What Works and equality?

Michael Sanders, Ella Whelan, Louise Jones, Rosie Campbell & Omar Khan   September 2021

Introduction

8.9%
gender pay gap for full-time workers

45%
of earnings inequalities are passed generationally

12.9%
unemployment rate for ethnic minorities

Inequality is a blight on society, and despite decades of activism and campaigning, significant gaps remain between groups in a number of important areas. Changes in the law have recognised the fundamental equality of different groups, making discrimination against people on grounds of race, sex, sexual orientation, or gender identity illegal, and the state has gradually removed its own explicit discrimination, for example through the equalisation of marriage rights.

However, inequality continues; both as a legacy of previous evils, and the consequence of current structural and individual factors.

The gender pay gap among full-time employees now stands at 8.9 per cent, and the gender pay gap is higher for all employees than full-time employees and part-time employees separately. This is because women fill more part-time jobs, which have lower hourly median pay than full-time jobs, and are more likely to be in lower-paid occupations. Friedman et al (2017) found that 45 per cent of earnings inequalities are passed across generations.

"Inequality continues; both as a legacy of previous evils, and the consequence of current structural and individual factors"

Unemployment rates are also significantly higher for people from ethnic minory backgrounds at 12.9 per cent compared with 6.3 per cent for White people. Just 6 per cent of Black school leavers attend a Russell Group university, compared with 12 per cent of mixed ethnicity and Asian school leavers and 11 per cent of White school leavers. Black Caribbean and Mixed White/Black Caribbean children have rates

of permanent exclusion about three times that of the pupil population as a whole.

The arc of history may bend towards equality, but that bending requires the expenditure of energy, intellect and money by society, policymakers, and individual allies to its cause.
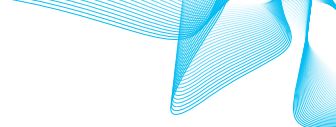
A review by The Child Safeguarding Practice Review Panel found that boys from black and minority ethnic backgrounds appear to be more vulnerable to harm from criminal exploitation. Children born into high deprivation local authorities had a greater chance of entering care in their first week than children born in less deprived authorities.

Reducing inequality, and ultimately removing the relationship between a person's characteristics and their outcomes, cannot be taken for granted as something that will solve itself given enough time. The arc of history may bend towards equality, but that bending requires the expenditure of energy, intellect and money by society, policymakers, and individual allies to its cause.

The goal of reducing inequality is shared by many across the UK's What Works movement. In some cases, this is made more explicit; like in the Education Endowment Foundation's focus on reducing attainment gaps between poorer and richer students – those receiving free school meals and their peers – and in others it is more implicit, but nonetheless real.

For all this desire to narrow gaps in outcomes, the What Works Centres have spent little of the substantial government investment in evidence-based policy into work that aims to reduce gaps across important dimensions of

inequality – across racial, gender, sexual orientation or gender identity lines.

> "To ignore inequality is to ignore an important structural component of the problems that What Works Centres try to solve"

This lack of focus on equalities has a few implications. First, to ignore inequality is to ignore an important structural component of the problems that What Works Centres try to solve. Crime, mental health and education do not exist in a vacuum, and many have their roots in inequality between the groups we describe. Fighting negative outcomes without considering these causes likely reduces the impact of the interventions considered, and limits us to treating symptoms, not diseases.

Second, there is a risk that those that are not counted do not count. This itself might take one of two forms. One, if people are perceived to be "hard to reach" and hence do not get included in trials or other research programmes in the first place, average effects reported for these interventions either don't capture the impacts on these people or group, meaning that policy recommendations cannot take into account these groups and their needs – something that is effective for those included in the sample might be ineffective for those excluded, or vice versa.

The other risk is that, even if people are included in trials, if we do not look at their data by itself, we run the risk of exacerbating inequalities. The lack of this data analysis means that it is difficult to be confident whether there are differences in effects of interventions – or the direction that these differences might point in – based on people's characteristics or identities. The small sample sizes in this research may also cause researchers to avoid conducting this kind of analysis. However, there are good theoretical reasons to suspect that people struggling with alienation,

the colonisation of their educational experience, or other challenges associated with or caused by inequalities, might respond differently to interventions than those of the majority, or privileged, group. There is also some data to support this idea around economic inequality, with 26 of 63 evaluations of interventions funded by the Education Endowment Foundation reviewed by Sanders et al (2019), found to be "gap-wideners" – those that increase the attainment gap between richer and poorer students.
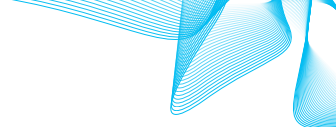
Moreover, when we do not think about inequality, in all its various forms, as part of the work of the What Works movement, we risk excluding voices who need to be heard, and creating a sense among women, people from ethnic minority backgrounds, LGBTQ+ people, and others that our movement is "not for people like me". This alienation means that we likely fail to grapple with issues of equality from a standpoint informed by lived experiences, and that our organisations are in a general sense weaker.

....

"By not taking inequalities seriously... we support the argument that there is something inherently, White, masculine, straight and cis gendered about quantitative research"

....

By not taking inequalities seriously as quantitative, causal researchers, we quit the field. By giving too much credence to arguments about erasure of minoritised voices, we support the argument that there is something inherently, White, masculine, straight and cis gendered about quantitative research, and hence, that the only way to conduct research credibly with other groups is to do so qualitatively. We believe that to do so is to do a great disservice to disadvantaged and minoritised groups and that the struggle for equality must draw on all tools and methods available to it.

In this paper, we'll discuss three key areas for What Works and equality: reducing gaps in success, including **improving attainment**; **discrimination**; and **tackling structural or systemic inequalities**. We'll argue that reducing inequalities with a What Works methodology requires a cross-cutting programme or initiative (or several), rather than a Centre of its own, and suggest next steps.

What is a What Works Centre?

A What Works Centre is established to use, share, and create robust evidence to improve a policy area or social issue. There are currently 13 What Works Centres, affiliates and associates covering policy areas across all aspects of life. The What Works Network was launched by the UK Government in 2013 to ensure that the best available evidence on "what works" is available to the people who make decisions about public services. The members of the network are listed below.

FIGURE 1: THE MEMBERS OF THE WHAT WORKS CENTRE NETWORK

What Works Centre	Policy area
National Institute for Health and Care Excellence (NICE)	Health and social care
Education Endowment Foundation (EEF)	Educational achievement
College of Policing What Works Centre for Crime Reduction	Crime reduction
Early Intervention Foundation	Early intervention
What Works Centre for Local Economic Growth (hosted by LSE, Arup, Centre for Cities)	Local economic growth
Centre for Ageing Better	Improved quality of life for older people
What Works Centre for Wellbeing	Wellbeing
Centre for Homelessness Impact	Homelessness
What Works for Children's Social Care	Children's social care
Affiliate: Youth Endowment Fund	Youth offending
Affiliate: Youth Futures Foundation	Youth employment
Affiliate: Centre for Transforming Access and Student Outcomes in Higher Education	Higher Education
Associate: Wales Centre for Public Policy	General

Interventions to reduce gaps

The problem

As we've seen, inequality manifests itself in numerous ways. Women, on average, do better in education than men up until at least age 18, but then gaps in university attendance, employment, and post qualifying outcomes emerge over time. Some of this is driven by both childbirth and differential caring responsibilities for both children and elderly relatives, but this inequality emerges from both societal and policy levers set to reinforce these gaps.

....

"Women, on average, do better in education than men up until at least age 18, but then gaps in university attendance, employment, and post qualifying outcomes emerge over time"

....

Educational attainment and university progression differs substantially by race, with many minoritised ethnic groups achieving better grades than their White peers, while others, including, children from Gypsy/Roma, Black Caribbean or Pakistani backgrounds, typically do worse. These gaps again widen over time and are manifest in the labour market. For example, just 6 per cent of Black school leavers attended a Russell Group university, compared with 12 per cent of mixed ethnicity and Asian school leavers and 11 per cent of White school leavers.

Of course, inequalities exist in different areas as well, including in health. For example, there is a significant disproportionate number of people from ethnic minority

backgrounds detained under mental health legislation in hospitals in England and Wales. Furthermore, Black African women are seven times more likely to be detained than White British women. Those from Gypsy, Roma, or Traveller backgrounds were found to suffer poorer mental health than the rest of the population in Britain and they were also more likely to suffer from anxiety and depression.

Another prominent, and relevant, example of health inequality is young people experiencing mental health challenges. People who identify as LGBTQ+, and particularly trans men and women, are more likely to experience these kinds of challenges than their peers. Indeed, Stonewall found in 2017 that 52 per cent of those who identify as LGBTQ+ had experienced depression.
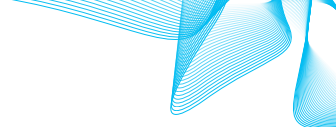
Mental health inequalities But are often linked with wider cultural and societal systems of disadvantage which impact a person's wellbeing, including (but not limited to) stigma, discrimination and environment – including housing security. Although all mental illness requires treatment, we should be particularly concerned that specific groups are LGdisproportionately experiencing harm.

In these cases, we are concerned with narrowing gaps between groups: we are trying to make the outcomes of one group more closely resemble another.

Methodological challenges

There are, naturally, methodological challenges when considering these issues. These are twofold and push us in opposite directions when considering how to improve our practice and our research.

The first issue is around the division of samples into smaller groups for subgroup analysis. Division into smaller groups carries risk both to statistical power – the likelihood of us

detecting an effect statistically, if it exists –, and to our false discovery rate.

Statistical power is determined by a number of factors in an experiment, but the main one that we are concerned with here is sample size. The smaller the sample we are looking at, the larger an effect we are able to detect statistically. This means that, for a given burden of proof –typically set at a 5 per cent probability of an observed effect occurring if there were to be no effect – we are less likely to detect an effect, the smaller our sample size.

This is also true of the smallness of the size of a subgroup to be analysed. If we are interested in the question "Does an intervention, X, have a larger effect on group Y than group Z?", the smaller groups Y and Z are, the harder the question is to answer reliably.

Moreover, recent developments in the analysis of randomised trials work against us here. The ongoing replication crisis in social science is in part caused by too much subgroup analysis, with researchers able to run as many tests as they wanted, on as many outcomes, subgroups, and combinations of these two, as they wanted, until they found a combination that produced a statistically significant result. Because the threshold for significance (5 per cent probability), implies that if we run 20 tests, we'll find *something*, and for years, this is exactly what people were doing.

In response, a new set of conventions have emerged that require researchers to pre-specify the analysis they're going to undertake and penalise running a large number of tests. There are a few ways of doing this, but at its simplest, it involves upping the threshold for success with each additional test so that the more tests you run, the harder it is to pass any of them. Taken together, these penalties and the realities of statistical power make finding effects for subgroups much harder.

Pushing in the opposite direction is the second issue. Statistical analysis fundamentally flattens the experiences of individuals in a sample in favour of the mean. This flattening has its uses – it's how we're able to use randomised controlled trials (RCTs) and quasi-experimental designs to ascertain causality – but it is problematic. As we've seen above, this approach favours running as few analyses as possible, on the largest groups you can find.
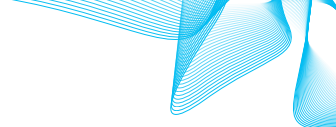
This means that, even when you are going to conduct subgroup analysis, statistical rigour steers you towards the kind of reductive thinking that, societally, we increasingly reject; grouping together large numbers of heterogeneous people under a single (data) label. White vs BAME, Cis-gendered and straight vs LGBTQAI+; these are the comparisons that make the most statistical sense, allowing few analyses on large(r) groups.

"Statistical rigour steers you towards the kind of reductive thinking that, societally, we increasingly reject; grouping together large numbers of heterogeneous people under a single (data) label"

But look beyond the statistics and these approaches have two further consequences. First, they erase the differences between groups, allowing no distinction between Afro-Caribbean and Black African participants, for example, and the circumstances, cultures and communities that contribute to their lives and the way they respond to intervention. By treating lesbians the same as gay men, not only are the important differences between these two groups in terms of sexual preference ignored, so too are gender differences. Erasure has a negative moral element to it, but also means we may miss out on important learning. If, for reasons of culture, history, community or anything else, an intervention has a positive effect for lesbians, and an equally sized,

negative effect for gay men, and these two groups are about as prevalent as each other in our data, then analysis that looks at LGBTQ+ participants will conclude that the intervention has no overall effect, whereas in fact it's helpful for some and harmful for others.

Where do we go from here?

....

"The rules of statistics are also not such that they can be easily ignored because they do not suit the way we wish to see the world"

....

These two tensions – between on the one hand trying to establish a nuanced picture of the impacts of interventions on people with different characteristics, and on the other, the constraints of statistical power and the avoidance of multiple comparisons – are seemingly insoluble. It is very different to be a straight trans woman than to be a cis-gendered lesbian, and conflating the two misses something important. The rules of statistics are also not such that they can be easily ignored because they do not suit the way we wish to see the world.

These tensions need not be impossible to overcome if we really care enough to try and tackle them, however.

First, for many of the trials carried out in social policy, the argument about reduced statistical power could well be overblown. This is because many of the trials are what is known as "cluster randomised controlled trials", that is, randomisation does not occur at the level of the individual participant, but instead at a high level. This could be at the level of the school, or the team or directorate at a workplace, or some geographical area, for example a Lower Super Output Area (LSOA) in the UK.

While randomising at a higher level (clustered randomisation), itself affects statistical power negatively, it is often chosen because it makes the most sense in terms of the delivery and implementation of an intervention. To take an example from the workplace, a new training programme might naturally be delivered to everyone in a particular team. Other interventions – like a whole system change that aims to alter the culture – cannot be delivered to some individuals but not others.
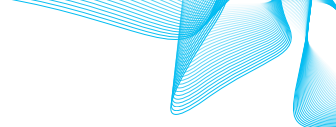
"People interested in equalities should consider oversampling minoritised or disadvantaged groups to further increase statistical power here"

Given that the interventions we're testing force us into the lower-powered world of cluster randomised trials, there is a silver lining. The consequences of cutting the data into subgroups on power are much smaller, meaning that we can test impacts on more diverse groups more rigorously. The figure below shows the rate at which power declines in two studies – one cluster and one not – as we increase the number of subgroups that we're analysing. As we showed in a recent paper, the loss of statistical power is much smaller in the case of clustering, allowing more rigorous subgroup analyses to be conducted. Alongside this methodological quirk, people interested in inequalities should consider oversampling minoritised or disadvantaged groups to further increase statistical power here.

The second approach to ensuring that we're rigorously evaluating impacts on people from different ethnicities, or with different sexual orientations or gender identities, is to make sure that these analyses are not an afterthought. When we decide, after receiving our data, to look for differences in outcomes for, say, British Asians, and find one, then it is fair

to be concerned that this may be the result of what's known as p-hacking – running multiple analyses until you find the one that gives you an interesting result.

If instead, we specify in advance not just that we're going to look for subgroup effects, but which subgroups we are interested in, much of this criticism can be avoided. This is why, at What Works for Children's Social Care, including subgroup analysis by race is now the default setting for studies, and evaluators are able to apply to an additional pot of money to allow this analysis to be conducted.

"The fight for statistical and methodological purity has been a good thing, but to the extent that it holds back progress on equality, it is unjust and should be put to one side"

This leaves only the issue of multi-comparison adjustments eating into the power of each and every test. This cannot be reconciled, but in our view, it must be forgiven. We argue that pre-specified analysis of subgroups, which aims to uncover differential effects for minoritised or otherwise disadvantaged groups, should be excluded from consideration for multiple-comparisons, or at least should be treated separately to other analyses. The fight for statistical and methodological purity has been a good thing, but to the extent that it holds back progress on equality, it is unjust and should be put to one side.
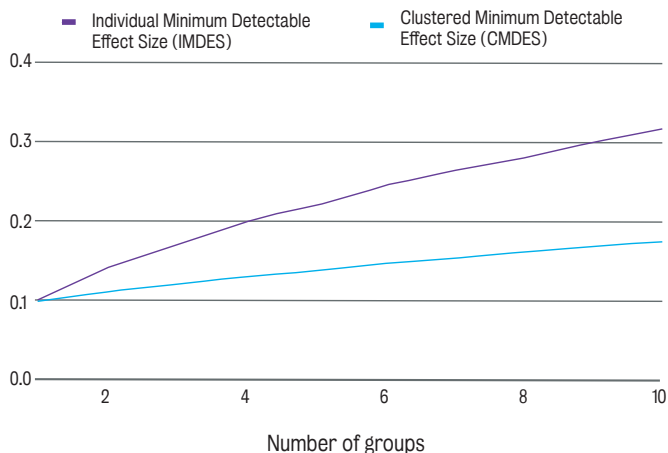
WWLGBTQ+

What Works LGBTQ+, was established in 2020 with the aim of bringing together the expertise that exists across the What Works Network to make the best possible evidence on "what works" for LGBTQ+ communities available to decision-makers.

What Works LGBTQ+ ultimately seeks to:

1. Better understand the **evidence base** around LGBTQ+ issues

2. Support in the efforts to **improve data**, including analysis and linkage

3. Support the e**valuation of interventions** for LGBTQ+ communities

4. Build a **focus on LGBTQ+ issues** into the wider work of What Works Centres

FIGURE 2:
COMPARING A CLUSTER AND NON-CLUSTER STUDY

- Individual Minimum Detectable Effect Size (IMDES)
- Clustered Minimum Detectable Effect Size (CMDES)

Number of groups

Interventions to reduce discrimination

It is against the law to discriminate against anyone because of:

* age
* gender reassignment
* being married or in a civil partnership
* being pregnant or on maternity leave
* disability
* race including colour, nationality, ethnic or national origin
* religion or belief
* sex
* sexual orientation

However, despite this, in 2017: one third of trans people (34 per cent) were discriminated against because of their gender identity when visiting a café, restaurant, bar or nightclub ; rates of prosecution and sentencing for Black people were three times higher than for White people (18 per thousand population compared with six per thousand population for White people); and race hate crimes on Britain's railway networks rose by 37 per cent .

Two in five trans people have experienced a hate crime or incident because of their gender identity, however, four in five anti-LGBT hate crimes and incidents go unreported, with younger LGBT people particularly reluctant to go to the police. Reducing discrimination requires a different approach methodologically, and brings with it different challenges. When attempting to reduce discrimination, tacit or otherwise, the focus changes compared to where we are attempting to reduce gaps in outcomes.

First, if we are looking at reducing discrimination, our sample will mostly be focused on the discriminators rather than the discriminated. As such, our sample is likely to vary between a general population, and some selection on White, cis-gendered, straight men (or some combination of these factors). This means that we are less likely to be interested in subgroup analyses, and more interested in overall effects.
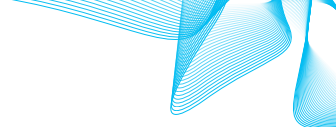
Second, our outcome measures are likely to be different. These might be attitudes, as in for example Josh Kalla and David Broockman's tests of the impacts of interpersonal conversations of people's support for trans rights, immigration, and gay marriage.

Where we are looking at more behavioural outcomes, these may be different to those routinely recorded in administrative datasets. Discrimination is often difficult to measure, or not measured at all, at the level of our participants. Instead, it might be measured in the outcomes of the people they interact with, as in List's (2004) study of discrimination in sports cards markets , or in the outcomes of (fake) applicants to jobs in Bertrand and Mullainathan's (2004) CV audit studies , which have spawned many further studies.

There are a great many interventions that have been developed, or are being actively implemented in the field, to reduce discrimination. These include blinding of participants characteristics when applying for jobs , anti-bullying training to reduce discrimination against LGBTQ+ people in schools, and a host of interventions based on the "contact hypothesis", which states that it is hard to hate a group if we know members of that group .

Methodological challenges

There are different methodological challenges to contend with when considering reducing discrimination, tacit or otherwise.

While larger samples are available if we are seeking to change the behaviour or either anyone, or of the majoritarian group, there are challenges around outcome measure collection. The outcomes that one would wish to measure – such as attitudes to gay marriage – are much harder to collect and isn't a measure that is routinely collected. This leaves trials reliant on survey outcome data collection, which may be subject to either classical attrition – that is, random attrition that effect statistical power but not the causal inference of the study – or non-classical attrition, which may affect both statistical power and our ability to make causal claims.

For example, if we were evaluating the impact of an anti-bullying training programme in a school, the intervention itself might energise participants and make them more likely to complete the endline survey. These "compliers" – those who would complete the survey if treated, but not if not – are likely to have started off less enthusiastic, and hence maybe have worse attitudes to the group we're trying to reduce bullying of. So, including them in the sample only for the treated group might make the intervention look ineffective or, worse, harmful because the intervention is having a positive effect by driving up survey completion.

There is also a risk of social desirability bias – the tendency to respond in ways that we feel are more appropriate or socially acceptable to others – when people have received training that gives them a sense of what the "correct" answer to survey questions are, even if their underlying beliefs and behaviours have not changed. These biases are more common in research on issues that participants find sensitive or controversial, and in situations where there are widely accepted attitudes, behaviors, or norms. Research has also shown that there is a positive association between educational attainment and socially tolerant attitudes.

A similar challenge exists with more concrete behavioural measures. Most discrimination, bullying, or abusive

behaviour is only recorded if someone reports it. The assumption most datasets make is that if abuse hasn't been reported, it doesn't exist. We also know that many crimes and behaviours of this type are systematically under-reported, because people do not feel comfortable or empowered to report them or believe that there will be negative consequences for them of reporting.

....

"Most discrimination, bullying, or abusive behaviour is only recorded if someone reports it. The assumption most datasets make is that if abuse hasn't been reported, it doesn't exist"

....

In these cases, an intervention that reduces the stigma around being a victim or survivor of abuse, for example, might lead to more reporting of abuse, as people are able to be more truthful. This increase in reporting is, we believe, a good thing, as it allows more people's voices to be heard, and uncovers the reality of a problem that is now, still sadly buried. In the context of an impact evaluation, however, if the treatment group feel more empowered and report more harassment, for example, we would say that the intervention is making things better (because it is reducing stigma), but the data would tell us that things are getting worse, because more cases were reported in the treatment group than the control group.

There is also a challenge around attribution either of the cause of the outcome, or the status of the victim. This is because there is often ambiguity around the cause of an action. For example, if a Black man is attacked, it may not be because of racism, even if it often will be. If we are looking for evidence of hate crimes and the 999 operator who records the attack does not interpret as a hate crime, but instead as "just" a mugging, then we might not pick it up.

The status of the victim is also a challenge to analysis, and one that we will return to later.

Where do we go from here?

The extent of the evidence base around how to reduce discrimiation – rather than, as List (2004) and Betrrand and Mullaintahan (2004) both do – measuring it more precisely, is currently limited. There are a hence several things that need to be done at the same time.

First, there is a clear need for more studies, to allow for us to begin to build a better picture of what works in this area. These studies may be uncomfortable for majoritarian groups. Indeed, our standard approach to consenting participants into studies and giving them an accurate sense of what the study is about, may drive those who are most needed away from participating.

There are two implications of this adverse selection, where the people whose behaviour most needs to change do not participate in research on how it should be changed. First, ethics committees need, in our view, to consider the societal benefit of research into reducing discrimination, harassment, abuse, and harm, especially in cases where unlike, for example, the typical medical model, the beneficiary of a treatment is not the treated themselves, but rather those they will interact with later on. Given the distribution of power between participants and those who are the victims of discrimination, it may be that the harm from a lack of informed consent might be outweighed by the benefit to others.

…………………………….

"Given the distribution of power between participants and those who are the victims of discrimination, it may be that the harm from a lack of informed consent might be outweighed by the benefit to others"
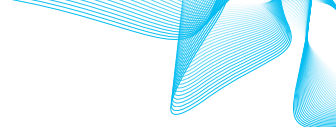
…………………………….

Second, researchers should consider post-study debrief – a comparative rarity in social policy trials – as opposed to baseline informed consent. This would be tricky logistically in some cases, but again, the benefits might exceed the cost.

Next, the issue of attrition, and in particular asymmetric attrition between treatment and control groups, may change the calculus of using attitudinal outcome measures rather than behavioural ones. There is a tendency, particularly in early stage, trials, to make use of survey measures of outcomes that are more proximate to the intervention themselves. So, for example, making use of a highly specific literacy test, rather than Key Stage 2 English results, or using attitudes to homosexuality as opposed to the incidence of homophobic behaviours. If attrition to data collection is so high, or so one-sided, then the calculus of this might change. Yes, less proximate outcomes may be harder to move with an intervention, but if they are collected administratively for the whole sample, even these smaller effects might be easier to *detect*.

Third, there is the issue that spurious negative effects occur because reported incidence – but perhaps not real incidence – rises when stigma is reduced by intervention. This underlines the argument for not just a quantitative approach to RCTs, but that these trials be associated with implementation and process evaluations, which allow us to pick up a more detailed qualitative sense of what is going on.

These spurious negative effects, which are driven by systematic under-reporting, stigma, and potentially biased data inputting, make the opposite case than the issue of attrition. That we should be less reliant, in some instances, on administrative data. For example, reporting of rape and sexual assault is famously low compared to the incidence in reality. To overcome this bias, we need large scale randomly sampled surveys, like the crime survey of England and Wales to give us a clearer picture. Currently, these surveys are used to identify the levels of things, and not the impact

of interventions on them. However, as with gap-narrowing interventions, large scale, cluster randomised controlled trials – with randomisation occurring at the same or similar levels as that at which data is identifiable through these large surveys, eg at school or LSOA level – could allow us to have our cake and eat it. To make use of routinely collected administrative-type data, but to avoid the issues of under-reporting.

Finally, the need for this kind of research, and the complexities around data collection and reliability, the need to conduct in depth implementation and process evaluation, and the likely need to cluster for reasons of both implementation and data integrity, mean that these studies cannot be conducted "on the cheap". While it is possible, through better administrative data access, to drive down the price of randomised trials generally, in matters as sensitive as tackling discrimination – and the intricacies of testing this – proper investment is needed.

GIWL

At the Global Institute for Women's Leadership King's College London, we work towards a world in which women of all backgrounds have fair and equal access to leadership.

Chaired by Julia Gillard, the only woman to have served as Prime Minister of Australia, the Institute brings together rigorous research, practice and advocacy to break down the barriers to women becoming leaders, while challenging ideas of what leadership looks like.

The Institute undertakes a range of activities designed to strengthen:

Research – drawing together existing findings and undertaking new studies.

Practice – using research to deliver evidence-based training and teaching.

Advocacy and engagement – bringing together experts and stakeholders from across the world.

GIWL has four intersecting areas of research to turn all the dials to ramp up the pace of change:

1. **Gender equality at work: building inclusive cultures**

2. **Gender and data: measuring and mapping gender inequalities**

3. **The representation of women in the media**

4. **The impact of women political leaders: from grassroots to world leaders**

Systemic issues

Much has been written about structural and systemic inequalities of many varieties. There is an abundance of these writings, and data on these phenomena; we encourage readers to consume this, and to draw their own conclusions.

Here, we will not argue about the existence or otherwise of systematic inequalities, but instead consider three things.

First, it is important to acknowledge a limitation to the "What Works" methodology in tackling structural or systemic factors. The What Works approach is easier when we are changing something discrete and well defined, and where an intervention – and a problem – has got "edges". What we mean by edges is that we can tell who is treated, and who is not. Assignment to treatment needn't be random – quasi-experimental approaches can do a good job of producing reliable causal estimates of impacts – but we must be able to tell who is who, and who got what, with at least some reliability. For the problem to have edges, we must be able to tell when it has improved for one person or group.

 If structural racism, or discrimination against LGBTQ+ people, or sexism, is like the air around us – ubiquitous but largely invisible – RCTs, or quasi-experimental approaches, cannot usefully evaluate interventions to change it. This does not mean that RCTs have no place in equality research – as we have argued above, we need more, not fewer of them – but it does mean that there are meaningful limitations to the approach we take, and it is important for us to acknowledge that.

Second, as we have touched on in the previous section, inequality shapes our understanding of what research is important and deserves funding. It is a cliche that there is more money and interest in funding research on Viagra than on all female cancers put together, but nonetheless a true

one. Until we take inequalities research seriously, and fund both the research and those who conduct it, the promise of using a What Works approach to reducing inequalities will be unmet. More than this, we must narrow the gulf between the researched and the researchers, by promoting and supporting the careers of people who are from the backgrounds and groups that we hope will benefit from this research.

Third, and most practically in the short term, we must be conscious of the way that the structure of the data we use in our analysis reflects and reinforces inequalities. In her bestselling book, *Invisible Women*, Caroline Criado-Perez draws a clear picture where the prominence of men in data leads to decisions that favour them over women, and perpetuate a world built for men.

There are other issues in this vein. We lack basic data on the lives of LGBTQ+ people, and still less to identify the impact of interventions on this group as a whole or of its constituent parts. While we do not routinely record matters of sexual orientation or gender identity, and while collection of data on race is patchy and unreliable, we are unable to help. Simply trying to gather more data will not be enough, however. We must be sensitive to the fact that many people may be reluctant to hand over their data to government or its associated agencies, for fear of the consequences. The human task of building trust, and the technical task of building databases, must go hand in hand.

Conclusions and recommendations

Although considerable progress has been made towards equalities along gender identity, racial, and sexual orientation lines in recent decades, it is clear that there is still a long way to go.

Given the distance to be travelled between where we are and where we wish to be, it is important that our energies are deployed as efficiently and as effectively as possible. This means that, as we have seen in medicine over the last century, and education during this one, there is value to be gained in systematising our approach.

One element of that systematisation, as we have argued here, is the use of a What Works methodology. This methodology sees us test ideas about how we can reduce inequality and see whether we are making a measurable impact on people's lives. Systematically caring about "what works" to reduce inequality means taking a steady, methodical approach to seeing if each new idea moves us closer to that goal and discarding those that do not.

....

"Interventions can unintentionally increase gaps between groups, much as they can passively decrease them"

....

How do we do this?

As we have seen, interventions can unintentionally increase gaps between groups, much as they can passively decrease them. When setting out to improve grades, it is difficult to imagine that intervention developers in education intended to increase the attainment gap between rich and poor

students; just as surely as none of them intended to develop interventions that were actively harmful to grades.

Given this, when trying to find out what works to reduce gaps in outcomes between majoritarian and minoritised groups, we do not believe that there is a compelling case for a standalone What Works Centre. We would need to replicate the infrastructure of existing Centres – for example to create a What Works Centre for equality in education – to re-analyse or replicate existing trials or others, and to answer questions that could already be answered by existing centres, if the will existed.
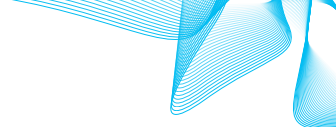
We have also argued that a What Works approach could be taken in reducing discrimination and bias. This might occur by testing, for example, anti-bullying training programmes. On the face of it, these types of interventions could be conducted or evaluated by a "What Works Centre for equality" separate and distinct from the other What Works Centres.

"If there is a case for not setting up a new Centre, there is also a case that the status quo is not working"

Even here though, the argument is overblown. Training interventions, as well as many other things that aim to reduce discrimination, must be conducted within a context, for example, a workplace. Existing What Works Centres operate across the workplaces of the public sector already, through their roles in schools, social care, policing, health, and mental health, and so on. What Works Wellbeing – the What Works Centre focused on various forms of wellbeing – also works across private sector employers. If we want a captive audience, the ability to capture meaningful outcome data, and to encourage reluctant participants into

our studies, the existing What Works Centres already have many of the needed relationships and levers.

If there is a case for not setting up a new Centre, there is also a case that the status quo is not working. The much-needed work on gaps in outcomes and on discrimination is systematically not happening. What, then, do we do?

Initiatives, not centres

The alternative that presents itself to creating a new Centre is to create a convening and coordinating force that cares about inequality.

This is what we have seen with the establishment of the What Works LGBTQ+ initiative, which is housed within the Evidence Quarter – a physical home for evidence-based policy organisations that also aims to incubate and support better evidence in a number of domains – and is supported by King's College London.

WWLGBTQ+ aims to convene and coordinate efforts across What Works Centres to build evidence that builds equality for LGBTQ+ people. This approach has a few appeals when attempting to tackle inequalities. First, it recognises that there is no one "domain" in which inequality exists, and therefore looks to disseminate it across different policy domains. Second, it aims to avoid siloing that can occur within individual domains. If an intervention to reduce implicit bias works with social workers, we might expect it to work with other professions. By having a central initiative with links to each What Works Centre, this knowledge can be shared, and hopefully replicated, more efficiently and effectively.

Next, sitting within these two organisations, as well as being secular across various evidence building operations, means that there are not the additional administrative and operational costs associated with setting up and running a

new Centre. This means that any and all income is devoted purely to the research agenda. This is not an attack on so-called core costs of charities, but a recognition that, in this case, the operational structures that we need to advance this agenda already exist within the extant Centres.

Although we are keen that this initiative is cost-effective, we have also seen the perils of attempts to convene and coordinate joint working across multiple Centres where the topic is of interest to multiple Centres, and to some individuals in particular, but is not anyone's specific job. The creation of an initiative, even with a modest amount of resource allocated, allows time and energy to be marshalled to support this goal.

There is also a history in the What Works network of such initiatives being supported financially, not least by the Economic and Social Research Council's What Works fund, which What Works Centres could bid into for collaborative projects. Using a similar structure in the future could ensure that initiatives around equality could represent good value for money by driving forward work across Centres.

....

"By avoiding the creation of a Centre for equality, we avoid equality becoming 'their problem', and instead recognise that this is everybody's responsibility"

....

Finally, the creation of an initiative rather than a dedicated Centre has the beneficial effect that it prevents creating a silo, or an abdication of responsibility, around matters of equality. By avoiding the creation of a Centre for equality, we avoid equality becoming "their problem", and instead recognise that this is everybody's responsibility.

References

1. Smith, R. (2019). *Gender pay gap in the UK*. [online] Ons.gov.uk. Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/genderpaygapintheuk/2019 [Accessed 1 Jun. 2021].

2. Friedman, S., Laurison, D. and Macmillan, L. (2017). *Social Mobility, the Class Pay Gap and Intergenerational Worklessness: New Insights from The Labour Force Survey*.

3. Equalityhumanrights.com. (2020). *Race report statistics | Equality and Human Rights Commission*. [online] Available at: https://www.equalityhumanrights.com/en/race-report-statistics [Accessed 1 Jun. 2021]

4. The Child Safeguarding Practice Review Panel. (2020). *It was hard to escape: Safeguarding children at risk from criminal exploitation.* [online] . Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/870035/Safeguarding_children_at_risk_from_criminal_exploitation_review.pdf.

5. Bilson, A. and Bywaters, P. (2020.) Born into care: Evidence of a failed state. *Children and Youth Services Review*, *116*, p.105164.

6. Equalityhumanrights.com. (2020). Race report statistics | Equality and Human Rights Commission. [online] Available at: https://www.equalityhumanrights.com/en/race-report-statistics

[Accessed 20 Aug. 2021].

7. Equalityhumanrights.com. (2020). Race report statistics | Equality and Human Rights Commission. [online] Available at: https://www.equalityhumanrights.com/en/race-report-statistics [Accessed 20 Aug. 2021].

8. Ibid.

9. Stonewall. (2018). LGBT in Britain - Health. [online] Available at: https://www.stonewall.org.uk/lgbt-britain-health [Accessed 20 Aug. 2021].

10. NHS England (n.d.). NHS England - Advancing Mental Health Equalities. [online] England.nhs.uk. Available at: https://www.england.nhs.uk/ltphimenu/mental-health/advancing-mental-health-equalities/ [Accessed 20 Aug. 2021].

11. Stonewall. (2018). LGBT in Britain: Trans Report. [online] . Available at: https://www.stonewall.org.uk/system/files/lgbt_in_britain_-_trans_report_final.pdf

12. Equalityhumanrights.com. (2020). Race report statistics | Equality and Human Rights Commission. [online] Available at: https://www.equalityhumanrights.com/en/race-report-statistics [Accessed 20 Aug. 2021].

13. Ibid.

14. Stonewall. (2018). LGBT in Britain: Trans Report. [online] . Available at: https://www.stonewall.org.uk/system/files/lgbt_in_britain_-_trans_report_final.pdf

15. Kalla, J. and Broockman, D. (2020). Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field
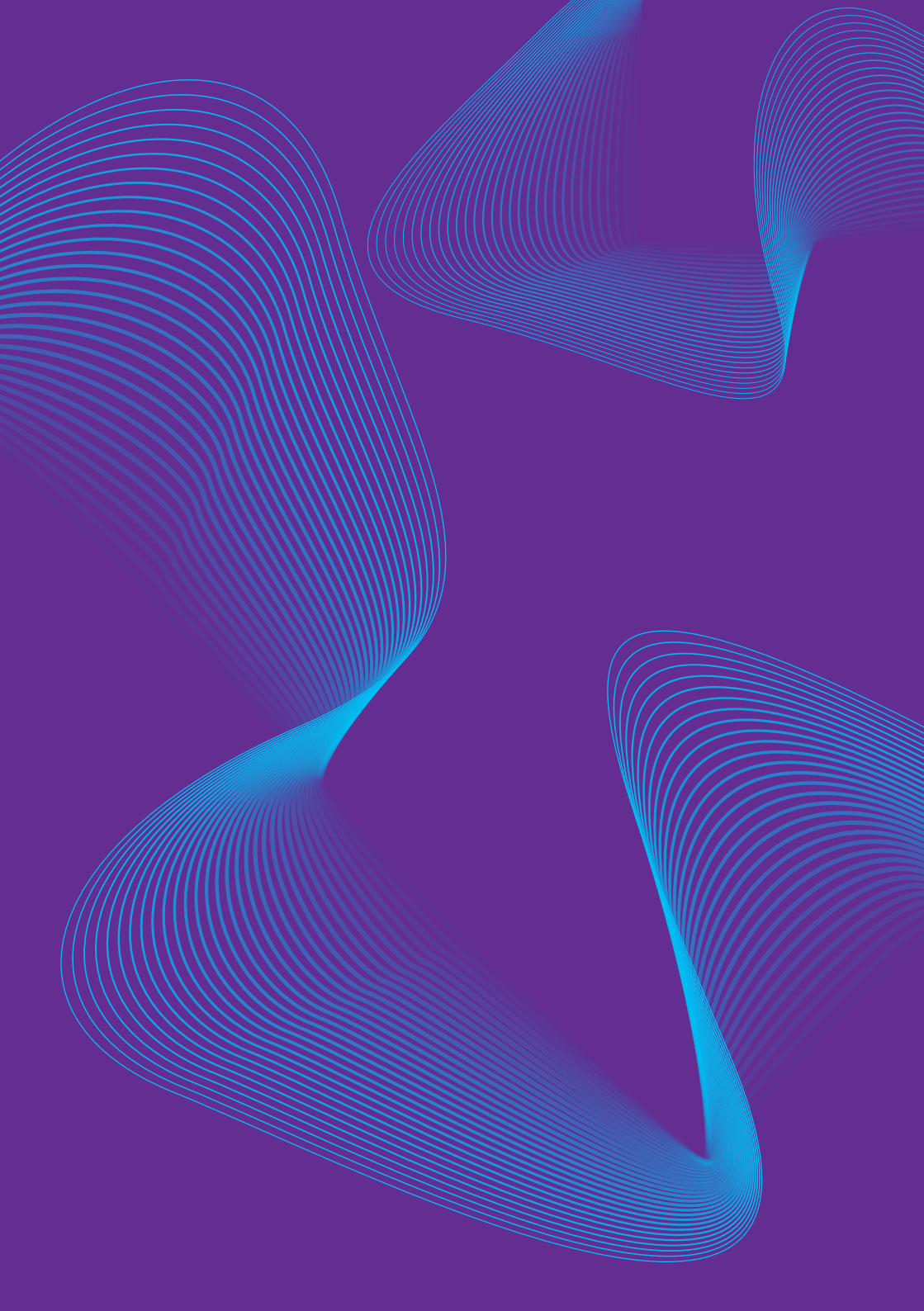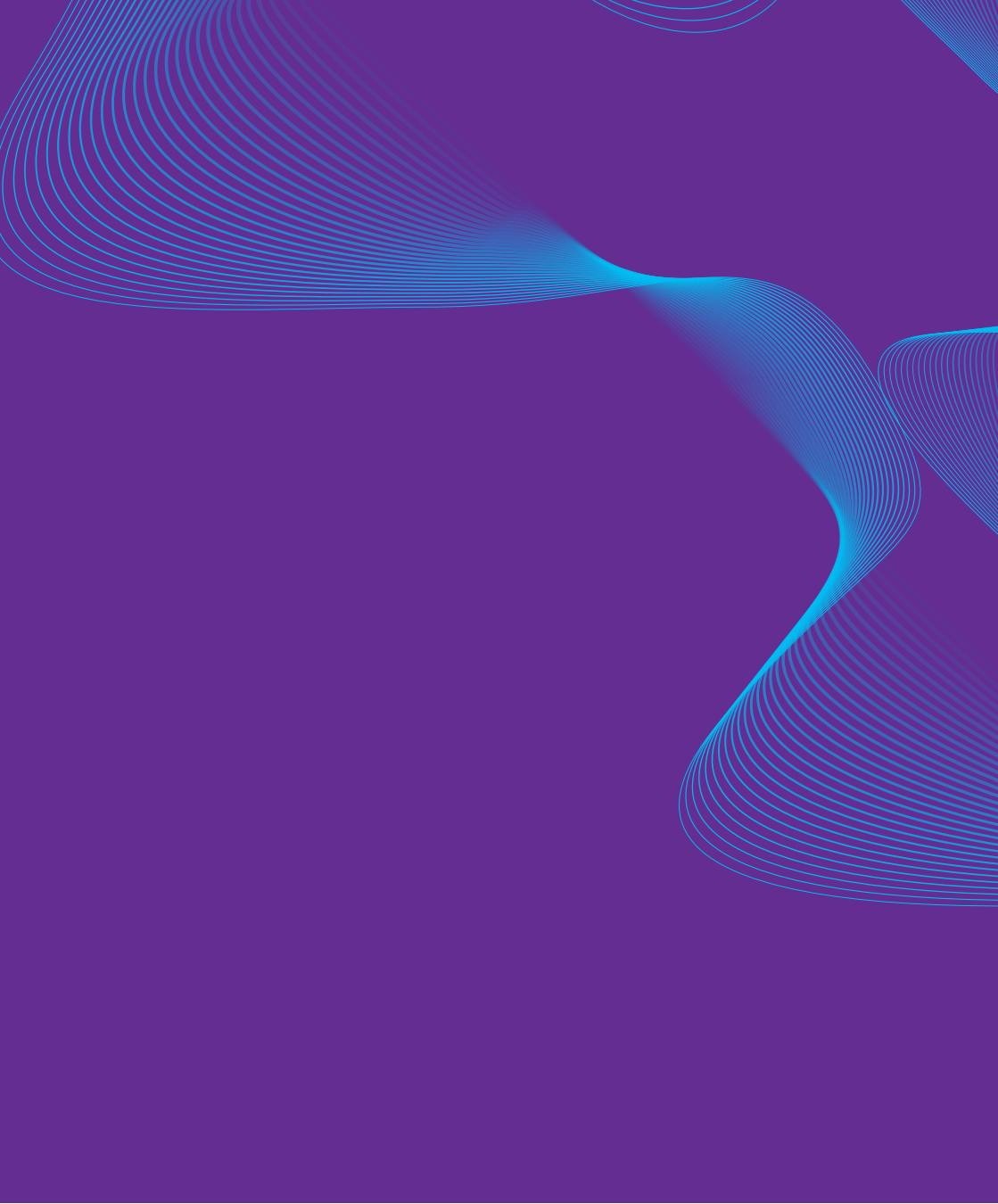
Experiments. *American Political Science Review*, [online] 114(2), pp.410–425. Available at: https://www.cambridge.org/core/journals/american-political-science-review/article/abs/reducing-exclusionary-attitudes-through-interpersonal-conversation-evidence-from-three-field-experiments/4AA5B97806A4CAFBAB0651F5DAD8F223 [Accessed 23 Aug. 2021].

16. List, J. (2004). The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field. The Quarterly Journal of Economics, [online] 119(1), pp.49–89. Available at: https://academic.oup.com/qje/article-abstract/119/1/49/1876078 [Accessed 23 Aug. 2021].

17. Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American Economic Review, [online] 94(4), pp.991–1013. Available at: https://www.aeaweb.org/articles?id=10.1257/0002828042002561 [Accessed 23 Aug. 2021].

18. Goldin, C. and Rouse, C. (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. American Economic Review, [online] 90(4), pp.715–741. Available at: https://www.aeaweb.org/articles?id=10.1257/aer.90.4.715 [Accessed 24 Aug. 2021].

19. Enos, R.D. (2014). Causal effect of intergroup contact on exclusionary attitudes. Proceedings of the National Academy of Sciences, [online] 111(10), pp.3699–3704. Available at: https://www.pnas.org/content/111/10/3699 [Accessed 24 Aug. 2021].

20. Heerwig, J. A., & McCabe, B. J. (2009). Education

and social desirability bias: The case of a black presidential candidate. *Social Science Quarterly*, 90(3), 674-686.

21. Grimm, P. (2010). Social desirability bias. In J. Sheth & N.Malhotra (Eds.), Wiley international encyclopedia of marketing. https://doi.org/10.1002/9781444316568.wiem02057