What is the external validity of sentencing research?

A multi-level meta-analysis of race and gender disparities

Jose Pina-Sánchez & Ian Brunton-Smith

Sentencing research is rarely cross-jurisdictional. More problematically, most quantitative sentencing research is based on a limited number of American jurisdictions where court data is available. As a result, it is difficult to assess the extent to which key findings from the sentencing literature apply universally. We build on the recent growth of sentencing research outside the US to explore the external validity of studies reporting the conditional association of offenders' race and gender with sentence length. To do so, we conduct two multi-level meta-analyses, distinguishing the proportion of between-study heterogeneity attributable to differences at the study and jurisdiction levels. Our findings reveal that while race disparities in sentencing are statistically significant, they are minimal in magnitude (a 3% penalty for racial minorities) and remarkably consistent across jurisdictions. In contrast, gender disparities are more pronounced (a 13% penalty against men) but highly variable, with some jurisdictions showing parity. Both analyses uncover substantial variability due to sample and modelling choices, highlighting the limited generalisability of existing sentencing research. We urge caution in interpreting findings from single studies on sentencing disparities and advocate for the pre-registration of analytical strategies to mitigate researcher bias.

1 Introduction

Jurisdictions worldwide generally adhere to a core set of sentencing goals: punishing wrongdoing, deterring crime, rehabilitating offenders, protecting the public, and restoring victims. However, the relative emphasis placed on achieving these goals and the legal frameworks through which they are pursued vary substantially across jurisdictions. This variation results in markedly distinct sentencing practices, shaped by both formal legal statutes and informal judicial norms.

Even within a specific jurisdiction, criminal courts can differ significantly depending on their specialisation in certain types of crimes and offenders. For instance, youth courts prioritise rehabilitation over other sentencing goals, whereas lower-tier courts are typically constrained in the severity of the sentences they are permitted to impose. In fact, it is well documented that, even within the same court type, notable variations in culture emerge across court actors based in different locations, reflecting diverse judicial philosophies, interpretative approaches, and procedural norms (Nardulli, Flemming, & Eisenstein, 1988; Ulmer & Johnson, 2004; Ulmer & Kramer, 1996).

Instead of acknowledging this fragmented reality, most sentencing studies focus on a single jurisdiction. The result is a highly atomised literature in which it remains unclear whether key findings apply universally, across countries with similar legal traditions (e.g., common law or civil law), or even across different jurisdictions within the same country. For example, should we expect the more lenient sentencing of female offenders (Pina Sánchez & Harris, 2020; Starr, 2015; Steffensmeier, Kramer, & Streifel, 1993), to be more pronounced in jurisdictions where judges have greater discretion, compared to those with highly prescriptive sentencing guidelines? Or, perhaps, should we expect such disparities to be smaller in societies with greater gender equality?

The external validity of the sentencing evidence base is further constrained by data availability, as most large-scale empirical studies have traditionally relied on datasets from a limited number of jurisdictions. Primarily, these include the United States Sentencing Commission and select state sentencing authorities, such as those in Pennsylvania, Minnesota, Arizona and Florida, where court data has been

readily accessible. Findings from this handful of American jurisdictions have disproportionately shaped the academic discussion on sentencing, and, in turn, influenced the types of sentencing reforms implemented worldwide[1].

Here, we draw on the growth of sentencing research outside the US in recent years to document the between jurisdiction variability in the literature exploring gender and race/ethnic[2] disparities in sentence length. We focus on these two extra-legal factors due to their academic and policy significance. Race and gender are the two most widely studied factors in the sentencing literature, as evidenced by the fact they remain the only two factors whose effects on sentencing has been explored through meta-analysis or systematic reviews (see Mitchell, 2005; Pratt, 1998; and Ferguson & Smith, 2024 for the case of race disparities; and Bontrager, Barrick, & Stupi, 2013; and Daly & Bordt, 1995 for the case gender disparities). Race disparities are also frequently cited as a justification for sentencing reform (see, for example, the First Step Act passed by the US Congress in 2018), underscoring their relevance beyond academic discourse.

Specifically, we estimate a series of multi-level meta-analyses, considering differences at the estimate, study, and jurisdiction levels. This enables us to assess the overall generalisability of findings from the sentencing literature, identifying whether between-study variability is due to genuine differences in sentencing practices across jurisdictions, or to other study-specific decisions such as the modelling choices or the offence groups considered.

Furthermore, building on the opportunities provided by meta-analysis, we expand our exploration of the external validity of sentencing research more widely by considering the presence of selective reporting and referencing. We do so by considering whether the effect sizes reported in the literature affect how studies are framed and their influence, as measured by the number of citations received, standardised by year of publication.

[1] See for example the 2019 Sentencing Review published by the Northern Ireland Department of Justice, or the response from the Judiciary in England and Wales in regarding the recommendation to establish a Sentencing Council in England and Wales.

[2] For simplicity, in this study we do not distinguish between race and ethnicity, using the terms interchangeably.

2 Methods

In this section we outline the criteria used to create a pool of potentially eligible studies, select those deemed eligible, record relevant information, harmonise effect sizes across studies, and specify our

multilevel-models. This analytic strategy was pre-registered before data collection. The pre-registration, along with an explanation of any deviations from our plan, is available on the project's the Open Science Foundation site: osf.io/y4gpf). In addition, to facilitate the reproducibility and expansion of our findings, we have integrated the code in the write-up of our article using Quarto V1.6.39 and R V4.4.2, and uploaded both the code and data to the project's Open Science Foundation site.

2.1 Inclusion Criteria

We identified 1,024 potentially relevant studies to be screened. This pool was created using Scopus. The search criteria were academic articles, written in English, published since 2000, containing the following string of terms either in the article's title, abstract or keywords: "sentencing" AND ("data" OR "quantitative" OR "regress*" OR "model*" OR "multilevel" OR "multi-level") AND ("decisions" OR "outcome*" OR "length" OR "*prison*" OR "custod*").

The screening process was based on the following inclusion criteria:

- Studies based on real sentences derived from a population of convicted adult offenders.
- Reporting the conditional effect (and respective standard errors) of gender or race on prison sentence length after controlling for legal factors (using either regression or matching methods).
- The conditional effect of gender or race on sentence length is reported as a main effect, without an interaction term.

We chose to focus on the length of prison sentences because this is the most explored outcome in the sentencing literature (Bontrager et al., 2013), but also because of methodological considerations. Other commonly explored sentence outcomes such as whether a custodial sentence is imposed, or whether the sentence implies a departure from the guidelines, were rejected because these are typically specified using logistic models, which use log-odds as measures of effect size. While log-odds are frequently employed in meta-analyses in the criminal justice literature (Mitchell, 2005; Petrich, Pratt, Jonson, & Cullen, 2021), they have important limitations that hinder

their comparability across studies[3]. Additionally, we chose not to use the partial (or semi-partial) correlation coefficient as our effect size, even though these are the most common effect sizes in similar meta-analyses pooling coefficients from linear models (Aloe, 2014; Aloe & Becker, 2012). Our decision is based on recent evidence highlighting bias in the estimated variance of these effect sizes (Aert, 2023), as well as our desire to maintain the intuitive interpretation of our pooled estimates as regression coefficients, reflecting the difference in sentence length between male and female offenders, and minority and white offenders.

Lastly, for the first part of our analysis, we applied another exclusion criteria by removing estimates based on overlapping samples. Specifically, we employ a greedy interval scheduling algorithm designed to maximise the number of studies retained while simultaneously avoiding overlapping time intervals. In practice this involves listing all studies that use the same dataset, ordered by the final year of data used, and then iterating through the list to select studies with the shortest coverage period that do not overlap with the coverage period of previously selected studies. When more than one eligible study is identified, we select the study that includes the most legal factors and the fewest extra-legal factors[4].

This exclusion criteria is applied to the estimation of the pooled effects for race and gender disparities, which would otherwise unduly overweight estimates from studies based on the same population. For the rest of our analysis, where our focus shifts to the estimation of the between-study heterogeneity in the sentencing literature, we will rely on the entire sample of estimates captured through our screening and inclusion criteria. Using the full sample of estimates provides much-needed precision to the estimation of jurisdiction-level random effects and helps avoid selection bias. For example, if we selected just one study based on the US District Courts, which is the dataset most commonly used by econometricians working on sentencing research (Fischman & Schanzenbach, 2012; Starr, 2015; Yang, 2015), we could unduly eliminate an important source of between-study heterogeneity. Specifically, Hofer (2019) and Holmes & Feldmeyer (2024) document how criminologists - and sociologists - tend to control for all major legal factors, even those that are influenced by the judge (such as the presumptive sentence), whereas econometricians tend to favour more parsimonious models.

[3] First, log-odds are influenced by the baseline rate, which may obscure differences in reported effects across jurisdictions with varying levels of punitiveness or across studies utilising different samples of offenses. For example, if the conditional probability of imprisonment in a given jusridiction is twice as high for black than for white offenders across all offence types, a model examining shoplifting offences would show a lower log-odds of race than we would see in another model focusing on burglary offences. This occurs because the baseline custody rate for shoplifting is lower than for burglary. Second, log-odds estimated from regression models are not collapsible, meaning their calculations are influenced by the variables controlled within the model. This influence persists regardless of the proportion of outcome variance explained by the set of controls. In other words, this issue extends beyond the mere failure to adjust for confounding variables (Uanhoro, Wang, & O'Connell, 2021; Xiao et al., 2022).

[4] Results were consistent when selection was based on an alternative interval scheduling algorithm designed to maximise the coverage period while retaining the fewest studies.

Table 1: Eligibility status

	Race		Gender	
	N	%	N	%
Screening pool	1024	100	1024	100
Ineligible	895	87.4	882	86.1
not a sentencing study	337		337	
no empirical analysis	64		64	
no multivariate analysis	44		44	
does not model sentence length	206		206	
no main effect or uncertainty	123		110	
young offenders	35		35	
no real sentences	58		58	
historical sample	8		8	
not in English	9		9	
other	11		11	
Eligible	127	12.4	140	13.7
repeated sample	39		44	
uncodeable	22		22	
Unable to retrieve	2	0.2	2	0.2
Eligible and codeable	105	10.3	118	11.5
Eligible, not repeated and codeable	66	6.4	74	7.2

Ultimately, the choice between using the full sample or a restricted sample of non-overlapping studies depends on the population we aim to generalise to. We use the trimmed sample to estimate the pooled effects of race and gender disparities since our target population in this case is all adult offenders. We use the full sample for the remainder of the study, specifically for estimating between-study heterogeneity, as our target population there is the sentencing literature itself.

As shown in Table 1, from our initial pool of 1024, we identified 105 as eligible and codeable for our meta-analysis of race disparities and 118 for that on gender disparities. After removing studies based on repeated samples, these numbers were further reduced to 66 and 74, respectively.

Table 2 presents the jurisdictions covered within the selected studies

before discarding repeated samples. The dominance of US-based research remains evident, with studies from the US comprising 87.6% and 83.1% of the total in our race and gender meta-analyses, respectively. Notable, a single jurisdiction - the US District Courts - accounts for 42.9% and 38.1% of all studies recorded, underscoring the importance of our research question.

2.2 Data Collection

From the pool of eligible studies, we extracted 32 variables covering four broad categories:

- Study metadata: Including author names, title, publication year, and citation count.
- Sample details: Such as jurisdiction[5], offence types, and study time-frame.
- Model specifications: Including adjustments for selection bias, and the number/type of legal and extra-legal controls (e.g., whether the offender was placed on remand or their socio-economic status).
- Effect sizes: The estimated regression coefficients for gender and race, along with their standard errors.

[5] We differentiate between jurisdictions within the US but not within other countries. For instance, estimates from the lower and higher courts of New South Wales were recorded simply as Australia, whereas for the US we distinguish between federal and state jurisdictions. This strikes a balance between jurisdictional granularity and sample size per jurisdiction.

All eligible studies were independently assessed by both authors. Where assessments differed, the paper was re-examined and inconsistencies resolved through consensus.

A detailed meta-data file describing the full list of coded variables and coding rules is available here: osf.io/y4gpf, where we have also published the resulting datasets. The following are some of the more consequential coding rules that we established:

- Race classifications: If multiple racial minority groups are considered in the same model, their effect sizes are recorded as separate observations.
- Model selection: When multiple models are reported, we prioritise those controlling for the highest number of legal factors to minimise potential confounding bias. If multiple models control for the same number of legal factors, we select the one with the fewest extra-legal factors to avoid overfitting and facilitate comparability.

Table 2: Selected studies per jurisdiction

Jurisdiction	Race	Gender
US District Courts	45	45
US State Courts	14	12
Pennsylvania	9	12
Florida	6	8
Minnesota	3	3
Arizona	2	2
Delaware	2	2
New York State	2	2
Ohio	2	2
Texas	2	2
England & Wales	1	3
Canada	1	2
China	1	2
Hong Kong	1	2
South Carolina	1	2
Spain	1	2
Arkansas	1	1
Australia	1	1
Belgium	1	1
Brazil	1	1
Iowa	1	1
Kentucky	1	1
Maryland	1	1
Missouri	1	1
Nebraska	1	1
Netherlands	1	1
North Carolina	1	1
Oregon	1	1
California	1	0
Georgia (US)	1	0
Russia	0	3
Czech Republic	0	1
South Korea	0	1

Note:
US State Courts refers to a sample of courts located in the 75 most populated counties in the US.

- Subgroup reporting: If effect sizes are reported separately by offence type, legal disposition (plea or trial), or time period, each is recorded as a distinct observation.

In total, the first part of our study, which excludes repeated samples, is based on 163 estimates of race disparities and 90 estimates of gender disparities. The remainder of the study draws on 268 and 145 estimates, respectively.

2.2.1 Data Analysis

We use the conditional multiplicative change in sentence length as our effect size, which we denote as β^*. To harmonise estimates from studies reporting their results using different effect sizes, we apply the following transformations:

- Log-transformed dependent variable. If the dependent variable is log-transformed, we derive β^* by exponentiating the regression coefficient of interest: $\beta^* = \exp(\beta)$.
- Standard error for log-transformed models. To approximate the standard error associated with β^*, denoted as $\text{SE}(\beta^*)$, when only the standard error from a log-linear model ($\text{SE}(\beta)$) is available, we apply the Delta method: $\text{SE}(\beta^*) = \beta^* \cdot \text{SE}(\beta)$.
- Linear models with non-log-transformed dependent variables. When a linear model is used and sentence length is not log-transformed, so the regression coefficient (β) and its standard error ($\text{SE}(\beta)$) are expressed as average differences between groups, we derive the effect size in multiplicative terms as follows: $\beta^* = 1 + \frac{\beta}{\bar{y}}$, where $\bar{y}$ represents the sample mean sentence length. The corresponding standard error in multiplicative terms is calculated as: $\text{SE}(\beta^*) = \frac{\text{SE}(\beta)}{\bar{y}}$.

After harmonising all estimates, we pool them using hierarchical meta-anaysis models. We specify a three-level hierarchical model, with effect sizes at level-1, studies at level-2, and jurisdictions as level-3. Study-reported standard errors are incorporated to properly reflect the within-study variance, ensuring each estimate is weighted proportionally to its precision. All our meta-analyses are based on log-linear models, as multiplicative effect sizes can range from zero to infinity, leading to a right-skewed distribution.

We present our findings using a three-stage modelling strategy. First, we use the dataset excluding repeated samples to estimate the pooled estimate of race and gender disparities. Second, we use the full dataset to estimate the variability at the jurisdiction level. Finally, we specify five meta-regression models that introduce a range of explanatory variables - including legal and non-legal factors, citation data and study titles - to explore their potential moderating effects and the presence of selective reporting and citation bias.

The first two meta-regression models include an explanatory variable indicating whether the estimate stems from an US jurisdiction or not, and separately, whether the estimate is based on US Districts data or not. The former tests whether race and gender disparities are systematically more or less pronounced in the US compared to non-US jurisdictions. The latter serves as a robustness check to assess the extent to which including the most heavily studied jurisdiction in the literature influences our findings.

Our third set of meta-regression models examines sample characteristics and types of controls used in the models from which sentencing disparities estimates were derived. Regarding sample characteristics, we consider offence type and whether non-prison sentences were included. In terms of control variables, we incorporate legal and extra-legal factors. The latter includes whether the offender is college-educated, unemployed, a foreign citizen, or has caring responsibilities. The former consists of legal factors commonly used across jurisdictions, including criminal history, guilty plea status, pre-trial detention (remand), and the level of offence-type specificity (e.g., whether the model controls for domestic, commercial, and aggravated burglary separately rather than grouping all property offences together).

The fourth and fifth meta-regression models test for selective reporting and citation bias. Citation bias refers to the tendency for stronger effect sizes to attract a higher number of citations (Barto & Rillig, 2012; Vries et al., 2018). To test the presence of such bias in sentencing research, we use the number of citations received by the study where the estimate is reported, standardised by year of publication[6].

Selective reporting refers to the practice of choosing to publish only certain results, typically those that are statistically significant or align with researchers' expectations (Chan, Hróbjartsson, Haahr, Gøtzsche,

[6] Specifically, we calculate the standardised number of citations (C^*) as follows: $C^* = C/(year - 2000)$; where year refers to the year of publication, and C to the number of citations at the time we compiled the pool of studies (16th of May 2024).

& Altman, 2004; Ioannidis, 2005; Vries et al., 2018). Since pre-registrations in sentencing research are practically non-existent[7], establishing clear evidence of selective reporting is challenging. Here, we approximate it by determining whether the estimate of sentencing disparities comes from a study in which *race*, *gender*, or any related terminology (e.g. *ethnicity*, *sex*), as well as their constituent categories (e.g., *Hispanic*, *female*) are mentioned in the title. This distinction allows us to assess whether an estimate was derived from a study framed as a *disparities* study as opposed to any other type of sentencing study where offender's race/gender were simply used as a control.

[7] See notable exception in Ferguson & Smith (2024).

For the estimation of pooled effects and between-jurisdiction variability (steps one and two of our analysis), we use Bayesian statistics to leverage its precision in estimating higher-level random effects (Harrer, Cuijpers, Furukawa, & Ebert, 2021). However, Bayesian multilevel meta-analysis models are computationally intensive. Therefore, for the final and more exploratory part of our study - where we examine multiple moderating effects - we rely on frequentist statistics[8].

[8] We conducted robustness tests to compare Bayesian and frequentist specification of our meta-analyses. We observed identical results up to the second decimal in the fixed-effects part of our models, which is the primary focus of our moderation analysis.

3 Findings

The pooled effect for race disparities is 1.02, while for gender disparities, it is 0.87. This indicates that minority ethnic offenders receive sentences that are 3% longer than those of offenders from the majority ethnic group. In contrast, female offenders receive sentences that are 13% shorter compared to male offenders.

The credible intervals for these two pooled effects are (1.01, 1.03) for race disparities and (0.83, 0.91) for gender disparities, confirming that both effects are statistically significant. However, only gender disparities could be considered substantively significant. Race disparities appear to be negligible when focusing on sentence length and considering all ethnic minority groups together.

This finding is consistent with previous meta-analysis from the US literature, which suggest that race disparities in sentence length are either small or undetectable (Ferguson & Smith, 2024; Mitchell, 2005; Pratt, 1998). No attempts have been carried out to pool estimates of gender disparities, but previous reviews of the literature indicate

that most estimates suggest greater leniency towards female offenders (Bontrager et al., 2013; Daly & Bordt, 1995), which is also corroborated in our analysis.

The above findings were derived from the pool of estimates where repeated samples were discarded. Notably, replicating the analysis using the full dataset (i.e. including all estimates reported in the literature) yields practically identical results, with pooled effects of 1.03 for race disparities and 0.87 for gender disparities.

3.1 Estimating External Validity

Rather than referring to between-effect heterogeneity as a single entity, our multi-level meta-analysis allows us to differentiate the extent of that heterogeneity attributable to variations across jurisdictions from other sources of heterogeneity at the estimate and study level - such as differences in sample composition or model specification. This distinction is illustrated in Figure 1, from which several key insights can be drawn.

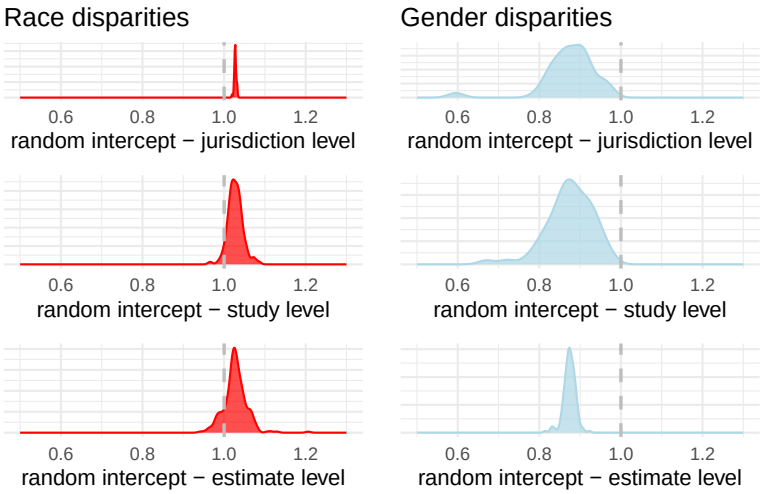

Figure 1: Distribution of random intercepts at the jurisdiction, study and estimate level

Race disparities are practically invariable at the jurisdiction level. The 95% credible interval for the distribution of jurisdiction-level

random intercepts is (1.02, 1.03), indicating that the previously observed negligible race disparities appear to be a global sentencing feature - at least across countries included in our meta-analysis. There is, however, a much more sizable amount of variability stemming from differences between studies (and different effect sizes within studies), accounting for 88% of the total between-study heterogeneity. Specifically, the 95% credible interval of random intercepts at the estimate level covers (0.97, 1.08). This means that, depending on analytic choices - such as the set of controls, years covered, offence types, or minority groups - studies can reach contradictory conclusions about whether minority offenders are penalised or treated more leniently than those from the majority ethnic group. Similarly, the 95% credible interval of random intercepts at the study level spans (0.99, 1.07).

A different pattern emerges when examining gender disparities. Here, jurisdiction-level variability is far greater, accounting for 45% of the total between-study heterogeneity and resulting in a 95% credible interval of (0.75, 0.96). This indicates that, in some jurisdictions, female offenders are treated considerably more leniently than the pooled effect suggested - roughly twice as leniently - whereas in others, gender disparities are effectively non-existent. As with race disparities, we also observe a large share of variability stemming from the study and estimate levels, highlighting how analytic choices can lead to contradictory findings.

To inspect the jurisdiction-level variability in greater detail, we plot the posterior distribution of random intercepts for each jurisdiction in Figure 2 and Figure 3.

Beyond the consistent absence of race disparities across jurisdictions, we also observe that the distribution of US jurisdictions (listed in the lower section of the plot) closely mirrors that of the jurisdictions from the rest of the world (upper section). Both distributions approximately cover the same range, suggesting no systematic difference between them. This finding is formally corroborated by our first meta-regression model, which indicates no appreciable difference in race disparities between US and non-US jurisdictions. Specifically, their respective pooled effects are indistinguishable at the second decimal place (1.03).

Given that a substantial proportion of studies rely on data from US District Courts (42.9%), we further estimated a meta-regression

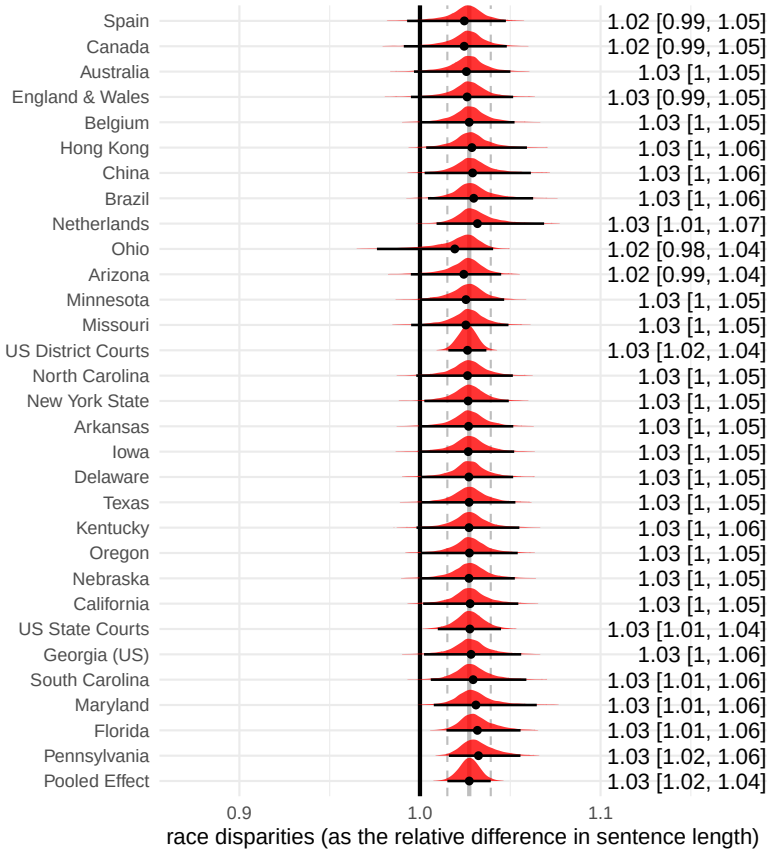

Figure 2: Forest plot of race disparities across jurisdictions

model to assess the potential influence of estimates from this specific jurisdiction. Once again, we found no significant difference, with a pooled effect of 1.03 for estimates derived from US District Courts and 1.03 for all other jurisdictions.

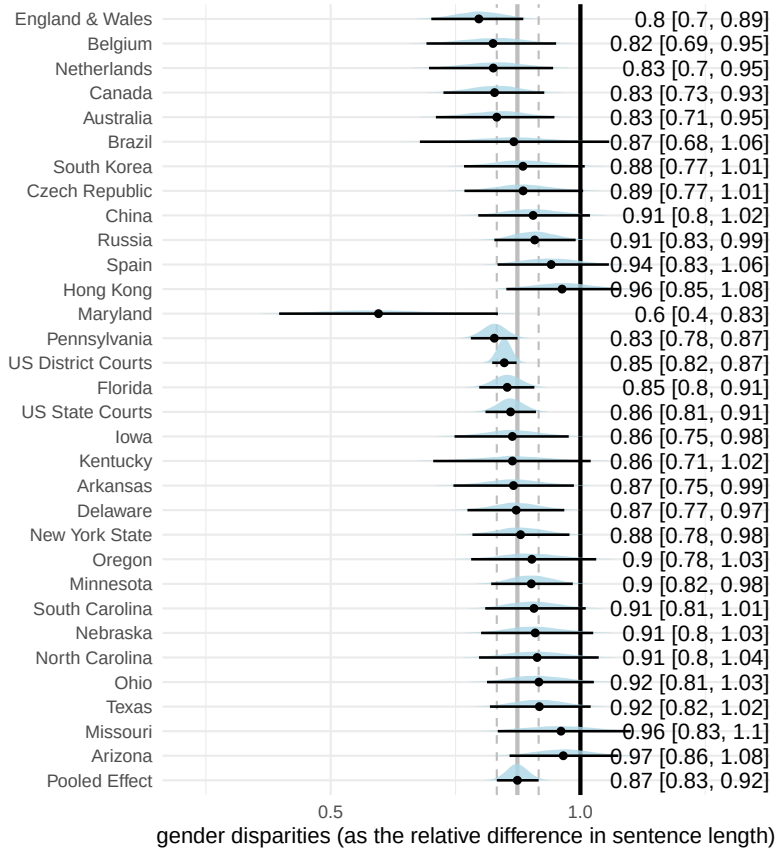

Figure 3: Forest plot of gender disparities across jurisdictions

The much wider between jurisdiction variability detected for gender disparities is clearly appreciable in Figure 3. In some locations, such as Spain, Hong Kong, and certain US states like Arizona and Missouri, sentencing appears to be largely uniform between male and female offenders. However, at the other end of the spectrum, we observe notable disparities - for instance, in the state of Maryland and the combined jurisdiction of England & Wales, where female offenders are estimated to receive sentences that are 45% and 20% more lenient than those of male offenders, respectively.

Despite this variation, we find little difference in overall gender disparities between the US (0.88) and the rest of the world (0.87). Similarly, when comparing pooled effects from US District Courts with those from all other jurisdictions, we detect only a three percentage point difference, which is not statistically significant.

3.2 Further Moderation Analysis

Previous meta-analysis on racial disparities have identified a range of analytic choices that appear to explain some of the between-study heterogeneity. Many of those analytic choices, however, pertain to legal factors that are primarily relevant to American jurisdictions, such as whether the study controlled for the presumptive sentence or whether the judge departed from sentencing guidelines. Here, we conduct a similar moderation analysis, focusing on broader analytic choices applicable to sentencing research worldwide.

Table 3 presents the results of our meta-regressions for race and gender disparities. It is evident that race disparities do not differ significantly based on the analytic choices examined here. Contrary to the recent meta-analysis by Ferguson & Smith (2024), which found that disparities were more pronounced among drug offenders - an effect attributed to the harsher treatment of substances more commonly used by ethnic minorities, such as crack-cocaine - our findings indicate that disparities are not more prevalent for specific types of offence.

The importance of controlling for key legal and non-legal factors becomes more apparent when considering gender disparities. Here, we observe that gender disparities vary strongly by offence type. Specifically, disparities are smaller for drug-related and immigration offences. In the case of immigration offences, gender disparities amount just four percentage points - effectively negligible. Conversely, gender disparities are most pronounced in terrorism-related offences, where male offenders receive sentences that are 38% longer.

Beyond analytic choices, we detect another factor contributing to the observed between study heterogeneity - namely, the potential influence of questionable research practices such as selective reporting. In our dataset on race disparities, 31% of estimates originate from studies where *race* was explicitly mentioned in the title. When comparing

Table 3: Selected studies per jurisdiction

	Race			Gender		
	Estimate	CI.LB	CI.UB	Estimate	CI.LB	CI.UB
Intercept	**1.06**	1.03	1.10	**0.84**	0.77	0.92
Offence: drugs	0.02	-0.01	0.05	**0.08**	0.02	0.13
Offence: firearm	-0.06	-0.16	0.05	0.23	-0.02	0.47
Offence: homicide	0	-0.12	0.11	0.06	-0.10	0.23
Offence: immigration	-0.01	-0.07	0.04	**0.13**	0.03	0.24
Offence: property	-0.06	-0.12	0.01	0.05	-0.01	0.11
Offence: sex	0	-0.05	0.05	0.03	-0.05	0.11
Offence: terrorism	0.11	-0.16	0.38	**-0.23**	-0.39	-0.06
Offence: violence	0.04	-0.01	0.09	-0.01	-0.07	0.05
Specific offence	-0.02	-0.05	0.01	-0.02	-0.07	0.04
Probation	0.02	0.00	0.04	-0.02	-0.06	0.02
Criminal history	-0.01	-0.04	0.02	-0.01	-0.06	0.05
Guilty agreement	-0.02	-0.05	0.00	0.05	-0.01	0.10
Pretrial detention	-0.01	-0.03	0.01	-0.01	-0.05	0.03
Education	0	-0.02	0.02	0.04	-0.01	0.09
Unemployed	-0.02	-0.05	0.02	0	-0.08	0.07
Citizen	0.01	-0.01	0.03	-0.03	-0.08	0.01
Dependents	-0.02	-0.05	0.01	-0.01	-0.06	0.04

Notes:

CI.LB and CI.UB refer to the 95% confidence interval lower and upper bounds.

The reference category for offence type is a mix of all offence types.

Estimates in bold are statistically significant.

pooled effects, we find a statistically significant difference: estimates from studies framed as examining race disparities are twice as strong (1.04) as those from studies where race is merely included as a control (1.02).

When disaggregating by ethnic group, this effect appears to primarily driven by the selective reporting of coefficients for Native Americans. Specifically, when no reference to race is mentioned in the title of the study, the average pooled effect for Native Americans is 0.9. However, when the study is explicitly framed as race disparities research, their pooled effect increases to 1.08. This suggests that studies explicitly investigating race disparities conclude that Native Americans are over-penalised, whereas studies merely including race as a control variable suggest the opposite that Native American offenders receive more lenient treatment.

Notably, no similar selective reporting effect was detected in our meta-analysis of gender disparities. Furthermore, we found no evidence of citation bias - the number of citations a study received (standardised by year) was not associated with the magnitude of either race or gender disparities.

4 Discussion

This study aimed to assess the external validity of sentencing research, specifically examining whether findings from the American literature can be generalised to other jurisdictions across the world. To achieve this we conducted two meta-analyses - the first on race disparities and the second on gender disparities, the two most frequently studied topics in sentencing research.

Our findings reveal small but consistent racial disparities, with minority offenders receiving slightly longer prison sentences (2% to 3% on average). This aligns with previous meta-analyses from the American literature (Ferguson & Smith, 2024; Mitchell, 2005), though stronger disparities have been reported in decisions of disposal type and other discretionary outcomes not considered here. However, in relation to our research question, the more interesting - and unexpected - finding is the cross-jurisdictional uniformity of these racial disparities.

Given the US' unique history of racial conflict, we anticipated that racial disparities in sentencing would be more pronounced there than in other jurisdictions. Contrary to this expectation, our results indicate remarkably similar disparities across a wide range of legal systems, including jurisdictions as divers as Brazil, Belgium, China, and individual US states. The variation in racial disparities across these jurisdictions is minimal — only a few percentage points — suggesting a potential universal sentencing pattern in which minority offenders face a small, albeit statistically significant, penalty regardless of location.

Gender disparities, on the other hand, are far more pronounced. As the first meta-analysis on this subject, our study provides a crucial benchmark, estimating that female offenders receive, on average, 13% shorter sentences than male offenders. However, the most compelling insight emerges when shifting focus from average effects to variance. Unlike racial disparities, gender disparities are far from uniform across jurisdictions, with a few showing gender parity, while others penalise male offenders with over 20% longer sentences.

We attribute this substantial cross-jurisdictional variability to the more conflicting legal interpretation of offenders' gender. In some jurisdictions - apparently in those where sentencing is more strictly codified and consequently judicial discretion more constraint, such as Spain or Hong Kong - sentencing seems to be gender neutral. That is, offender's gender, just like their race, is not taking into account, and consequently exerts little to no influence in the sentence outcome. However, in many jurisdictions, even where legally recognised as a protected characteristic, gender also serves as a proxy variable for different considerations underlying core sentencing goals such as rehabilitation, retribution, and public safety.

For example, female offenders tend to have lower reoffending rates (National Offender Management Service, 2015), pose less risk to public safety (Maden et al., 2006; Sapouna, Bisset, Conlong, & Matthews, 2015), and exhibit higher rates of self-harm in prison (Gauke, 2018; Player, 2014). These factors likely influence sentencing decisions differently across jurisdictions, depending on how they balance the principles of individualisation and consistency. The result is a sentencing landscape in which gender disparities fluctuate

widely depending on the jurisdiction's legal framework and judicial philosophy.

Notably, this variation persists even within the United States. While states like Arizona and Missouri exhibit gender-neutral sentencing, jurisdictions such as Pennsylvania, Florida, and the US District Courts impose significantly longer sentences on male offenders (15% to 17%). This finding raises questions about the external validity of sentencing research. Even when the scope of a study is clearly restricted to the sentencing practices within a given country, if the study examines only a specific jurisdiction or court type, its findings should not be assumed to generalise to the broader national context. For instance, a quick literature review on gender disparities in the US would likely over-represent studies based on the US District Courts, leading to the conclusion that strong gender disparities are prevalent across the country. However, this would overlook the more moderate disparities observed across the entire US, or the fact that,in certain states sentencing is effectively gender neutral.

Beyond cross-jurisdictional variation, our study also reveals substantial uncertainty at the estimate level. Our meta-regression analyses indicate that gender disparities are particularly pronounced in cases involving terrorism offenses — likely due to heightened concerns about dangerousness and public safety — but nearly non-existent for immigration offenses, which may involve more standardised sentencing procedures. Even after accounting for such factors, a considerable degree of unexplained variation remains, with female offenders receiving anywhere from 2% to 18% shorter sentences depending on model specifications.

Racial disparities show no systematic differences across offense types or case characteristics. However, substantial variability between estimates remains, with racial minority offenders receiving anywhere from 11% longer to 5% shorter sentences depending on researchers' methodological choices.

This large between estimate variability - observed for both racial and gender disparities - could be treated as an indication of sentencing research being highly unreliable as a result of researchers' degrees of freedom. This is particularly problematic if we consider how such model uncertainty is rarely acknowledged in the sentencing literature

leading to unduly overconfident interpretations. More concerningly, our findings suggest that this uncertainty is not entirely random.

Specifically, we identify a pattern of selective reporting in studies framed explicitly as investigations of racial disparities. When race (or related terms such as ethnicity) appears in a study's title, estimates of racial disparities are, on average, twice as large as those in studies where race is merely included as a control variable. This discrepancy suggests that some researchers may be leveraging analytical flexibility to produce stronger findings when their study is explicitly positioned as an examination of racial disparities.

While this pattern does not appear in gender disparity research, its presence in race disparity studies has broader implications for the field. Not only does it introduce bias into individual studies, but it also distorts the overall literature by inflating perceived disparities. Importantly, different manifestations of the same underlying problem have been detected in past meta-analyses. Mitchell (2005) found ethnic disparities published in academic journals or books are twice as large as those that did not follow that route (e.g. doctoral dissertations). Similarly, Ferguson & Smith (2024) shows that studies selectively citing evidence that supports their hypotheses - while ignoring to cite evidence that contradicts them - also appear to report twice stronger racial disparities. Consequently, literature reviews and meta-analyses — including this one — are likely over-estimating the presence of racial disparities in sentencing.

4.1 Way Forward

Despite the concerning state of sentencing research, we believe that most of its problems can be addressed through the collective adoption of open science practices. Incidentally, these practices also have the potential to mitigate key limitations of this study and, in doing so, improve its external validity.

One of the primary limitations of our study is its scope. We restricted our analysis to: i) a specific judicial decision—sentence length; ii) two research questions—race and gender disparities; and iii) studies published in English. These choices were made for practical reasons. However, to assess the robustness of our findings, future research should expand its scope to consider additional judicial decisions, such

as whether to convict, grant bail, or depart from sentencing guidelines; other sentencing-related estimates, such as the effects of criminal history or guilty pleas; and studies published in other languages, such as Spanish (Domínguez, 2024; Páez-Mérida & Montero Molera, 2022), German (Kaiser & Leibetseder, 2024), Czech (Drápal, 2017), or Chinese. To facilitate this expanded analysis and ensure the reproducibility of our findings, we have made all our materials publicly available and integrated our R code directly into the article's text using Quarto.

Given our partial exploration of the literature, it is possible that we are underestimating the true extent of estimate and - especially - cross-jurisdictional variability. Even if that is not the case, one clear takeaway from our study is the need to promote cross-jurisdictional sentencing research. Most sentencing studies focus on a single jurisdiction and set of courts, but as we have demonstrated, this can lead to misleading conclusions when findings are generalised to other jurisdictions or even to different courts within the same system. For example, sentencing research in England and Wales is almost exclusively focused on the Crown Court, despite the fact that over 90% of sentences are imposed in the lower magistrates' courts (Ministry of Justice, 2024). This raises serious concerns about the generalisability of findings in that jurisdiction.

Conducting more cross-jurisdictional research will enhance the reliability of sentencing research by providing a clearer picture of between-jurisdiction variability. Likewise, we should aim to represent the model uncertainty affecting our findings more transparently. To achieve this, we recommend moving away from the traditional practice of reporting point estimates followed by a significance test, as this approach only accounts for uncertainty stemming from sampling error. Instead, we advocate for the use of specification curves (Simonsohn, Simmons, & Nelson, 2020) to empirically capture the model uncertainty introduced by analytical choices for which there is no clear consensus — such as which legal factors should be controlled for in studies of sentencing disparities (Hofer, 2019; Holmes & Feldmeyer, 2024; Pina-Sánchez, Hamilton, & Tennant, 2024).

Beyond making uncertainty more transparent, adopting specific open science initiatives — such as the pre-registration of analytical strategies and the publication of registered reports — could help address

many of the problems affecting sentencing research. Pre-registration would reduce outcome bias by ensuring that analytical strategies are not altered after results are known (Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). Registered reports would minimise researchers' degrees of freedom by subjecting analytical strategies to peer review before data collection and analysis; additionally, since studies would be accepted for publication based on their methodology rather than their findings, publication bias would be substantially reduced (Chambers & Tzavella, 2022; Lakens, Mesquida, Rasti, & Ditroilo, 2024)

While open science initiatives should be embraced as a matter of good research practice, their adoption becomes especially urgent given the high degree of model uncertainty and the evidence of selective reporting uncovered in this study. There is still a long road ahead, but embedding these practices into sentencing research is the only way forward if we want to ensure that the field produces robust and trustworthy findings.

5 Conclusion

What is the external validity of sentencing research? In short, quite low. However, further context is required.

We found that racial disparities in sentencing are remarkably consistent across jurisdictions. Ethnic minority offenders tend to receive slightly longer sentences, and this effect appears to be universal, with penalties ranging narrowly from 2% to 4% longer sentences across jurisdictions. Similarly, our analysis indicates that, on average, estimates of gender disparities in US jurisdictions do not significantly differ from those observed elsewhere. As a whole, then, the US literature appears to be broadly representative of sentencing practices worldwide.

However, when we consider the full extent of between-jurisdiction variability — rather than just comparing average effects — substantial differences in gender disparities emerge. While sentencing appears to be truly gender-neutral in jurisdictions such as Hong Kong, Spain, Missouri, and Arizona, in others — including England

& Wales, Belgium, the Netherlands, Australia, Canada, and Pennsylvania — female offenders receive approximately 20% shorter prison sentences.

Beyond jurisdictional differences, we also uncovered considerable uncertainty stemming from the diverse analytical choices made by researchers. This variability was substantial in both of our meta-analyses, suggesting that sentencing outcomes are highly sensitive to researchers' degrees of freedom. More worryingly, we documented how this leeway appears to have been exploited to exaggerate the effect size of racial disparities.

These findings serve as a cautionary note for researchers, practitioners, policymakers, and sentencing reform advocates. A more measured approach is needed when interpreting sentencing research. Findings from one jurisdiction should not be assumed to generalise to others. Even within a specific jurisdiction, it is crucial to avoid over-relying on a single study and instead assess the broader body of evidence. Furthermore, even when considering the literature as a whole, it is important to remain moderately skeptical and recognise how the misalignment between career incentives and research best practices may have influenced the evidence base.

References

Aert, R. C. M. van. (2023). Meta-analyzing partial correlation coefficients using Fisher's z transformation. *Research Synthesis Methods*, *14*(5), 768–773. https://doi.org/10.1002/jrsm.1654

Aloe, A. M. (2014). An empirical investigation of partial effect sizes in meta-analysis of correlational data. *The Journal of General Psychology*, *141*(1), 47–64. https://doi.org/10.1080/00221309.2013.853021

Aloe, A. M., & Becker, B. J. (2012). An Effect Size for Regression Predictors in Meta-Analysis. *Journal of Educational and Behavioral Statistics*, *37*(2), 278–297. https://doi.org/10.3102/1076998610396901

Barto, E. K., & Rillig, M. C. (2012). Dissemination biases in ecology: effect sizes matter more than quality. *Oikos*, *121*(2), 228–235. https://doi.org/10.1111/j.1600-0706.2011.19401.x

Bontrager, S., Barrick, K., & Stupi, E. (2013). Gender and sentencing: A meta-analysis of contemporary research. *Journal of Gender, Race & Justice*, *16*, 349. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/jgrj16&id=371&div=&collection=

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized TrialsComparison of protocols to published articles. *JAMA*, *291*(20), 2457–2465. https://doi.org/10.1001/jama.291.20.2457

Daly, K., & Bordt, R. L. (1995). Sex effects and sentencing: An analysis of the statistical literature. *Justice Quarterly*, *12*(1), 141–175. https://doi.org/10.1080/07418829500092601

Domínguez, I. G. (2024). Revisión de sentencias de personas en situación de sinhogarismo en los tribunales penales españoles (años 2016-2020). *Revista Española de Investigación Criminológica*, *22*(1). https://doi.org/10.46381/reic.v22i1.878

Drápal, J. (2017). Vede větší užití trestních příkazů k ukládání méně nepodmíněných trestů odnětí svobody? *Jurisprudence - Časopis Právnické Fakulty Univerzity Karlovy*, *5*, 3–17. Retrieved from https://www.jurisprudence.cz/cz/casopis/vede-vetsi-uziti-trestnich-prikazu-k-ukladani-mene-nepodminenych-trestu-odneti-svobody.m-264.html

Ferguson, C. J., & Smith, S. (2024). Race, class, and criminal adjudication: Is the US criminal justice system as biased as is often assumed? A meta-analytic review. *Aggression and Violent Behavior*, *75*, 101905. https://doi.org/10.1016/j.avb.2023.101905

Fischman, J. B., & Schanzenbach, M. M. (2012). Racial Disparities Under the Federal Sentencing Guidelines: The Role of Judicial Discretion and Mandatory Minimums. *Journal of Empirical Legal Studies*, *9*(4), 729–764. https://doi.org/10.1111/j.1740-1461.2012.01266.x

Gauke, D. (2018). *Female Offender Strategy*. Retrieved from https://assets.publishing.service.gov.uk/media/5b3349c4e5274a55d7a54abe/female-offender-strategy.pdf

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2021). *Doing meta-analysis with r: A hands-on guide*. New York: Chapman; Hall/CRC. https://doi.org/10.1201/9781003107347

Hofer, P. J. (2019). Federal sentencing after booker. *Crime and Justice*, *48*, 137–186. https://doi.org/10.1086/701712

Holmes, B., & Feldmeyer, B. (2024). Modeling Matters: Comparing the Presumptive Sentence Versus Base Offense Level Approaches for Estimating Racial/Ethnic Effects on Federal Sentencing. *Journal of Quantitative Criminology*, *40*(2), 395–420. https://doi.org/10.1007/s10940-023-09573-0

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kaiser, N., & Leibetseder, I. (2024). Spezialprävention in der Praxis: Zum Entscheidungsverhalten von Richter:innen und Staatsanwält:innen. *Journal für Strafrecht*, *11*(2), 125–130. Retrieved from https://www.verlagoesterreich.at/spezialpraevention-in-der-praxis-zum-entscheidungsverhalten-von-richter-innen-und-staatsanwaelt-innen/99.105005-jst202402012501

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and registered reports. *Evidence-Based Toxicology*, *2*(1), 2376046. https://doi.org/10.1080/2833373X.2024.2376046

Maden, A., Skapinakis, P., Lewis, G., Scott, F., Burnett, R., & Jamieson, E. (2006). Gender differences in reoffending after discharge from medium-secure units: National cohort study in England and Wales. *The British Journal of Psychiatry*, *189*(2), 168–172. https://doi.org/10.1192/bjp.bp.105.014613

Ministry of Justice. (2024). *Criminal Justice System statistics quarterly: March 2024*. Retrieved from https://www.gov.uk/government/statistics/criminal-justice-system-statistics-quarterly-march-2024

Mitchell, O. (2005). A Meta-Analysis of Race and Sentencing Research: Explaining the Inconsistencies. *Journal of Quantitative Criminology*, *21*(4), 439–466. https://doi.org/10.1007/s10940-005-7362-7

Nardulli, P. F., Flemming, R. B., & Eisenstein, J. (1988). *The tenor of justice: Criminal courts and the guilty plea process*. Urbana: University of Illinois Press.

National Offender Management Service. (2015). *Better outcomes for women offenders*. Retrieved from https://assets.publishing.service.gov.uk/media/5a81a29ded915d74e33ff44f/Better_Outcomes_for_Women_Offenders_September_2015.pdf

Páez-Mérida, A., & Montero Molera, A. (2022). ¿Cómo se juzga a

las chicas en el sistema de justicia juvenil español? Un estudio exploratorio a partir de datos primarios. *Revista Española de Investigación Criminológica*, *20*(2), e691–e691. https://doi.org/10.46381/reic.v20i2.691

Petrich, D. M., Pratt, T. C., Jonson, C. L., & Cullen, F. T. (2021). Custodial sanctions and reoffending: A meta-analytic review. *Crime and Justice*, *50*, 353–424. https://doi.org/10.1086/715100

Pina Sánchez, J., & Harris, L. (2020). Sentencing gender? Investigating the presence of gender disparities in Crown Court sentences. *Criminal Law Review*, *2020*(1), 3–28. Retrieved from https://eprints.whiterose.ac.uk/154388/

Pina-Sánchez, J., Hamilton, M., & Tennant, P. W. G. (2024). *Modelling unwarranted disparities in sentencing: Distinguishing between good and bad controls*. https://doi.org/https://doi.org/10.31235/osf.io/ymzsv

Player, E. (2014). Women in the criminal justice system: The triumph of inertia. *Criminology & Criminal Justice*, *14*(3), 276–297. https://doi.org/10.1177/1748895813495218

Pratt, T. C. (1998). Race and sentencing: A meta-analysis of conflicting empirical research results. *Journal of Criminal Justice*, *26*(6), 513–523. https://doi.org/10.1016/S0047-2352(98)00028-2

Sapouna, D. M., Bisset, C., Conlong, A.-M., & Matthews, B. (2015). *What Works to Reduce Reoffending: A Summary of the Evidence*. Retrieved from https://www.gov.scot/binaries/content/documents/govscot/publications/research-and-analysis/2015/05/works-reduce-reoffending-summary-evidence/documents/works-reduce-reoffending-summary-evidence/works-reduce-reoffending-summary-evidence/govscot%3Adocument/00476574.pdf

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Starr, S. B. (2015). Estimating gender disparities in federal criminal cases. *American Law and Economics Review*, *17*(1), 127–159. https://doi.org/10.1093/aler/ahu010

Steffensmeier, D., Kramer, J., & Streifel, C. (1993). Gender and Imprisonment Decisions. *Criminology*, *31*(3), 411–446. https:

//doi.org/10.1111/j.1745-9125.1993.tb01136.x

Uanhoro, J. O., Wang, Y., & O'Connell, A. A. (2021). Problems with using odds ratios as effect sizes in binary logistic regression and alternative approaches. *The Journal of Experimental Education*, *89*(4), 670–689. https://doi.org/10.1080/00220973.2019.1693328

Ulmer, J. T., & Johnson, B. (2004). Sentencing in Context: A Multilevel Analysis. *Criminology*, *42*(1), 137–178. https://doi.org/10.1111/j.1745-9125.2004.tb00516.x

Ulmer, J. T., & Kramer, J. H. (1996). Court Communities Under Sentencing Guidelines: Dilemmas of Formal Rationality and Sentencing Disparity. *Criminology*, *34*(3), 383–408. https://doi.org/10.1111/j.1745-9125.1996.tb01212.x

Vries, Y. A. de, Roest, A. M., Jonge, P. de, Cuijpers, P., Munafò, M. R., & Bastiaansen, J. A. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. *Psychological Medicine*, *48*(15), 2453–2455. https://doi.org/10.1017/S0033291718001873

Xiao, M., Chu, H., Cole, S. R., Chen, Y., MacLehose, R. F., Richardson, D. B., & Greenland, S. (2022). Controversy and Debate : Questionable utility of the relative risk in clinical research: Paper 4 :Odds Ratios are far from "portable" - A call to use realistic models for effect variation in meta-analysis. *Journal of Clinical Epidemiology*, *142*, 294–304. https://doi.org/10.1016/j.jclinepi.2021.08.002

Yang, C. S. (2015). Free at last? Judicial discretion and racial disparities in federal sentencing. *The Journal of Legal Studies*, *44*(1), 75–111. https://doi.org/10.1086/680989