Bias is persistent: Sequencing case information does not protect against contextual bias in criminal risk assessment

Oberlader, V.; Verschuere, B.

DOI
10.1111/lcrp.12279

Publication date
2025

Document Version
Final published version

Published in
Legal and Criminological Psychology

Received: 26 April 2024 | Accepted: 7 October 2024

DOI: 10.1111/lcrp.12279

ARTICLE

the british psychological society
promoting excellence in psychology

Bias is persistent: Sequencing case information does not protect against contextual bias in criminal risk assessment

Verena Oberlader[1] | Bruno Verschuere[2]

[1]Johannes Gutenberg University Mainz, Mainz, Germany

[2]University of Amsterdam, Amsterdam, The Netherlands

Correspondence
Verena Oberlader, Johannes Gutenberg University Mainz, Binger Straße 14-16, 55122 Mainz, Germany.
Email: verena.oberlader@uni-mainz.de

Abstract

Purpose: A large body of research indicates that bias is an inherent part of human information processing. This way, bias affects all disciplines that rely on human judgements, such as forensic psychological assessment, including criminal risk evaluation. Although there is a lack of empirical studies, scholars recommend considering case information sequentially beginning with the most relevant information to reduce the effect of potentially biasing task-irrelevant contextual information.

Methods: We ran a preregistered experimental study to test, first, whether task-irrelevant information results in bias effects when people use criminal risk assessment tools, and second, whether such bias could be reduced by sequencing case information according to its prognostic relevance. We collected data of 308 informed lay participants instructed to apply an empirical actuarial risk scale based on a case vignette.

Results: Results showed that task-irrelevant information biased risk assessment. Yet, sequencing case information did not protect against it.

Conclusions: Considering various boundary conditions (e.g., overconfidence in the accuracy of one's own assessment and other sources of bias), we discuss challenges to mitigate the biasing effect of task-irrelevant information.

KEYWORDS

bias, debiasing strategy, forensic psychological assessment, risk assessment, sequencing of case information

DEBIASING IN RISK ASSESSMENT

The impact of biases on human information processing is a well-established and central finding in psychological research. Pointing to common principles underlying different forms of bias, Oeberst and Imhoff (2023) define bias as the product of two inherent human tendencies: The tendency to form beliefs once information is processed, and the tendency to process further information in the light of these beliefs. This process-oriented bias approach can help to understand why expertise does not protect against contextual bias based on task-irrelevant information in forensic psychological assessments. For example, Bogaard et al. (2014, Study 1) showed that police officers rated the quality of a statement higher when task-irrelevant contextual information supported the credibility of a testifying person than when contextual information challenged a person's credibility.

Although experts should be able to correctly identify when information is task-irrelevant (see the performance-based approach to expertise by Shanteau et al., 2002), the processing of this information itself to identify it as task-irrelevant may lead to biasing beliefs. Numerous empirical studies have shown that people typically fail to prevent or correct the biasing influence of beliefs even when they are aware of it or explicitly motivated to avoid bias (e.g., Harley, 2007; Lieberman & Arndt, 2000). For example, in a meta-analysis ($k = 48$, $N = 8474$), Steblay et al. (2006) showed that jurors' verdicts were influenced by inadmissible evidence, even though they were instructed to ignore it. Given the extensive consequences of forensic psychological assessments on individuals' lives, action must be taken to mitigate such bias. Yet, empirical research on bias reduction in forensic psychological assessment is in its infancy (for a review see Neal et al., 2022; for proposed guidelines to minimize bias see Vredeveldt et al., 2022).

Although there is a lack of empirical research on the effectiveness of debiasing strategies in forensic psychological assessment, scholars recommend sequencing case information according to its diagnostic or prognostic relevance to reduce contextual bias (e.g., *Linear Sequential Unmasking-Expanded*, Dror & Kukucka, 2021). Ideally, irrelevant information could be completely withheld throughout the assessment process. However, this is not possible in forensic psychological assessments as they require the assessor to analyse all the information in the file. It is therefore only possible to sort information according to its relevance and introduce it into the assessment process at different times. Accordingly, rather than providing an expert with all available case information at the very beginning of the assessment process, a case manager should first share information with the highest diagnostic or prognostic value and withhold all other information, especially task-irrelevant information. Elaborate information management protocols, such as the LSU-E approach, include further instructions to reduce bias, for example in cases where the relevance of case information cannot be assessed (e.g., use of an independent expert, explicit consideration of alternative hypotheses; see Quigley-McBride et al., 2022). In this study, however, we focused on the role of sequencing case information as a key element of the LSU-E approach. In terms of the process-oriented bias framework, sequencing case information according to its relevance aims to prevent that the initial beliefs formed by an expert forms when confronted with case information are based on task-irrelevant information. We present how the debiasing strategy of sequencing case information could be applied in the context of criminal risk assessment and investigate its effectiveness in a preregistered experimental study.

SEQUENCING CASE INFORMATION AS DEBIASING STRATEGY IN RISK ASSESSMENT

In the context of criminal risk assessment, a large body of research indicates that empirical actuarial risk scales generally outperform other, especially non-standardized, risk assessment procedures in terms of objectivity, reliability and validity (e.g., Ægisdóttir et al., 2006; Hanson & Morton-Bourgon, 2009). It is therefore recommended that the risk assessment process should begin with applying empirical actuarial risk scales before a case-specific, non-standardized assessment of reoffending risk is made (e.g., Helmus, 2021). Following this, adopting the debiasing strategy of *sequencing case information according to its prognostic relevance* to criminal risk assessment implies that a case manager should initially share only the case information necessary to apply an empirical actuarial risk scale, while withholding other case information irrelevant to that task. The call to prioritize empirical actuarial risk scales in the process of risk assessment is consistent with existing recommendations (e.g., Helmus, 2021). However, the present approach adds the crucial detail that experts should be blinded to task-irrelevant case information before applying such a procedure; as opposed to the typical current practice, where experts work through the entire case file at the very beginning (see Figure 1).

Because they are highly standardized, empirical actuarial risk scales already provide protection against bias (see a study by Murrie et al., 2013, where an empirical actuarial risk scale was less prone to allegiance bias than a less standardized tool). Once it is clear how to apply a risk scale and interpret the result, the biasing effects of beliefs and belief-consistent information processing are limited. However, the use of an empirical actuarial risk scale does not provide complete protection against bias, as there is always some scope for subjective decision making (e.g., selecting from different norm tables or scoring items with incomplete file information; see Chevalier et al., 2015). However, when using empirical actuarial risk scales, the degrees of freedom and thus the scope for bias are kept as low as possible (see Oeberst & Oberlader, 2024, on the role of degrees of freedom regarding bias in forensic psychological reports). The question therefore arises as to how much bias can be mitigated by blinding assessors to task-irrelevant case information prior to the application of empirical actuarial risk scales?

Although there is little scope for reducing bias in the application and interpretation of an empirical actuarial risk scale, the use of sequencing case information could still reduce bias in the weighting of its results within the final risk assessment. To arrive at a final risk assessment, experts need not only to apply empirical actuarial risk scale scores but also to assess, interpret and integrate other data (e.g., the results of structured professional judgement or idiographic risk assessment; see for example Helmus, 2021). Empirical research has shown that within the process of data integration, the results of empirical actuarial risk scales are regularly overridden by an unstandardized professional assessment of case information reducing the predictive validity of the final risk assessment (e.g., Guay & Parent, 2018; Hanson et al., 2015; Schmidt et al., 2016; Wormith et al., 2012).

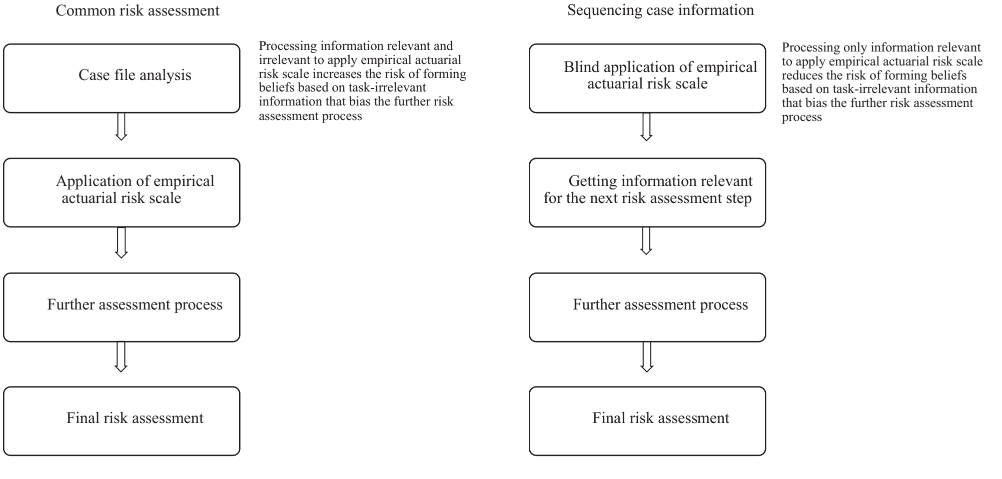

FIGURE 1 Common risk assessment procedure versus sequencing case information and corresponding information processing on the right of each string. This simplified illustration includes only information processing specification for the initial risk assessment step.

Present study

We conducted a preregistered experimental study to investigate the effectiveness of sequencing case information according to its prognostic relevance to reduce bias in the context of risk assessment. Do participants align their risk assessment more closely with the result of an empirical actuarial risk scale when they are initially blinded to task-irrelevant information?

We instructed informed laypersons to apply an empirical actuarial risk scale to a fictitious case and to decide how strongly they would align their final risk assessment with the risk score. As empirical actuarial risk scale, we used the Brief Assessment for Recidivism Risk-2002R (BARR-2002R; Babchishin et al., 2013). To test whether the decision of how strongly participants aligned their risk assessment with the BARR-2002R risk score was biased by task-irrelevant information, and if so, whether this bias could be reduced by sequencing case information, participants were randomly assigned to one of three conditions. These conditions differed in whether and at what stage task-irrelevant case information was provided. In the *fully blind-condition*, participants received only task-relevant case information before applying the BARR-2002R and making their risk assessment. In the *complete information-condition*, participants received task-relevant and task-irrelevant information before applying the BARR-2002R and then making their risk assessment. In the *sequential information-condition*, participants received task-relevant case information before applying the BARR-2002R and task-irrelevant case information afterwards, but before making their risk assessment (see Figure 2).

First, we tested the biasing effect of task-irrelevant information on the risk assessment. In Hypothesis 1, we supposed that participants' risk assessment would deviate more from the BARR-2002R risk score when they received task-irrelevant case information in addition to task-relevant case information in the complete information-condition than when they received only task-relevant case information in the fully blind-condition.

Second, we tested the debiasing effect of sequencing case information on risk assessment. In Hypothesis 2, we supposed that participants align their risk assessment more closely with the BARR-2002R risk score when they received task-irrelevant information after applying the BARR-2002R in the sequencing information-condition than when they received task-irrelevant information before applying it in the complete information-condition.

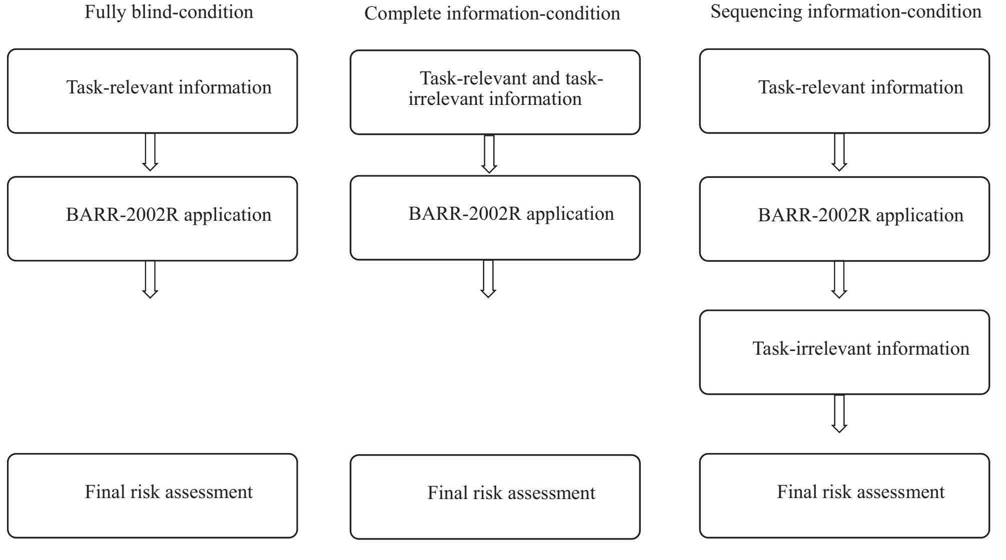

FIGURE 2 Simplified illustration of three experimental conditions. Not all steps of the study procedure are shown in this simplified illustration.

MATERIALS AND METHODS

Participants

We preregistered our intention to stop data collection once conclusive evidence for the alternative hypotheses ($BF_{10} > 5$) or the null hypotheses ($BF_{01} > 5$) of Hypothesis 1 and 2 was found, with a minimum of 200 and a maximum of 400 participants. After running 317 participants, we obtained conclusive evidence for the alternative over the null hypothesis for Hypothesis 1 testing the biasing effect of task-irrelevant information ($BF_{10} = 22.76$), and for the null over the alternative hypothesis for Hypothesis 2 testing the debiasing effect of sequencing ($BF_{01} = 6.45$), and stopped data collection.[1] Following our preregistered rules, we excluded nine participants who reported low data quality and two participants who answered at least two control questions incorrectly, resulting in a final sample of $N = 308$ participants (231 women, 74 men, two miscellaneous individuals, and one missing) with a mean age of 29.77 years ($SD = 13.22$).

Most of the sample had a high school diploma ($n = 151$) or higher education (bachelor's degree $n = 94$, master's degree/diploma $n = 47$, PhD $n = 2$). Thirteen participants had completed secondary education, and one participant reported no school diploma. Most participants ($n = 259$) reported that they had no previous experience with real-life risk assessment in personal, educational or professional context. Fifteen participants reported rarely dealing with risk assessment cases, 13 occasionally, nine frequently, and 12 very frequently. In addition, 263 participants stated that they were not familiar with the BARR-2002R prior to participating in the study, while 45 participants stated that they were familiar with it.

Procedure

Participants were recruited using convenience online sampling via by direct email, social networks and a university intern study network. Participants gave informed consent and were free to stop the experiment at any time. Student participants were offered course credit for their participation in the study. In addition, all participants were offered to receive a digital lesson on the challenges of risk assessment.

First, participants watched a self-made and subtitled introductory video on risk assessment focusing on the BARR-2002R (video length 01:28 mm:ss). Participants were instructed to watch the video carefully as they would then be asked to answer two multiple-choice quiz questions afterward. In addition, they were informed that they would be shown the video a second time if they answered at least one of these questions incorrectly. All participants answered both multiple-choice quiz questions correctly.

Participants were then randomly assigned to one of three experimental conditions (see Figure 2). They were instructed to apply the BARR-2002R to a fictitious case in which a man was sentenced to 4 years in prison for raping a female colleague. In the fully blind-condition, participants were given only task-relevant information necessary to apply the BARR-2002R. Participants were asked to read the presented case information carefully and then answer four BARR-2002R items. After that participants were informed that if the BARR-2002R was completed correctly, the fictitious character would have a below-average risk of recidivism (i.e., 12% risk of recidivism for violent offences including sexual offences or 20% general risk of recidivism within the next 5 years).

Next, participants were instructed to select a norm value for the BARR-2002R risk score. Afterwards, participants indicated how strongly they would align their risk assessment on the BARR-2002R risk score. In addition, participants were asked to justify their risk assessment in an open text box. Participants were then asked to respond to moderator variables (Experiential Engagement subscale of the Rational-Experiential Inventory-40, Keaton, 2017; Attitudes Toward Sexual Offenders scale, Hogue

[1]Following the preregistered sampling strategy, we could have terminated data collection at 250 participants, as we reached Bayes factors of at least five for Hypotheses 1 and 2 at this sample size. However, by the time we made this observation, 319 participants had already taken part in the study.

& Harper, 2018; Questionnaire on Implicit Theories about Sexual Offenders, Harper & Bartels, 2018; for detailed information see Moderator Variables). Afterwards, three multiple choice control questions were posed to ensure processing of the case information. These questions related to case information available to all three groups (e.g., age at release, frequency of prior charges, convictions and incarcerations). Finally, demographic variables (gender, age and education) and prior knowledge of risk assessment were assessed. At the end of the study, participants were asked to indicate whether or not their data should be used in terms of data quality. We pointed out that participants are sometimes distracted or unfocused, so answers are more likely to be random. No other variables were collected. The study was programmed in SoSciSurvey (Leiner, 2021).

The complete information-condition and the sequencing information-condition differed from the fully blind-condition only in that additional case information was presented that was irrelevant to the application of BARR-2002R, but, at different times. In the complete information-condition, participants received additional task-irrelevant case information at the same time as they received task-relevant case information, that is, before they applied the BARR-2002R. In sequencing information-condition, participants received task-relevant case information before applying the BARR-2002R and task-irrelevant case information after applying the BARR-2002R and after receiving feedback on the BARR-2002R result.

Participants did not differ statistically significant between the three conditions regarding age, F (2, 305) = 1.56, p = .213, η^2 = 0.01, gender, χ^2 (6) = 7.26, p = .297, φ = .15, prior experience in risk assessment, F (2, 305) = 1.46, p = .234, η^2 = 0.01, or prior knowledge of the BARR-2002R, χ^2 (2) = 1.75, p = .416, φ = .08.

Measures and material

Manipulating bias

To examine the biasing effect of task-irrelevant case information and the debiasing effect of sequencing case information on risk assessment, we selected task-relevant and task-irrelevant information to generate discrepant beliefs about the recidivism risk of the fictitious character. Task-relevant case information, that is, information required to apply the BARR-2002R, related to the offender's age at release and his previous involvement with the justice system. This information was chosen so that the correct application of the BARR-2002R would result in a below-average risk of recidivism. Task-irrelevant information, that is, information that was not necessary to apply the BARR-2002R, was chosen to reinforce or trigger the belief that the sexual offenders is high risk (e.g., morally unacceptable behaviour such as lack of empathy with victim or denial; Craissati, 2015; Huls et al., 2018).

Barr-2002R

The BARR-2002R is a highly standardized instrument for predicting the risk of violent including sexual and general recidivism of sexual offenders. It contains six items relating to the age of sexual offenders on release and their criminal history. We only used BARR-2002R items 1 to 4, because a pilot study showed that lay people (N = 10) had problems understanding BARR-2002R item 5 (i.e., years free prior to index sex offence) and BARR-2002R item 6 (i.e., any prior non-sexual violence sentencing occasion). The items were presented in the original multiple-choice format and participants were asked to select the correct item response.

Dependent variable risk assessment

To test the biasing effect of task-irrelevant case information and the debiasing effect of the sequencing of case information, participants indicated how strongly they would base their risk assessment on the

BARR-2002R risk score. They indicated on a five-point scale whether their risk assessment would be strongly below, below, equal to, above, or strongly above the BARR-2002R score.

To achieve interval scale level, the dependent variable 'risk assessment' was recoded into a three-point scale, with downward and upward deviations from the BARR-2002R risk score combined to indicate deviations regardless of the direction (i.e., 1 = risk assessment equal to BARR-2002R risk score, 2 = risk assessment below/above BARR-2002R risk score, 3 = risk assessment strongly below/strongly above BARR-2002R risk score).[2]

Moderator variables

In addition, we assessed intuitive decision making (four selected items from the Experiential Engagement subscale of the Rational-Experiential Inventory-40, Keaton, 2017) and attitudes toward sexual offenders (four selected items from the 21-item short form of the Attitudes Toward Sexual Offenders scale, ATS-21, Hogue & Harper, 2018; three-item questionnaire on implicit theories about sexual offenders IT-SO, Harper & Bartels, 2018). All items were answered on a five-point Likert scale ranging from 'strongly disagree' to 'strongly agree'.

However, as the Cronbach's α of the ATS-21 and IT-SO were not acceptable (ATS-21 Cronbach's $\alpha = .66$, IT-SO Cronbach's $\alpha = .54$), we did not examine these moderating effects. The selected items, hypotheses, results and discussion for the moderator analysis of intuitively decision making are available on the *Open Science Framework* (OSF, https://osf.io/6xg8b/?view_only=76b00669ce9e4292813289171f7e58ae).

Deviations from preregistration

Unlike preregistered, due to a programming error (i.e., the Likert scale was presented in a wrong order), we could not use the responses to the item 'norm selection' as a second dependent variable. In addition, we decided to recode the dependent variable risk assessment to achieve interval scale level.

Data and materials sharing

The study design and the analysis plan were preregistered at OSF. All study materials, data, and analysis scripts, as well as the approval by the Ethics Committee of the Department of Psychology at the University of Mainz, Germany, are available on OSF.

RESULTS

Preregistered statistical analyses

Biasing effect of task-irrelevant information on risk assessment

The biasing effect of task-irrelevant case information on the risk assessment was tested with a one-tailed Bayesian independent samples *t*-test. The grouping factor was fully blind-condition versus complete information-condition and the dependent variable was risk assessment. We compared the null hypothesis H0: $\delta = 0$ (i.e., no difference in risk assessment between the two conditions) with a one-sided

[2] Alternatively, we could have excluded the participants who deviated downwards in their risk assessment. A reanalysis excluding these participants showed that the results of the confirmatory hypothesis tests did not change significantly (Hypothesis 1 $BF_{10} = 14.75$, $n_{\text{Fully blind condition}} = 101$, $n_{\text{complete information condition}} = 95$; Hypothesis 2 $BF_{01} = 6.10$, $n_{\text{sequencing information condition}} = 99$, $n_{\text{complete information condition}} = 95$).

alternative hypothesis H1: $\delta > 0$ (i.e., greater deviation from the BARR-2002R score in risk assessment in the complete information- vs. fully blind-condition), where δ is the standardized effect size (i.e., the population version of Cohen's *d*).

When applying Bayesian statistics, one needs to define a prior that represents the prior knowledge about the parameters of a distribution. Alternatively, one can use an uninformed prior to incorporate minimal prior knowledge in a Bayesian analysis allowing the data to be the primary source of information in estimating parameters. We selected an uninformed prior option for the independent samples *t*-test, that is, a Cauchy distribution with spread *r* set to .707. Since we specified a one-sided alternative hypothesis, the prior distributions were truncated at zero, so that only positive effect sizes values were allowed. The robustness of the Bayes factor to this prior specification was assessed in Bayes factor robustness plots, taking into account different prior widths. This specification and description apply to all further *t*-tests reported. We ran statistical analyses in JASP (version 0.17.2.1; JASP Team, 2023).

Figure 3 shows that the Bayes factor of a one-tailed independent *t*-test strongly favours H1 over H0. That is, participants who received task-irrelevant case information in addition to task-relevant case information in the complete information-condition deviated more from the BARR-2002R score in their risk assessment than participants who received only task-relevant case information in the fully blind-condition (see Table 1 for descriptive statistics). Specifically, $BF_{10} = 21.53$ means that the data are approximately 21.5 times more likely under H1 than under H0 ($BF_{01} = 0.05$). The error percentage was $\sim 1.540 \times 10^{-4}$%, which indicates great numerical stability of the result. The Bayes factor robustness check showed high stability of Bayes factors over a wide range of prior widths *r* with most values indicating strong evidence (see Figure 3). The median of the resulting posterior distribution for δ was moderate with a 95% credible interval including small to moderate parameter estimations, $\delta = 0.40$, 95% CI [0.14; 0.68].

In the complete information-condition, 76% of the participants deviated from the BARR-2002R score, as compared to 48% of participants in the fully blind-condition. In both conditions, the deviations were almost exclusively upward (see Table 1).

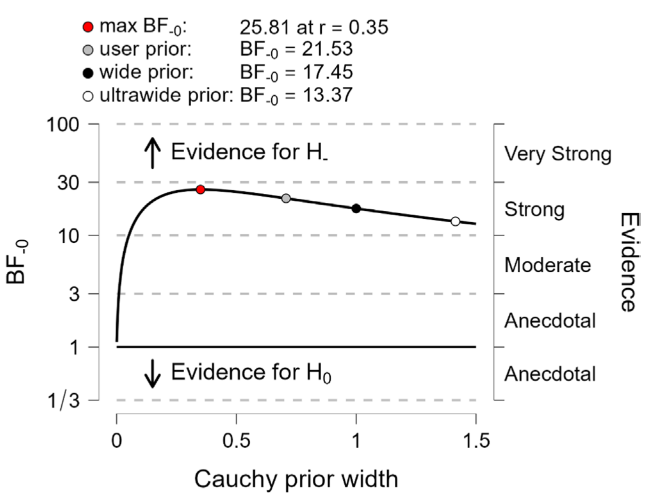

FIGURE 3 Bayes Factor Robustness Check for Biasing Effect on Risk Assessment. The figure shows a Bayes factor robustness plot for a one-sided Bayesian independent samples *t*-test with the grouping variable fully blind-condition versus complete information-condition and the dependent variable risk assessment. The plot indicates BF_{10} for the user specified prior ($r = 1/\sqrt{2}$), wide prior ($r = 1$), and ultrawide prior ($r = \sqrt{2}$). Figure from JASP.

TABLE 1 The extent of deviation from BARR-2002R risk score in risk assessment in experimental conditions.

Experimental condition	(in)correct BARR-2002R application	*n*	*M*	*SD*	SE	% of participants deviating from BARR-2002R risk score	% of deviation going upward	% of deviation going downward
Fully blind	All	106	1.58	0.63	0.06	48	96	4
	Only correct	86	1.53	0.64	0.07	41	94	6
Complete information	All	99	1.82	0.55	0.06	76	96	4
	Only correct	67	1.91	0.57	0.07	82	96	4
Sequencing information	All	103	1.85	0.62	0.06	72	95	5
	Only correct	79	1.84	0.59	0.07	76	97	3

Note: Higher mean values in the risk assessment indicate stronger deviation from the BARR-2002R risk score. The second column represents all participants, that is, participants that applied the BARR-2002R incorrectly and correctly or only participants that applied it correctly.

Debiasing effect of sequencing information on risk assessment

The debiasing effect of sequencing case information on the risk assessment was tested with a one-tailed Bayesian independent samples *t*-test. The grouping factor was the sequencing information-condition versus complete information-condition and the dependent variable was risk assessment. We compared the null hypothesis H0: $\delta = 0$ (i.e., no difference in the risk assessment between the two conditions) with a one-sided alternative hypothesis H1: $\delta < 0$ (i.e., lower deviation from the BARR-2002R result in risk assessment in the sequencing information- vs. the complete information-condition), where δ is the standardized effect size (i.e., the population version of Cohen's *d*).

Figure 4 shows that the Bayes factor of a one-tailed independent *t*-test moderately favours H0 over H1, that is, the deviation from the BARR-2002R score in the risk assessment did not differ between participants in the sequencing information-condition versus participants in the complete information-condition (see Table 1 for descriptive statistics). Specifically, $BF_{01} = 7.55$ means that the data are approximately 7.5 times more likely under H0 than under H1. The error percentage was ~0.008%, which indicates great numerical stability of the result. The Bayes factor robustness check revealed low stability of the Bayes factors over a wide range of prior widths *r* with values indicating moderate-to-strong evidence (see Figure 4). The median of the resulting posterior distribution for δ was close to zero with a 95% credible interval including parameter estimations between almost zero and small parameter estimations, $\delta = -0.08$, 95% CI [−0.29; −0.01].

In the complete information-condition, 76% of the participants deviated from the BARR-2002R score, as compared to 72% of participants in the sequencing information-condition. In both conditions, the deviations were almost exclusively upward (see Table 1).

Exploratory analyses

Biasing and debiasing effect excluding incorrect BARR-2002R results

We reran the two one-tailed Bayesian independent *t*-tests to examine the biasing and debiasing effects as described above, excluding 76 participants (25% of the total sample) who did not correctly apply

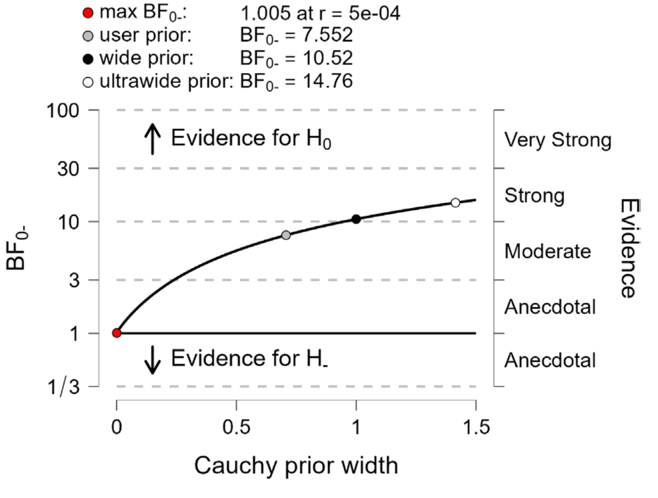

FIGURE 4 Bayes factor robustness check for debiasing effect on risk assessment. The figure shows a Bayes factor robustness plot for a one-sided Bayesian independent samples *t*-test with the grouping variable sequencing information-condition versus complete information-condition and the dependent variable risk assessment. The plot indicates BF_{10} for the user specified prior ($r = 1/\sqrt{2}$), wide prior ($r = 1$), and ultrawide prior ($r = \sqrt{2}$). Figure from JASP.

the BARR-2002R. Descriptive results are presented in Table 1. Excluding these participants had no substantial effects on our results. The evidence was still strong (even stronger) for the biasing effect of task-irrelevant case information on the risk assessment increased $BF_{10} = 189.74$. For the debiasing effect of sequencing case information on the risk assessment, the results still favoured H0 over H1, but the Bayes factor decreased from $BF_{01} = 7.55$ to $BF_{01} = 2.75$.

Classification of justification of risk assessment

Justifications of risk assessment were classified into seven categories by a student assistant blind to the hypotheses: (1) Use of case-specific task-irrelevant information to justify upward deviation in risk assessment, (2) case-unspecific arguments to justify upward deviation in risk assessment, (3) use of case-specific task-irrelevant information to justify downward deviation in risk assessment, (4) case-unspecific arguments to justify downward deviation in risk assessment, (5) reporting confidence in BARR-2002R to justify no deviation, (6) using other arguments to justify no deviation and (7) residual category for responses that were inconclusive or could not be coded.

Table 2 shows the distribution of risk assessment justifications across the seven categories for the three experimental conditions. When participants received task-irrelevant information, either in the complete information-condition or the sequencing information-condition, about three-quarters of participants used case-specific arguments to justify their upward deviation (e.g., no victim empathy). When participants in the fully blind-condition received only task-relevant information, they used only case-unspecific arguments to justify their upward deviation (e.g., expecting of higher base rates for sexual recidivism). Thus, when task-irrelevant information was presented, it was used more often than task-relevant information to justify the risk assessment.

DISCUSSION

To our knowledge, this was the first empirical study to examine the effectiveness of sequencing case information according to its prognostic relevance in reducing bias in the context of criminal risk assessment. We randomly assigned participants to three conditions that differed in whether task-irrelevant information was provided or not and at what stage. We assessed the extent to which task-irrelevant information caused the risk assessment to deviate from the outcome of an empirical actuarial risk scale, and whether sequencing protected against such bias.

Biasing effect of task-irrelevant information

As supposed in Hypothesis 1, we found substantial evidence, with a moderate effect size, that participants' risk assessment was biased by task-irrelevant case information. In the complete information-condition, about 75% of participants deviated from the BARR-2002R risk score, almost invariably upwards and on average moderately strongly. In comparison, in the fully blind-condition, about 50% deviated from the BARR-2002R risk score, again, almost invariably upward, but to a lesser extent. How do these results relate to the process-oriented bias approach, which points to the role of beliefs plus belief-consistent information processing?

Overconfidence in accuracy of one's own assessment

In the complete information condition, we experimentally manipulated discrepant beliefs about the fictitious offender's recidivism risk: While the BARR-2002R resulted in a recidivism risk

TABLE 2 Proportion of risk assessment justifications across seven categories (absolute numbers).

	Down-ward deviation		No deviation		Upward deviation			
Condition	**Case-specific arguments**	**Non-case-specific arguments**	**Trust in BARR-2002R**	**Other reasons**	**Non-case-specific arguments**	**Case-specific arguments**	**No classification**	**Total**
Fully blind	2	0	28	25	47	0	4	106
Complete information	1	2	14	5	14	56	7	99
Sequencing information	1	3	9	14	17	53	6	103
Total	4	5	51	44	78	109	17	308

below-average, task-irrelevant information was chosen to indicate high risk. As there was no instruction on how to evaluate task-irrelevant information, participants could have resolved these discrepant beliefs in manifold ways. Yet, participants systematically resolved them in favour of beliefs based on task-irrelevant information resulting in an upward bias. This finding is consistent to with the widespread phenomenon of professional override in the context of criminal risk assessment. As stated in the Introduction, empirical studies have shown that experts regularly override the results of empirical actuarial risk scales by an unstandardized professional assessment of case information, almost always upward (e.g., Guay & Parent, 2018; Hanson et al., 2015; Schmidt et al., 2016; Wormith et al., 2012).

One explanation for the systematic upward adjustment is based on peoples' tendency to overestimate the accuracy of their own assessments (e.g., *bias blind spot*; Pronin et al., 2002; for the forensic context see Zappala et al., 2018). If people believe that they are making correct assessments, we should expect them to give their own assessment a higher wight when it conflicts with data that based on a highly standardized risk assessment tool (on the role of fundamental beliefs see Oeberst & Imhoff, 2023). This general overconfidence in the accuracy of one's own assessment may have been reinforced by the specific manipulation of task-irrelevant information as morally unacceptable behaviour. Empirical research indicates that people prefer moral justifications when explaining deviant behaviour (e.g., on the role of moral judgements when explaining suboptimal behaviour of others see De Freitas & Johnson, 2018). This human tendency is consistent with the finding that approximately three-quarters of participants used task-irrelevant information to justify their upward deviation in risk assessment.

In contrast, about 50% of the participants who did not deviate from the BARR-2002R risk score based their decision on trusting the risk scale. Empirical studies in other applied contexts show that standardized procedures or automated systems only reach their full potential when users have an appropriate level of trust (e.g., for highly automated vehicles see Kraus et al., 2020). The same interaction could apply to criminal risk assessment: People need to trust in empirical actuarial risk scales to base their risk assessment on their results. However, they must also be able to recognize case-specific exceptions (Helmus, 2021).

Against the background of these consideration, future studies should investigate the role of confidence in the accuracy of one's own unstandardized risk assessment versus empirical actuarial risk scales. It could be a set screw to increase the bias-buffering effect of applying standardized procedures blind for task-irrelevant information.

This discussion highlights that the use of standardized risk assessment tools alone may buffer but not fully protect against bias. This is also shown in a study by Murrie et al. (2013): here trained practitioners rated the likelihood of recidivism of a person who had committed a sexual offence as higher if they thought that they had been instructed by the prosecution rather than by the defence. Although this allegiance bias was more pronounced when participants used a partially standardized procedure to assess the likelihood of reoffending (PCL-R; Hare, 2003), it also occurred when they used a highly standardized risk assessment tool (Static-99R; Helmus et al., 2012).

Other bias sources

In addition, the high proportion of participants in the fully blind-condition that deviated from the BARR-2002R risk score indicated that bias is not limited to task-irrelevant information. Approximately 50% of participants in this condition made an upward adjustment without receiving task-irrelevant information. This finding is consistent with the observation that people generally overestimate the recidivism risk of sexual offenders (e.g., Jung et al., 2014; Olver & Barlow, 2010). Moreover, it underscores that bias in forensic decision-making can have both case-specific but also case-unspecific sources (Dror, 2020).

Sequencing case-information did not reduce bias

Unfortunately, and in contrast to Hypothesis 2, the results showed that providing task-irrelevant information only after participants had applied an empirical actuarial risk scale did not protect against contextual bias in risk assessment. Specifically, participants deviated comparably strong from the BARR-2002R risk score regardless of whether task-irrelevant information was provided before or after application.

This null effect could have several causes. First, it could be due to a limited statistical power. However, as the means of the risk assessment deviation in both conditions differed only to the second decimal place, the question of statistical power is of little interest. Even if Bayes factors had supported the alternative hypothesis, such a small mean difference would have no practical relevance. Second, the null effect could also be due to an ineffective manipulation. However, since the results of Hypothesis 1 demonstrated that bias by task-irrelevant information was successfully induced this is highly unlikely. Third, the null effect could be due to the way in which the dependent variable was operationalized, namely as the extent of deviation from the BARR-2002R risk score and not as an independent risk assessment. In further studies, the biased deviation from the risk score could be measured on a percentage scale from 0 to 100%.

Therefore, we suppose that sequencing case information, at least as implemented in the present study, is not effective in reducing contextual bias if the process of data integration is not standardized. The results of our study suggest that in the absence of decision rules on how to integrate the result of an empirical actuarial risk scale with the non-standardized assessment of case information, the latter may still bias the final risk assessment. Initial beliefs based on the results of an empirical actuarial risk scale and free of task-irrelevant information could not protect against their later biasing effects. Against this consideration, standards for when to adhere to the outcome of an empirical actuarial risk scale and when and how much to deviate from it (Helmus, 2021), could help the present debiasing strategy to work.

Limitations

The limitations of the present study relate primarily to its ecological validity. First, given the difficulty in accessing to the limited sample of forensic psychological experts, we conducted this first study of the effectiveness of this debiasing strategy in a sample of informed lay participants. Of course, to make any generalizations for the context of risk assessment, the results need to be replicated in a representative sample of experts using a complete risk assessment tool. However, since the underlying processes of bias are inherent human, we do not expect experts to behave differently. Second, we simulated only a very limited part of the comprehensive and complex process of risk assessment. To make the risk assessment feasible for lay participants, we have chosen an easy-to-use empirical actuarial risk scale and provided straightforward information on how to use it.

CONCLUSION

To our knowledge, this is the first empirical study of the effectiveness of sequencing as a debiasing strategy in the context of risk assessment. The results suggest that sequencing case information according to its prognostic relevance does not protect against bias. Yet, further research is needed to investigate whether this finding is replicated in an expert sample. This research should delve deeper into the role of confidence in the accuracy of one's own assessment versus highly standardized risk assessment tools and consider several boundary conditions like standardization. In addition, future research could also address the combination of different debiasing approaches, as suggested in the LSU-E approach (e.g., considering alternative hypotheses, requesting a second independent expert). Even if the empirical focus should be on effectiveness, practical aspects of feasibility should also be discussed (e.g., cost and time aspects of involving a second independent expert). For now, the present findings underscore the difficulty of reducing the impact of contextual bias introduced by task-irrelevant information in criminal risk assessment.

AUTHOR CONTRIBUTIONS

Verena Oberlader: Conceptualization; writing – original draft; formal analysis; investigation; methodology; data curation. **Bruno Verschuere:** Writing – review and editing; conceptualization.

CONFLICT OF INTEREST STATEMENT

Both authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at Open Science Framework at https://osf.io/6xg8b/?view_only=76b00669ce9e4292813289171f7e58ae.

ORCID

Verena Oberlader https://orcid.org/0000-0003-1902-1330
Bruno Verschuere https://orcid.org/0000-0002-6161-4415

REFERENCES

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Babchishin, K. M., Hanson, R. K., & Blais, J. (2013). *User guide for the Brief Assessment for Recidivism Risk–2002R (BARR-2002R)*. [Unpublished manual] www.static99.org

Bogaard, G., Meijer, E. H., Vrij, A., Broers, N. J., & Merckelbach, H. (2014). Contextual bias in verbal credibility assessment: Criteria-based content analysis, reality monitoring and scientific content analysis. *Applied Cognitive Psychology*, *28*(1), 79–90. https://doi.org/10.1002/acp.2959

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C., & Varela, J. G. (2015). Static-99R reporting practices in sexually violent predator cases: Does norm selection reflect adversarial allegiance? *Law and Human Behavior*, *39*(3), 209–218. https://doi.org/10.1037/lhb0000114

Craissati, J. (2015). Should we worry about sex offenders who deny their offences? *Probation Journal*, *62*(4), 395–405. https://doi.org/10.1177/0264550515600543

De Freitas, J., & Johnson, S. G. (2018). Optimality bias in moral judgment. *Journal of Experimental Social Psychology*, *79*, 149–163. https://doi.org/10.1016/j.jesp.2018.07.011

Dror, I. E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical Chemistry*, *92*(12), 7998–8004. https://doi.org/10.1021/acs.analchem.0c00704

Dror, I. E., & Kukucka, J. (2021). Linear sequential unmasking–expanded (LSU-E): A general approach for improving decision making as well as minimizing noise and bias. *Forensic Science International: Synergy*, *3*, 100161. https://doi.org/10.1016/j.fsisyn.2021.100161

Guay, J.-P., & Parent, G. (2018). Broken legs, clinical overrides, and recidivism risk: An analysis of decisions to adjust risk levels with the LS/CMI. *Criminal Justice and Behavior*, *45*, 82–100. https://doi.org/10.1177/0093854817719482

Hanson, R. K., Helmus, L. M., & Harris, A. J. R. (2015). Assessing the risk and needs of supervised sexual offenders: A prospective study using STABLE-2007, static-99R, and static-2002R. *Criminal Justice and Behavior*, *42*, 1205–1224. https://doi.org/10.1177/0093854815602094

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, *21*, 1–21. https://doi.org/10.1037/a0014421

Hare, R. D. (2003). *The hare psychopathy checklist-revised* (2nd ed.). Toronto: Multi-Health Systems.

Harley, E. M. (2007). Hindsight bias in legal decision making. *Social Cognition*, *25*(1), 48–63. https://doi.org/10.1521/soco.2007.25.1.48

Harper, C. A., & Bartels, R. M. (2018). Implicit theories and offender representativeness in judgments about sexual crime. *Sexual Abuse*, *30*(3), 276–295. https://doi.org/10.1177/1079063216658019

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the Predictive Accuracy of Static-99 and Static-2002 With Older Sex Offenders: Revised Age Weights. *Sexual Abuse*, *24*(1), 64–101. https://doi.org/10.1177/1079063211409951

Helmus, L. M. (2021). Estimating the probability of sexual recidivism among men charged or convicted of sexual offences: Evidence-based guidance for applied evaluators. *Sexual Offending: Theory, Research, and Prevention*, *16*, 1–24. https://doi.org/10.5964/sotrap.4283

Hogue, T. E., & Harper, C. A. (2018). *Attitudes to sexual offenders scale—Short form (ATS, ATS-21)*. APA PsycTests. https://doi.org/10.1037/t70539-000

Huls, L., Nentjes, L., Rinne, T., & Verschuere, B. (2018). Gebruikte versus werkelijke voorspellers van seksuele recidive: invloed van morele verwerpelijkheid op oordeel pro Justitia-rapporteurs. *Tijdschrift voor Psychiatrie*, *2*, 78–86.

JASP Team. (2023). *JASP (Version 0.17.2.1) [Computer software]*. https://jasp-stats.org/

Jung, S., Ahn-Redding, H., & Allison, M. (2014). Crimes and punishment: Understanding of the criminal code. *Canadian Journal of Criminology and Criminal Justice*, *56*(3), 341–366. https://doi.org/10.3138/cjccj.2013.E17

Keaton, S. A. (2017). Rational-experiential inventory-40 (REI-40). In D. L. Worthington & D. B. Graham (Eds.), *The sourcebook of listening research: Methodology and measures* (pp. 530–536). Wiley Online Library. https://doi.org/10.1002/9781119102991.ch59

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors*, *62*(5), 718–736. https://doi.org/10.1177/0018720819853686

Leiner, D. J. (2021). *SoSci Survey (3.1.06)*. [Software]. https://www.soscisurvey.de

Lieberman, J. D., & Arndt, J. (2000). Understanding the limits of limiting instructions: Social psychological explanations for the failures of instructions to disregard pretrial publicity and other inadmissible evidence. *Psychology, Public Policy, and Law*, *6*(3), 677–711. https://doi.org/10.1037/1076-8971.6.3.677

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science*, *24*(10), 1889–1897. https://doi.org/10.1177/0956797613481812

Neal, T. M., Martire, K. A., Johan, J. L., Mathers, E. M., & Otto, R. K. (2022). The law meets psychological expertise: Eight best practices to improve forensic psychological assessment. *Annual Review of Law and Social Science*, *18*, 169–192. https://doi.org/10.1146/annurev-lawsocsci-050420-010148

Oeberst, A., & Imhoff, R. (2023). Toward parsimony in bias research: A proposed common framework of belief-consistent information processing for a set of biases. *Perspectives on Psychological Science*, *18*, 17456916221148147. https://doi.org/10.1177/17456916221148147

Oeberst, A., & Oberlader, V. (2024). Degrees of freedom as breeding ground for biases—A threat to forensic practice. *Law and Human Behavior*. https://doi.org/10.1037/lhb0000579

Olver, M. E., & Barlow, A. A. (2010). Public attitudes toward sex offenders and their relationship to personality traits and demographic characteristics. *Behavioral Sciences & the Law*, *28*(6), 832–849. https://doi.org/10.1002/bsl.959

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381.

Quigley-McBride, A., Dror, I. E., Roy, T., Garrett, B. L., & Kukucka, J. (2022). A practical tool for information management in forensic decisions: Using linear sequential unmasking-expanded (LSU-E) in casework. *Forensic Science International: Synergy*, *4*, 100216. https://doi.org/10.1016/j.fsisyn.2022.100216

Schmidt, F., Sinclair, S. M., & Thomasdóttir, S. (2016). Predictive validity of the youth level of service/case management inventory with youth who have committed sexual and non-sexual offences. *Criminal Justice and Behavior*, *43*, 413–430. https://doi.org/10.1177/0093854815603389

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, *136*(2), 253–263. https://doi.org/10.1016/S0377-2217(01)00113-8

Steblay, N., Hosch, H. M., Culhane, S. E., & McWethy, A. (2006). The impact on juror verdicts of judicial instruction to disregard inadmissible evidence: A meta-analysis. *Law and Human Behavior*, *30*, 469–492. https://doi.org/10.1007/s10979-006-9039-7

Vredeveldt, A., van Rosmalen, E. A., Van Koppen, P. J., Dror, I. E., & Otgaar, H. (2022). Legal psychologists as experts: Guidelines for minimizing bias. *Psychology, Crime & Law*, *30*, 705–729. https://doi.org/10.1080/1068316X.2022.2114476

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior*, *39*, 1511–1538. https://doi.org/10.1177/0093854812455741

Zappala, M., Reed, A. L., Beltrani, A., Zapf, P. A., & Otto, R. K. (2018). Anything you can do, I can do better: Bias awareness in forensic evaluators. *Journal of Forensic Psychology Research and Practice*, *18*(1), 45–56. https://doi.org/10.1080/24732850.2017.1413532

How to cite this article: Oberlader, V., & Verschuere, B. (2025). Bias is persistent: Sequencing case information does not protect against contextual bias in criminal risk assessment. *Legal and Criminological Psychology*, *30*, 143–158. https://doi.org/10.1111/lcrp.12279