79	0.28	2.85	0.43	0.005	0.024	2.64	184.91	144.22	0.008	**
BM	FT	0.36	0.28	1.32	0.20	0.189	0.553	0.78	201.96	190.04	0.433	
	MT	1.14	0.28	4.10	0.62	<0.001	<0.001	3.75	201.96	144.22	<0.001	***
FT	MT	0.78	0.28	2.80	0.43	0.005	0.027	2.95	190.04	144.22	0.003	**

Flags refer to the higher out of the two *p*-values: $max(p_t, p_D)$; *$p \leq 0.05$; **$p \leq 0.01$; ***$p < 0.001$

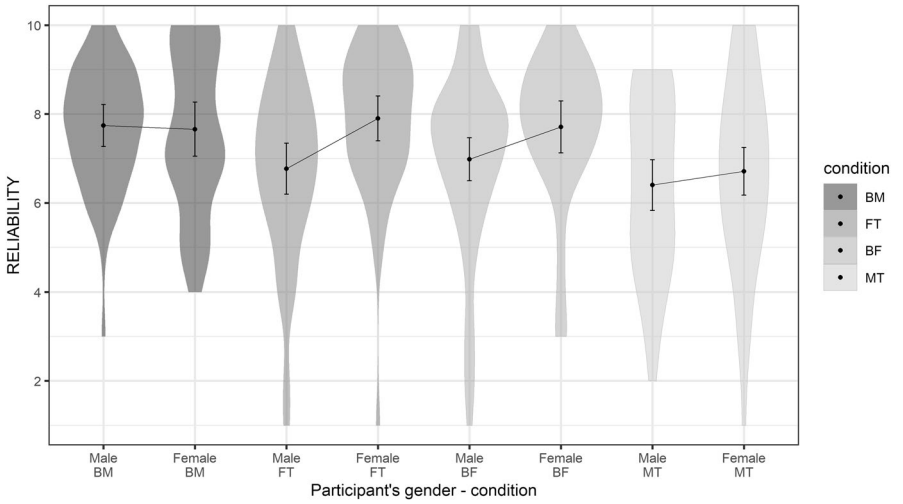

Fig. 2 Violin plots and comparisons of means for *RELIABILITY*, by *Gender-condition*, grouped by *condition*. Points represent the means, error bars denote the 95% confidence intervals calculated using ANOVA model-based standard errors

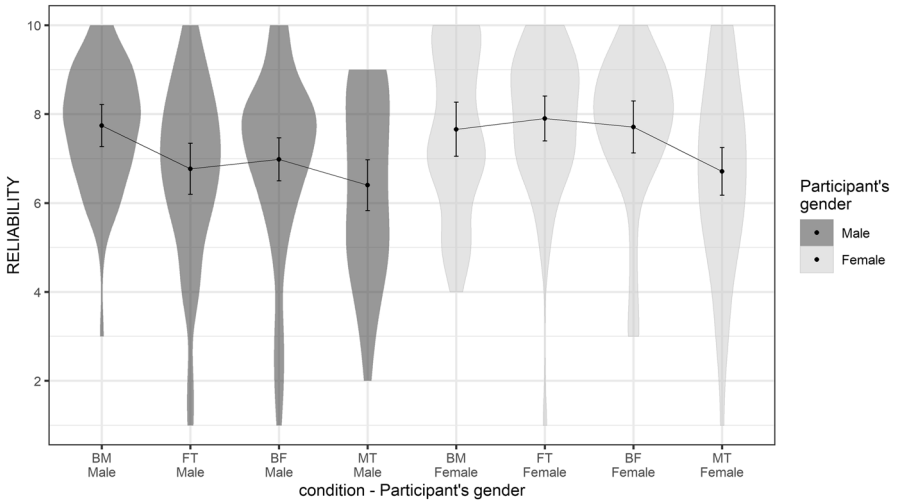

Fig. 3 Violin plots and comparisons of means for *RELIABILITY*, by *condition-Gender*, grouped by *Gender*. Points represent the means, error bars denote the 95% confidence intervals calculated using ANOVA model-based standard errors

$|d|=0.39$). These effects are moderately strong. There is no basis for rejecting the hypothesis of equality across groups for the BM ($t(91)=0.25$, $p_t=0.799$; $W=1036$, $p_W=0.871$; $|d|=0.05$) and MT ($t(83)=-0.69$, $p_t=0.489$; $W=818$, $p_W=0.468$; $|d|=0.15$) conditions.

Table 6 Central tendency comparisons between gender groups for *RELIABILITY* with the splitting by *condition*

Condition	A. Student's *t*-tests							B. Wilcoxon tests		Flag
	Mean Male	Mean Female	Mean Diff	t	df	$\|d\|$	p_t	W	p_W	
BF	6.98	7.71	−0.73	−1.83	91	0.39	0.071	782	0.036	○
BM	7.74	7.66	0.08	0.25	91	0.05	0.799	1036	0.871	
FT	6.77	7.90	−1.13	−2.95	88	0.62	0.004	638	0.003	**
MT	6.40	6.71	−0.31	−0.69	83	0.15	0.489	818	0.468	

Flags refer to the higher out of the two *p*-values: $max(p_t, p_W)$; ○$p \leq 0.10$; **$p \leq 0.01$

Table 7 One-way ANOVA for *RELIABILITY*, for the male participants group only

Cases	Sum of Sq	df	Mean Sq	F	p		η^2
Condition	47.95	3	15.98	4.78	0.003	**	0.071
Residuals	628.63	188	3.34				

**$p \leq 0.01$

Turning to the differences by participants' gender group, an one-way ANOVA displayed in Table 7 suggests a significant difference between the conditions in the male participants group ($F(3, 188) = 4.78$, $p = 0.003$, $\eta^2 = 0.071$). Assessments of *RELIABILITY* in BM condition were highly significantly higher than MT ($t = 3.57$, $p_t < 0.001$, $|d| = 0.73$) and significantly higher than FT ($t = 2.57$, $p_t = 0.011$, $|d| = 0.53$) and BF (in BF vs BM: $t = -2.21$, $p = 0.029$, $|d| = 0.42$), see Table 8. Note that the effect size of the BM *vs* MT difference is quite a large ($|d| = 0.73$). Other differences are rather not significant ($p_t > 0.126$, $p_D > 0.121$). This suggests that the higher assessments in the BM condition are responsible for the significant differences detected by the one-way ANOVA, see Table 7.

In the case of the female's group, on the other hand, the significant difference between the groups, as indicated by the one-way ANOVA (presented in Table 9; $F(3, 165) = 3.85$, $p = 0.011$, $\eta^2 = 0.065$), is due to the difference in scores for MT. More precisely, ascriptions of *RELIABILITY* in the MT condition were highly significantly lower than in FT (in FT *vs* MT: $t = 3.18$, $p_t = 0.002$, $|d| = 0.65$) and significantly lower than in BF (in BF *vs* MT: $t = 2.48$, $p_t = 0.014$, $|d| = 0.55$) and BM (in BM *vs* MT: $t = 2.29$, $p_t = 0.023$, $|d| = 0.52$). Differences in other conditions are not significant ($p_t > 0.542$, $p_D > 0.450$). Dunn's tests provide nearly same results. The results of the post-hoc tests are provided in Table 10.

In both gender groups there is no basis to assume inter-group differences after excluding the BM and MT conditions, respectively ($F(2, 121) = 0.25$, $p = 0.779$ for females with MT excluded; $F(2, 131) = 0.98$, $p = 0.376$ for males with BM excluded). This confirms the earlier comments about the conditions responsible for the intergroup differences noted in the ANOVAs.

Table 8 Post-hoc comparisons for *RELIABILITY*, by *condition*, for the male participants group only

Group 1	Group 2	A. Student's *t*- and Tukey's HSD tests						B. Dunn's tests				Flag
		Mean Diff	SE	*t*	|*d*|	p_t	p_{tukey}	*Z*	W_i	W_j	p_D	
BF	BM	−0.76	0.34	−2.21	0.42	0.029	0.125	−1.96	95.64	115.82	0.050	*
	FT	0.21	0.38	0.56	0.12	0.579	0.945	0.68	95.64	87.87	0.497	
	MT	0.58	0.38	1.53	0.32	0.127	0.421	1.55	95.64	78.09	0.122	
BM	FT	0.97	0.38	2.57	0.53	0.011	0.053	2.47	115.82	87.87	0.013	*
	MT	1.34	0.38	3.57	0.73	< 0.001	0.003	3.36	115.82	78.09	< 0.001	***
FT	MT	0.37	0.41	0.90	0.20	0.371	0.806	0.80	87.87	78.09	0.426	

Flags refer to the higher out of the two *p*-values: $max(p_t, p_D)$; $^{*}p \leq 0.05$; $^{***}p < 0.001$

Table 9 One-way ANOVA for *RELIABILITY*, for the female participants group only

Cases	Sum of Sq	*df*	Mean Sq	*F*	*p*		η^2
Condition	38.78	3	12.93	3.85	0.011	*	0.065
Residuals	553.46	165	3.35				

*$p \leq 0.05$

Table 10 Post-hoc comparisons for *RELIABILITY*, by *condition*, for the female participants group only

Group 1	Group 2	A. Student's *t*- and Tukey's HSD tests						B. Dunn's tests				Flag
		Mean Diff	SE	*t*	|*d*|	p_t	p_{tukey}	*Z*	W_i	W_j	p_D	
BF	BM	0.05	0.43	0.12	0.03	0.901	0.999	0.35	91.39	87.46	0.727	
	FT	−0.19	0.39	−0.49	0.10	0.626	0.962	−0.39	91.39	95.42	0.696	
	MT	1.00	0.40	2.48	0.55	0.014	0.067	2.41	91.39	65.88	0.016	*
BM	FT	−0.24	0.40	−0.61	0.13	0.543	0.929	−0.75	87.46	95.42	0.451	
	MT	0.95	0.41	2.29	0.52	0.023	0.104	1.99	87.46	65.88	0.047	*
FT	MT	1.19	0.37	3.18	0.65	0.002	0.009	3.00	95.42	65.88	0.003	**

Flags refer to the higher out of the two *p*-values: $max(p_t, p_D)$; *$p \leq 0.05$; **$p \leq 0.01$

3.3 The *Any Legal Expertise* Variable in the Model

After adding the *Any legal expertise* variable into the model (which then included *Gender*, *condition*, *Any legal expertise* variables and their three first order interactions), the most significant term was *Gender* * *Any legal expertise* ($F(1, 348) = 9.68$, $p = 0.002$, $\eta^2 = 0.026$) interaction. Taking into account very small group sizes of lawyers in the sample, such a significant term could be interpreted as the existence of an important (mediation) effect of *Any legal expertise* on the relation between gender and *RELIABILITY* assessments. The table with precise results is presented in the appendix (Table A2 in the online Appendix).

4 General Discussion

The experiments do not allow to conclude, what conditions, or which gender, deviate from the norm, because due to the experiment design, setting the norm is not possible. Thus, we will not formulate our results in statements claiming that one gender is more biased than the other, because there is no objective reference point. Additionally, when participants are split based on their gender, the reference point for female participants' answers is generally higher than for males: female answer depict higher reliability ascriptions.

However, the analysis of reliability ascriptions to testifiers allows to formulate five main conclusions.[3]

1. A man who claims that he has been assaulted by a woman is generally perceived as the least reliable.
2. Males tend to perceive FT as much less reliable, as compared to females.
3. It is likely that males tend to perceive BF as less reliable, compared to females.
4. In the group of female participants, the assessments of reliability were lower than in other conditions only in the MT case.
5. In the group of male participants, the assessments of reliability were higher as compared to other conditions only in the BM case.

In the discussion, we adopt four interpretations, each formulated from a different point of view, which may allow to explain the results of the experiments. The interpretations, however employing different optics, are not mutually excluding, and may be treated as complementary. The first one will allow to compare results of both genders studied together. The second will investigate differences in inter-gender beliefs. The third one will examine what we call intra-gender beliefs about gender. The fourth one is an attempt to reconcile conclusions from the former three paths, with an employment of an explanation based on gender roles and gender norms.

4.1 Disbelief Towards Men Testifying to Have Been Assaulted by Women

Typically, female participants tended to ascribe a higher level of reliability to testifiers than males. However, we found a substantially reduced trust toward males who claimed that they had been victims of female perpetrators.

One potential explanation of such a low assessment of male testifiers' reliability is a commonly held stereotype that men cannot be, or at least very rarely are, victims of female violence [4, 24]. Perception of male testifiers as least reliable could be caused by a low frequency of such violence, as compared to the higher incidence of male violence towards men or women. On the other hand, this effect may be explained by an untrue belief about a low frequency of female violence towards men. Benatar [5] argues that empirical studies have shown that male and female violence against a partner of a different gender is equally, or almost equally frequent. At the same time, the prevailing stereotype is that most violence between partners is that of males against females. Benatar claims that the often-used expression "gender violence" is commonly misunderstood as an equidistant for male violence against women. In his line of argumentation, the misperception of violence against men may have its roots in gender stereotyping and prevailing gender norms. Some

[3] Further in the discussion, analogous to the description of the statistical analysis, we will use the shorter term "male testifier" or MT (and, likewise, FT, BF for "both female", and BM), because the scenarios have not specified whether the person, whose reliability was assessed, has really been a victim. The scenarios were formulated without this information on purpose, as it was the perceived reliability that had been measured.

beliefs about men and women, according to Benatar, are harmful to both genders. For instance, the perception of men as strong, active agents and women as weaker, passive and in need for extending protection over them, effects both in disadvantaging women from many aspects of social and professional life and downplays the weight of harm and psychological trauma suffered by men [5].

A more traditional perception of gender roles influences practices of victim blaming for violence and sexual violence when the gender non-conforming behaviour is followed by sexual violence [16]. This may indicate that holding the traditional, stereotypical views on gender norms may influence the disbelief towards less typical instances of gender violence, such as violence towards men committed by women, accompanied by a view that men, as the "tougher" gender are less likely to suffer such violence, or that they should be able to defend themselves, hence shifting the weight of their testimony from endured harm to expected behaviour. These beliefs have been proven counterfactual. Men suffering mobbing at work are more prone to depression, anxiety, and even paranoia [1]. On the other hand, a majority of men, about two thirds, share a belief that mobbing behaviour is normal, contrasted with about one third of women [14].

It may be the case that due to commonly held stereotypes about masculinity and femininity as well as the social stigma for male victims of female violence, report rate for such crimes is low. If that were true and the "dark number" of such crimes was large, these crimes would be perceived as infrequent, and the victims less reliable. The role of stereotype, social stigma, low reporting, and perceived infrequency would create a "looping-effect" [23] for the male victims of female violence and their perception as unreliable.

The low reliability for male testifiers is of particular interest, because it does not differ among the experimental groups. This suggests the internalization of gender stereotypes by both men and women. Given the social stigma, the situation of male victims of female violence is very difficult. They may be labeled as weak, ridiculous, or not manly enough. For this reason, a testifier that decides to speak up could be perceived as more reliable, while confessing despite the social stigma can be seen as an additional proof of the truthfulness and determination of the speaker. Yet, the reliability of the male testifiers has been ranked as the lowest for both experimental groups.

4.2 Inter-Gender Differences

The focus on low reliability in the MT scenario is not the only possible way to look at the experimental data. By contrast, inter-gender differences in specific conditions could be analysed. We discovered that males tend to perceive FT as much less reliable and, probably, BF as less reliable, compared to females.

There may be at least a few possible explanations of the observed differences in ascribed reliability of female testifiers (here, both FT and BF) between gender participants' groups. The first one could be called "gender solidarity", or "gender interest". It may be the case, that women may believe women more than men do, because they would prefer to be believed themselves, when they become victims of violence

or assault. Similarly, men may distrust women more, because they can be influenced by the fear of false accusations of violence against women. Of course, such fear does not need to be based on the actual numbers showing such behaviour to be frequent; it may arise as successfully on the basis of the availability heuristic, i.e., a cognitive shortcut that is based on attaching greater weight to information that are easily recalled [27]. This would explain both, the major difference in ascription of reliability inter gender in FT, and the smaller difference in BF scenario.

Another explanation may be on par with Ferzan's [17] claims about the double standard applied to reliability of male and female victims of violence. According to this view, female testifiers would suffer from testimonial injustice [18, 19], because women tend to be perceived by men as unreliable in this kind of situations. This may be grounded in unjust gender stereotypes and biases and could indicate the perception of women by men as lesser epistemic agents.

Thirdly, the higher belief in women's testimonies by women may be explained by women's superior knowledge about violence against women, be it from experience, witnessing, or hearing reported by other women—friends, family members, etc. Women may thus have greater knowledge and be more sensitized for violence against women, because their experience of living as a woman provides such knowledge, and their personal safety often depends on it [17].

4.3 Intra-Gender Beliefs About Gender

If we look at females' and males' answers separately, we can see that participants of each gender ranked almost all scenarios with the same level of reliability, and for each gender, one scenario clearly stands out. For female participants, FT, BF and BM are perceived as almost identically reliable, and MT's reliability is seen to be significantly lower. In general, males have shown a significant distrust to the testimonies of all testifiers, compared to women, with only one exception. In the scenario, where a man had potentially been assaulted by another man (BM), male participants tended to ascribe higher levels of reliability. Female's assessment of the only differing scenario (MT) is so low, it equalizes with men's assessment of this same scenario. And males' assessment of the only differing scenario (BM) is so high, it equalizes with women's assessment of this scenario.

One way to explain these results is to assume that there are some stereotypes or beliefs about gender that are local for each gender and specific to it. It may be the case that these stereotypes influence the level of ascribed reliability, which falls or rises depending on how the described situation is viewed by members of this gender. "Intra-gender beliefs about gender" would be beliefs about men and women that are specific to either men, or women, and are shared by members of the group. This approach would assume that in certain situations, men and women hold different opinions about the frequency of certain actions depending on gender, or about typical gender behaviors, or even typical attitudes and convictions for men or women.

In general, the female participant's average tendency to ascribe reliability has been higher, than males'. The high reliability ascriptions by males in the BM condition may be explained by their experience with male-on-male fights, and generally,

male violence against other men. It may be the case that such situations are best known to them or considered more frequent. On the other hand, men give less credit to both, women, and men who claim to be victims of women. The latter situation is specific, because a male testifier may be perceived as having a stereotypically feminine role—a man, who is hurt by a woman may be viewed as weak, unmanly, and equalized with a woman by the male participant. If that were true, these findings would support Ferzan's claim that men tend to employ a different epistemic standard towards women and believe them less. Curiously enough, men also distrust other men, when they are being perceived in a feminine social role (MT). This situation shows similarity to the findings of the research of Wasarhaley et al. [45], in which a masculine presentation of lesbians influenced their perceived reliability in comparison with ones who presented femininely. Gender stereotypes about masculinity may explain the low assessment of reliability in the MT condition, high trust in BM, and also low reliability of female victims of male aggressors. Men may perceive the former scenario as unreliable, because if they employ the stereotypical view on men and masculinity, they may be more prone to see men as strong, righteous and chivalrous—ones who use their physical strength to protect women, not harm them. Low reliability of FT may be then explained by an incompatibility of the aggressor with the stereotypical image of relations of men and women. Lastly, low reliability in BF, compared to BM, may be an effect of stereotypes of femininity in action. Traditional gender roles for women and the shared stereotypes do not include aggression, and women are rather seen as calm, caring, delicate, and opposite to violent.

When the answers of female participants are considered, firstly, the generally high reliability in comparison to men's answers is noticeable. Possibly, this difference emerges due to women's experience with aggressors of both genders. Male violence towards women is frequent both in the private sphere, family, or romantic relationships, and in the public realm. Women's personal experience and reports of male violence frequently present in the media could explain the high reliability assessments in FT and BM. Similarly high outcome in BF possibly emerges because women may have better knowledge of women-on-women violence, its specific characteristics and frequency, also gained from personal experience. The low assessment of MT though is rather puzzling. It may be the case that women, similarly to men, have internalized the gender stereotypes and perceive this kind of victim as weaker than a "regular", model man. Sexism and unjust gender stereotypes are in fact a reality for members of all the genders, who are socialized to gender norms and roles from very young age. For this reason, the presence of the stereotype among women too, would not be very surprising.

4.4 Gender Roles and Norms

The last path of interpreting the experimental data is an attempt to integrate the three preceding paths. The results show that for two of the scenarios, BM and MT, the perceived reliability is equal for both male and female participants. FT and BF on the other hand differ substantially.

Possibly, when it comes to BM and MT, men and women share the same, or similar gender stereotypes that influence the assessment of reliability. This stereotype seems to be connected to traditional understandings of gender roles, especially masculinity. The normative model of a man that emerges from this stereotype is one that is rather strong, active and even aggressive, rather than passive and vulnerable. Such stereotype is, of course, harmful, especially to the MT.

The differences in the assessment of reliability of FT and BF by male and female participants may be explained by different gender stereotypes the participant groups hold. Men may be more prone to perceive the other men as guardians, gentlemen, protective of women, and generally have a better opinion on their own gender, while women themselves may be more inclined to perceive the male as an aggressor. This perception of males may be grounded in women's experience and, consequent of it, higher caution. Similarly, for BF, men may not perceive women as aggressive because of the cultural image of femininity. Women's life experience with woman-on-woman violence, and generally more true understanding of womanhood, may on the other hand be a factor neutralizing the gender stereotype, hence higher assessment of reliability by women.

The partial compatibility of assessments (BM and MT) may be explained by partial compatibility of gender stereotypes and perception of gender norms and roles—male victims are perceived similarly in both scenarios, by all participants. Incompatibility of other assessments (FT and BF) could be an effect of discrepancy of beliefs held about females by women and men.

4.5 Semiotic Aspects

The study can provide interesting insights into the discussion of language as a way to prevent stereotypes. In the previous part of the discussion, we have commented on the results pertaining to the gender differences in participants' answers to the question on the testifier's reliability. As discussed, the reliability assessment was jointly dependent on combination of three factors: the gender configuration of (1) testifier and (2) potential perpetrator in the scenario and (3) the gender of the study participant. This configuration is significantly more complex than the testifier's gender alone, which did not systematically affect the reliability assessment in the same direction. The possible stereotype therefore did not refer to gender in general, but to gender-related social configurations. The names "male" or "female" alone (regardless of the gender of the accused) did not have the effect of systematically increasing or decreasing the reliability of the testifier. This provides a rationale against the hypothesis of a mainly semantic explanation of stereotypes, since the decrease or increase in reliability was provided by the gender configuration of the protagonists of the story.

However, we also asked another question, namely, how guilty did participants think that the suspect was (see Appendix). We did not find strong and highly significant differences in participant's answers to the question on guilt, regardless of gender of the characters in the scenarios evaluated (only very slight differences were found). This is certainly good news for the current shape of the criminal legal

procedure. Moreover, the significant differences in answers on the reliability question, depending on the gender configuration in the scenario, contrasted with the lack of substantive differences in answering the question on guilt can provide an interesting insight into the mechanics of the gender bias. In the vignettes we employed gender non-neutral language, especially binary adjectives such as "male" or "female". If the bias was only the result of using the non-neutral terms, gender conditions effects on responses should have been observable in both credibility and guilt questions, yet this was not the case. In the case of the concept of guilt, the differences have disappeared, suggesting that some concepts are gender-free and therefore: that bias-free perceptions of concepts can also be produced when non-neutral terms are used. Consequently, we think that the gender bias stems from some deeper stereotypical representation than the linguistically triggered representation. As a result, to fight gender bias, one needs to fight stereotypes themselves. Trying to bypass non-neutral terms will not contribute to avoiding stereotypes if stereotypes written at a deeper level remain in people. Therefore, the focus should be on fighting stereotypes at their roots and forming an understanding of legal terms that is not susceptible to stereotyping regardless of whether the language used is gender-neutral or not. Such a fight can be effective, as shown by the almost completely even perception of guilt regardless of gender configuration. Our results are in line with Mooney's [31, 32] observations about the nature of stereotypes and the fight against them, which must not focus only on the linguistic level. (see e.g., Fernandez-Blanco, K. Kristan, V. M. Mind the Gap. The Power of Social Norms in Gender Inequality in Europe: When Law is Not Enough, accepted in the European Law Journal for a description of the location of the bias at a deeper level of social norms, [18, 19]).

4.6 Limitations

The most salient limitation of the study stems from the fact that the distribution of legal expertise in the sample was not representative. This variable has not been controlled well enough during the study, and for this reason, it could not be included in the main analysis. At the same time, additional analysis has unanimously indicated to the possibility of an influence of legal expertise to assessments and reliability (with an interaction with the gender of a participant). This suggests that a different percentage of people with legal expertise among the groups may have partly influenced the differences between the conditions. Thus, in the research we have detected a potentially explanatory variable that was not included in the analysis. It remains an avenue for future research to perform the experiment with a large, professional lawyers' sample. Furthermore, it is important to be aware that the sample was made of US residents. The issue under study is highly culturally dependent, hence in a different population the results could be significantly different. Therefore, the results of the study are not generalisable to all Western populations.

The presented research is burdened with a number of limitations stemming from the characteristics of the design of the experiments. First of all, the concept of "reliability" can be understood in different ways. To be precise, the participants may

have employed differing definitions of what is reliability. Reliability may have been assessed by different standards, for example a standard of everyday usage of the word, or a standard of the criminal trial. These potentially different understandings of the concept employed, introduce an uncontrollable element of variation.

Secondly, the design of the study assumed an absolute abstraction from the justifications of the assessments the participants may have formulated. Our aim was to detect a general trend, but for this reason, no data has been collected about the reasons participants have provided with the particular assessments of reliability. For this reason, the differences we interpret as a systemic gender bias, may in reality, at least in some cases, stem from reasonable premises. We are unable to determine what cognitive intuitions or personal experiences may have motivated the answers, and because of that, explanations of the observed differences that do not include gender stereotypes cannot be excluded.

Thirdly, as we have indicated earlier in the discussion, the design of the study makes it very hard, if not impossible to designate natural points of reference, against which obtained results could be juxtaposed. It may not be possible to specify, what is the place of our results with some model benchmark, mostly because there is no data on how on average is reliability rated in uncontroversial situations. Because of that, we cannot determine, whose reliability is enhanced, and whose is belittled, even though we have found dependencies of reliability assessments, and gender. It is very hard to conclude with strong confidence, how the assessments would alter, should the gender bias not exist.

During the course of the study, the subjective assessment of the probability of the occurrence of the events described, has not been controlled. It means that in the study it was not possible to determine, how was the dissimilarity of assessed probabilities responsible for the scope of the observed differences. It is thus possible that part of the impact of the differences in assessments of reliability comes from (potentially false) belief about the probability of event frequency.

5 Conclusions

During the study, we have observed dependencies of gender and ascription of reliability. In particular, we have found that the reliability of men testifying that they had been assaulted by a woman has been ranked as uncontroversially lower. The reliability of men who testified that they had been assaulted by another man, has been assessed as uncontroversially higher. The instances of testifying women have been ranked differently by participants of different genders.

The interpretation of the results that integrates different points of view is one that indicates that differences in reliability ascription arise not only in virtue of expected event frequency, but they can also be explained by gender stereotype, or a gender role, through which a testifier is perceived.

The results of our research may constitute an argument for the existence of different epistemic status endowed on people depending on their gender and the existing gender stereotypes. Due to a, with high probability, populational

character of the discovered dependencies, it is plausible to assume that a gender bias lies at the core of these differences.

5.1 Further Research

Conclusions on the potential strong influence of legal education on both reliability assessments remains an avenue for future research to construct an experiment with a sample consisting exclusively of practicing lawyers, recruited through online blogs, emails or contacting the bar association and providing its members with a link to the survey. Another line of further research could be based on using different types of crimes in surveys (e.g., financial, clerical), to check if differences found in this paper are generalizable. Yet another line of further research could enrich the project of the experiment by controlling the subjective assessment of probability of events by the participants, and defining, how relevant a predictor of the observed differences could this subjective assessment be, or whether this justification could be considered as indicative of lesser gender bias. In the future, another research should be conducted, extending the experiment beyond the gender binary. Specifically, conducting the experiments on all queer participant groups could allow for establishing, whether the gender stereotypes have similar effect in these groups. Moreover, another experimental design should be developed, one that would allow for researching reliability assessments for testifiers of different than binary gender identities, but without excessive grow of the number of experimental conditions, to enable a reliable evaluation of results.

5.2 Practical Implications

Regardless of the genuine cause, the experiment shows that there is a general tendency to ascribe a lower level of reliability to some testimonies when only the gender of the victim changes—in other words, in some circumstances people are perceived as less trustworthy only because of their gender. Nevertheless, these findings may be connected with both, gender stereotypes operating within society, and smaller exposure of assault with certain gender arrangements (such as male victim of a woman) to the common imagination. For this reason, recipients of such testimonies, including professional judges, should exercise increased vigilance in such cases, to avoid bias caused by the stereotype and gender norms present in the culture and society. Officials receiving such testimonies should be advised to remain alert of the bias, and they should proceed to gather, possibly fast, all the relevant and advisably, detailed, information about the given case, to override the emerging interpretation based on gender. The gender of the testifier should at least not be the main, most important factor determining the decision, even though in certain cases it may be loaded with additional, relevant information (one example of such relevant information may be the frequency of sexual violence towards women). The practical implication of the experiment is that officials and other recipients of testimonies, prone to be affected by this bias, are especially exposed to the modification of

their assessment of reliability. In these situations, they should be extra attentive not to allow the bias to influence their assessment and be vigilant about the unjustified activation of stereotypes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11196-023-10055-6.

Acknowledgements We would like to thank Markus Kneer for helpful discussions. We would also like to express gratitude the Chair of Legal Theory at the Jagiellonian University in Kraków for funding one experiment.

Author contributions Klaudyna Horniczak - 40%; Andrzej Porębski - 40%; Izabela Skoczeń - 20%.

References

1. Alfano, Vincenzo, Tiziana Ramaci, Alfonso Landolfi, Alessandro Lo Presti, and Massimiliano Barattucci. 2021. Gender patterns in mobbing victims: Differences in negative act perceptions, MMPI personality profile, perceived quality of life, and suicide risk. *International Journal of Environmental Research and Public Health* 18 (4): 2192. https://doi.org/10.3390/ijerph18042192.
2. American Bar Association. 2022. ABA survey finds 1.3M lawyers in the U.S. 20 June. https://www.americanbar.org/news/abanews/aba-news-archives/2022/06/aba-lawyers-survey/.
3. Ask, Karl, and Pär Anders. Granhag. 2007. Motivational bias in criminal investigators' judgments of witness reliability. *Journal of Applied Social Psychology* 37: 561–591. https://doi.org/10.1111/j.1559-1816.2007.00175.x.
4. Barber, Christopher F. 2008. Domestic violence against men. *Nursing Standard* 51 (22): 35–39. https://doi.org/10.7748/ns2008.08.22.51.35.c6644.
5. Benatar, David. 2012. *The second sexism: discrimination against men and boys*. Malden, MA: Wiley-Blackwell.
6. Borysenko, Karlyn. 2020. The Dark Side Of #MeToo: What Happens When Men Are Falsely Accused. *Forbes*, 12 February. https://www.forbes.com/sites/karlynborysenko/2020/02/12/the-dark-side-of-metoo-what-happens-when-men-are-falsely-accused/?sh=.
7. Bressan, Dino. 2006. The role of women in Italian Legislation. *International Journal for the Semiotics of Law* 19 (1): 25–38. https://doi.org/10.1007/s11196-005-9009-2.
8. Castelli, Paola, Gail S. Goodman, and Simona Ghetti. 2005. Effects of interview style and witness age on perceptions of children's credibility in sexual abuse cases. *Journal of Applied Social Psychology* 35: 297–317. https://doi.org/10.1111/j.1559-1816.2005.tb02122.x.
9. Dent, Helen R., and Geoffrey M. Stephenson. 1979. Identification Evidence: Experimental Investigations of Factors Affecting the Reliability of Juvenile and Adult Witnesses. In *Psychology, Law and Legal Processes*, ed. David P. Farrington, Keith Hawkins, and Sally M. Lloyd-Bostock, 195–206. London: Palgrave Macmillan. https://doi.org/10.1007/978-1-349-04248-7_13

10. van Doorn, Janne, and N. Koster Nathalie. 2019. Emotional victims and the impact on credibility: A systematic review. *Aggression and Violent Behavior* 47: 74–89. https://doi.org/10.1016/j.avb.2019.03.00.
11. Eagly, Alice H., and Madeline E. Heilman. 2016. Gender and Leadership: Introduction to the Special Issue. *The Leadership Quarterly* 27 (3): 349–353. https://doi.org/10.1016/j.leaqua.2016.04.002.
12. Eagly, Alice H., and Wendy Wood. 1999. The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist* 54 (6): 408–423. https://doi.org/10.1037/0003-066X.54.6.408.
13. Edwards, Hailey Sweetland. 2018. How Christine Blasey Ford's Testimony Changed America. *TIME*, 15 October. https://time.com/5415027/christine-blasey-ford-testimony/
14. Ertürk, Abbas. 2013. Mobbing behaviour: Victims and the affected. *Educational Sciences: Theory & Practice* 13 (1): 169–173.
15. Faul, Franz, Edgar Erdfelder, Albert-Georg. Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39: 175–191.
16. Felson, Richard B., and Christopher Palmore. 2021. Traditionalism and victim blaming. *The Journal of Social Psychology* 161 (4): 492–507. https://doi.org/10.1080/00224545.2021.1896466.
17. Ferzan, Kimberly Kessler. 2021. #BelieveWomen and the Presumption of Innocence: Clarifying the Questions for Law and Life. In *Truth and Evidence*, ed. Melissa Schwartzberg and Philip Kitcher, 1–36. New York, NY: New York University Press.
18. Fricker, Miranda. 2007. Testimonial Injustice. In *Epistemic Injustice: Power and the Ethics of Knowing,* ed. Miranda Fricker, , 9–29. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.003.0002.
19. Fricker, Miranda. 2007. Hermeneutical Injustice. In *Epistemic Injustice: Power and the Ethics of Knowing,* ed. Miranda Fricker, , 147–175. Oxford University Press https://doi.org/10.1093/acprof:oso/9780198237907.003.0008.
20. Gardiner, Georgi. 2021. The "She Said, He Said" Paradox and the Proof Paradox. In *The Social Epistemology of Legal Trials*, ed. Zachary Hoskins and Jon Robson, 124–143. New York, NY: Routledge.
21. Gardiner, Georgi. 2021. Relevance and risk: How the relevant alternatives framework models the epistemology of risk. *Synthese* 199 (1–2): 481–511. https://doi.org/10.1007/s11229-020-02668-2.
22. Goodman, Steven. 2008. A dirty dozen: Twelve P-Value misconceptions. *Seminars in Hematology* 45 (3): 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003.
23. Hacking, Ian. 1995. The looping effects of human kinds. In *Causal cognition: A multidisciplinary debate*, ed. Dan Sperber, David Premack, and Ann James Premack, 351–383. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524021.003.0012.
24. Hines, Denise A., and Emily M. Douglas. 2009. Women's use of intimate partner violence against men: Prevalence, implications, and consequences. *Journal of Aggression, Maltreatment & Trauma* 18 (6): 572–586. https://doi.org/10.1080/10926770903103099.
25. Hildebrand-Edgar, Nicole, and Susan Ehrlich. 2017. "She was quite capable of asserting herself": Powerful speech styles and assessments of credibility in a sexual assault trial. *Language and Law.* 4 (2): 89–107.
26. Huiping, Wu., and Leung Shing-On. 2017. Can likert scales be treated as interval scales?–A simulation study. *Journal of Social Service Research* 43 (4): 527–532. https://doi.org/10.1080/01488376.2017.1329775.
27. Kahneman, Daniel. 2012. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
28. Kahneman, Daniel, and Amos Tversky. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185 (4157): 1124–1131. https://doi.org/10.1126/science.185.4157.1124.
29. Kassambara, Alboukadel. 2022. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. R package version 0.7.1. https://CRAN.R-project.org/package=rstatix.
30. Levon, Erez, and Yang Ye. 2020. Language, indexicality and gender ideologies: Contextual effects on the perceived credibility of women. *Gender and Language* 14 (2): 123–151. https://doi.org/10.1558/genl.39235.
31. Mooney, Annabelle. 2006. When a woman needs to be seen, heard and written as a woman: Rape, law and an argument against gender neutral language. *International Journal for the Semiotics of Law* 19: 39–68. https://doi.org/10.1007/s11196-005-9010-9.

32. Mooney, Annabelle. 2008. A Response to a response: Gender neutrality, rape and trial talk. *International Journal for the Semiotics of Law* 21: 157–160. https://doi.org/10.1007/s11196-008-9072-6.
33. Neal, Tess, and Stanley L. Brodsky. 2008. Expert witness credibility as a function of eye contact behavior and gender. *Criminal Justice and Behavior* 35 (12): 1515–1526. https://doi.org/10.1177/0093854808325405.
34. Norman, Geoffrey. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* 15 (5): 625–632. https://doi.org/10.1007/s10459-010-9222-y.
35. R Core Team. 2022. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.
36. Recanati, François. 2012. *Mental Files*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199659982.001.0001.
37. Rothman, Kenneth J. 1990. No adjustments are needed for multiple comparisons. *Epidemiology* 1 (1): 43–46.
38. RStudio Team. 2022. *RStudio: Integrated Development Environment for R. RStudio.* PBC, Boston, MA. http://www.rstudio.com/.
39. Rumney, Philip N.S.. 2008. Gender neutrality, rape and trial talk. *International Journal for the Semiotics of Law* 21: 139–155. https://doi.org/10.1007/s11196-008-9071-7.
40. Russell, Bertrand. 1910. Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society.* 11: 108–128.
41. Singmann, Henrik, Ben Bolker, Jake Westfall, Frederik Aus, and Mattan S. Ben-Shachar. 2022. *afex: Analysis of Factorial Experiments*. R package version 1.2–0. https://CRAN.R-project.org/package=afex.
42. Starr, Sonja. 2015. Estimating gender disparities in federal criminal cases. *American Law and Economics Review* 17 (1): 127–159. https://doi.org/10.1093/aler/ahu010.
43. Stepnick, Andrea, and James D. Orcutt. 1996. Conflicting testimony: Judges' and attorneys' perceptions of gender bias in legal settings. *Sex Roles* 34 (7–8): 567–579. https://doi.org/10.1007/BF01545033.
44. Voogt, Ashmyra, Bianca Klettke, and Angela Crossman. 2019. Measurement of victim credibility in child sexual assault cases: A systematic review. *Trauma, Violence, & Abuse* 20 (1): 51–66. https://doi.org/10.1177/1524838016683460.
45. Wasarhaley, Nesa E., Kellie R. Lynch, Jonathan M. Golding, and Claire M. Renzetti. 2015. The impact of gender stereotypes on legal perceptions of lesbian intimate partner violence. *Journal of Interpersonal Violence* 32 (5): 635–658. https://doi.org/10.1177/0886260515586370.
46. Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a World Beyond "p < 0.05". *The American Statistician* 73(S1): 1–19. https://doi.org/10.1080/00031305.2019.1583913.
47. Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino. McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.