Te Puna Haumaru
New Zealand Institute for Security and Crime Science

THE UNIVERSITY OF WAIKATO

Systematic evidence map of disparities in police outcomes: final report

October 2021

Dr Lisa Tompson, Dr Simon Davies and Professor Devon Polaschek

With research assistance from Mallory Dobner, Taryn Farr, Daniel Jones, Hillary Pinion and Michaela Sibbald.

University of Waikato

Contact details:
Dr Lisa Tompson
Te Puna Haumaru NZ Institute for Security and Crime Science
Te Whare Wānanga o Waikato The University of Waikato
Private Bag 3105
Hamilton 3240
New Zealand

Email: lisa.tompson@waikato.ac.nz

Executive summary

To support the New Zealand Police Understanding Policing Delivery (UPD) project, we were tasked with completing an academic literature review on bias in policing. However, the vastness and complexity of the relevant research, and the short timeframe for completion meant that it was not possible to complete a single literature review that accurately reflected the evidence base; indeed, it would be difficult to do justice to this literature in a single review under any circumstances. Thus, rather than attempting a traditional literature review, we developed an *evidence and gap map* of the international research on bias in policing. In this report, we present the rationale, methods used, and key features of the studies on the evidence and gap map (referred to as the 'evidence map').

This evidence map serves several purposes. Its overarching purpose is to be a resource for more targeted work on specific areas of apparent disparity in policing, including areas prioritised by New Zealand Police as part of the UPD project: decisions about stops, charges, and the use of force. Researchers will be able to use the map to easily identify a) the extent of research in the area of interest, and b) the key features of the research that has been conducted in that area, as a basis. The map will guide rapid evidence reviews and future primary research, both of which can be used to design police practice-based interventions to reduce police-generated disparities in New Zealand.

To create the map, we were provided with the results of a systematic search of the Global Policing Database from 2000 to 2018 (the latest date of records in the database) which identified over 10,000 potentially eligible records. We subsequently conducted a two-stage screening process using explicit inclusion criteria and inter-rater reliability tests, resulting in 403 studies that were judged to be eligible for inclusion on the evidence map, which is thus a summary of a large body of research.

The map, and this report summarise key features of those 403 studies. We categorised key features in each study using a predefined coding instrument which covered the location of the study, the policing agency involved, the police actions examined, the dimensions of bias investigated, the theories tested in the study, and the methodological approach taken. The final evidence map can be found here along with a video showing how to use the map.

For reasons we explain in section 1, studying biased police decision-making is fraught with methodological difficulties. Instead, researchers have looked for disparities (disproportionalities) in police outcomes, and then sought to explain those found. Examination of evidence map key features reveals several important points about the nature and extent of disparities in police outcomes:

- Studies examining disparities in police outcomes are increasingly being published year-on-year.
- Most of the evidence (86%, 329 studies) comes from studies conducted in North America, predominantly the United States. Three-quarters of the entire map examined racial bias in the US (299 studies). Only 3 studies from New Zealand and 14 from Australia are included in the evidence map, raising questions about evidence transferability for policing in New Zealand.
- A very wide range of police actions is covered. A few studies have examined deployment (e.g., decisions about where to patrol or dispatch officers), investigation (e.g., decisions about whether to proceed with an investigation), and professional conduct (e.g., verbal communication used with citizens, investigation of complaints against police). The most evidence lies within the central areas identified for the UPD project: acting on suspicion (e.g., decisions to stop citizens, post-stop outcomes, and use of force) and charging decisions (e.g., decisions about whether to make an arrest). In particular, we identified 116 studies on stops of citizens; 123 studies on threat of, or use of force; and 129 studies on decisions about arrests.
- The three most studied dimensions of bias are race (351 studies), gender (166 studies), and age (143 studies). Language, appearance, and income/education level have been rarely studied.

There is also a small but meaningful body of evidence on structural (e.g., legislation), institutional (e.g., the racial composition of the police force), and ecological (e.g., place or time characteristics that could affect decision-making) dimensions of bias.

- Studies commonly examined more than one police action or dimension of bias. This is important because one dimension of bias may explain disparity in another dimension (e.g., differences in age profile of different racial groups may explain race disparity).
- Over 20 theories framed studies on the map or explained the disparities in police outcomes (when found). There is little theoretical consensus on why disparities exist, and studies often lacked any theoretical framework from which to make sense of the findings.
- The methods used are also highly varied. Many studies rely on police data, but many others use other sources, including survey and observational data. Although some studies simply use the residential population as a benchmark to assess for disparity, more commonly studies use a more rigorous benchmark that accounts for the likelihood of coming into contact with police (e.g., an estimate of the driving population in a study examining police stops). Over three quarters of the studies in the evidence map used robust inferential statistics.

Time constraints prevented us from systematically appraising the design quality of studies in the evidence map; quality assessments are time-consuming for dozens of studies, let alone several hundred. So, we are not suitably positioned to judge the strength or reliability of the findings presented here. For this reason, we have not quantified how frequently disparities in police outcomes, as an indicator of bias, were found in the studies on the evidence map. Presenting headline findings uncritically risks oversimplifying and proliferating the limitations of the primary studies (as documented in section 2.7), and potentially does a disservice to more reliable studies. We leave those judgements to researchers who use the evidence map to form their own literature reviews and conclusions. Systematic, quantifiable evaluations of quality are feasible on smaller pools of studies.

To offer some insight about whether/where bias may exist, we provide a brief summary of eight meta-analyses and systematic reviews (i.e., evidence syntheses) that cover 253 primary studies in section 2.8. Six of these eight aggregate studies have built in tests of research quality, to varying extents. The evidence syntheses on search, arrest, and use of force decisions consistently found disparities based on race and gender. This pattern remained—although the effect was often reduced—after controlling for rival explanations for disparities, confirming that there may be layers of determinants of disparities. However, these evidence syntheses cover only a handful of police actions, only parts of the literature we reviewed, and were dominated by North American studies, so it is unclear how these findings apply to jurisdictions such as New Zealand. A further literature review on the nature and impact of police bias is presented in section 4, along with recommendations on how to think about designing future research in this area.

Overall, the evidence and gap map we have developed indicates there is a vast body of evidence relevant to the topic of bias in policing. This evidence encompasses a wide range of different police actions and dimensions of bias, albeit that it is dominated by US research on racial bias. Unfortunately, very little of this research has been conducted in New Zealand or even in the similar jurisdiction of Australia. Consequently, although this evidence map can be used to quickly synthesise evidence relevant to a particular area of policing, its main use will be to provide guidance on theory and methodology for future research commissioned by New Zealand Police to fill identified knowledge gaps.

Contents

1. Introduction

To support the Understanding Policing Delivery project, a research team at the University of Waikato was tasked with doing an academic literature review on the topic of bias in policing. Because this topic covers a vast and complex area that has not been brought together *en masse* previously, we decided that an evidence and gap map would be the best method to survey the evidence base.

An evidence and gap map is a systematic evidence synthesis product that displays the evidence relevant to a specific research question. The aim is generally breadth rather than depth (which might instead be the focus of a systematic review). The systematic nature of the methods used to assemble an evidence and gap map protects against the usual researcher inclination to just select studies that are easy to find, confirm a particular viewpoint, or are based on some other idiosyncrasy. Reducing such selection bias in the data collection process - the assembling of studies - enables an objective account of the evidence base to be compiled. Given that policing research spans criminology (often with a strong sociological approach), law, public health and economics, not to mention think tanks and government bodies that produce statistics, an evidence map was considered to be the best way of bringing together the widest range of research in the timescales of the project.

By mapping the contours of a research topic and displaying the evidence visually, it becomes clear where there are knowledge gaps that need to be addressed through new research. Evidence and gap maps are also useful for scoping out what is known on a topic, and they facilitate the synthesis of subsets of studies. They are therefore not a traditional literature review but have many more practical uses. For simplicity, from this point on we refer to the product described in this report as the 'evidence map'.

Turning now to the focus of the evidence map: Bias in policing is a perennially controversial research topic with a long history. The powers of arrest and detention given to police officers are unparalleled, and in a democratic society it is only right that they be scrutinised. However, it is fair to say that this is also an incendiary topic, with forceful views typically expressed from both the perspective of the police and the communities that criticise them or hold them accountable.

The precursor to much of the literature on the evidence map was legislative changes in the US after lawsuits against states regarding overt racial discrimination practices in policing. These were mirrored in the UK with the 1999 MacPherson report that was the product of an Inquiry into the death of a young Black man - Stephen Lawrence - in Southeast London, of which the most newsworthy conclusion was that the Metropolitan Police were 'institutionally racist'. Other Western police agencies have experienced similar scrutiny in the 21st century and the events of 2020 and the Black Lives Matter movement indicate that this issue is far from resolved in the public consciousness.

To conclusively evidence that the police are biased would require charting a decision from its antecedents (e.g., police policies, individuals' pre-employment upbringing and exposure, learning "on the job" from more senior staff, contributions to a police officer's attitudes and belief systems) through to the situational components (including the place, and the citizen's behaviour), to the outcome. To our knowledge, no studies have been able to achieve this to date, and it is difficult to imagine that the data needed for charting such a process would ever be accessible for field research. And, even if it was, such a study would need to generate a representative sample of police officers (without invoking any observer bias effects) to be able to claim that the police are indeed biased in their decision-making. What researchers have done instead is something quite different. They have looked for disparities in police outcomes —sometimes finding them and sometimes not—and when they have found them, they have tried to explain where they are coming from. A lot of research on the evidence map has only looked for correlations between variables and disparities, rather than

testing formal theory, which is the primary way of explaining empirical patterns. Without doing that, the field struggles to move forward.

It is also worth stating very clearly at this point that disparities in police outcomes do not equate to unequivocal evidence that the police agency in question is biased. There are many legally relevant reasons why the police may be justified in targeting particular social groups. Disproportionality, or disparities (both terms are used interchangeably in this report) in outcomes need to be explained with reference to causal processes to plausibly suggest that police bias is operating (see section 2.6.1.). Disentangling the legal factors that influence police decision-making from the extra-legal (the legally irrelevant factors) is at the heart of much of the higher quality evidence on the map.

Before we proceed with the summary of the evidence map, it is important to note a few limitations. The first is that the literature only goes up to 2018; this was the latest date in the source used to give us a head start on identifying relevant studies. Due to the trajectory of publishing on this topic (see section 2.1.), there are likely to be many more studies. The second is that although the search was as extensive as possible in the timescales of the project, it nevertheless omits some important studies (e.g., Eberhardt, 2016; MacPherson report, 1999) that were not discoverable within the search strategy because they are grey literature and not easily retrievable in databases. However, the evidence map does capture similar studies or those that could be considered to follow up these studies (see Miller, 2010 for example).

This report should be read in conjunction with the interactive evidence map. The remainder of this report covers the methods used to produce the evidence map in section 2; the findings on several dimensions considered relevant to police bias research in section 3; and a discussion of observations and an overall summary (which is more akin to a literature review) in section 4.

2. Methods used to produce evidence map

This research aims to answer the question: "What is known about the nature and impact of possible bias in policing policies and practices internationally?". The search strategy involved a keyword search of the Global Policing Database (GPD), developed by researchers from the University of Queensland[1]. This database contains published and unpublished research on policing interventions from 1950-2018, in 12 languages, acquired from 42 academic databases and grey literature sources. Further details on how the database has been compiled are in the GPD protocol document.

Since the GPD had already searched for international policing research, our keywords were designed to identify studies on bias in policies and practices (the other components in the research question). Relevant synonyms for the concept of 'bias' were harvested from known studies on police bias, and these were subsequently checked in electronic bibliographic databases, using index terms and the thesauri functionality. The final search syntax contained 26 phrases or keywords[2].

[1] Higginson, A., Eggins, E., Mazerolle, L., & Stanko, E. (2015). The Global Policing Database [Database and Protocol]. Retrieved from http://www.gpd.uq.edu.au/search.php

[2] The syntax used to search the GPD was as follows: (discrimina* OR bias OR disparity OR disproportion* OR discretion OR minorit* OR ethnic* OR "racial profiling" OR "racial bias" OR racism OR racist OR "race relations" OR stereotyp* OR "hate crime" OR "use of force" OR "lethal force" OR "deadly force" OR brutality OR mistreatment OR unfair OR subconscious OR prejudice OR improper OR inequality OR inequity OR "social class*")

2.1. Inclusion criteria

We applied the following inclusion criteria when screening records for eligibility in this evidence map:

1. The study was available in English, or translatable using Google translate or other online tools.
2. The study was published after 1999 .
3. The study explicitly focused on the police as an organisation. Police here refer to 'sworn' officers or public police as an executive arm of the government providing a service at the local, county, state or federal level. This includes police staff (e.g., scenes of crimes officers, CSI, other civilian employees).
4. The study was tasked with examining, assessing or evaluating disparate outcomes that may indicate bias from police decisions or actions on different social groups external to the organisation.
    a. Behavioural outcomes in scope can be enacted or simulated (e.g., in 'shoot/don't shoot' experiments).
    b. All forms of bias should be included. Bias covers:
        i. *Structural* - The interplay of policies, practices and programmes at a societal level which lead to disproportionate outcomes and conditions for some communities.
        ii. *Institutional* – Policies, practices, and procedures that work to the benefit of one group of people and the detriment of others, whether intentionally or inadvertently.
        iii. *Individual/Interpersonal* – Pre-judgment, bias (implicit or explicit), stereotypes, or generalisations about an individual or group. May result in discrimination.
    c. Bias may be directed to the following human characteristics (and categories may interact):
        i. Race or ethnic groups
        ii. Age
        iii. Skin colour
        iv. Gender or gender identity
        v. Dress or appearance
        vi. Disability or health issue
        vii. Accent/language/nation of origin
        viii. Religious beliefs
        ix. Sexual orientation
        x. Substance misuse
        xi. Sex workers
        xii. Victim behaviour/credibility
        xiii. Income/education
    d. Police actions should be able to be reliably determined (e.g., not just 'interaction/ contact with police'). Therefore, one of the following actions must be mentioned in the study:
        i. Communications (e.g., Body worn cameras, 111/911/999 calls for service)
        ii. Dispatching police vehicles to incident reports

iii. Stops of citizens
iv. Searches of citizens
v. Handcuffing
vi. Referrals and arrests/apprehension
vii. Infringements and fines
viii. Charges
ix. Recommending prosecution
x. The use or threat of the use-of-force
xi. Granting bail
xii. Investigating crimes committed by citizens
xiii. Investigating crimes committed by police officers
xiv. Investigating citizen complaints of police officers
xv. Interviewing suspects
xvi. Clearance rates
xvii. Patrolling/presence in specific areas

5. The study reports original, empirical findings.

6. The study uses a comparator to assess disparities in police actions[3].

No restrictions were placed on research designs or regions of the world from which studies came.

2.2. Screening process

The 10,223 citation results from the GPD search were imported into EppiReviewer Web[4] software and duplicate records (n=114) were removed. A conventional two-stage screening process was adopted. The first stage involved screening records on the title and abstract provided by the citation information. In the second stage we consulted the full texts of studies, where these were available electronically or through university libraries.

Priority screening, using machine learning algorithms, was used within the software[5]. This involved creating a 'training set' of studies from which the software learned which studies would be eligible and which would be excluded. We created the training set by searching for studies that we already knew were eligible for inclusion from our preliminary reading on the topic. Because these studies primarily related to racial bias, we supplemented the training set with studies on other dimensions of bias (e.g., gender, age, sexual orientation, mental health) to ensure that the software was prioritising the full spectrum of studies we sought. These records were subsequently coded as 'include' or 'exclude' to train the algorithm to prioritise other records with similar combinations of keywords.

We screened records on title and abstract until the 'hit rate', that is, the number of studies that were included out of a sample of records, dropped below 3 per cent (see Figure 1). This meant we

[3] Disproportionality can only be robustly assessed through the use of a comparator. The study should therefore attempt to compare police decision/action for one group (e.g., men) against police decision/action for a comparison group (e.g., women), and not just action within one group.

[4] Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. (2020). *EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis*. EPPI-Centre Software. London: UCL Social Research Institute.

[5] This method has been validated and compared to similar machine learning approaches to prioritising records in the following publication: Tsou, A. Y., Treadwell, J. R., Erinoff, E., & Schoelles, K. (2020). Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. *Systematic reviews*, 9(1), 1-14.

screened 4,517 records, at which point we were confident we had identified 95 per cent of eligible studies in the GPD[6].

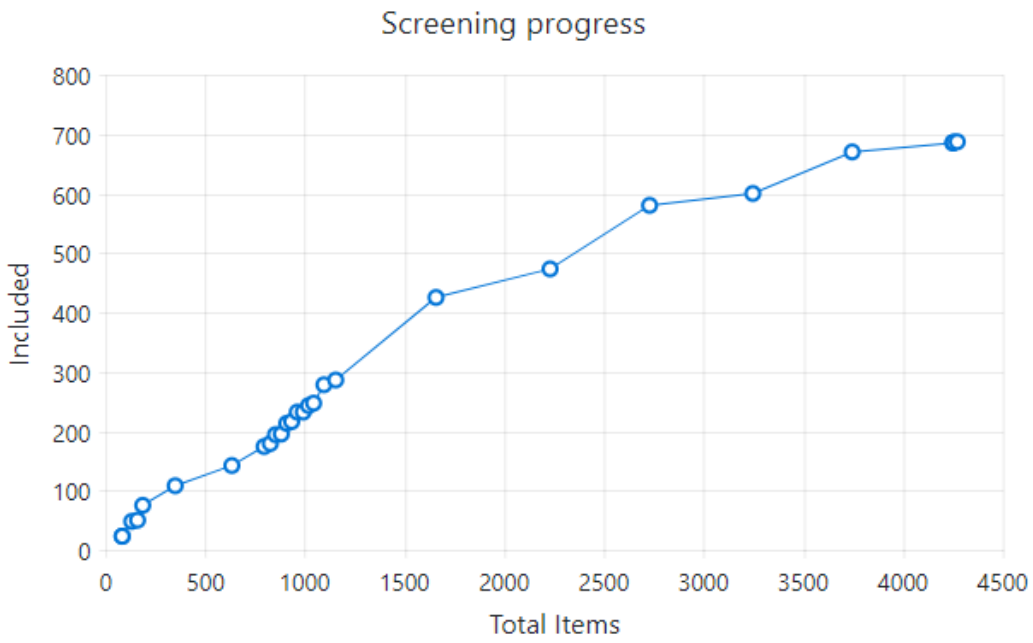

Figure 1 - the screening 'hit rate' over the course of the priority screening process

2.3. Inter-rater reliability, data extraction and sense-checking

Studies were screened and coded by a team of seven researchers and inter-rater reliability (IRR) tests were conducted to resolve discrepancies and ensure consistency and quality of coding. This process served to check whether coders shared a common understanding of the inclusion criteria. Researcher screening training commenced with an IRR test at each stage, to ensure coding behaviour was consistent across the team. The training was supported by a bespoke codebook which was updated and refined after weekly team meetings that discussed records that were deemed borderline (which were initially coded as 'Include for second opinion').

The inter-rater reliability exercises provided an opportunity to discuss the reasons why researchers made certain decisions. For each stage, this discussion was done in two batches, because the maximum number of coders the software could process was 3 at a time. The seventh team member was the lead researcher who had created the codebook and all disagreements were discussed as a team.

The include/exclude agreement rates can be calculated in two ways: 1) counting 'include for second opinion' as an include, and 2) removing 'include for second opinion' from the disagreements. The former is the main proportion reported below, with the latter in parentheses. For the IRR test for screening on title and abstract (n=101 randomly selected studies) the agreement rates were:

- Researcher 1 vs. 2 = 92% (87%)
- Researcher 2 vs. 3 = 97% (89%)
- Researcher 1 vs. 3 = 96% (90%)
- Researcher 4 vs. 5 = 96% (89%)
- Researcher 5 vs. 6 = 85% (79%)

[6] This was due to the hit rate dropping throughout the process. So, we could expect the 3% hit rate to progressively decrease as we screened the remaining 5,593 records.

- Researcher 4 vs. 6 = 89% (83%)

As can be seen from these proportions, a sizeable number of the records that caused disagreements at this stage were coded 'include for second opinion' by researchers. This was to be expected at an early stage of familiarisation with the inclusion criteria, and when these are discounted, the agreement rates were all 85 per cent or above, which is considered acceptable according to research standards[7]. Given these results, screening on title and abstract then progressed with single coding (i.e., one researcher screening each record), with any records that were deemed borderline were flagged for a second opinion and were discussed in team meetings, or with the lead researcher so as to arrive at a collective decision.

For the training IRR test on screening on full texts (n = 18 randomly selected studies) the agreement rates were:

- Researcher 1 vs. 2 = 94% (83%)
- Researcher 2 vs. 3 = 89% (83%)
- Researcher 1 vs. 3 = 83% (78%)
- Researcher 4 vs. 5 = 83% (83%)
- Researcher 5 vs. 6 = 94% (83%)
- Researcher 4 vs. 6 = 89% (78%)

To ensure that decision-making at this crucial stage was consistent we aimed for 90 per cent agreement rates across the research team. For this reason, researchers undertook a pairwise second IRR test. The agreement rates for this exercise were:

- Researcher 1 vs. 2 = 97% (91%)
- Researcher 3 vs. 4 = 97% (91%)
- Researcher 5 vs. 6 = 100% (100%).

After this point, screening on full texts was changed to single coding (i.e., one coder), with the option to flag up borderline records to the team for discussion using a similar 'include for second opinion' code. All such records were discussed between the coder and the lead researcher at a minimum, with other team members offering opinions when appropriate to maintain consistency across coding decisions.

2.4. Coding procedure

Studies that were deemed eligible after the two-stage screening process were then coded for essential information relating to the research question. Sense checking exercises were conducted at the beginning and end of the coding process. The team met weekly to discuss specific studies and difficult coding decisions as well as to ensure whether codes were being applied consistently. Due to the tight timescales of the project, and the volume of eligible studies, no formal evidence appraisal was undertaken for the studies included in the evidence map.

The coding instrument was heavily influenced by the terms of reference given to us by NZ Police. We supplemented this with themes that we thought would be useful for understanding the international literature (e.g., region of the world). Consequently, codes were prospectively developed, but inductively added to when they represented a recurring theme and reformulated at a quality assurance stage when the studies were viewed collectively.

[7] For example, see: Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociological methods & research, 50(2), 837-865.

3. Results

A total of 403 studies were judged as eligible at the end of the screening process. Figure 2 shows the breakdown of studies through this process[8]. This shows that after 114 records were identified as duplicates, 4,517 records were screened using the priority screening machine learning algorithm (see section 2.2.). Out of these, and based on the title and abstract:

- eight were excluded for not being in English,
- 180 had too little information (there was too little information in the abstract section - often because the record was not formatted in an academic style)
- 1,244 records were not on police as an organisation (e.g., were on the wider criminal justice system rather than police decisions)
- 1,785 were not an assessment of disparity of police action, and
- 549 were not empirical research.

Therefore, a total of 3,765 records were excluded in the first round of screening, leaving 752 to be screened by consulting the full texts (when these could be sourced). Of these:

- One study was published pre-2000 (our cut-off, chosen to align with the GPD data collection period).
- Ten were not in English and could not easily be translated (although may have had an English abstract).
- 57 studies were not retrievable. These studies largely related to conference abstracts and unpublished documents.
- 27 records were not on police (we had refined our understanding of police at this stage in consultation with NZ Police to not include drug enforcement agents and immigration professionals and off-duty police officers).
- 125 were not explicitly assessing disparity of police outcomes.
- 62 did not report empirical data, or if they did, they were data reproduced from other publications rather than being original data.
- 28 records did not use a comparator.
- 39 records were documents linked to eligible studies (e.g., a report/thesis and a journal article on the same study, by the same authors).

[8] This figure is known as a PRISMA diagram in systematic review terminology.

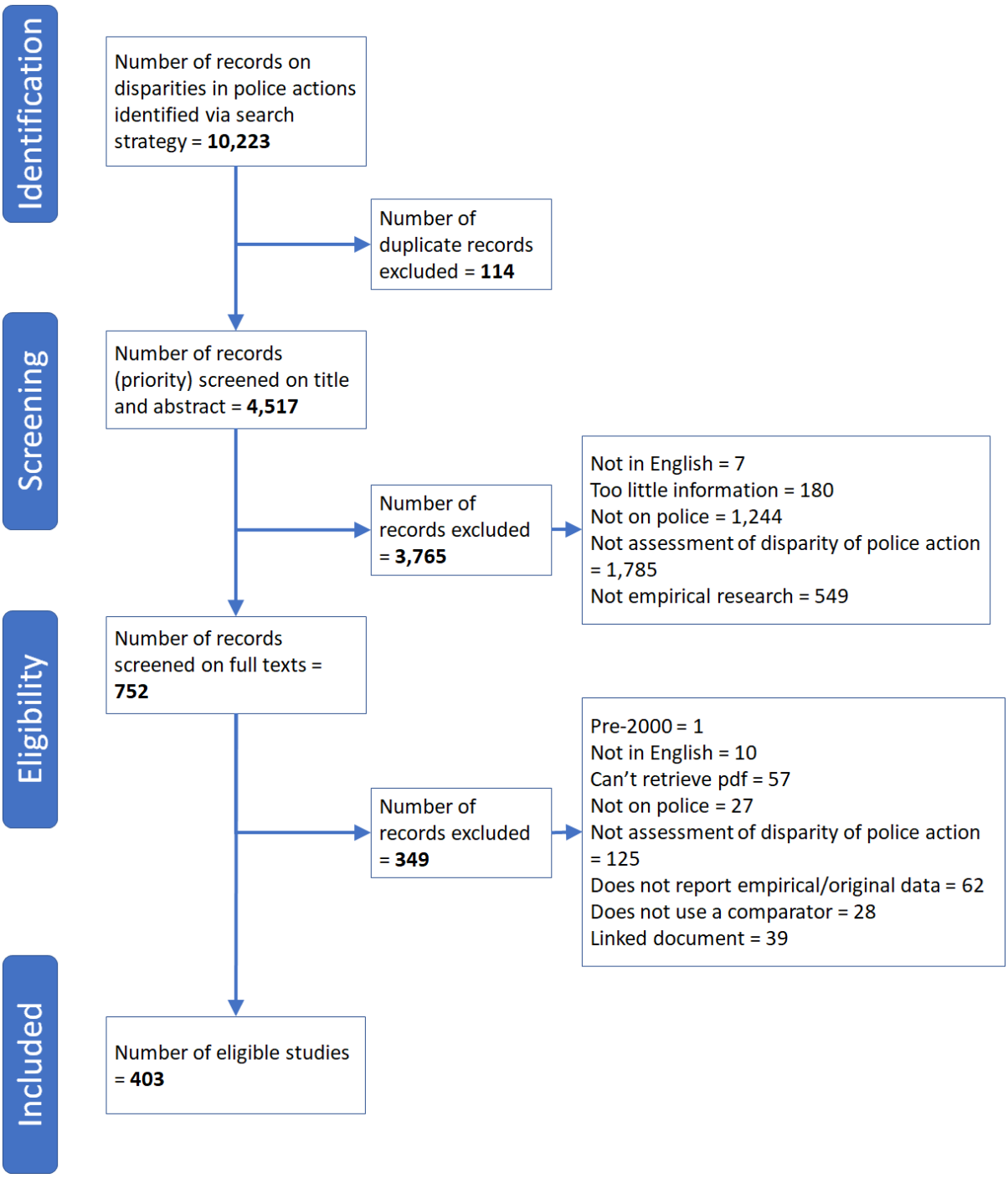

Figure 2: PRISMA diagram of study selection

3.1. Year of publication

Starting with a temporal profile of the evidence map, we can see from Figure 3 that publication of studies on disparities in police outcomes started to gather pace in 2004. For the next decade there was a steady state of around 20 studies a year published, before a sharp increase in 2016 to 41 studies. This trend appears to continue, with 54 eligible studies being published in 2018. We anticipate that this trend will continue.

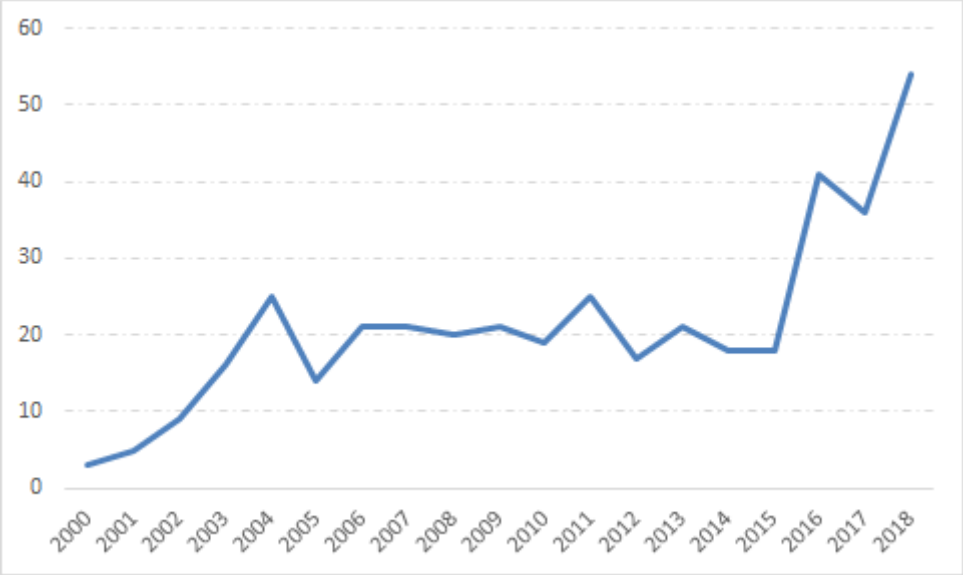

Figure 3 - publication date of studies on the evidence map

3.2. Regions of the world

Before summarising the geographical distribution of studies, it is worth noting that resource constraints prohibited the translation of studies published in languages other than English, and hence this aspect of the inclusion criteria biases studies on the evidence map towards western countries. This is a common limitation of evidence syntheses, but one that is not easily overcome in the absence of significant resources for translation.

As figure 4 illustrates the breakdown of studies by continent (inner circle) and, where multiple studies existed, by countries (outer circle). This shows most of the research on the evidence map comes from North America (86%), with the bulk of this coming from the United States (329 studies). In addition, there were 15 studies from Canada and two from Mexico.

The second most represented continent was Europe (8.1%), with 23 British studies and 10 studies from continental Europe. Australasia comprised another 4.2%, with 14 studies from Australia and 3 from New Zealand. A smaller proportion was seen from South and Central American countries (1.2%), with 3 studies from Brazil and 2 from other South American countries. Lastly, there was one study from Africa and three from Asia that were published in English.

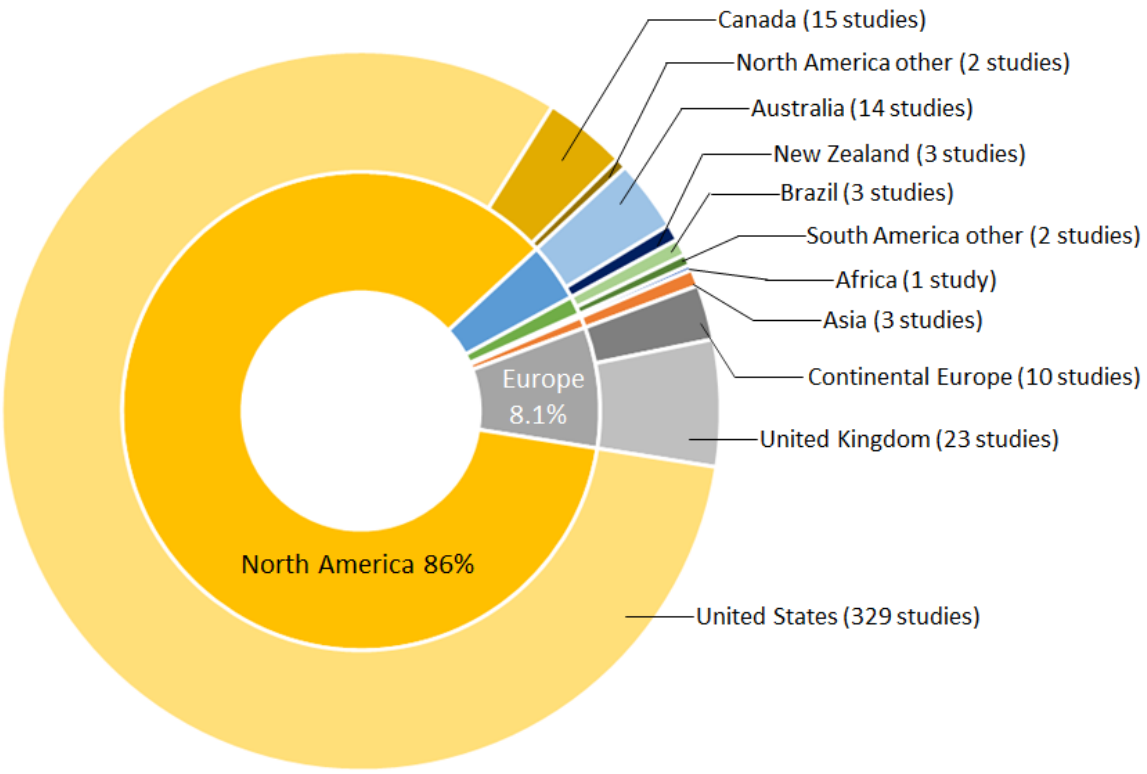

Figure 4 - the geographical breakdown of studies on the evidence map

3.3. Type of policing agency

The structure of policing varies internationally, with the relationship between the police and the state, and the philosophy underpinning the function of the police taking many forms. We captured the type of policing agency in the studies on the evidence map so that assessments could be made about the generalisability of the findings to particular policing jurisdictions. As above, these are biased towards western-style policing agencies and can be summarised as:

- Most studies on the evidence map covered urban areas and/or were city-based (57% or 228 studies).
- A notable proportion of studies covered a large geographical area (29% or 119 studies). These often used nationally collected data, or the police agency covered an entire state (e.g. in the US or Australia).
- 24 studies (6%) related to highway patrol in the US.
- Rural police agencies were under-represented on the evidence map, with just 2 studies.
- One study related to a militarised police agency (SWAT teams).
- 39 studies did not contain enough information to apply such judgements (or spanned multiple police agencies that were dissimilar).

3.4. Police actions

A wide range of police actions have been the focus of research attention in the studies on the evidence map. Due to these studies primarily hailing from the US, in places we have retained American policing terms to describe the actions (e.g., 'frisks'). But otherwise, wherever possible we

have sought to group similar actions together from different regions of the world, so that different audiences can readily understand where they fit in the police workflow process (e.g., citation / warning / caution are grouped together). Lastly, for a New Zealand audience, we have noted where charging decisions are widely understood to fall under the category of 'alternative resolutions'. Categories are not mutually exclusive so when a study examined multiple police actions, we coded all the actions. The numbers in table 1 therefore do not add up to the total number of studies (n= 403).

Visual inspection of table 1 reveals the five themes under which we organised the police actions[9]. Taking these in turn, under 'deployment', we see that there were 15 studies that examined patrolling or police presence in specific areas, with a further seven studies looking at dispatching police once an incident report had been called in (usually by citizens).

Under 'acting on suspicion' there were 116 studies that assessed police stops of citizens, with a number of these also specifying post-stop outcomes. Four studies looked at record checks, three were on asking citizens to move on or disperse, 21 included frisks, and 106 covered post-stop searches. Of note, some studies covered post-stop outcomes without assessing the stops themselves (i.e., the data used was all citizens stopped, so an assessment of bias was not being made at that with regard to the decision to stop itself). Actions that could occur independently of police stops included threat of, or the use of, force which was covered by 123 studies. One study looked at pursuit of vehicles and two investigated breathalyser tests.

With respect to 'investigation', four studies looked at the process of deciding whether an incident should be recorded as a crime; 21 studies examined an aspect of the investigative process (e.g., factors associated with clearance rates) and two studies focused on interviewing suspects.

A large proportion of the literature examined bias in relation to 'charging decisions'. Here, 46 studies looked at the alternative resolution of warnings or cautions (known as 'citations' in the US). A focus on fines comprised another 26 studies. Two studies looked at the severity of a crime recorded, so decisions related to whether to record an incident as more or less severe/serious. Many studies (n=129) investigated disproportionality in relation to arrests, with a further eight studies looking at clearance rates (which exclusively used arrests as the outcome measure). Three studies focused on disparities in referring people to other agencies or services, which again falls under the alternative resolution umbrella, and five studies looked at diversion from court processes. Five studies examined whether pre-trial detection (i.e., remand) was equitable. Finally, 14 studies compared multiple charging decisions within the same study.

A smaller body of literature looked at issues relating to professional conduct. For example, four studies examined procedural justice (or the lack of it); eight studies covered aspects of police verbal communication with citizens, including one study on de-escalation techniques. Three studies focused on whether there were disparities with regards to complaints against police being upheld. Finally, nine studies did not fit any of the above categories. These included studies on forming suspicion (n=2), soliciting bribes (n=1), recording race and ethnicity data (n=1), a SWAT team executing warrants (n=1), strip searching in custody (n=1), sexual abuse of citizens (n=1) and general 'police misconduct' (n=2) which was a composite variable made up of multiple behaviours.

[9] To organise these into something easier to understand we consulted with Inspector Scott Gemmell who chairs the operational reference group for the Understanding Policing Delivery project. He assisted in highlighting that police actions are often not undertaken in a linear manner, but they can be iterative and complex and reflect the variety of situations and incidents encountered.

Police actions studied		n
Deployment	Patrolling/presence in specific areas	15
	Dispatching police to incident reports	7
Acting on suspicion	Stops	116
	Post-stop record checks	4
	Post-stop asked to move on/disperse	3
	Post-stop frisks	21
	Post-stop searches	106
	Threat of, or the use of, force	123
	Pursuit of vehicles	1
	Breathalyser tests	2
Investigation	Recording an incident as a crime	4
	Investigating crimes committed by citizens	21
	Interviewing suspects	2
Charging decisions	Citation / warning / caution (alternative resolution)	46
	Infringements and fines (alternative resolution)	26
	Seriousness of offence	2
	Arrests	129
	Clearance by arrest	8
	Referred to other agencies/services (alternative resolution)	3
	Diversion from court process (alternative resolution)	5
	Pre-trial detention	5
	Multiple charging decisions compared	14
Professional conduct	Procedural justice/injustice	4
	Police verbal communication with citizens	8
	Complaints against Police	3
Other	Not categorised elsewhere	9

Table 1 - police actions studied across the studies on the evidence map

It should be noted that studies where the police action could not be reliably determined, that is, they discussed or examined 'interactions' or 'contact', were excluded from the map. This meant that some research on under-studied populations (e.g., see Steele et al., 2018) and using interesting methods (e.g., see Crutchfield et al., 2012) is not represented on the evidence map

3.5. Dimensions of bias

Bias, or discrimination, can take many forms in policing as in life. It is widely acknowledged that it can be structural in nature (e.g., poverty), institutional (e.g., police policies or culture specific to a particular agency) and can separately be perpetuated by individuals (who are implicitly or explicitly biased). It is also possible, and even likely, that these interact to compound bias that leads to disproportionate outcomes and conditions for minority groups.

When coding the evidence map studies, we strived to categorise the dimensions of bias along the types of human characteristics that might shape unwelcome stereotypes. These needed to be social, rather than behavioural. For example, suspect demeanour regularly came up in studies on police stops, but this is a behaviour rather than a social characteristic.

It is worth saying that many studies looked at multiple dimensions of bias - so they may have (say) looked at race, age and gender. Or studies may have examined individual dimensions (e.g., officer race) alongside training standards or some other characteristic of the police agency.

Lastly, some studies attempted to look at ecological dimensions - these encompass place characteristics (e.g., high crime areas) and time characteristics (e.g., late at night). Ecological explanations for bias typically look at the interactions between human behaviour and the environment/situation. Therefore, they expect the environment (e.g., a police beat) to shape police behaviour and decision-making. When the study setting was disaggregated into smaller areas such as neighbourhoods or police beats with the express purpose of controlling for place-based effects we coded these as 'place characteristics'. The same applied when studies examined the impact of the time of day or day of the week. For example, some studies proposed that officers' decision-making might be different at night-time or when it was dark. We coded these as 'time characteristics'.

As we see from table 2, 86.9% of studies covered race as a dimension of bias (n=351), with gender (n=166) and age (n=143) being the next most studied characteristics. A range of other individual-level characteristics were also studied but in smaller numbers of studies. Also noticeable in table 2 is the fact that only a minority of studies covered structural dimensions of bias, and these largely related to legislation (n=14). With regards to institutional dimensions, 36 studies examined the diversity composition of the police agency, with a further four studies looking at training. One study investigated the nature of bias in police data collection itself[10]. Ecological dimensions of bias were assessed in 78 studies, with nine of those looking at both place and time characteristics. Overall, we see a large range of dimensions covered by the evidence map, albeit race remains dominant, which is wholly unsurprising given that most of the literature comes from the US where there is a long history of strained race relations in society as within policing.

[10] Other studies touch upon this too. See, for example: Schlosberg (2002) and Lundman (2010).

Dimensions of bias		n
Structural	Legislation	14
	Racial composition of elected officials	1
Institutional	Diversity composition of police agency	36
	Training (standards or continued professional development)	4
	Bias in data collection	1
Individual	Race or ethnicity	351
	Skin colour	8
	Accent/language/nation of origin	17
	Religious beliefs	3
	Immigrant	3
	Place of residence	1
	Dress or appearance	12
	Age	143
	Gender or gender identity	166
	Income / education level	74
	Disability or health issue	24
	Sexual orientation	5
	Substance misuse	34
	Victim behaviour	15
	Suspect known to police	19
Ecological	Place characteristics	78
	Time characteristics	18

Table 2 - dimensions of bias covered by the evidence map studies (note that one study may examine multiple dimensions)

3.6. Dominant theories in this research area

Analyses of police behaviour have consistently revealed differences between different social groups (Alvarado, 2016; Bolger, 2015; Hollis and Jennings, 2018; Lytle, 2014; Mekawi and Bresin, 2015). Disparities in police treatment of majority (usually white) and minority citizens have been observed across many police functions, including, but not limited to, stops and searches of citizens, arrests, investigations, and charging decisions. There is substantial evidence of the negative effects of this differential treatment on minority groups (Desai et al, 2012), particularly because it is often occurring on top of layers of other types of disadvantage. Determining the sources of, causes of, or motivations underlying, these differences however, has been very challenging from traditional criminological perspectives.

Disparity is a difference between two comparable groups. The difference may be due to legally relevant factors or, in contrast, contextual factors that have no bearing on the legal factors. These are known as legal and extra-legal in the literature (Dabney et al., 2017). Extra-legal factors relating to a citizen's social characteristics (e.g., age, gender, race, ethnicity, appearance) are often seen as illegitimate factors on which to base a police action and is often equated with discrimination.

Research that identifies a disparity in police outcomes for different social groups cannot explain why those disparities exist unless they use theory. Theories are based on a set of concepts and assumptions and make the relational links between them explicit (Kraska, 2004). Research that does

not test theory is arguably not scientific (Bernard and Ritti, 1990) since science depends on the assessment of evidence in support or otherwise of a hypothesis.

In early research on police bias, Engel et al (2002) highlighted that the evidence base was largely devoid of theory, and thus failed to create insight into possible causes of disparities that were found. Since then, many theories that might explain why disparities in police outcomes exist have been proposed or elaborated. Some of these are complementary to each other while others are in competition with each other.

There are many ways to categorise theories about disparities in police outcomes. A crude distinction is between those theories that assume that there is *differential offending* (i.e., different social groups commit crime at different rates) and those that assume there is *differential scrutiny* (police treat different groups differently, regardless of criminality). Many sociological theories suggest the former is plausible: for example, strain theory, control theory and subcultural theory (among others) and there is evidence to support that minority groups do commit some forms of crime more than majority groups (Smith and Alpert, 2007). Importantly, the Casey Foundation (2003) talked about the greater number of risk factors for criminality that minority youths face growing up in poverty, such as underperforming schools, poor healthcare, food insecurity, violence in the home and many more besides. Empirical patterns in Western countries certainly support that minority groups are disproportionately represented in the criminal justice system, although they are often not assessed alongside which of the above assumptions are more plausible. Traditional criminological theories explaining differential offending are rarely drawn from in the research on police bias, which instead more heavily focus on differential enforcement—which is also called differential scrutiny or differential selection and processing.

Some scholars have argued that both differential offending and differential scrutiny may be simultaneously occurring, and have offered integrated explanations (e.g., see Piquero, 2008). Others, like Owusu-Bempah (2016), emphasise potential causal links between these two theoretical traditions, and argue that differential scrutiny can undermine the legitimacy of police for minority communities which can promote law-breaking and differential offending (Tyler and Fagan, 2008). It is also conceivable that individuals who have experienced differential scrutiny may not call the police when they are victimised, due to mistrust of police.

Theories also differ according to what disciplinary tradition they hail from. Many theories are *sociological*, proposing that structural conditions and power differentials cause disparities in police outcomes. *Psychological* theories focus more on individual decision-making - usually, but not always from the perspective of the police officer. *Economic* approaches use game theory to propose the factors influencing officer and citizen behaviour. Other researchers have attempted to examine relationships between individuals and the situations they find themselves in - which we call here *ecological* approaches. As in many fields, most theories rely on a single explanatory factor. With the possible exception of ecological approaches, the field has yet to develop multifactorial theories that can better accommodate the likely complexity of the phenomena involved.

3.6.1. Summary of theories that meaningfully influenced studies on the evidence map

To contribute to answering the research question guiding this evidence map we extracted information from the studies on the evidence map regarding the theoretical framework used to frame the research. These theories had to be used meaningfully in a study; that is, they had to be used to motivate the study or the operationalisation of the measurements. For a theoretical framework to be coded as present in a study, the authors had to describe the theory (even if briefly), not just reference it in relation to an empirical finding or use it in the discussion to explain results. The theories did not have to be tested directly to be coded as meaningfully used, and indeed most

studies did not formally test theories in full. Here we summarise the three most common categories found for theoretical frameworks in studies on the evidence map.

As can be seen in Table 3, and consistent with Engel's observations almost 20 years ago, most studies in the evidence map (n=227) mentioned no explicit theoretical framework. In other words, a lot of research is still descriptive in nature and does not formally test theoretically derived hypotheses. Typically, when theory is not formally driving research, it is nevertheless motivated by researchers' implicit assumptions, and it was rare for studies not to discuss *any* of these. Implicit assumptions are problematic because they are typically not formalised in a conceptual framework that makes the relationships between entities clear, which readily leads to confusion and misunderstanding. For example, the notion of the 'symbolic assailant' (Skolnick, 1994) was mentioned in 35 studies. This term represents individuals whose appearance, gestures, or language might be judged by police to be markers of a criminal lifestyle. Another example is the 'veil of darkness' hypothesis that predicts that police are less likely to know the race of a motorist before making a stop after dark than they are during daylight hours (Grogger & Ridgeway, 2006). This idea is implicitly drawn from an ecological perspective because it examines the interaction of environmental conditions (e.g., darkness) with officer decision-making. However, studies on the 'veil of darkness' did not always specify the assumptions associated with individual decision-making, or the situational factors that related to decision-making.

The most common framework used by studies on the evidence map was minority group threat theory, which featured in 49 studies (12% - see Table 5). This is a sociological theory that is an outgrowth of racial threat theory (Blalock, 1967) and proposes that as the proportion of minority groups increases, the dominant group will feel threatened. The dominant group, with their social capital, will then put pressure on the authorities to 'control' this threat, by policing the minority groups more stringently. The prediction is that as the minority group becomes more dominant (in number, not necessarily power) the perceived threat levels subside and police outcomes for different social groups equalise. Minority group threat theory was first applied to large geographical units such as US States or counties. It has since been used at more sensitive levels of geography with ecological considerations. Overall, the evidence base has been mixed for this theory, with some studies finding evidence for the theory and others finding no support for the theoretical propositions.

The next most common framework used is as much methodological as it is theoretical. It comes from economics and has been variously called the KPT test (after it's proponents, Knowles, Persico and Todd, 2001), the outcome test, or 'statistical discrimination'. The reasoning underpinning this approach centres on the idea that racial prejudice is different from the police acting as utility-maximising agents. That is, police officers may make decisions based on what will produce the greatest rewards, defined in operational terms, for the least effort. In this way of thinking, police believe that targeting minorities will yield better outcomes in terms of detecting crime. So, this relates back to police officers assuming that different social (racial) groups commit crime at different rates.

Both of these explanations may explain disparities in (say) stops of vehicles or pedestrians, but only racial prejudice will explain disparities in 'hit rates' of searches of citizens. In other words, if more people from minority groups are searched, but not found to have contraband on their person, this may suggest that they are being unfairly targeted. This approach, used exclusively in US studies, uses complicated economic statistical formulae to test whether the 'hit rate' of searches can be used to determine if the disparities can be explained by unfair targeting when other variables are accounted for. Such models typically make their assumptions about officer decision-making explicit. In keeping with the broader evidence base, evidence has been found to both support and undermine the racial prejudice argument using this methodological approach, depending on the study setting, data and assumptions made.

A further 18 theoretical frameworks were used in studies we reviewed (see Table 5), covering a range of police actions and putative causal processes. Some share assumptions (e.g., ecological theory of patrol and organisational theory), others are unique in their propositions (e.g., focal concerns theory). Most are formulated in relation to over-policing - the frequent and unfair stopping of members of minority groups for instance. However benign neglect theory deals with under-policing; it proposes that police vary in their beliefs about the extent to which some individuals/communities deserve police support. This theme also pervades other theories, such as the ecological theory of patrolling, which focuses on the 'vigor' with which police exercise formal authority across different communities. Another theme that emerges from some of the ecological studies looking at place characteristics is people who are 'out of place' attract greater police attention (e.g., Dabney et al., 2017). In other words, if an individual from a minority group is in a neighbourhood where the predominant residents are minorities, then they are not conspicuous, but if they are in a predominantly dominant group area then they are more likely to attract police scrutiny.

An additional twenty-eight studies covered different theories or hypotheses. These did not occur frequently enough over the evidence base to warrant their own category, so were coded as 'other'. These included: expectancy theory, feminist theory, attribution theory, the spatial opportunity hypothesis, concepts from cognitive psychology, representative bureaucracy, deference exchange theory, learning theories, and out-group salience bias.

This summary cannot do justice to the many nuances within the theories that are drawn from in the studies on the evidence map but suffice to say that there are many common threads that link different theories. It should, hopefully, be clear that there exist many causal explanations for disparities in police outcomes. Determining which one has the greatest explanatory power is challenging in such a complex research topic and research to date has not commonly tested mechanisms that explain individual officer decision-making. In addition, police decision-making happens over different timescales; some decisions, where there is threat to life, must be made very quickly. Other decisions can be taken in a slower, more considered way (e.g., collating intelligence about a suspect or incident before deciding how to proceed). Consideration of this in the studies on the evidence base was rare.

However, if we wish to enact change for more equitable policing, we need to understand the mechanisms driving any disparities in police outcomes. Theory can help us to do this, and also provide a roadmap to the sorts of data that need to be collected and the methods that are appropriate to test the theory. And to conclude, any high-quality research should look to explicate their theoretical assumptions so it is clear which body of evidence the findings can be usefully compared to.

Name of theory	n	Orientation	Key principles, assumptions, predictions
Minority group threat theory (Blalock, 1967)	49	Sociological, conflict, institutional/ individual	● As the minority population ratio increases, the majority group fear of crime increases. The majority group put pressure on authorities (including police) to exert social control over minorities to maintain their social status. ● Proposes a quadratic (curved) relationship between % minority population and police scrutiny of minority groups. When the minority population reaches a numerical majority, disparities in outcomes diminish.
Statistical discrimination outcome test (Knowles, Persico and Todd, 2001)	31	Economic, institutional/ individual	● Statistical discrimination is the term given to the police trying to maximise the success of their actions (e.g., stops) by targeting particular social groups (e.g., youths). This is different from racial prejudice and the 'outcomes test' proposes to determine which of these motivations underlies disparities in police stops in an economic model that compares 'hit rates' of searches or other actions taken by police (e.g., warnings or fines issued) across different social groups. ● Many studies have since challenged and/or extended the original model proposed in 2001.
Theory of law (Donald Black, 1976)	28	Sociological, conflict, institutional/ individual	● Suggests that a citizen's social status will determine how much 'law' (i.e., police scrutiny or action) is used against them. ● Proposes that incidents that appear to involve less informal social control will receive stricter enforcement from the police, which can escalate formal social control. Informal social control is estimated indirectly from situational or neighbourhood characteristics.
Conflict theory	21	Sociological, conflict, structural	● Generic. Views police as maintaining the status quo to uphold the social status of the majority group by targeting minority groups.
Ecological theory of police patrol (Klinger, 1997)	11	Ecological, individual	● Officer norms (rules and standards that are understood by a group) can be driven by: 1) subgroup effects from operational teams, and 2) places. These norms can affect officer decisions. ● Argues that social and ecological aspects of patrol beats can influence a patrol officer's perceptions of how "deserving" those places are of service and, in turn, influence outcomes of response.

Place hypothesis (Crank 1998)	10	Ecological	● Assumes that structural disadvantage and social characteristics interact (e.g., disadvantaged minority communities). Police may come to associate such communities with crime and threats to their personal safety through repeated exposure and vicarious experiences. ● Through the process of 'ecological contamination' or 'ecological attribution bias' all individuals engaged with in certain communities (e.g., high crime) will be perceived as a potential (real or symbolic) threat.
Social Disorganisation Theory	10	Sociological, ecological	● Neighbourhoods that are characterised by economic deprivation, racial dissimilarity (i.e., a mix of different people) and instability (people frequently moving in and out) are considered socially disorganised and these are the places where criminality flourishes. ● Proposes that neighbourhood context affects police decision-making. That is, more stops, arrests and general scrutiny occurs in socially disorganised neighbourhoods.
Focal concerns theory	8	Psychological, individual	● Like all humans, police officers depend on cognitive shortcuts to make decisions, and these are influenced by: 1) the 'blameworthiness' of the offender, 2) protection of the community, and 3) practical considerations. ● Such cognitive shortcuts are influenced by previous interactions with citizens, media portrayals of criminals, social identities and feedback loops that come from policies and organisational culture (among others). This can result in unconscious bias against minorities.
Racially biased policing theory	8	Sociological, conflict, institutional	● Rooted in conflict theories, this argues that police officers have explicit or implicit racial biases that shape their development of policies and practices in a way that disadvantages minority groups. ● Generic, no proposed relationships.
Social conditioning theory	7	Psychological, individual	● Explains racial bias primarily as an unconscious function of social conditioning and stereotyping. ● Implicit stereotypes associating minorities with crime and violence are developed through both direct and vicarious experience. This makes it more likely that police officers will process new situations through the filter of their scripts, which can result in assumptions about an individual being made based on perceived group attributes. ● Proposes that individuals with higher levels of conformity (e.g., police officers) may have hidden animosity towards individuals that do not conform to expectations. Consequently, negative perceptions about the non-conformist behaviour develop and are associated with all members of the group.

Benign neglect theory (Liska & Chamlin, 1984)	6	Sociological, individual	● Argues that government representatives (including police) do not place equal value on citizens - central theme is who is 'deserving' of police resources. ● Crimes in minority communities will be expected to involve a minority victim & assailant, and therefore do not threaten the majority group and will not receive police attention and/or support. ● Has also been used to explain differences in services to sexual assault victims (who is 'credible' or deserving).
Police organisational theory (Wilson, 1968)	6	Organisational, institutional	● Contains a variety of assumptions, often not framed in a relational way (e.g., representativeness of police agency). Generic. ● 'Formal organisational theory' - three major styles of policing organisations: 1) the law enforcer, 2) the social agent, and 3) the watchman. Believed that these styles translate into police priorities which may foster differential enforcement. Assumes most police officers will act similarly based on organisational culture. ● 'Informal organisational theory' - informal structures at the localised unit level (e.g., department, police specialism etc) are more influential on disparities in police outcomes than organisational factors (e.g., policies).
Differential drug involvement theory (Baumer 1994)	5	Sociological, individual	● Proposes that minorities are more likely to use and sell drugs and, as a result, minorities are disproportionately arrested and punished because they are disproportionately involved in drugs. ● A related assumption is that minorities are less likely to have access to private space and therefore be on the streets more, and consequently come to the attention of the police. ● Minorities are more likely to use and sell drugs as a response to the stressors of economic inequality.
Differential scrutiny theory	5	Organisational	● Proposes that police presence is greater in communities characterised by high crime rates, particularly violent crime, and large volumes of citizen complaints. Minority groups are more likely to live in these communities than majority groups. Consequently, the heavy deployment of police officers to communities where minorities are more likely to live increases the risk for police scrutiny. ● Also known as the 'deployment hypothesis'.
Community accountability	4	Sociological, organisational	● Argues that the organisational characteristics of police agencies promote excessive scrutiny of minority groups.

theory			● The closed nature of the police subculture is suggested to diminish the accountability to the communities they are supposed to serve. ● Proposed solutions are policies that hold the police accountable (e.g., diversity scrutiny boards, diverse workforces, publishing of disparity data, community policing).
Procedural justice theory	4	Sociological, organisational	● Procedural justice is a form of police discretion that encourages public belief in the legitimacy of their authoritative powers. It refers to the fairness of the procedures police use to deal with a situation. Four elements capture its meaning: participation, neutrality, dignity, and trustworthy motives. ● Studies typically investigate if procedural justice is distributed equally across citizens with different social status, situational context and/or citizen's behaviour during an incident.
Critical Race Theory	3	Sociological	● Generic set of assumptions that argue that existing power structures impact and oppress marginalised communities at the intersection of race, sex, class and other characteristics.
Consensus theory	3	Sociological, structural	● The police are seen as an institution that works to ensure the collective good, primarily by commanding compliance with laws that govern social behaviors. Police help to maintain order by their authority or by exerting force in situations where it is demanded
Shooter bias	3	Psychological, individual	● Shooter bias is the term given to the empirical finding that citizens are more likely to shoot armed targets more quickly and more frequently when those targets are Black, rather than White. This is usually explained in relation to unconscious (or conscious) bias on the part of the citizen.

Table 3 - summary of theories used in evidence map studies

3.7. Methods and methodological challenges

3.7.1. Data used for the police action

The data used to examine police bias is an important consideration for researchers. Given that this area of research represents a test of whether police have been acting fairly, there is arguably a conflict of interest when the data used to test that proposition is provided by police. Data from police may be biased or inaccurate for a variety of reasons (see Lundman, 2010; Schlosberg, 2002). This bias may be intentional, or it may be inadvertent, such as in the common example of police records of race or ethnicity, which may reflect an individual officer's (mistaken) perception during an interaction where they did not explicitly ask for that information[11]. Alternative sources of data (e.g., direct observations by a researcher, crime surveys, etc.) come with their own advantages and disadvantages, some of which mirror the issues with police data (e.g., mistaken perception of race or ethnicity by researchers). At the very least, they provide an alternative viewpoint, making them an important factor to consider when evaluating research on police bias.

Most studies in the evidence map use data provided by police. There were 105 studies that used police-recorded crime or arrest data, 137 studies that used police records of stops of citizens (including both vehicle and pedestrian stops), and 50 studies that used police records of use of force incidents. By comparison, there were 93 studies that used some form of qualitative data provided from another source, usually survey, interviews or observational data. A further 62 studies were categorised as using yet another source of data (e.g., studies that simulated tasks such as shoot/don't shoot exercises (n=16), calls for service, complaints about officers, crowd-sourced data on police brutality, media sources, death certificates, to name a few).

3.7.2. Denominators

The denominator used to assess disparity, frequently referred to as the benchmark, is an important feature of research examining police bias (Fridell, 2004). Without a benchmark, it would be unclear whether a disparity is present. In its simplest form, the benchmark provides a point of comparison for the rates at which a police action occurred across a dimension of interest. For example, the number of individuals of different ethnicities stopped by police can be compared against the number of individuals of those ethnicities within the population where the study occurred, as measured by census data. In that example, the benchmark would be the residential population.

Benchmarks need to be chosen carefully. For several reasons, the residential population may be a misleading benchmark in some circumstances. Using the example of a traffic stop study again, the residential population may not reflect the population who: a) drive at all, b) drive in that particular area at that time of day, or c) commit driving violations. If there are important differences between the benchmark residential population and the traffic stop sample used, a disparity may reflect the difference between the actual driving population and the resident population. A better indication of possible bias may come from using an estimate of the driving population in that area or an estimate of individuals who commit driving violations. The care with which benchmarks are chosen may therefore reflect the thought or effort put into establishing the most appropriate comparator. Unrepresentative benchmarks introduce errors into the research design and can produce misleading findings.

Due to the wide range of possible benchmarks available to researchers, rather than coding each individual type of benchmark, we categorised the benchmarks used into six domains: 1) residential

[11] Although it is important to note that if actions *are* based on misconceptions, then police data may be the best source of information to examine that.

population, 2) available/at-risk population, 3) stopped population, 4) suspect population, 5) offender population, and (6) victim population. In Table 4, we provide definitions of each of these domains and examples of common approaches within each category. Broadly, the categories fall along a continuum from encompassing the entire population (of interest in the study), with no attempt to determine whether individuals in that population might be the target of police action, through to estimating a very specific group shown to have either committed a crime or been the victim of a crime, both of whom would be expected to be the target of police action. Where possible, we attempted to create categories that were both consistent with language used in this research (i.e., studies that used the term suspect were sorted into the suspect population category) and conceptually similar (i.e., falling under a single definition). One important note: although people who are stopped are conceptually similar to suspects, because police should stop individuals they suspect have committed a crime or a driving violation, rather than stopping individuals at random, we chose to include the stopped population and the suspect population as separate benchmarks. The high proportion of studies examining police stops and post-stop outcomes suggested they were deserving of their own category, and stopped individuals are often not described as suspects in the literature.

We defined the term benchmark broadly. Traditionally, a benchmark might be thought of as a rate that can be compared against the frequency of occurrence of the police action being examined. Alongside those traditional types of studies, we also included studies where the rate was implicitly evident through the sample group used for a particular study (e.g., studies that examined the likelihood of specific races/ethnicities being searched within a sample of stopped drivers), rather than simply comparing racial/ethnicity proportions in the stopped and searched populations. We also included studies where aggregate data was used (e.g., proportion of the population that is black in the area where the arrests occurred) as an independent variable in statistical models. Studies could have multiple benchmarks, most commonly when they examined two different police actions. For instance, stop rates might be benchmarked against an estimate of the driving population, whereas search rates might be benchmarked against the stopped population. Or sometimes studies compared the same action across multiple benchmarks to examine how benchmark choice influenced the findings.

In total, there were 126 studies that used the residential population as a benchmark; 66 that used the available population; 116 that used the stopped population; 103 that used the suspect population; 73 that used the offender population; and 36 that used the victim population. These findings show that many studies did not rely on the residential population—arguably the least rigorous methodological approach—to assess disparity. In fact, of the 126 studies categorised as using the residential population, there were 30 that also used the stopped population and 25 that also used the offender population. This is further evidence that it is relatively uncommon to rely solely on the residential population.

A major limitation of the approach we used to categorise the benchmarks is that we do not capture instances where, even though the benchmark may not represent the studied population, additional non-police variables that might explain a disparity were controlled for. For example, we did not record whether a study had controlled for whether an individual stopped by police was subsequently found to have engaged in a driving violation or criminal offence, unless those populations were used as the sample group. Given the range of different police actions included in the evidence map, we determined that it would not be possible to systematically code all possible explanatory or control variables in the limited time available. Anecdotally, coders observed that most studies used at least some control variables, most commonly a control related to whether an individual had committed a criminal offence or driving violation. Therefore, we would estimate that only a few studies examined disparities using solely a residential population benchmark.

Because so many studies examined multiple police actions and used multiple benchmarks, it is difficult to isolate the benchmark used most for each police action. Some trends are evident, including that the majority (78 out of 116) of the studies examining post-stop searches used the stopped population as the benchmark, but the map is otherwise hard to interpret on an aggregate level in that respect. Relatedly, several studies appear to use the same numerator and denominator. For example, 57 studies that examined stops appear to have used the stopped population as the denominator. In reality, though, this figure is an artefact produced by studies examining multiple police actions and using multiple benchmarks to assess disproportionality. There were, however, a few studies examining stops that did use the stopped population as the benchmark, including Grogger and Ridgeway (2006), who used stops at night (where race is not as visible) as the benchmark, and Bricker (2002), who examined the predictors of being stopped multiple times among the population of individuals who have been stopped at least once.

One of the primary aims of evidence maps is to facilitate the synthesis of subsets of studies. For the benchmarks, the evidence map provides a helpful resource to researchers who might be interested in methods of examining disproportionality of a particular police action. Perhaps the most methodologically challenging benchmark to identify is the available/at-risk population, most used in studies examining police stops. For these studies, we have provided a brief notation in the “Info” box in Table 4 highlighting the method used to estimate the driving population. Other benchmarks can also be challenging to estimate. A study by Beckett and colleagues (2006) stood out for their unique approach of surveying needle exchange users and observing “open air drug markets” to develop their offender population benchmark against which to compare drug arrest rates across race.

Benchmark	Definition	Common examples
Residential population	Measures of the entire population of interest, not restricted in any way to individuals who might be more likely to be the target of police action.	● Census data ● Survey data from sample intended to be representative of wider population
Available/at-risk population	Measures of the population of individuals who are available to or at risk of coming into contact with police, but otherwise not restricted to those who might have or did commit a driving violation or criminal offence.	● Adjusted census data (e.g., driving age population) ● Observational data of the available individuals (e.g., systematic observation of drivers in the area of interest) ● Survey data of sample intended to be representative of available population (e.g., survey of licensed drivers, or asking surveyed population how frequently they drive)
Stopped population	Measures of individuals, whether on foot or in a vehicle, who are stopped by police, where no restriction is made to those suspected or known to have committed a driving violation or criminal offence.	● Police data on individuals stopped ● Survey data from sample with experience of being stopped by police
Suspect population	Measures of individuals suspected to have committed a driving violation or criminal offence.	● Individual data from observed encounters with police where the individual was identified as being in the role of a suspect ● Individual data from use of force incidents ● Police incident data or crime victim survey data where information about alleged perpetrator was included
Offender population	Measures of individuals who have either received a formal sanction for a driving violation or criminal offence or who were observed by researchers to have committed a driving violation or criminal offence. Arrest was used as the threshold for a formal sanction.	● Observational data on driving violations (e.g., speeding drivers observed by researchers) ● Police data on individuals arrested ● Survey data from prisoners ● Arrest or conviction rates in the relevant area*

Victim population	Measures of individuals who reported being the victim of a criminal offence.	● Surveys of victims of crime ● Police incident data where demographic or behavioural information about victims was included

* This benchmark only applied to studies where the unit of analysis was rates of police action in a particular area rather than individual interactions with police.

Table 4 - Definitions of benchmark categories used to assess disparity and examples of types of data falling within each category

3.7.3. Analytic methods used in eligible studies

As our inclusion criteria did not eliminate any research designs, a range of methods were used across the studies on the evidence map. However, methods were heavily skewed towards inferential statistics (n=304) which often involved regression models (see Figure 5). Other types of robust statistical methods used were propensity score matching (n=8), economical statistical models (n=4), and spatial statistics (n=1). These studies often also presented descriptive statistics to contextualise the data. Where studies presented multiple statistics, we only coded the most robust method.

110 studies only presented descriptive statistics. 47 of these used univariate statistics (e.g., proportions, means and standard deviations) and 56 used multivariate statistics of more than one variable (e.g., cross-tabulations, correlation statistics).

22 studies used qualitative methods to analyse their data. These ranged from approaches using grounded theory (n=3), to narrative analysis (n=4), thematic analysis (n=2) and content analysis (n=2). However, most studies were not explicit about the qualitative technique they used to analyse their data. Finally, five studies were evidence syntheses. That is, they were systematic reviews, or meta-analyses of multiple primary studies (for a summary of the findings of these see section 2.8).

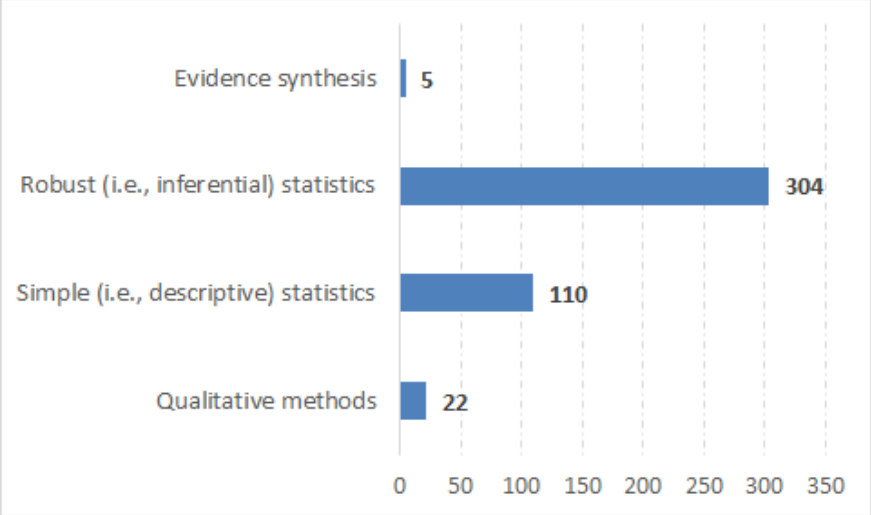

Figure 5 - methods used by studies on the evidence map. N.B. where multiple statistics were present, the most robust method was coded. Where studies collected qualitative data and then analysed those data quantitatively, we coded both. Evidence syntheses refer to systematic reviews and meta-analyses

3.8. Summary of aggregate studies of empirical findings on key police actions

Summarising all the studies on the evidence map would be a herculean task, and not one that is in keeping with the spirit of producing an evidence map (see introduction section). Since we did not undertake systematic quality appraisal of the studies on the evidence map, presenting a summary of findings based on unappraised studies may lead to erroneous conclusions (i.e., because the methods used to arrive at the findings are weak). For this reason, we present here findings from systematic reviews and meta-analyses, which are usually considered highly reliable as they use systematic methods to collect data (i.e., primary studies), use transparent criteria to select studies, use quality appraisal methods to rate the studies and synthesise the findings using the most appropriate conceptual or statistical frameworks. However, like primary studies, meta-analyses (i.e., evidence syntheses) can be executed on a continuum of weak to strong. In what follows, we provide a summary of the portrait of evidence offered by the evidence syntheses on topics relating to police

bias but point out where we have concerns about the robustness of the methods used to synthesise the evidence[12].

Given the large number of studies we included in our map, it was surprising how few evidence syntheses have been conducted in this area. We found only two meta-analyses on arrest decisions; one focused solely on race (Kochel et al., 2011) and one that looked more broadly at all suspect characteristics (Lytle, 2014). An unpublished master's thesis (Alvarado, 2016)—on search decisions during traffic stops—was the only evidence synthesis to examine an aspect of police stop decisions, even though this is the most common police action on the evidence map. The Alvarado study was though weak because the decisions made about study eligibility were not well reported, and it was not clear which 38 studies had been aggregated in the meta-analysis. Hence, caution is needed when understanding the findings from this study.

There was one meta-analysis (Bolger, 2015) and one narrative meta-review (Hollis & Jennings, 2018) on use of force. Finally, there was a single meta-analysis on performance in racial bias shooting task studies (Mekawi & Bresin, 2015).[13] This latter meta-analysis used very limited search tactics (one electronic database, plus Google scholar, which is not recommended – see Tompson and Belur, 2016) and thus the findings from this study may not be representative of the wider evidence base on shooting task studies.

As we have already noted, the absence of studies published in the almost 3 years between 2018 and the time of writing in June 2021 is a limitation of our evidence map. To supplement the six studies on the evidence map, we searched the literature for systematic reviews and meta-analyses that were published between 2018 and June 2021. We found two additional ones: one on the effect of suspect demographics on search decisions for both drivers and pedestrians (Bolger & Lytle, 2018). The other meta-analysis was on arrest decisions, but this one focused specifically on arrest decision making in sexual assault cases (Lapsey et al., 2021). None of these evidence syntheses included more than a fraction of studies we reviewed on the evidence map —the largest number was 42—and therefore these syntheses do not comprehensively cover the area we have mapped out in this report.

In Table 5, we provide a summary of the key features of the eight evidence syntheses examining disparities in police outcomes conducted between 2000 and 2021. Reflecting the broader evidence map, these studies synthesising evidence focused predominantly on racial bias in North America. Information about the methods used in individual studies to assess bias were often not provided in these aggregate studies. In other words, it was difficult to ascertain whether the control or moderator variables that might explain a finding of disproportionality were tested in the same models as the individual characteristics or whether they were tested separately (c.f., Kochel et al., 2011; Lytle, 2014). Similarly, it was often unclear what the denominator was in the calculation of disproportionality. The term 'suspects' was used in almost every study, but this was never explicitly defined. Therefore, it was often unclear whether the term referred to individuals known or suspected to have been engaged in criminal behaviour, or whether it referred more broadly to all individuals involved in encounters with police.

The most notable feature of Table 5 is the consistency of the findings from these evidence syntheses. In almost every study, some form of disparity was found. On the issue of race, these studies suggest minority suspects are more likely to be searched (Alvarado, 2016; Bolger & Lytle,

[12] Whilst we have not undertaken a formal evidence appraisal of these evidence syntheses due to time constraints, a preliminary review of the quality of such studies was undertaken. Where we do not raise concerns about the methods used, we consider these satisfactory from a methodological perspective.

[13] An earlier systematic review on police use of improper force (Harris, 2009) was excluded from the map because it only included three studies, all of which were published prior to 2000 and no attempt was made to aggregate the findings of the three studies in that review.

2018), arrested (Kochel et al., 2011; Lytle, 2014), have force used against them (Bolger, 2015), and be shot (Mekawi & Bresin, 2015) than white suspects. Although Hollis and Jennings (2018) concluded that there is mixed evidence of the relationship between race and use of force, most studies in their synthesis that focused specifically on police action (as opposed to perceptions of disproportionality) appeared to demonstrate evidence of disproportionality. The only evidence synthesis study that did not find evidence of racial bias was the recent meta-analysis by Lapsey and colleagues (2021) that focused on arrests in sexual assault cases using focal concerns theory. In fact, these researchers found arrests of perpetrators were more likely when the victim was non-White.

Each of the three meta-analyses that examined gender (Bolger, 2015; Bolger & Lytle, 2018; Lytle, 2014) found evidence of disproportionality, but this was not similarly seen when age was examined. Bolger (2015) did, however, find that suspects who were lower class, intoxicated, or demonstrating mental health issues were more likely to have force used against them.

Author(s)	Publication date range [*Search range*]	Police action examined	Location of studies	*n* studies	Individual characteristics	Control/moderator variables*	Author conclusions
Studies included in the evidence map							
Kochel et al. (2011)	1968 – 2006 [*Not specified*]	Arrests	USA	40	Race	Amount of evidence, crime type, crime during encounter, demeanour, offense seriousness, intoxication, criminal record, victim requested arrest, witness	"The meta-analysis shows with strong consistency that minority suspects are more likely to be arrested than White suspects[…] The significant race effect persists when taking into account the studies' variations in research methods and the nature of explanatory models used in the studies"
Lytle (2014)	1977 – 2010 [*Not specified*]	Arrests	USA	42	Race and ethnicity, gender, and age	Demeanour, offence seriousness, amount of evidence, intoxication, weapon use	"Black individuals, males, and Hispanic individuals were significantly more likely to be arrested than white individuals, females, and non-Hispanic individuals. These effects persisted across the majority of moderator categories"
Bolger (2015)	1998 – 2011 [*1995 – 2013*]	Use of force by patrol officers	USA	19	Race, sex, age, social class, mental illness, and intoxication	All effect sizes came from multivariate analyses but not further specified	"Suspects who are minorities, males, and/or lower class are more likely to have force used against them."
Mekawi & Bresin (2015)	2002 – 2012 [*Not specified*]	Performance in experimental shooter tasks	USA and Canada	42**	Race	Not applicable	"Relative to White targets, participants were quicker to shoot armed Black targets, slower to not shoot unarmed Black targets, and more likely to have a liberal shooting threshold for Black targets."

Alvarado (2016)	2000 – 2015 [– *Jan 2016*]	Searches during traffic stops	USA and Canada	38	Race and ethnicity	Not specified	Black motorists were 1.64 times more likely to be searched compared to White motorists.
Hollis & Jennings (2018)	1994 – 2017 [– *Aug 2017*]	Use of force	USA	41	Race and ethnicity	Not specified	"The findings were generally inconsistent across studies revealing that more high-quality research relying on more comparable operationalizations of variables and methodologies is needed"
Studies not included in the evidence map (i.e., published after 2018)							
Bolger & Lytle (2018)	2004 – 2014 [*1960 – 2017*]	Searches	USA	17	Race, gender, and age	Offence seriousness, presence of evidence, weapon, resistance, post-search arrest, conflict at scene, suspect intoxication	"Suspect race and gender appear to influence search decisions, while the age of a suspect appears to be of little consequence"
Lapsey et al. (2021)	2000 – 2019 [– *Dec 2019*]	Arrests in sexual assault cases	USA and Canada	14	Victim and suspect age and race	Suspect resistance, victim injury, presence of a weapon, physical evidence, report time, report time, witness, victim cooperation,	"Except for victim race, which increased the arrest odds by 1.49 when the victim was non-White, no other victim or suspect demographic variable impacted the magnitude of effects"

* Listed in this column are variables that explain, or arguably justify (to some extent), a finding of disproportionality; we have not listed methodological variables (e.g., year of publication) that are frequently tested as moderator variables in meta-analyses

** Not all studies used Police officers in their samples; however, moderator analyses found no difference in the performance of police officers (and recruits) compared to either undergraduates or other community members

Table 5 - Meta-Analyses and Systematic Reviews Examining Disproportionality in Police Action Published Since 2000

4. Discussion

This evidence and gap map was commissioned by the New Zealand Police to inform the Understanding Policing Delivery project. The research question guiding the work was: "What is known about the nature and impact of possible bias in policing policies and practices internationally?" Because the literature in this research area is vast and scattered across many domains, we decided to develop an *evidence map* of the existing international research on bias in policing. This resource can be used to understand the scope of research on this topic and where the gaps are in existing knowledge, and to enable literature reviews or syntheses of findings on smaller, more defined, subsets of studies.

4.1. Key findings

In this section we discuss what are, in our subjective view, key themes and findings across the 403 studies as an entire evidence base. What we hope has been made clear in section 3, is that bias – intimated through disparities in police outcomes – has been studied across many different police practices and policies. Often disparities are indeed found, but they belie any attempt to reduce them to simple headline findings.

Due to time constraints, we could not perform a systematic appraisal of the design quality of the studies in the evidence map (this is time consuming for dozens of studies, let alone several hundred). This means we are not suitably positioned to make judgements on the strength or reliability of the findings presented in this report. For this reason, we have not quantified how frequently disparities in police outcomes, as an indicator of bias, were found in the studies on the evidence map. Presenting headline findings uncritically would have risked oversimplifying and proliferating the limitations of the primary studies (as documented in section 2.7) and potentially doing a disservice to the reliability of the evidence base. Instead, we leave those judgements up to researchers who may make use of the evidence map to form their own literature reviews and conclusions. Systematic, quantifiable evaluations of quality can feasibly be conducted on smaller pools of studies, without requiring the very considerable resources such an appraisal would have required here.

Taking the key words in the research question, 'nature' and 'impact', we turn to the nature of potential bias first. One of our key findings is that this is a vastly complex area with many layers of nuance. A study could find a disparity on one dimension of bias (e.g., race), but not another (e.g., gender), or within one aspect of a dimension of bias (e.g., between Blacks and whites), but not another (e.g., between Hispanics and whites). Similarly, a study could find a disparity on one police action (e.g., searches), but not another (e.g., stops - see Alpert et al., 2007). Additionally, disparities were sometimes found to be operating in the opposite direction to the hypothesis - in other words, they found the majority group were disadvantaged (for an individual example see Helfers, 2016 or an aggregate example see Lapsey et al. 2021), or police officers from minority backgrounds were not more equitable than their white counterparts (for one example of this see Brown and Frank, 2006). Clearly, summarising whether bias exists is intricate and sensitive to the study setting, methods used, and rival explanations controlled for.

Perhaps the most straightforward finding is that three-quarters of the studies (n = 299) on the evidence map examine aspects of racial bias in US police agencies. This finding is unsurprising given the dominance of the US in social science research, and historical prominence of concerns about the mistreatment of African Americans, but it raises important questions about the generalisability or transferability of the findings. There are several reasons to think that aspects of that body of evidence may be unique to the US. In particular, the strained history of race relations, the vast range of different police agencies, and the preponderance of gun ownership in US society are features that may not be unique to the US but, arguably, are more extreme or dominant than in other jurisdictions. Perhaps most importantly, findings may differ in jurisdictions where policing by consent

is at the heart of the policing philosophy, which is not the case in the US. The bulk of the evidence map concentrates on studies of police stops of citizens (*n* = 116), the threat of or use of force (*n* = 123) and decisions around arrests (*n* = 129), which are of interest to the New Zealand Police.

New Zealand shares more similarity with Australia for its colonial past (which brings its own unique racial issues), and the United Kingdom for legislation and policing practices. However, studies from these countries are less voluminous than those from the US. Assessing the generalisability of the findings from the US is about the external validity of the studies; the extent to which you can generalise the findings of a study to other situations, people, settings and measures. For these studies to have a high level of external validity, the same mechanisms found to be operating in one setting, would be shown to be operating in a in a different but comparable setting. However, as noted in section 2.6.1. mechanisms are not commonly articulated in the studies on the evidence map, although they can sometimes be inferred from the hypotheses and the theoretical reasoning. Importantly, without knowledge about mechanisms it is difficult to design interventions that are likely to be effective since we don't know what we are seeking to change and therefore how an intervention should work. A potent future research agenda would be to mine the putative mechanisms from the US studies and ascertain which ones can be tested in other contexts (such as New Zealand). If the findings are corroborated across different jurisdictions, it lends strong weight to the theories used and provides a crucial basis for designing policy and practice interventions.

The processes that underpin decision-making conceivably are the mechanisms causing disparities in police outcomes and therefore are crucially important to advance this research area. One of our observations from surveying the literature on the evidence map is that such processes are difficult to chart, and accounting for the myriad influences that can be in play is fraught with methodological difficulties. Police officers themselves will be all too aware of the manifold ways in which incidents develop, even if they do lead to the same outcome (e.g., arrest). Police-citizen interactions will be influenced by the social and environmental conditions at the scene, what else is going on, and the citizen's behaviour, among many other features. But at a level above these concerns are the potential influences that an officer may bring to a particular scene. For example, if we consider racial threat theory to be an accurate picture of reality for a moment, how is this sense of threat transmitted to the officer making a traffic stop or deciding on arrest. Are they coming from policies espoused by management? From the ethos of the team they police within? Or from their own sense of threat? These differences have different implications. Aspirations to capture all this complexity in a study are rarely achieved, for understandable reasons.

The broad range of theories proposed to explain disparities in police outcomes incorporate dozens of mechanisms that might be causing police officers to make decisions that disadvantage particular social groups. Some are structural – insofar that the police are perceived to uphold the status quo of white dominance. Others are institutional and relate to the culture in a police agency, or the design of policies and practices that perpetuate inequities. And police agencies themselves are not monolithic but have subcultures that develop within different teams and specialist units. More still look at individual biases that are formed over the lifespan – that begin early in life and can be reinforced by community-level attitudes, media portrayal of minorities and the police role, and the nature of police work which frequently bring officers into the lives of structurally disadvantaged communities which can easily foster illusory correlations and strengthen stereotypes. As we already noted the evidence base to date has not been wholly successful at surfacing which mechanisms are more and less plausible, let alone in what circumstances each is most relevant.

To offer some insight into the question of whether and where bias exists, we provided a brief summary of the five meta-analyses and systematic reviews (i.e., evidence syntheses) that are included in our evidence map, and three more that have been published since 2018. Collectively, these cover 253 primary studies, many of which are included on the evidence map. Evidence syntheses are considered to be the pinnacle of frequently used 'evidence hierarchies' because of

their robust methods and, when executed well, their ability to transcend findings from individual studies and offer a composite portrait of evidence across a body of studies. Evidence syntheses on search, arrest, and use of force decisions consistently found that disparities in race and gender are evident, even after controlling for variables that may otherwise explain the disparity, although where rival explanations for disparities were controlled for, the magnitude of the disparities found often reduced, confirming that there may be layers of determinants of disparities. However, these evidence syntheses were dominated by North American studies, so it is unclear to what extent these findings apply to other jurisdictions including New Zealand.

Now turning to the 'impact' of police bias; this is a term that can be interpreted in various ways. One interpretation relates to how many social groups might be affected by biased decision-making in policing. Whilst race is the most studied dimension of bias, disparities have been found in many other areas, such as mental health, gender, age, skin colour, immigrant generation and many more. Ben Bowling and colleagues, known for their well-respected work on race and policing, note that: "Police abuse of force also occurs in countries where social divisions are based not on race or ethnicity, but on class and political affiliation (such as Jamaica), religious sectarianism (such as Northern Ireland) and tribal heritage (such as Rwanda)" Bowling et. al. (2004: 2). If this is true, it is conceivable that there is potential for biased decision-making by police officers that might uphold and exacerbate a variety of social divisions. The impact is thus widespread. The corpus of studies on the evidence map document disparities across many layers of identity characteristics, lending weight to the supposition that the impact of disparate policing may be experienced by many citizens.

Relevant to the New Zealand context, a handful of studies examined Indigenous populations. For instance, from Australia, Walsh (2017) documents that Indigenous people in Queensland are over-represented among those who are charged for using offensive language directed at a police officer. McCarthy et al (2018) look at police use of force in Australia, the use of which may be a consequence of perceptions of offensive behaviour. These scholars found a significant positive relationship between Indigenous populations and use of force in the absence of rival explanations; however, when these explanatory variables were included in the final model, this effect disappeared and hence the relationship was believed to be mediated by other socio-economic community and crime factors.

In another study, Snowball (2008) examined diversion rates for Indigenous and non-Indigenous offenders and found that the former were diverted at a lower rate than the latter. Adding controls into the statistical model (age, sex, current offence etc) the disparities between the two cohorts diminished but remained strong and significant. Other Australian studies were documented by Fitzgerald and Carrington (2011) in their study of disproportionate minority (police) contact in Indigenous populations[14]. On the same topic of diversion on Indigenous people in Australia, Allard et al (2010) found that Indigenous people were less likely to be diverted following their first offence recorded by police.

In her study of recorded rape offences in Victoria, Australia, Heenan and Murray (2006) found that 1.9 per cent the victim and offender population she studied were Aboriginal or Torres Strait Islanders, despite them comprising 0.5 per cent of the Victorian population. Therefore, Indigenous people were over-represented as both victims and offenders. Heenan and Murray further suggested that this was likely underestimated as in more than half of the case samples the discretion afforded

[14] Australian Bureau of Statistics 2008; Walker and McDonald 1995; Weatherburn, Snowball, and Hunter 2008) and New Zealand (Fergusson, Horwood, and Swain-Campbell 2003; New Zealand Department of Corrections 2007).

to police in recording Indigenous status meant that it was not recorded. She noted several studies on how Indigenous women are alienated from support provided by police[15].

On a similar topic, Bachman et al. (2010) examined the sexual violence victimisation experiences of American Indian and Alaska Native (AIAN) women. This study found that although victimisations against AIAN women are more likely to come to the attention of police, they are much less likely to result in an arrest compared to attacks against either White or African American victims. Whilst an important finding, the evidence base sorely needs more studies looking at how indigenous populations experience policing outcomes.

O'Brien et al (2011) conducted a study about the use of tasers on people with mental illness in New Zealand. They found that Maori and Pasifika people were over-represented in the incident data where ethnicity was recorded. Maori accounted for 28% of the sample and Pasifika people for 25%, despite their representation in the population being much lower. However, involvement of Maori and Pasifika people with the police was less likely to be attributed to mental health emergencies than for other ethnicities. O'Brien and colleagues suggest that this might be because the police are more inclined to manage such incidents as criminal ones, rather than mental health incidents.

Biased policing might also have downstream consequences. The articulation of such consequences was not a feature of our inclusion criteria, but nevertheless some primary studies on the map discussed this in some way and we review them here. For example, Levchak (2016) notes that unproductive stops of citizens carry a significant social cost. They can be seen as intrusive or illegitimate and undermine perceptions of police legitimacy. Substantiating this, Hitchens et al. (2018) found, in their qualitative study of young Black and Latina women, that the *quality* of interactions between police and young women from low-income communities differed according to race. Black and Latina women reported experiencing more negative encounters, 'punitive chauvinism', and a lack of support when under threat. These personal experiences coalesced with vicarious experiences of others in their communities to contribute to expressions of legal cynicism. Such cynicism is intertwined with police legitimacy.

In her interviews with 30 male juveniles in a correctional facility, Feinstein (2015) documents how discretionary police decision-making feeds into perceptions of unfair treatment and ill will towards the police. Themes that resonated throughout the accounts provided by youths from minority groups were that police were more lenient with White youths. Some youths from minority backgrounds were repeatedly arrested by the same officer, whereas this rarely happened with White youths. Police were said to factor family reputation into their decisions and often used unnecessary force against minority youth.

Being treated unfairly can foster attitudes to police that can attract even more police attention. A recurring theme in studies in the evidence map is that suspects' demeanour is one of the strongest predictors of police use of force (Worden et al. 1996, Engel et al. 2000). Put differently, discrimination and demeanour can be mutually reinforcing (Rosenfeld et al, 2012). In communities where there is longstanding mistrust of police, individuals are likely to start an interaction with a police officer from an anxious, defensive and/or hostile position (Brunson, 2007). Taken together with officer pre-conceptions, this reaction may exacerbate the potential for being perceived by officers as uncooperative and/or suspicious and reinforce pre-existing stereotypes. It is easy to see

[15] Heenan and Murray (2006: 11) note that "The willingness of Indigenous women to report sexual assault is inhibited by a distrust of, and alienation from, the criminal justice system (Lievore 2003). Research has documented that complaints made by Indigenous women are often inadequately investigated, leaving women feeling both disbelieved and vulnerable to re-victimisation (Lievore 2003, 2005; Thorpe, Solomon & Dimopoulos 2004)."

how resulting police actions then reinforce an understanding that police are unfair, which may spread throughout communities.

In addition, perceptions of 'respectful' demeanours may be culturally (mis)understood – for example, Kaldenbach (2011), as cited by Svensson and Saharso (2015), suggested that immigrant youths have yet to master the Dutch art of expressing 'sincere regret', which may see them treated more harshly than their Dutch peers. Other scholars have noted nonverbal communication cues, such as avoiding direct eye contact may be perceived as untrustworthiness, when it may simply not be a cultural custom or may actually be a sign of respect (Vrij & Winkel, 1992; Winkel, Koppelaar, & Vrij, 1998). Arguably, the police need to especially prioritise using procedural justice principles in dealing with individuals who are at risk of pre-emptively expecting unfair treatment, as the potential for the incident to escalate is conceivably greater than with individuals from a majority group, while fair treatment may have a helpful impact.

Further impacts of biased decision-making in police work are outlined by Meng (2014). These are that, first, it diverts police attention away from those that legally justify scrutiny – the 'real criminals'. It also diminishes the resources that are available to provide a police service to victims of crime. A lack of support—or under-policing—was evident in interviews with Roma immigrants across Bulgaria, Hungary and Spain (Miller and Gounev, 2007), in addition to over-policing.

Meng (2014) also argues that biased policing heightens anxiety and reduces feelings of safety and belonging in society in individuals affected by it. In turn these consequences may effect mental health, as was found in interviews with young men in Scotland about being stopped by police (Reid-Howie Associates, 2002).

To summarise the key findings, the only clear aspect of the evidence represented on the map is that the 'nature' of police bias is equivocal. It depends on the police agency being studied, the context in which they operate—which includes demographic, economic, political, and other features—the police action being examined, the policies and leadership within the organisation, among a whole host of other things. The sheer range of police actions studied suggest that biased decision-making can pervade the whole police mission, whether that relates to deployment, acting on suspicion, investigating crimes committed by citizens (and supporting victims), deciding what charge to lay against a suspect, or professional conduct more generally. This observation suggests to us that the way forward is more specific studies of police actions or specialist police functions to unravel what is happening, where it is happening and why.

With regards to impact, biased decision-making has been documented as affecting a wide range of social groups on the evidence map. The consequences of inequitable treatment of minority groups by police are far-reaching and overwhelmingly negative. They also undermine police legitimacy which is crucial to the police mission in the New Zealand Police.

4.1.1. Knowledge gaps in the evidence map

As noted in section 1, evidence maps are useful for determining where there are gaps in knowledge. Here we outline gaps that have not previously been mentioned (e.g., the lack of non-US research, the paucity of research looking at disparities in outcomes for Indigenous populations). For instance, few studies attempted to examine decision-making as a process. One notable exception was Alpert et al's (2005) observational study of the role of race in explaining how discretionary suspicion is formed in police officers in Savannah, Georgia (see also, Dunham et al., 2005). This started to disentangle some of the complexities with regards to how suspicion is formed. They conclude that police are more likely to form non-behavioural suspicions for individuals from minority groups, which is consistent with psychological theories on how cognitive schema operate. However, their results also suggest that this does not influence the ultimate decision to stop and question people that evoke suspicions.

Another exception was Lum's (2011) study on police decision pathways when recording an incident as a crime. Incorporating ecological considerations into her analysis – such as the crime rates of the places where incidents occurred, as well as community characteristics of such locations – she found that "places with a greater proportion of Black or wealthy residents significantly influenced officers' decisions to downgrade crime classifications and actions taken on incidents reported to the police" (Lum, 2011: 2). Lum offers some reasoning behind these seemingly contradictory findings. First, that informal social control likely plays out at a local level in both of these disparate communities (wealthy Black communities in the study area were uncommon). Wealthy communities may be able to 'handle things themselves' with their economic capital, whereas Black communities may be able to internally resolve disputes through their cultural capital and ties. Lum notes that police may exert less social control overall in Black communities, but when they do decide to use their authoritative powers, may do so more strenuously. These two examples serve to highlight that when research is designed to get at the heart of the decision-making process and at the level of specific communities, it reveals both intuitive and counter-intuitive findings. We need many more of these studies before we arrive at some consensus on how biased decision-making arises and operates within policing.

Another gap was the scant evidence on police verbal communication. Many studies highlighted that the 'interaction' between police and citizens was crucial but did not unpack the nature of the communication – ostensibly because data did not exist on this. Studies of body worn camera footage (e.g., Eberhardt, 2016; Willits and Makin, 2018) demonstrate potential to plug this gap. Other studies might consider calls for service dialogues or use systematic social observation or virtual reality simulations to better understand the verbal dynamics that lead up to police decision-making. The procedural justice subset of studies on the evidence map can also provide insight into such approaches.

Finally, the evidence base overwhelmingly focuses on racial bias. Whilst this is of paramount importance, other individual identity characteristics are important too. We were surprised not to find more studies on disability, religion or sexual orientation, for example.

4.2. Methodological complexities

The methodological approaches used in the studies on the evidence map are highly varied. In section 2.7, we summarised some of the key methodological challenges facing researchers. Our aim was to highlight several of the important decisions that researchers must make when designing studies that aim to examine police bias and provide some guidance about the options available to researchers.

One of the first decisions that confronts researchers in this area is where to obtain data on police actions. The evidence map indicates the most common decision is to use data provided by the police. This decision is understandable and most likely reflects the pragmatic reality that police hold the most extensive data on their own activities. Nonetheless, there are several issues and limitations with the use of police data. Lundman (2010) highlighted some of these issues in their study looking at whether systematic bias is present in how police record ethnicity data. Many other authors noted the limitation that they were restricted to using data that police recorded when they were aware this did not encompass the full extent of the relevant police-citizen interactions. For researchers wanting to make use of alternative data sources, the evidence map includes a lot of studies that used survey or observational data. Of course, these approaches have limitations of their own, perhaps most notably the possible influence of having a researcher present during an interaction. It is likely that a combination of these different methods will produce the most robust evidence following the well-established procedure of triangulation of evidence sources.

The evidence highlights the variety of different methods that are used in this area to assess disparities. It is widely recognised that using the residential population as a benchmark will be

inadequate in most situations, although there were still several studies that used that approach exclusively. More commonly, though, studies assessed disparity using a more rigorous benchmark that accounted for the likelihood either of coming into contact with police, or engaging in criminal behaviour. This indicates that researchers are actively grappling with the methodological challenges inherent to this area of research; however, in many cases, they appear to be reaching different conclusions about the most appropriate way to proceed. It appears highly likely that the most appropriate approach will depend on the police action being investigated. Future researchers would be advised to review the different approaches taken when examining a particular action to determine the most appropriate benchmark to use. By contrast, there is much greater agreement around the optimal analytical approach, with over three quarters of the studies in the evidence map using robust inferential statistics.

Probably the greatest challenge for researchers of police bias is explicating all the interrelated components that may be producing patterns of disparities in outcomes, and striving to control for them in analyses. For example, many factors that raise an individual's risk of police scrutiny co-occur. For example, looking at traffic stops, people from low-income communities are likely to have older vehicles in poorer condition than others. They may also be more likely to travel at night-time if they have jobs that require shift working. Particular roads that police commonly patrol (e.g., high crime areas) may be routes into minority neighbourhoods. All these factors might make it more likely that such individuals come to police attention during traffic stops, and these factors may be operating independently from race/ethnicity. There is overwhelming evidence that people in minority social groups are more likely to come from low-income communities, and hence there are several explanations for racial disparities in stop outcomes.

It is also obvious from consulting the evidence map that bias that is influenced by race is multifaceted. It can vary within racial groups based on skin colour. For instance, Alcalá and Montoya (2018) found that darker skin colour was associated with higher odds of arrest, but only for the second-generation Latino immigrants. It can also vary between different heights; in analysing over one million stops in New York, Hester and Gray (2018) find that tall Black men are more likely than shorter Black men to be the subject of police stops because, they argue, they are perceived to be more of a threat. Using similar data from New York, Kwate and Threadcraft (2015) found that black women were more likely than white women to be labelled heavy in police stop records, even when BMI and other factors were controlled for. There are also cultural markers that police may use as cognitive shortcuts – for example Dabney et al. (2017) found that appearance associated with contemporary hip-hop culture (i.e., dreadlocks, gold teeth, saggy pants) predicted more severe police outcomes in a largely African American metropolitan jurisdiction. Similarly, in simulated shooting tasks, Ma and Correll (2011) found that although police showed no racial bias on average, target 'prototypicality' (especially among Whites) significantly influenced shooting judgments.

Clearly perceptions of race are multifaceted. They are also socially constructed by citizens as well as police officers. An individual may self-identify as one race but be perceived to be another by a police officer. Penner and Saperstein (2015) dissect this in their study on differences in arrest rates using data from the U.S. National Longitudinal Study of Adolescent Health. They found that the likelihood of arrest was significantly higher for people who were classified by others as Black, even if they themselves did not identify as Black. Conversely, there were no statistical differences between people who self-identified as Black, but who were not seen by others as being Black. The UK Home Office statistical report on stops and searches recommended using self-identified race/ethnicity as this more closely approximates census data (if that is being used in analysis), however this may not represent the information that is being used in police decision-making.

4.3. Limitations

As we note elsewhere in this report, at the time of our search, the Global Policing Database only contained records from 2000-2018. With the exception of the two recent meta-analyses we identified (Bolger & Lytle, 2018; Lapsey et al., 2021), we therefore do not represent any literature that falls outside this timeframe. Of course, evidence syntheses have short shelf-lives because research is continually being published. However, given the trajectory with which research is being published on police bias, the evidence map may exclude a number of valuable recent studies.

Due to how we developed our inclusion criteria, we excluded studies that looked at perceptions of bias (which can fall on a wildly accurate to inaccurate continuum) when they did not define the police action and/or include a comparison group. In other words, if a marginalised population were interviewed but their views were not compared to a majority group the study was excluded. Without a comparator, it is unclear to what extent experiences of the interviewees may differ from other groups, so this was a necessary decision. However, it does mean that the evidence map does not benefit from the insight of a lot of studies that captured lived experiences—for example, Boppre's (2018) dissertation on women's experiences in the Oregon Criminal Justice System, which was illuminating but did not separate out the police actions referred to by interviewees. There is far more evidence on the negative impact of inequitable police treatment from lived experience than is currently represented on the evidence map.

An important limitation is that due to time constraints, we could not perform a systematic appraisal of the design quality of the studies in the evidence map. This means we are not suitably positioned to make judgements on the strength or reliability of the findings presented in this report. Instead, we leave those judgements up to researchers who may make use of the evidence map to form their own literature reviews and conclusions. Systematic, quantifiable evaluations of quality can feasibly be conducted on smaller pools of studies, without requiring the very considerable resources such an appraisal would have required here.

4.4. Recommendations for future research

Many 'solutions' have been proposed to reduce the disparities in police outcomes that are pervasively found across the literature. Sometimes these solutions appear to have face-validity (they seem intuitively appealing). For example, in the first decade of the millennium there were calls in the UK and some parts of the US to recruit more ethnic minority officers, so that they might better represent the communities they served. It sounds like a good idea, but does it reduce racial disparities? The short answer is no (see Brown and Frank, 2006; Donohue and Levitt, 2001; Gilliard-Matthews, 2017; Helfers, 2016). The longer answer is it is more complicated than it first seems. Officers are inculcated into the police culture. The strength and depth of that culture, through the structure of the organisation, is unlikely to afford minority officers the power to make the structural changes that are perhaps required in the institution. Interestingly however, the proportion of women in a Canadian police agency did correlate with lower levels of police killings (Carmichael and Kent, 2015).

Another popular idea posed in recent years is unconscious bias training. Again, it sounds plausible that being alert to your own biases would enable you to stop acting unconsciously and engage in reflection. However, evidence is emerging that it does not work in the intended way to reduce disparities in street-level police behaviours (e.g., see Miller et al., 2020), and may even create unhelpful counterbiases. So why do these seemingly good ideas fail to enact the desired change? We would argue it is because the mechanisms are not well understood – and have not been considered in relation to whether they are activated, neutralised, or even backfire in different contexts. In fact,

these interventions often lack the necessary intervention logic that links them to theory and evidence of mechanism. So, when they fail, it isn't clear why.

What we have hopefully demonstrated across this section is that doing research on the topic of police bias is complex and convoluted and should not be undertaken without a thorough consideration of all the possible mechanisms that could be operating to cause disparities in police outcomes. However, surfacing mechanisms from the messy social world is difficult. Here we offer some suggestions on how to approach the design of future scientific research in this area.

To understand what mechanisms are causing a crime problem, Crime Scientists will typically break down a problem into very narrowly defined crimes with common features. That is, instead of studying (say) burglary, they will do some preliminary analysis that breaks the problem down into its constituent parts (for example, burglaries using keys obtained from breaking into residents' cars). This helps to reveal the 'opportunity structure' – those aspects of the social, natural or built environment that might be enabling the crime to occur. They may also do this sort of breakdown in specific neighbourhoods or environments.

The same approach might fruitfully be adopted in this area of police research. So that, rather than looking at all stop and searches, it might prove useful to consider what are meaningful ways to break data into the different subsets that make up the entire set of stops. It may be that there is a high volume of stops in 'hotspot' areas. Or it might be thought that police officers make decisions differently when the perceived threat level from citizens is elevated. Or it might be that particular units, because of the expectancy placed on them by senior officers or external agencies, will be motivated to be productive. There are many conceivable subsets. Disaggregating the bulk of the police activity down in this way enables the 'opportunity structure' for stops to be better articulated. From this, officers can be invited to propose their working theories on what might be going on. In our experience, the people making the decisions on a regular basis are valuable sources of practical knowledge. These might be similar or different from the theories identified in section 2.6. From these theoretical assumptions, hypotheses can be created to guide what data need to be collected and what methods are most appropriate to test the hypothesis.

We have been surprised by the lack of studies on the evidence map that investigate decision-making *as a process*. Policing is complex. Decisions are not unvarying. They depend on all sorts of factors. It seems to us that there is tremendous scope in examining how decision-making is done from the perspective of those making the decisions. Giving police officers a voice to be heard. And that can then be triangulated with other evidence, that may come from systematic social observation or simulation exercises. Of course, police data can be used to study disparities in police outcomes, but these studies should not be done blind to the issues with the sources of bias that are likely within such data.

With respect to the methods used in future research, these will very much depend on the research question or hypotheses posed, since that is how methods should be selected. However, on the back of section 2.7.2. we recommend that the population available to be targeted for the police action being studied be carefully considered for benchmarking purposes. Minority populations are typically younger and more economically disadvantaged than the general resident population. Going hand in hand with this is that people with a lack of private space in their accommodation will spend more time on the street with likeminded people, or may spend more time driving. These are all issues that increase their risk of being scrutinised (rightly or wrongly) by police. Similarly, rival explanations for disparities in police outcomes need to be thoughtfully incorporated into the method used, and are particularly important in future evidence syntheses. Control variables are generally fitted to the specific police action being studied, and therefore are not universal (although using structural measures such as % of population living in poverty or % population in minority group are common).

For example, when examining whether there is disparity in use of force against people displaying mentally disordered behaviour, you would need to control for other factors such as whether the suspect resisted an arrest or were intoxicated, as these can predict use of force and may present at the same time as mentally disordered behaviour and thus muddy the waters.

Economic models, that explicitly lay out their assumptions about behavioural decision-making or hierarchical linear models (also known as multilevel models) are strong in controlling for ecological/situational effects that may be influencing empirical patterns in police outcomes. Theoretically motivated regression models that control for rival explanations are also suitably robust.

It is our view that ecological approaches are the future of police bias research. Ecological approaches are gaining traction in this area of research for good reason. So many situational features of a scene (e.g., time of day, group dynamics, community characteristics) may influence decision-making, it seems foolish not to account for these. And ecologically oriented studies help us to determine if place-based targeting of police resources (e.g., hotspots policing) are exacerbating inequities for minority groups[16]. In Hollis and Jenning's (2018) evidence synthesis on police use of force they emphasise that studies that fail to take the local context into consideration, fail to advance the evidence base.

4.5. Conclusion

To conclude: without theory researchers cannot go beyond describing the *what* to begin talking about the *why* and *how*. Research examining police bias needs to be context-specific to both a particular policing environment and the police action being studied. And multivariate methods need to be used to understand how decision-making works as a process, to reveal what mechanisms are causing the disparities. Finally, research needs to control for rival explanations (e.g., differential offending) to ensure that the findings are reliable. In combination, these principles offer the promise of linking theory and research through intervention logic, to effective ways of reducing disparity.

[16] On this topic see Barnes (2018) who argues that hotspots policing creates racial disparities in traffic stop data. See also Weisburd's (2016) counterargument as to why this is not inevitable. Briggs & Keimig (2017) also discuss this in relation to targeted vehicle stops.

5. References

Alcalá, H. E., & Montoya, M. F. (2018). Association of skin color and generation on arrests among Mexican-origin Latinos. *Race and Justice*, 8(2), 178-193.

Alpert, G. P., MacDonald, J. M., & Dunham, R. G. (2005). Police suspicion and discretionary decision making during citizen stops. *Criminology*, 43(2), 407-434.

Alpert, G. P., Dunham, R. G., & Smith, M. R. (2007). Investigating racial profiling by the Miami-Dade Police Department: A multimethod approach. *Criminology & public policy*, 6(1), 25-55.

Alvarado, E. J. (2016). *Racial profiling and traffic search: A meta-analysis* (Doctoral dissertation, San Diego State University).

Bachman, R., Zaykowski, H., Lanier, C., Poteyeva, M., & Kallmyer, R. (2010). Estimating the magnitude of rape and sexual assault against American Indian and Alaska Native (AIAN) women. *Australian & New Zealand Journal of Criminology*, 43(2), 199-222.

Barnes, S. F. (2018). *Police-Community Relations: A Study of Racial Disparity and the Effects of Hot Spots Policing Leadership Strategies* (Doctoral dissertation, North Carolina Agricultural and Technical State University).

Beckett, K., Nyrop, K., & Pfingst, L. (2006). Race, drugs, and policing: Understanding disparities in drug delivery arrests. *Criminology*, 44(1), 105-137.

Bernard, T.J. and R.R. Ritti (1990). The Philadelphia Birth Cohort and Selective Incapacitation. *Journal of Research in Crime and Delinquency.* 28(1): 33-54.

Blalock, H. (1967). *Toward a Theory of Minority-Group Relations*. New York: Capricorn Books.

Bolger, P. C. (2015). Just following orders: A meta-analysis of the correlates of American police officer use of force decisions. *American Journal of Criminal Justice*, 40(3), 466-492.

Bolger, P. C., & Lytle, D. (2018). A meta-analysis of suspect demographic characteristics and American police officer search decisions. *Criminology, Crim. Just. L & Soc'y*, 19, 1.

Boppre, B. L. (2018). *Intersections Between Gender, Race, and Justice-Involvement: A Mixed Methods Analysis of Women's Experiences in the Oregon Criminal Justice System.* PhD thesis. University of Nevada, Las Vegas.

Bowling, B., Phillips, C., Campbell, A., Docking, M., & UN Research Institute for Social Development. (2004). *Policing and human rights: Eliminating discrimination, xenophobia, intolerance and the abuse of power from police work*. Geneva: UN Research Institute for Social Development.

Black, D. J. (1976). *The behavior of law*. New York: Academic Press.

Bricker, T. E. (2002). *The enforcement of traffic offenses by the police: Exploring the issue of racial profiling.* [Doctoral thesis, Michigan State University]. http://search.proquest.com.libraryproxy.griffith.edu.au/docview/252292

Briggs, S. J., & Keimig, K. A. (2017). The impact of police deployment on racial disparities in discretionary searches. *Race and Justice*, 7(3), 256-275.

Brown, R. A., & Frank, J. (2006). Race and officer decision making: Examining differences in arrest outcomes between black and white officers. *Justice quarterly*, 23(1), 96-126.

Brunson, R. K. (2007). “Police don't like black people”: African-American young men's accumulated police experiences. *Criminology & public policy*, 6(1), 71-101.

Annie E. Casey Foundation. (2003). *Kids Count Data Book: State Profiles of Child Well-Being*. Baltimore: Author.

Carmichael, J. T., & Kent, S. L. (2015). The use of lethal force by Canadian police officers: Assessing the influence of female police officers and minority threat explanations on police shootings across large cities. *American Journal of Criminal Justice*, 40(4), 703-721.

Crutchfield, R. D., Skinner, M. L., Haggerty, K. P., McGlynn, A., & Catalano, R. F. (2012). Racial disparity in police contacts. *Race and Justice*, 2(3), 179-202.

Dabney, D. A., Teasdale, B., Ishoy, G. A., Gann, T., & Berry, B. (2017). Policing in a largely minority jurisdiction: The influence of appearance characteristics associated with contemporary hip-hop culture on police decision-making. *Justice Quarterly*, 34(7), 1310-1338.

Desai, R. A., Falzer, P. R., Chapman, J., & Borum, R. (2012). Mental illness, violence risk, and race in juvenile detention: Implications for disproportionate minority contact. *American Journal of Orthopsychiatry*, 82(1), 32.

Donohue III, J. J., & Levitt, S. D. (2001). The impact of race on policing and arrests. *The Journal of Law and Economics*, 44(2), 367-394.

Dunham, R. G., Alpert, G. P., Stroshine, M. S., & Bennett, K. (2005). Transforming citizens into suspects: Factors that influence the formation of police suspicion. *Police Quarterly, 8*(3), 366-393.

Eberhardt, J. L. (2016). *Strategies for change: Research initiatives and recommendations to improve police-community relations in Oakland, Calif*. Stanford University, CA: Stanford SPARQ

Engel, R. S., Tillyer, R., Klahm IV, C. F., & Frank, J. (2012). From the officer's perspective: A multilevel examination of citizens' demeanor during traffic stops. *Justice Quarterly*, 29(5), 650-683.

Engel, R. S., Calnon, J. M., & Bernard, T. J. (2002). Theory and racial profiling: Shortcomings and future directions in research. *Justice Quarterly*, 19(2), 249-273.

Feinstein, R. (2015). A qualitative analysis of police interactions and disproportionate minority contact. *Journal of Ethnicity in Criminal Justice*, 13(2), 159-178.

Fitzgerald, R. T., & Carrington, P. J. (2011). Disproportionate minority contact in Canada: Police and visible minority youth. *Canadian Journal of Criminology and Criminal Justice*, 53(4), 449-486.

Fridell, L. (2005). *By the numbers: A guide for analyzing race data from vehicle stops*. Washington, DC: US Department of Justice.

Gilliard-Matthews, S. (2017). Intersectional race effects on citizen-reported traffic ticket decisions by police in 1999 and 2008. *Race and Justice*, 7(4), 299-324.

Grogger, J., & Ridgeway, G. (2006) Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association,* 101(475),878-887. https://doi.org/10.1198/016214506000000168

Heenan, M., & Murray, S. (2006). Study of reported rapes in Victoria 2000–2003: Summary research report. Retrieved from the State of Victoria (Australia), Department of Human Services: http://www. dhs. vic. gov. au/__data/assets/pdf_file/0004/644152/StudyofReportedRapes.Pdf.

Helfers, R. C. (2016). Ethnic disparities in the issuance of multiple traffic citations to motorists in a southern suburban police agency. *Journal of Ethnicity in Criminal Justice*, 14(3), 213-229.

Hester, N., & Gray, K. (2018). For Black men, being tall increases threat stereotyping and police stops. *Proceedings of the National Academy of Sciences*, 115(11), 2711-2715.

Higgins, G. E., Vito, G. F., & Grossi, E. L. (2012). The impact of race on the police decision to search during a traffic stop: A focal concerns theory perspective. *Journal of Contemporary Criminal Justice*, 28(2), 166-183.

Hitchens, B. K., Carr, P. J., & Clampet-Lundquist, S. (2018). The context for legal cynicism: Urban young women's experiences with policing in low-income, high-crime neighborhoods. *Race and Justice*, 8(1), 27-50.

Hollis, M. E., & Jennings, W. G. (2018). Racial disparities in police use-of-force: a state-of-the-art review. *Policing: An International Journal, 41(2), 178-193.*

Horrace, W. C., & Rohlin, S. M. (2016). How dark is dark? Bright lights, big city, racial profiling. *Review of Economics and Statistics*, 98(2), 226-232.

Jetelina, K. K., Jennings, W. G., Bishopp, S. A., Piquero, A. R., & Reingle Gonzalez, J. M. (2017). Dissecting the complexities of the relationship between police officer–civilian race/ethnicity dyads and less-than-lethal use of force. *American journal of public health*, 107(7), 1164-1170.

Kaldenbach, H., 2011. *Act Normal, 99 tips for dealing with the Dutch*. Amsterdam: Prometheus.

Klinger, D. (1997). Negotiating order in patrol work: An ecological theory of police response to deviance. *Criminology*, 35(2), 277–306.

Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1), 203-229.

Kochel, T. R., Wilson, D. B., & Mastrofski, S. D. (2011). Effect of suspect race on officers' arrest decisions. *Criminology*, 49(2), 473-512.

Kraska, P.B. (2004). *Theorizing criminal justice: Eight essential orientations*. Long Grove, IL: Waveland.

Kwate, N. O. A., & Threadcraft, S. (2015). Perceiving the Black female body: Race and gender in police constructions of body weight. *Race and social problems*, 7(3), 213-226.

Lapsey Jr, D. S., Campbell, B. A., & Plumlee, B. T. (2021). Focal concerns and police decision making in sexual assault cases: a systematic review and meta-analysis. *Trauma, Violence, & Abuse*, doi: 1524838021991285.

Liska, A., & Chamlin, M. (1984). Social structure and crime control among macrosocial units. *American Journal of Sociology*, 90(2) 383–395.

Levchak, P. J. (2017). Do precinct characteristics influence stop-and-frisk in New York City? A multi-level analysis of post-stop outcomes. *Justice Quarterly*, 34(3), 377-406.

Liederbach, J., Trulson, C. R., Fritsch, E. J., Caeti, T. J., & Taylor, R. W. (2007). Racial profiling and the political demand for data: A pilot study designed to improve methodologies in Texas. *Criminal Justice Review*, 32(2), 101-120.

Lum, C. (2011). The influence of places on police decision pathways: From call for service to arrest. *Justice Quarterly*, 28(4), 631-665.

Lundman, R. J. (2010). Are police-reported driving while Black data a valid indicator of the race and ethnicity of the traffic law violators police stop? A negative answer with minor qualifications. *Journal of Criminal Justice*, 38(1), 77-87.

Lytle, D. J. (2014). The effects of suspect characteristics on arrest: A meta-analysis. *Journal of Criminal Justice*, 42(6), 589-597.

Ma, D. S., & Correll, J. (2011). Target prototypicality moderates racial bias in the decision to shoot. *Journal of Experimental Social Psychology*, 47(2), 391-396.

Macpherson, W. (1999). *The Stephen Lawrence inquiry* (Vol. 1). London: Stationery Office Limited.

McCarthy, M., Porter, L., Townsley, M., & Alpert, G. (2019). Influence of community characteristics on serious police use of force events in an Australian policing jurisdiction: a test of minority threat, social disorganisation, and ecological contamination theories. *Policing and society*, 29(9), 1091-1108.

Meehan, A. J., & Ponder, M. (2002). How roadway composition matters in analyzing police data on racial profiling. *Police Quarterly*, 5(3), 306-333.

Mekawi, Y., & Bresin, K. (2015). Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology, 61*, 120-130.

Meng, Y. (2014). Racially biased policing and neighborhood characteristics: A Case Study in Toronto, Canada. Cybergeo: *European Journal of Geography*.

Miller, J. (2010). Stop and search in England: A reformed tactic or business as usual? *The British Journal of Criminology*, 50(5), 954-974.

Miller, J., & Gounev, P. (2007). *"I Can Stop and Search Whoever I Want": Police Stops of Ethnic Minorities in Bulgaria, Hungary, and Spain.* Open Society Institute.

Miller, J., Quinton, P., Alexandrou, B., & Packham, D. (2020). Can police training reduce ethnic/racial disparities in stop and search? Evidence from a multisite UK trial. *Criminology & Public Policy*, 19(4), 1259-1287.

O'Brien, A. J., McKenna, B. G., Thom, K., Diesfeld, K., & Simpson, A. I. (2011). Use of Tasers on people with mental illness: A New Zealand database study. *International journal of law and psychiatry*, 34(1), 39-43.

Owusu-Bempah, A. (2014). *Black males' perceptions of and experiences with the police in Toronto.* University of Toronto (Canada).

Penner, A. M., & Saperstein, A. (2015). Disentangling the effects of racial self-identification and classification by others: the case of arrest. *Demography*, 52(3), 1017-1024.

Piquero, A. R. (2008) *Disproportionate minority contact*. Future Child 18:59–79

Reid-Howie Associates. (2002). *Police stop and search among white and minority ethnic young people in Scotland.* Stationery Office.

Rosenfeld, R., Rojek, J., & Decker, S. (2012). Age matters: Race differences in police searches of young and older male drivers. *Journal of research in crime and delinquency*, 49(1), 31-55.

Schlosberg, M. (2002). *A department in denial: The San Francisco Police Department's failure to address racial profiling*. American Civil Liberties Union Foundation of Northern California.

Skolnick, J.H. (1994). *Justice without trial: Law enforcement in democratic society*, 3rd edition. New York: Macmillan College Publishing Company.

Smith, M. R., & Alpert, G. P. (2007). Explaining police bias: A theory of social conditioning and illusory correlation. *Criminal justice and behavior*, 34(10), 1262-1283.

Snowball, L., & Snowball, L. (2008). *Juvenile diversion and Indigenous offenders: a study examining juvenile offenders in Western Australia, South Australia and New South Wales*. Criminology Research Council.

Steele, S. M., Collier, M., & Sumerau, J. E. (2018). Lesbian, gay, and bisexual contact with police in Chicago: Disparities across sexuality, race, and socioeconomic status. *Social Currents*, 5(4), 328-349

Svensson, J. S., & Saharso, S. (2015). Proactive policing and equal treatment of ethnic-minority youths. *Policing and society*, 25(4), 393-408.

Todak, N. (2017). *De-escalation in police-citizen encounters: A mixed methods study of a misunderstood policing strategy* (Doctoral dissertation, Arizona State University).

Tompson, L., & Belur, J. (2016). Information retrieval in systematic reviews: a case study of the crime prevention literature. *Journal of Experimental Criminology*, 12(2), 187-207.

Tyler, T., & Fagan, J. (2006). Legitimacy and cooperation: Why do people help the police fight crime in their communities? (Paper no. 06Á99). New York: Columbia Law School. *Public Law & Legal Theory Working Paper Group.*

Vrij, A., & Winkel, F. (1992). Cross-cultural police-citizen interactions: the influence of race, beliefs, and nonverbal communication on impression formation. *Journal of Applied Social Psychology*, 22(19), 1546-1559.

Walsh, T. (2017). Public nuisance, race and gender. *Griffith Law Review*, 26(3), 334-354.

Weisburd, D. (2016). Does hot spots policing inevitably lead to unfair and abusive police practices, or can we maximize both fairness and effectiveness in the new proactive policing. *University of Chicago Legal Forum*, 16.

Willits, D. W., & Makin, D. A. (2018). Show me what happened: Analyzing use of force through analysis of body-worn camera footage. *Journal of Research in Crime and Delinquency*, 55(1), 51-77.

Wilson, J.Q. (1968), *Varieties of Police Behavior: The Management of Law and Order in Eight Communities*, Atheneum, New York, NY.

Winkel, F.W., Koppelaar, L., & Vrij, A. (1998). Creating suspects in police-citizen encounters: Two studies on personal space and being suspect. *Social Behaviour*, 3, 307-318.