Behavioural Public Policy, Page 1 of 30
© *Cambridge University Press* This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited. doi:10.1017/bpp.2018.25

The contact hypothesis re-evaluated

ELIZABETH LEVY PALUCK

Department of Psychology and Public and International Affairs, Princeton University, Princeton, NJ, USA

SETH A. GREEN*

Code Ocean, New York, NY, USA

DONALD P. GREEN

Department of Political Science, Columbia University, New York, NY, USA

Abstract: This paper evaluates the state of contact hypothesis research from a policy perspective. Building on Pettigrew and Tropp's (2006) influential meta-analysis, we assemble all intergroup contact studies that feature random assignment and delayed outcome measures, of which there are 27 in total, nearly two-thirds of which were published following the original review. We find the evidence from this updated dataset to be consistent with Pettigrew and Tropp's (2006) conclusion that contact "typically reduces prejudice." At the same time, our meta-analysis suggests that contact's effects vary, with interventions directed at ethnic or racial prejudice generating substantially weaker effects. Moreover, our inventory of relevant studies reveals important gaps, most notably the absence of studies addressing adults' racial or ethnic prejudices, an important limitation for both theory and policy. We also call attention to the lack of research that systematically investigates the scope conditions suggested by Allport (1954) under which contact is most influential. We conclude that these gaps in contact research must be addressed empirically before this hypothesis can reliably guide policy.

Submitted 23 February 2018; accepted 28 February 2018

What we have learned since Pettigrew and Tropp: a 10-year retrospective and update

For more than a century, researchers have sought to understand what causes people to harbor and express prejudice against outgroups. Sustained attention

* Correspondence to: Seth Green, Code Ocean, 311 West 43rd Street, New York, NY 10036, USA. Email: sag2212@columbia.edu

Replication code for this article has also been published in Code Ocean, a computational reproducibility platform that enables users to run the code, and can be viewed here: https://doi.org/10.24433/CO.f152260c-bebb-4157-a640-44579452b4e4.v4.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

to the topic of prejudice reflects the fact that in every era and region, stereotyping, discrimination and xenophobia manifest themselves in ways that contribute to social inequality and sometimes erupt into intergroup violence.

Policy-makers have historically looked to social science for guidance about how to reduce prejudice (Myrdal, 1944), and social scientists themselves have sought to conduct research on prejudice that could inform programs and policies. In 2006, the American Psychological Association resolved to "call upon psychologists to use findings from relevant psychological research on prejudice, stereotyping and discrimination to inform their research, practice, training and education … [and] to inform anti-prejudice, anti-stereotyping and anti-discrimination positions in public and organizational policy" (APA, 2006, p. 308).

Among the many prominent theories in this domain (for a review, see Paluck & Green, 2009), the promotion of intergroup contact has arguably become the foremost strategy for reducing prejudice. The palliative effect of intergroup contact is a central theme of Gordon Allport's landmark book, *The Nature of Prejudice* (1954), which drew its inspiration from earlier studies suggesting that housing and workplace desegregation in the United States reduced prejudice toward black people (Williams, 1947; Mussen, 1950). Although skeptical of the notion that any form of contact diminishes prejudice, Allport conjectured that prejudice

> may be reduced by equal status contact between majority and minority groups in the pursuit of common goals. The effect is greatly enhanced if this contact is sanctioned by institutional supports (i.e., by law, custom, or local atmosphere), and provided it is of a sort that leads to the perception of common interests and common humanity between members of the two groups. (p. 281)

The so-called contact hypothesis set in motion decades of research assessing whether and under what conditions intergroup contact diminishes hostility toward outgroups.

A crucial turning point in the prejudice reduction literature came in 2006 with the publication of Pettigrew and Tropp's (2006) influential review of more than 500 studies of the effects of intergroup contact. Their widely cited meta-analysis provided evidence so decisive that the authors concluded "[t]here is little need to demonstrate further contact's general ability to lessen prejudice. Results from the meta-analysis conclusively show that intergroup contact can promote reductions in intergroup prejudice" (p. 751). As Hewstone (2003) put it, thanks to "the Herculean labors of Pettigrew and Tropp," we may now answer the question of whether contact reduces prejudice "with an emphatic 'yes'" (p. 352). This conclusion is echoed in an array of social psychology articles and textbooks that describe intergroup contact as a "clearly demonstrated" method for reducing intergroup hostility (Yablon, 2012, p. 250).

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

The Pettigrew and Tropp meta-analysis is also noteworthy for what it did not find. The four special conditions – equal status between the groups in the situation; common goals; intergroup cooperation; and the support of authorities, law or custom (p. 752) – that Allport believed made for propitious intergroup contact received relatively little empirical support. Evidently, "Allport's conditions are not essential for intergroup contact to achieve positive outcomes. In particular … samples with no claim to these key conditions still show significant relationships between contact and prejudice" (p. 766). These effects were said to extend beyond direct exposure, which further magnified the policy implications of intergroup contact:

> Indeed, the generalization of contact's effects appears to be far broader than what many past commentators have thought. Not only do attitudes toward the immediate participants usually become more favorable, but so do attitudes toward the entire outgroup, outgroup members in other situations, and even outgroups not involved in the contact. This result enhances the potential of intergroup contact to be a practical, applied means of improving intergroup relations. (p. 766)

A decade has passed since the publication of Pettigrew and Tropp's meta-analysis. Many new studies have been conducted in the meantime, some of them quite elegant and well-powered. How would the meta-analysis look today if the literature were brought up to date and expanded to include pertinent research conducted outside psychology? In this paper, we set ourselves the task of updating the original meta-analysis and broadening its disciplinary scope.

A second and central aim of this paper is to attend specifically to policy-relevant studies that speak to the practical applications of intergroup contact. Scholars who have looked to the prejudice-reduction literature as a guide to public policy have lamented its common design limitations. Non-experimental studies are prone to bias due to well-known threats to validity (Campbell & Stanley, 1963), and research has specifically found that the positive correlation between contact and non-prejudiced behavior in observational data can be explained by less-prejudiced individuals seeking contact (Bertrand & Duflo, 2017). Among experimental studies, another limitation is that researchers tend to focus on outcomes that can be measured immediately after an intervention. A policy-maker might reasonably ask whether the effects of contact endure for days, weeks, months or years; yet as Abrams (2010) observes in his report to the UK's Equality and Human Rights Commission, "there is a dearth of good-quality longitudinal research on prejudice or prejudice reduction" (p. 68).

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

When searching for policy-relevant findings among the hundreds of studies that comprise the contact literature to date, we used the following two critiques of the literature as criteria for determining whether a research study was capable of generating actionable research findings: first, did the study assign a contact intervention randomly, allowing for unbiased causal inference about the effects of intergroup contact? Second, were outcomes measured at least one day after the contact intervention began?[1] Testing whether intervention effects endure beyond the first day of engagement is a minimum policy standard of efficacy. This requirement also reflects the greater stock we put in studies that separate the experimental intervention process from the measurement process. Of the hundreds of studies we reviewed, only 27 experiments track post-intervention outcomes for at least one day. Notably, just 11 of these studies focus on contact across racial or ethnic lines, which has been a concern of courts and policy-makers from the start of intergroup contact research.

We review this select group of studies both qualitatively and quantitatively. After describing the evolution of the contact literature and our procedures for identifying relevant studies, our qualitative review attends to important design nuances, such as details of the contact interventions, the contexts in which they were launched and the outcomes measured. Our quantitative analysis assesses the statistical robustness of the meta-analytic estimates given various coding and estimation choices. (Our database, replication programs and archive of digitized reprints[2] are publicly available so that readers may retrace our steps.)

The results we obtain are much more equivocal than those presented by Pettigrew and Tropp. Our analysis reveals that effects vary significantly by the type of prejudice addressed. This finding runs counter to one of the main findings from Pettigrew and Tropp's (2006) meta-analysis: namely that racial and ethnic prejudices are affected to approximately the same degree as other prejudices (p. 762). Furthermore, the literature has some

1 Requiring outcomes to be measured at least one day after the intervention began is a more permissive filter than one that asks for outcomes to be measured at least one day after treatment *ends*. The latter requirement would eliminate ten studies from our sample. Some of those test the effects of a sustained intervention extending over many months and measured outcomes during or on the final day of the intervention. Specific examples are roommate studies where measurements are collected while students live with same- or other-race roommates or sustained diversity trainings for which outcomes are measured on the last day. Another filter we considered was setting a sample size minimum of 30 people per cell, which would have eliminated nine studies from this sample. We decided to retain the most permissive filter aimed at policy relevance.

2 Our manuscript, online appendices and the archive of digitized reprints are available on the Open Science Framework: https://dx.doi.org/10.17605/OSF.IO/TTPVY

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

important gaps; for example, not one study assesses the effects of interracial contact on people older than 25. Given the narrow scope and mixed findings of the policy-relevant contact literature, we conclude that the jury is still out regarding the contact hypothesis and its efficacy as a policy tool. In particular, we note that the scope conditions suggested by Allport (1954) under which contact is more likely to be effective have not been systematically investigated.

A timeline of developments in the contact literature

The birth of the contact hypothesis

The history of the contact hypothesis begins in the early 20th century, when American social scientists initiated empirical studies of the effects of intergroup contact. For example, Williams (1934) measured the effects of a series of activities involving white and black women in the Young Women's Christian Association aged 14–18, including a field trip and a buffet dinner. Smith (1943) arranged for a "four-day seminar in Negro Harlem" (p. 26) on black cultural life and accomplishments for white students at Teacher's College, followed by a social tea at which Harlem residents were guests and speakers.

Following the desegregation of the military and other institutions after World War II, the social psychologist Gordon Allport (1954) distilled ideas about the benefits of intergroup interaction and friendship into a testable proposition. His core idea was that contact reduces prejudice. But, wary that contact could merely affirm existing social group hierarchies, Allport also proposed the set of conditions described above, under which contact should be especially influential. His work grew in prominence with the United States' struggles with desegregation. Many social scientists took up the call to test the contact hypothesis in the service of legal and public policy questions about the effects of school, neighborhood and workplace desegregation (Cook, 1985).

Thus, at its inception, the contact hypothesis served both an academic and a policy agenda. In the service of theory, the contact hypothesis characterized prejudice as a product of fear, ignorance, hierarchy or a lack of shared life patterns and goals. In the service of policy, the contact hypothesis has been proposed as a rationale for desegregation policies (Mussen, 1950; Pettigrew, 1979), as a guide for designing peacebuilding interventions (Kelman, 1998; Maoz, 2010) and as a theoretical narrative for interpreting the persistence of discrimination and interracial conflict (for an overview of the literature, see Pettigrew, 2016). Hundreds of studies followed, gauging the relationship between intergroup prejudice and interaction across racial, ethnic, religious and other group lines.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

The canonical meta-analytic result describing the effect of contact on prejudice

When Pettigrew and Tropp (2006) assembled the contact literature, they counted 515 studies, dating from the 1940s through the year 2000, comprising "slightly more than 250,000 participants from 38 different countries" (Pettigrew *et al.*, 2011, p. 16). Approximately half of these studies focused on racial or ethnic divisions; the rest investigated prejudice against groups including the mentally and physically handicapped, the elderly, political partisans, and gays and lesbians.

While all studies resulted in some type of empirical estimate of the effect of contact, the studies varied widely in terms of their research designs. Seventy-one percent of the meta-analysis database consisted of observational surveys of broad populations. In one widely cited study, Pettigrew (1997) surveyed 3806 people in France, Great Britain, The Netherlands and West Germany in 1988. Controlling for seven covariates, Pettigrew found that self-reported contact with members of immigrant outgroups was strongly associated with more positive evaluations of those groups.

Another 24% of the meta-analysis database consisted of observational intervention studies designed to assess outcomes among those experiencing intergroup contact and comparison groups who did not experience contact. For example, Lazar *et al.* (1971) studied the effects of a four-week curriculum unit developed by one teacher about "creative Americans" for a class of high-IQ children. As part of the curriculum, students interacted with people with physical disabilities and with a special education teacher. The treatment class's scores on an Attitudes Toward Disabled Persons (ATDP) survey were compared to one control classroom whose students were similar in terms of age, IQ and prior attitudes.

Just 5% of the database employed an experimental design. Within that subset, contact interventions, target groups, outcomes and settings vary widely. In a college setting, Pagtolun-an and Clair (1986) had a gay man answer questions about homosexuality for 90 minutes in a "deviant behavior class" (p. 125) and then post-tested students within the hour. The 35 students who experienced this form of contact with the speaker displayed statistically significant reductions in homophobia vis-a-vis an untreated control group. In one of a handful of experimental studies conducted outside the laboratory or classroom, DiTullio (1982), a special education job coordinator for the school district of Philadelphia, studied the effects of a job-training program that integrated adolescents with intellectual disabilities into custodial positions in Philadelphia elementary schools. Among coworkers and supervisors of the adolescents, this experimentally induced contact induced more positive

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

attitudes toward individuals with intellectual disabilities across a battery of measures.

The next ten years: adding to the meta-analytic database

The value of Pettigrew and Tropp's (2006) monumental collection of studies is beyond dispute. Taken together, the studies provide rich descriptions of contact experiences, develop new approaches to measuring attitudes toward stigmatized versus dominant groups and illustrate the many contexts within which intergroup contact may occur (e.g., within a programmatic intervention setting, an exchange program, a school setting or an incidental encounter in a community). Analyzed as a whole, they provide evidence for an association between contact and reduced prejudice that is robust to substantial variation across time, place and subjects.

However, the value of this collection of studies is less clear in one particular respect: for understanding whether contact *causes* policy-relevant reductions in prejudice. The vast majority (95%) of studies do not randomly assign contact; of those that do, just eight measure outcomes at least a single day after treatment. Of those eight, three study interracial contact. Thus, evidence for whether contact's effects on racial prejudice persist – the focus of policy and legal work on intergroup contact research and advocacy – is sparse.

In the 10 years since Pettigrew and Tropp's meta-analysis, contact research has entered a more methodologically sophisticated era in which social scientists are paying attention to new and re-emerging issues of research design, analysis and transparency. None of Pettigrew and Tropp's experimental studies, for instance, feature a pre-analysis plan or open-access data. Subsequently, three studies featuring one or both have been published (Broockman & Kalla, 2016; Finseraas & Kotsadam, 2017; Scacco & Warren, 2018).

As an example of recent developments, consider two high-quality experiments conducted after Pettigrew and Tropp's (2006) meta-analysis. Scacco and Warren (2018) provided 16 weeks of small computer training classes for low-income Christian and Muslim men in northern Nigeria. Classes were randomly assigned to be religiously homogenous or mixed. The authors found that contact produced "no changes in prejudice," and that while subjects in heterogeneous classes discriminated less than those in homogeneous classes, this was attributable to "increased discrimination by homogeneous-class subjects" (p. 1) relative to those who had not taken the class. In an American university context, Page-Gould *et al.* (2008) brought white and Latinx students into an immersive laboratory friendship-building experience over the course of three consecutive weeks, randomly assigning students to work with a same- or cross-group student partner. In the 10 days following the final session, the authors found statistically insignificant and substantively small effects on participant likelihood of initiating

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

cross-group interaction, although the effects were somewhat larger among participants who scored high on a pre-treatment test of implicit prejudice. Overall, the authors found "benefits of cross-group friendship, particularly among people who are most likely to experience anxiety in intergroup contexts" (p. 1089).

We reassess two core propositions about intergroup contact in light of these and other studies. First, we assess whether contact reduces prejudice. Second, we assess whether Allport's original moderating conditions shape the extent to which contact reduces prejudice.

Assembling studies for meta-analysis

Following Pettigrew and Tropp (2006), we "define intergroup contact as actual face-to-face interaction between members of clearly defined groups" (p. 754). To update the universe of relevant studies, we sought all studies that met this definition, randomly assigned contact, had delayed outcome measures and were published (or available as working papers) by July 2016. Next, we summarized the resulting set of studies qualitatively and conducted a meta-analysis using methods similar to those of Pettigrew and Tropp (2006).

Assembling the collection of studies

First, we identified all studies in Pettigrew and Tropp's database that randomly assigned intergroup contact and measured outcomes more than a day after treatment. To do so, we cross-referenced each study that Pettigrew and Tropp classified as experimental with the bibliography provided in their subsequent book, *When Groups Meet: The Dynamics of Intergroup Contact* (2011). After removing studies that did not have over-time outcome measures, were mislabeled as randomly assigned, did not feature "actual face-to-face interaction" or did not have a non-contact control group, we were left with eight research reports comprising nine experiments on intergroup contact.

Second, we incorporated studies cited by other recent literature reviews and meta-analyses. For example, Lemmer and Wagner (2015) compiled every contact and "imagined contact" study taking place in the field through 2012 that measured outcomes more than one month after treatment; this collection furnished an additional four randomized controlled trials.[3] A literature review on anti-prejudice interventions (Paluck & Green, 2009) provided three studies

3 Additionally, the authors produce a wealth of supplementary information, making their work both transparent and particularly helpful for this project. We depart from their paper, however, in that we look exclusively at randomized controlled trials, perform sub-analyses by target of prejudice, and, in particular, attend to the relationship between effect sizes and precision of estimates.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

comprising four samples, and an unpublished literature review of interracial roommate pairings (Green, 2014) provided four studies. Lastly, a review of sexual discrimination (Tucker & Potocky-Tripodi, 2006) provided one study.

Third, we informally canvassed intergroup contact researchers, discussing our project with, among others, Linda Tropp, Thomas Pettigrew, Kristin Davies and Ryan Enos, as well as attending conferences and reading relevant journals. This furnished four additional studies; those studies' citations led to two more.

Fourth, we searched Google Scholar for all studies that cited Pettigrew and Tropp (2006) and had the words 'random', 'assign' and 'contact' somewhere in the text, which revealed one further study, bringing our final sample to 27 studies and 31 treatment arms.

Intergroup contact studies: who, what, where and when

Looking within the group of studies that comprise this meta-analysis, we now ask: who are the participants, what were the treatments, where did they take place and when? By answering these questions, we attempt a richer qualitative description of this evidence than the numbers alone can provide.

Who: participants and types of prejudice

First, whose prejudices are being studied? And who are the targets of the prejudice under study? Table 1 summarizes our universe of cases along these two dimensions.

Participants

Thirteen of the 27 experiments study college students, and all but one of these experiments took place in the USA. The exception is Burns *et al.* (2015), who studied students at the University of Cape Town. Scacco and Warren (2018) examined young adults in Nigeria of college age who were not in college.

Of the six studies of adults over 25 years of age, three took place outside of the USA: one in a housing complex in Hyderabad, India (Barnhardt, 2009) and two studies with Norwegian military recruits (Finseraas *et al.*, 2016; Finseraas & Kotsadam, 2017). In the USA, Dessel (2010) studied teachers in an evangelical Christian community; DiTullio (1982) studied custodial teams in Philadelphia; and Broockman and Kalla (2016) canvassed residents of Miami, Florida.

Elementary and middle-school students participated in studies in the USA (Katz & Zalk, 1978; Meshel & McGlynn, 2004) and Australia (Cluniess-Ross & O'Meara, 1989), and high-school students participated in Israel

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Table 1. Participants and targets of prejudice

Targets of prejudice	Study participants			
	Elementary and middle-school students ($n = 3$)	High-school students ($n = 4$)	College students ($n = 13$) and college-aged young adults ($n = 1$)	Adults over 25 years of age ($n = 6$)
Members of other racial and ethnic groups ($n = 11$)	1	1	9	0
Immigrants and foreign nationals ($n = 2$)	0	0	1	1
Members of other religious groups ($n = 3$)	0	1	1	1
LGBTQ ($n = 3$)	0	0	1	2
Women ($n = 1$)	0	0	0	1
Individuals with intellectual disabilities ($n = 4$)	1	1	1	1
Individuals with physical disabilities ($n = 2$)	0	1	1	0
Age ($n = 1$)	1	0	0	0

(Yablon, 2012), Germany (Krahe & Altwasser, 2006) and the USA (Sheare, 1974; Green & Wong, 2009).

Types of prejudice

Pettigrew and Tropp (2006) point out that the contact hypothesis was

> originally developed to address racial and ethnic prejudices, but recent decades have witnessed a massive use of the theory for a range of different target groups. Is this expansion of contact theory justified? And do these non-racial and nonethnic samples yield meta-analytic patterns that are similar to those for racial and ethnic samples? (p. 762)

Their meta-analysis seems to provide a resounding 'yes': average effects of contact are strikingly similar across a range of target groups, and confidence intervals always overlap (see their Table 11, p. 764). Like their database, ours features a preponderance of studies focusing on ethnic or racial prejudice: 11 out of 27, or 40%. Camargo *et al.* (2010), Marmaros and Sacerdote (2006), Katz and Zalk (1978) and Sayler (1969) addressed relations between blacks and whites in the USA, and Burns *et al.* (2015) addressed relations between whites and blacks in South Africa. Four other studies (Boisjoly *et al.*, 2006; Green & Wong, 2009; Markowicz, 2009; Sorensen, 2010) tested relations

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

among blacks, whites and members of other groups, such as Asians, Latinxs and "Native Hawaiian and Other Pacific Islander" (Markowicz, 2009, p. 66). Lastly, Page-Gould *et al.* (2008) assessed relations between whites and Latinxs, and Furuto and Furuto (1983) assessed relations between white and "Polynesian and Oriental" (p. 153) students at Brigham Young University – Hawaii. As Table 1 shows, all of these studies were conducted with populations from elementary school through college.

All other categories of prejudice, discrimination and stigma are addressed by four or fewer studies, and yet contain a great deal of demographic and geographic heterogeneity. One study addressed discrimination against foreign nationals in the USA (Hull, 1972), and another against immigrants in Norway (Finseraas & Kotsadam, 2017). Three studies tested prejudice, discrimination and stigma against LGBT individuals: either transgender people (Broockman & Kalla, 2016) or gays and lesbians (Grutzeck & Gidycz, 1997; Dessel, 2010).

The three studies examining contact between religious groups selected Hindus and Muslims in India (Barnhardt, 2009), Christians and Muslims in Nigeria (Scacco & Warren, 2018) and Jews and Arabs in Israel (Yablon, 2012).[4] Four studies targeted prejudice against the intellectually disabled; three of those took place in the USA (Hall, 1969; Sheare, 1974; DiTullio, 1982) and one in Australia (Clunies-Ross & O'Meara, 1989). Prejudice against people with physical disabilities was studied in the USA (Evans, 1976) and Germany (Krahe & Altwasser, 2006). Finally, we note one study targeting prejudice against the elderly (Meshel & McGlynn, 2004) and one targeting discrimination against women (Finseraas *et al.*, 2016).

What: interventions and measurements

What kind of contact did the study authors randomize across participants? Some contact was crafted by researchers, while other types of contact were more naturalistic; some contact was sustained, while other engagements were very brief. We also describe the variety of outcomes measured following the intervention. Most outcomes were self-reported attitudes and social evaluations, while behavioral outcomes mostly included observed interactions and measures of friendship with members of the other group. Outcome measures also varied in whether they focused on reduced prejudice toward the outgroup involved in the study or on general levels of social tolerance.

4 Readers might think of the divide between Jews and Arabs as an ethnic distinction. We classified this as religious. Classifying this study one way or the other does not affect our overall results.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

What type of contact?

The contact interventions can roughly be characterized as falling along two interrelated dimensions: *scriptedness*, or the degree to which experimenters control and direct the treatment content and whether they employ confederates as a means to steer the contact experience (e.g., Evans, 1976); and *duration*, which ranges from brief and impersonal exposure to sustained and intimate contact.

In an example of a brief, unscripted encounter, Hall (1969) examined University of Alabama students who were gathered together with residents of an institution for people with severe intellectual disabilities and encouraged to pair up or assemble groups to sing songs or practice skills like tying shoes. Far more common are brief, scripted interactions. These typically take place in a laboratory or classroom and involve a structured conversation or an activity with a member of a presumed outgroup. Evans (1976), for instance, randomly assigned college students ($n = 40$) to have one of two types of conversation with a blind woman. In one, they were asked to discuss their hometowns, majors and family; in the other, the woman who was blind explicitly invited questions about blindness. Katz and Zalk (1978) assigned interracial and racially homogenous groups of second and fifth graders to work on puzzles for 15 minutes under observation by their teachers.

Scripted and sustained interventions are typically designed around intergroup dialogue and excursion interventions. Yablon (2012) studied the effects of six monthly meetings between Palestinian and Israeli high-school students in which they discussed social issues and concluded with a joint trip to an amusement park. Dessel (2010) led discussion groups between straight teachers and LGB volunteers over the course of two months for nine hours in total; Sorensen (2010) and Markowicz (2009) each studied the effects of interracial dialogues held at universities.

Sustained, unscripted intergroup encounters featured extensive contact in a naturalistic environment that researchers cannot directly control or sometimes even monitor. In general, they follow Allport's (1954) argument that to reduce prejudice, intergroup contact experiences "should occur in ordinary purposeful pursuits, [and] avoid artificiality" (p. 489). Ten studies targeted intergroup living situations, such as interracial college roommates, ranging in duration from a weekend (Hull, 1972) to eight weeks (Finseraas *et al.*, 2016; Finseraas & Kotsadam, 2017) to a year (Boisjoly *et al.*, 2006; Camargo *et al.*, 2010) or more (Barnhardt, 2009).

Another way to describe the type of contact in these interventions is to ask whether they fit the conditions that Allport specified as critical for prejudice reduction. Very few interventions fit all four conditions. Most interventions,

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

because they needed approval to be launched, are characterized by authority approval of the contact (26 out of 27 – all but Broockman & Kalla, 2016). Seventeen studies could be characterized as featuring equal status contact, 14 feature cooperation and 12 have a common goal between groups in contact. Several of the interventions, however, are difficult to characterize according to Allport's classification scheme. Given a general lack of detailed description about the interventions, it was particularly difficult to determine whether an intervention involved equal status or a common goal. Roommate studies and more generally naturalistic and sustained interventions are also challenging, given that conditions of cooperation or equal status likely fluctuate over time. Naturalistic studies are also likely to involve some amount of negative contact experiences, like misunderstandings or outright conflict, which could affect outcomes.

What kinds of outcome measures?

Because our collection of studies spans six decades, four continents and a host of different demographic groups, it is not surprising that outcome measures range widely. We group outcome measurements into four broad categories: (1) explicit evaluations of the outgroup; (2) political and cultural attitudes commonly associated with prejudice (e.g., opinions about affirmative action); (3) behavioral measures of actions toward the outgroup, such as white subjects' numbers of black friends or the percentage of all emails that white subjects sent to black peers; and (4) indirect or projective measures of prejudices such as implicit attitude tests or the evaluation of hypothetical vignettes.

Explicit evaluations of the outgroup typically took the form of a series of evaluative questions. Such outcomes were common in studies of prejudice against people with intellectual disabilities; Hall (1969), for instance, asked participants to rate the "mentally retarded along a clean–dirty axis" (p. 31). Studies of ethnic, racial and religious prejudice also featured explicit outcome measurements in settings where it is (or was) more common to express outright hostility toward an outgroup, such as the USA in the 1960s (Sayler, 1969), contemporary Nigeria (between Christians and Muslims; Scacco & Warren, 2018) and India (between Hindus and Muslims; Barnhardt, 2009).

Experimenters sometimes used more oblique measures to elicit prejudiced attitudes. Some focused on political and cultural attitudes, soliciting opinions about affirmative action (Boisjoly *et al.*, 2006) or policies discriminating against transgender people (Broockman & Kalla, 2016). This category also includes measures of nationalism and world-mindedness (Hull, 1972) and general beliefs about the extent of racial privilege in the USA (Markowicz, 2009).

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Other experimenters used behavioral indicators to track interactions with outgroup members. Marmaros and Sacerdote (2006) unobtrusively tracked how many emails Dartmouth students sent to white and black peers, including and excluding their own roommates. Camargo *et al.* (2010) asked white students at Berea College with and without black roommates to report how many black friends they have, again with and without their roommates included. Page-Gould *et al.* (2008) included a daily diary report following intervention of how likely participants are to initiate a cross-group interaction. A novel behavioral measure of social distance between whites and blacks comes from Katz and Zalk (1978). After 15 minutes of cross-group interaction, children were asked to place a variety of felt objects on a flannel board, with either a black or white examiner standing to one side of the board; the outcome measure was "literally the average distance the subject placed the five forms from the examiner" (p. 451).

Finally, a minority of studies (four) gathered evidence typically tested in social scientific laboratories: behavioral games (Scacco & Warren, 2018), the implicit attitude test (IAT; Barnhardt, 2009; Burns *et al.,* 2015) and a vignette experiment (Finseraas *et al.*, 2016).[5] Scacco and Warren (2018) used behavioral games to assess cooperation and trust between Nigerian Christians and Muslims, specifically dictator and destruction games in which individuals allocated real money to their real partners. The vignette experiment in Finseraas *et al.* (2016) varied the qualifications of female and male officers in the Norwegian military to measure proclivity to discriminate based on gender.

An additional feature of outcome measurement is whether a dependent variable pertains to a particular outgroup or toward outgroups in general. For example, Green and Wong (2009) measured tolerance toward a variety of groups, and Markowicz (2009) investigated awareness of racial privilege in the USA. Most studies in our sample focus on how prejudice reduces discrimination toward the group to which the treatment subjects were directly exposed.

Where

Of the 27 studies, 19 (70%) took place in the USA. Of the remaining eight, two were located in Norway and one study was located in each of the following countries: India, South Africa, Australia, Germany, Nigeria and Israel.

Of the seven studies with child or adolescent subjects, five took place in a classroom or in the context of a school-required activity such as a field trip;

5 We include only a few studies that feature this type of evidence because it is typically collected during or immediately after intervention.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

the two in naturalistic settings took place during after-school activities with the elderly (Meshel & McGlynn, 2004) or during an outdoor hiking expedition (Green & Wong, 2009).

Of the 13 studies with college students, six took place in a naturalistic setting like a dormitory, a neighborhood center (Sayler, 1969) or a living facility for the severely intellectually disabled (Hall, 1969). Three others were structured, monitored intergroup discussions (Hull, 1972; Markowicz, 2009; Sorensen, 2010); two were studies that students enrolled as part of a class (Evans, 1976; Page-Gould *et al.*, 2008); and one took place in a normal class lecture (Grutzeck & Gidycz, 1997). The 13th study took place at Brigham Young University – Hawaii and consisted of 14 weekly "spiritual, cultural and social experiences" in integrated settings, both on and off campus (Furuto & Furuto, 1983, p. 150).

When

The breakdown by decade of study is shown in Table 2. The majority of evidence for our review was generated since 2000.

Meta-analytic methods

This section describes the criteria used to identify the key outcome variable in each study and the procedures used to transform each study's reported results into the inputs for our meta-analysis.

Selecting dependent variables

Some studies in our sample reported a single outcome (Hull, 1972), while others reported dozens (Burns *et al.*, 2015). Some tested multiple subgroups, such as Broockman and Kalla (2016), whose pre-analysis plan specified looking for heterogeneous treatment effects by party identification, and Scacco and Warren (2018), who tested for contact effects on both Christians and Muslims. Some experiments involve multiple, conceptually distinct treatment arms, such as Boisjoly *et al.* (2006), who measured separately the effects of having a black roommate and having a non-black minority roommate. Others varied the intensity of one treatment, such as Barnhardt (2009), who measured the effects of having one, two or three Muslim or Hindu households in one's four-household living unit.

Some outcome measurements are composites of multiple subscales (Sayler, 1969) or multiple items intended to evaluate feelings toward the outgroup (DiTullio, 1982). One study delineated a 'main outcome' (Finseraas & Kotsadam, 2017), while others present a collection of response variables

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Table 2. Contact studies by decade. Note that we place Meshel and McGlynn (2004) in the 1990s and Dessel (2010) in the 2000s, going by the initial publication of the relevant data in the authors' dissertations (Meshel, 1997; Dessel, 2008)

Decade	Number of studies
1960s	2
1970s	4
1980s	3
1990s	2
2000s	8
2010s	8

with no ranking system. We sought to apply consistent rules for choosing which outcomes to demarcate as representative of a study's overall findings so that we could condense each paper's findings down to a single estimate and accompanying standard error. We decided on the following rules:

- First, we chose estimates evaluating the highest dosage (experimentally varied intensity) of contact whenever possible.
- Second, when estimates are split by dominant versus subordinate or majority versus minority groups, we chose estimates evaluating prejudice reduction among the dominant or majority groups.
- Third, we considered studies to have multiple treatment arms if they met any of the following criteria: (a) featured one ingroup exposed to multiple distinct outgroups; (b) measured the effects of contact on multiple participant groups; or (c) featured one intervention across multiple, distinct settings. Our meta-analysis includes one effect size for each treatment arm.
- Fourth, when studies look at contact between two groups in conflict, in which neither is clearly dominant, we chose effect sizes that measure changes across both populations.
- Fifth, we chose the prejudice outcome on which the author(s) focused primarily.
- Sixth, if there were multiple post-tests, we chose the latest possible post-test.
- Seventh, when faced with a choice among estimators, we chose linear estimators so that we could express treatment effects in terms of standardized units.
- Eighth, when multiple econometric specifications were present, we chose the specification that estimated the treatment effect with the smallest apparent standard error.

After selecting our dependent variables, we next turned to converting them to a common framework.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Creating a common statistical framework

The most common analytic strategy for meta-analyses in the social sciences is to calculate standardized mean difference, commonly referred to as Cohen's *d*, defined in Cooper *et al.* (2009, p. 226) as

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

in which μ_1 and μ_2 represent the population averages of the treatment and control groups and σ is the sample standard deviation.

Estimating the numerator of equation (1) is straightforward, whereas there is some debate about how to estimate the denominator. Lemmer and Wagner (2015) follow the recommendation of Morris (2008) to pool standard deviations from the treatment and control groups at pre-treatment as

$$\sigma_{pre} = \sqrt{\frac{(n_t - 1)\sigma^2_{pre,T} + (n_c - 1)\sigma^2_{pre,C}}{n_t + n_c - 2}}.$$

Using pre-treatment information to estimate population standard deviation has the advantage of not making any additional distributional assumptions about the effects of treatment. However, pre-treatment standard deviations are not available for all of the studies in our sample. To keep comparisons constant across studies, we standardized all changes associated with treatment by the standard deviation of the control group, a statistic commonly called Glass's Δ.[6] After standardizing effect sizes for each study, we calculated standard errors for each, correcting for bias arising in small studies using Hedge's *G* correction factor (Cooper *et al.*, 2009).

Two studies (Hall, 1969; Katz & Zalk, 1978) and one treatment arm of a study (Sayler, 1969) did not provide enough information about sampling variability to compute standardized effect sizes. We exclude these studies, representing a total of four treatment arms, from our meta-analysis, although they remain relevant for the sign tests that we conduct below. This left us with 25 studies comprising 27 treatment arms.

Meta-analytic results

A graphical overview of results from our 27 comparisons can be found in Figure 1. This scatterplot depicts the relationship between the effect estimate

6 In practice, we find that all available methods of standardizing effect sizes produced substantively similar estimates.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

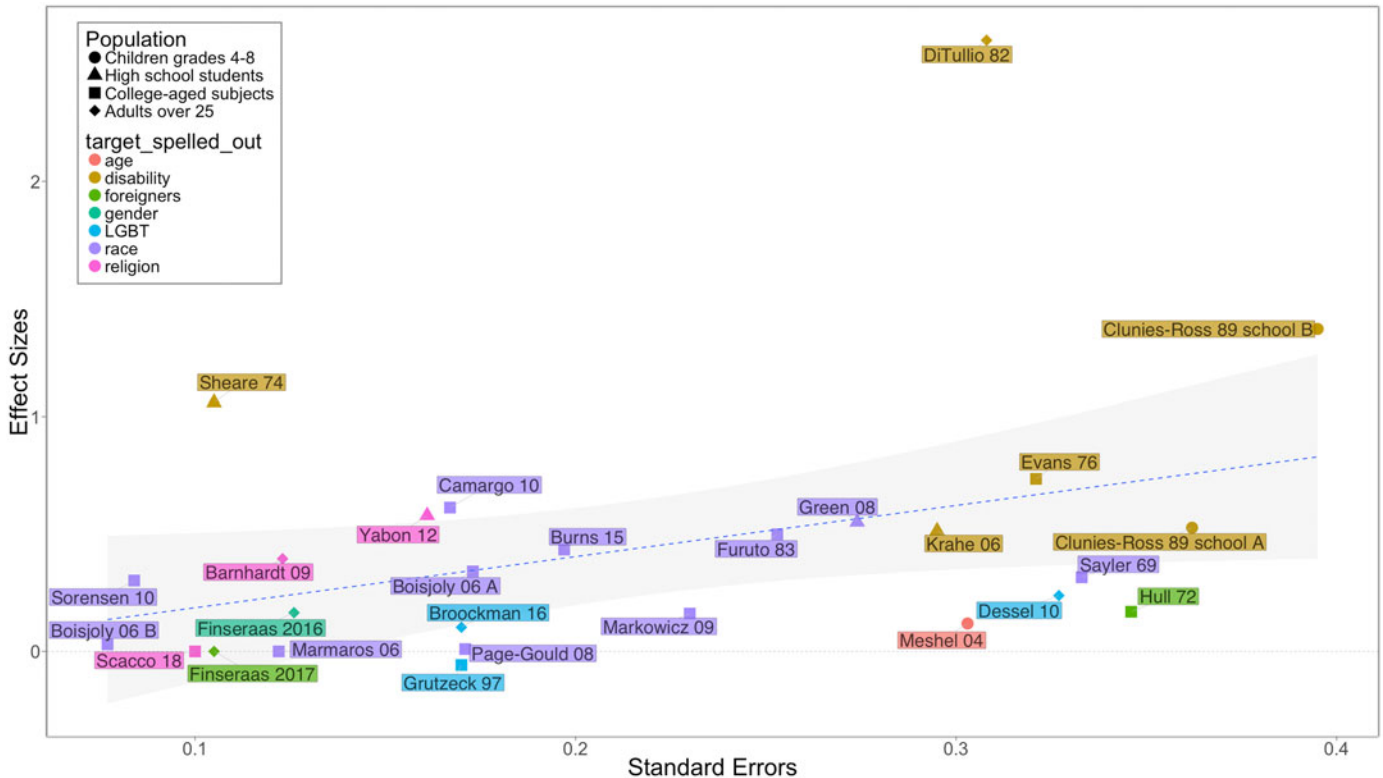

Figure 1. Standard error and effect size (Glass's Δ) of each experimental comparison. Colors of points and of labels correspond to target of intervention; shapes of plotted points correspond to population. Fitted line is ordinary least squares and gray bands are 95% confidence intervals. Boisjoly 'A' and 'B' refer, respectively, to effects associated with having black or 'other minority roommates'

(vertical axis) and its standard error (horizontal axis) for each of the 27 effect estimates in our analysis. To aid interpretation, we color-coded each study according to whether it focuses on prejudice against ethnic or racial groups, religious groups, immigrants, people with mental or physical disabilities, the elderly, women or LGBT individuals. The participant pool for each study is also indicated according to the polygon used to represent each observation. Study participants who are children are depicted with circles, teenagers with triangles, college students and young adults with squares and adults aged 25 years or older with diamonds.

This graphical overview of the core studies underscores several noteworthy features of the contact literature. First, effect sizes vary considerably by target group, with substantially larger effects observed in studies that target prejudice toward those with disabilities. Second, four out of six studies with adult subjects are clustered on the bottom left of the figure, reflecting both smaller effect sizes and standard errors than the collection of studies on average. Third, vertical positioning of the points indicates that the overwhelming majority of experiments (24 out of 27) report a positive effect of contact. The probability of observing 24 positive estimates out of 27 studies is less than 0.001 under the null hypothesis of no effect.[7] The distribution of estimated effects seems to offer strong support to the hypothesis that the types of contact facilitated in these studies led to reductions in prejudice.

However, Figure 1 also suggests that caution is warranted when summarizing the results via meta-analysis. The regression line that passes through the points calls attention to the fact that studies with smaller standard errors tend to report weaker effects than studies with larger standard errors. In other words, the larger the study, the smaller the standard error and the smaller the estimated effect. This pattern is symptomatic of a 'file drawer problem', in which studies are more likely to be reported when they show significant results (Rosenthal, 1979). In light of this pattern, our meta-analysis considers not only the pooled study average effect, but also the study average that would be forecasted as the standard errors tend toward zero.

We begin our quantitative analysis by assessing cross-study heterogeneity using Cochran's Q. The test decisively rejects the null hypothesis of homogeneity of effects across studies ($Q(26) = 173.178$, $p < 0.001$; $I^2 = 0.85$). We therefore reject the fixed-effects meta-analysis model in favor of a random-effects meta-analysis model, where the variance of the normal random component

7 When we conduct a sign test that includes the four studies for which we could calculate unstandardized but not standardized effect sizes, we find that 26 out of 31 studies show positive effects, the probability of which is lower than 0.001 under the null hypothesis of no effect.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

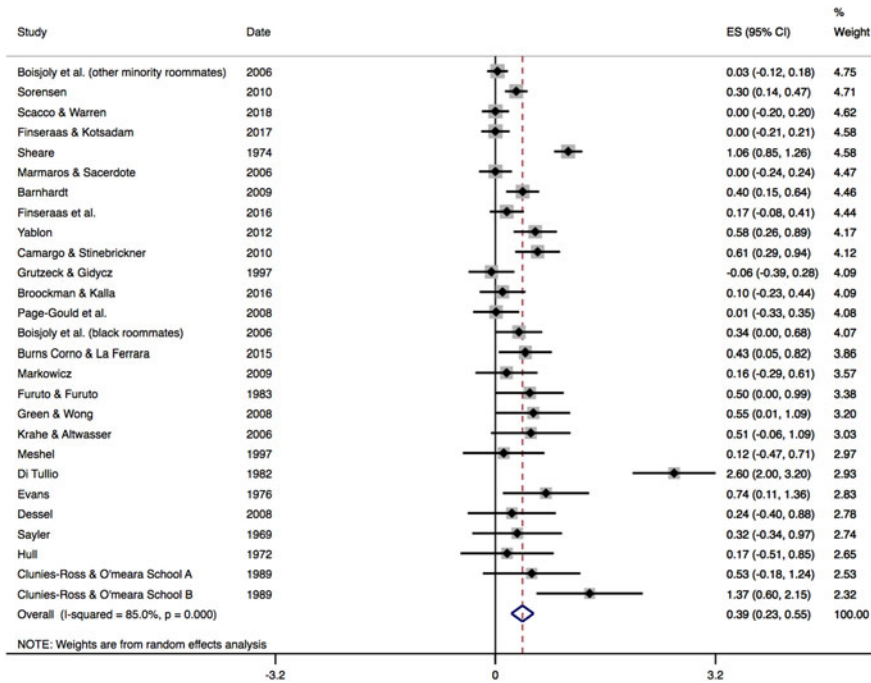

Figure 2. Forest plot of standard errors and effect sizes (ES). Areas of squares correspond to weight given to each study; lines represent 95% confidence intervals (CI). Dotted line is a random effects estimate of average effect ($\Delta = 0.39$); solid line is an effect size of 0. Studies are sorted by the inverse of their standard errors

is estimated using method of moments. The resulting estimate is 0.39, with a 95% confidence interval ranging from 0.231 to 0.554. This pooled estimate of the effect of contact on prejudice suggests that, on average, the contact induced by these experiments reduced prejudice by more than a third of the standard deviation in the control group.

The pooled average, however, glosses over the heterogeneous effects found in the experimental literature. This heterogeneity is illustrated in Figure 2, which displays the results of the meta-analysis in the form of a study-by-study forest plot, where the studies have been sorted by their estimated standard errors. If a single causal parameter were at work in these studies, one would expect 95% of the experiments' confidence intervals to overlap with 0.39. In fact, only 20 of the 27 studies produce 95% confidence intervals that overlap with this pooled estimate ($p < 0.001$). The problem of coverage is especially acute among larger studies, which tend to produce relatively precise estimates: of the 10 studies with the smallest standard errors, four have 95% confidence intervals that fall below the overall meta-analytic average effect.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Table 3. Relationship between standard error and effect size

Δ (effect size)	Coefficient	Standard error	*t*	*P* > |*t*|	95% confidence interval (lower)	95% confidence interval (upper)
Reported standard error	2.086	1.007	2.07	0.049	0.0103	4.16
Intercept	−0.014	0.216	−0.06	0.949	−0.458	0.430

One way to model between-study heterogeneity is to allow effects to vary as a function of their standard errors, on the grounds that publication bias inflates the average effects observed in smaller studies. In this meta-regression model, the slope represents the expected change in effect size when the standard error increases by one unit. The intercept is also of interest, as it represents the expected effect if the standard error were zero (i.e., if the study were of infinite size). The results presented in Table 3 suggest that a one-unit increase in standard error is associated with a 2.09-unit increase in effect size, although the pattern is of borderline significance (two-tailed *p*-value = 0.049). Notably, the intercept is –0.014 with a 95% confidence interval ranging from –0.46 to 0.43. The implication is that a very large study would be expected to produce a minuscule *increase* in prejudice. Like the tests for publication bias presented by Pettigrew and Tropp (2006, p. 758), our results are statistically equivocal. The same may be said for our analysis of *p*-hacking (Simonsohn *et al.*, 2014; Head *et al.*, 2015), which is symptomatic of research discretion that favors significant relationships. These tests may be found in the Online Supplementary Materials.

Another way to model effect heterogeneity is to focus on the targets of prejudice in these studies. Here, we regress effect size on indicator variables for disabilities, gender, LGBT status and age with racial, ethnic, religious and immigrant targets at the base category. In contrast to the rather muted degree of heterogeneity found by Pettigrew and Tropp (2006), we find some prejudices to be much more responsive than others. An *F*-test indicates significant heterogeneity in effects across target groups ($p = 0.01$). Especially significant is the contrast between disability and the base category, where $p = 0.001$. When ethnic, racial, religious and immigrant studies are considered on their own (i.e., the estimated intercept), the estimated effect is 0.25, with a 95% confidence interval ranging from 0.08 to 0.42. This remains a fairly strong and significant pooled effect, although it is subject to the proviso that four of the five largest studies come in below this average.

In sum, meta-analysis offers qualified support for the contact hypothesis. On the one hand, the overwhelming majority of studies report positive effects, and a random-effects model suggests that the true underlying effect is substantively

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

quite large. On the other hand, the collection of studies has three important limitations. First is the gap in coverage. We know little about the effects of contact on adults over 25 years of age. In particular, the meta-analysis furnishes no evidence about contact's effects on adults' racial or ethnic prejudices, which was the original policy-based motivation for this body of work. Second, the larger experiments tend to produce weaker effects, which suggests that a file drawer problem may be concealing smaller studies with more equivocal findings. Finally, effect sizes vary significantly according to the target of prejudice, suggesting that certain kinds of prejudice are more amenable to contact-based remediation.

Robustness check: including seven borderline studies

Many studies fall just short of the selection criteria we used to identify the most policy-relevant research.[8] Some studies, for instance, do not use a fully randomized design, but instead capitalize on quasi-experiments. van Laar *et al.* (2005) studied interracial roommate pairings at University of California, Los Angeles, which we did not include because of uncertainty about randomness of roommate assignment in this particular context. Other studies assign something similar, but not exactly identical, to intergroup contact; Enos (2014), for instance, randomly assigned physical proximity to outgroup members (Mexican nationals living in the USA) without assigning face-to-face interaction, and Fuegen (2000) varied whether an experimental confederate identifying as a feminist displayed stereotype-confirming or -disconfirming behaviors.

When we augment our meta-analysis using seven studies that fall in this category, our results remain largely unchanged. Random-effects meta-analysis renders an overall estimate of 0.373 (standard error = 0.075), which is very similar to what we obtained above. We continue to find significant heterogeneity in effects across target groups ($p = 0.0024$) and marginally significant evidence that treatment effects diminish as studies' standard errors decrease ($p = 0.058$).

Robustness check: studies with pre-analysis plans

Another way to test for the presence of publication bias is to look separately at studies that meet the very highest standards of experimental quality and research transparency. In our sample, three studies – Broockman and Kalla (2016), Scacco and Warren (2018) and Finseraas and Kotsadam (2017) – have pre-

8 For a detailed look at which studies we did not include and why, see Appendix A, 'An overview of excluded studies'.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

analysis plans[9]. As Olken (2015, p. 69) writes, "[f]or readers, referees, editors, and policy-makers, knowing that analysis was pre-specified offers reassurance that the result is not a choice among many plausible alternatives, which can increase confidence in results."

The relationship between effect size and study quality has played a central role in the assessment of contact's effects. Pettigrew and Tropp (2006) contend that "research rigor is routinely associated with larger effect sizes. Put differently, the less rigorous studies sharply reduce the overall relationships observed between contact and prejudice" (p. 759). Revisiting the same theme a decade later, Pettigrew (2016) writes: "the most rigorous studies tend to provide the largest effects. This phenomenon is repeated in 21st-century research. Recent work is more rigorously executed and yields larger contact effects than earlier work" (p. 14). In their meta-analysis of the effects of contact on sexual prejudice, Smith *et al.* (2009) hypothesized that "those studies with higher methodological quality will have more scientific rigor which will produce results with stronger effects than those studies with lower methodological quality" (p. 181).

In our sample, however, we find that studies conforming to the very highest standards of research quality[10] show much smaller effects on average than the sample as a whole. Studies with pre-analysis plans have a random effects estimate of 0.016. The studies without pre-analysis plans, by contrast, have a random effects estimate of 0.451. Given broader discussions of replicability in science (Ioannidis, 2005; Nosek *et al.*, 2015), this divergence is of particular concern for policy-makers interested in assessing the reliability of contact as a policy tool.

Conclusion

In reviewing the contact literature, previous authors have lamented a dearth of high-quality designs. Tucker and Potocky-Tripodi (2006), who reviewed the effects of contact on prejudice against gays and lesbians in the USA, concluded that "[n]o intervention met the criteria of a well established or probably efficacious treatment, as all studies had substantial methodological limitations" (p. 176). Yuker (1994) reviewed studies of discrimination against people

9 Broockman and Kalla (2016) and Scacco and Warren (2018) have publicly accessible code and data as well.

10 While a pre-analysis plan is not a sufficient condition for a high-quality study, it is a strong indicator of a high-quality replication effort; moreover, these three studies are exemplary in other regards, taking place in field settings and featuring some of the smallest standard errors among the studies we analyzed.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

with disabilities and argued that the "general quality of research … is not very high. Many studies suffer from faults such as inadequate sampling, the lack of adequate control groups, failure to randomly assign subjects to groups, the lack of pretests or retrospective pretests, etc." (p. 4). Speaking generally, Hopkins *et al.* (1997) wrote, "the initial hopes of contact theorists have failed to materialize" (p. 306).[11] Pettigrew and Tropp (2006, p. 752) explicitly challenged these earlier literature reviews on the grounds that they assembled and analyzed the literature in an unsystematic manner.

To prepare our review, we spent years attempting to gather all of the contact studies that used high-quality research designs. Our assessment of the policy-relevant contact literature falls somewhere between the pessimistic accounts and Pettigrew and Tropp's (2006) declaration that "meta-analytic results provide substantial evidence that intergroup contact can contribute meaningfully to reductions in prejudice across a broad range of groups and contexts" (p. 766).

On the one hand, the vast majority of these experiments do indeed show positive effects of contact. Of the 27 experimental comparisons that seem most policy relevant, 24 reveal positive effects. The average effect across these experiments is substantively large, diminishing measured prejudice by 0.39 standard deviations. Both results are statistically significant at the 0.001 level, and the inclusion of seven quasi-contact experiments studies does not materially affect the size or significance of the meta-analytic estimates.

On the other hand, five features of the contact literature give us pause. First, the set of policy-relevant studies has important gaps. What we know about prejudice reduction comes largely from studies of children or young adults. Few studies address prejudice in adults over 25 years of age. Notably, no studies of ethnic or racial contact include participants over 25 years of age.

Second, the extent to which contact diminishes prejudice seems to vary according to the target of prejudice. Contact seems to work especially well as a strategy for reducing prejudice toward people with mental or physical disabilities. Prejudice toward individuals with disabilities may differ from other types of prejudice due to the distinctive ways in which disabled people are perceived (Fiske, 2011). When studies involving disabilities are excluded, the meta-analytic estimate remains significant but diminishes to 0.20. This finding suggests a rather different theoretical interpretation from the one offered by Pettigrew and Tropp (2006), who found that "[c]omparisons across the racial and ethnic subsets and the nonracial and nonethnic subsets yield virtually identical mean estimates of contact-prejudice effect sizes"

11 See also McClendon (1974), Riordan (1978), Ford (1986), Stephan (1987) and Patchen (1999).

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

(p. 762). It now appears that some types of prejudice may be more malleable than others, or that some combinations of contact and prejudice mesh especially well.

Third, larger studies – those that estimate the effects of contact with greater precision – tend to reveal weaker effects. Weaker effects are also characteristic of studies that adhere to the highest standards of analytic transparency. Time will tell whether these correlations are statistical flukes or a genuine cause for concern.

Fourth, we know little about what happens within the contact interventions we are assessing. The authors of these research reports rarely describe the contact programs in sufficient detail to allow others to recreate the experience with other populations. In particular, few state explicitly whether their contact intervention meets one or more of Allport's four conditions for reducing prejudice. As a result, we learn little about what specific aspects of the contact are reducing participants' prejudice.

Fifth, and relatedly, no randomized study with over-time outcome measurement has systematically varied, as part of its experimental design, Allport's facilitating conditions. Without manipulating the features of group contact or the conditions under which it occurs, one can only speculate about whether divergent results reflect the treatments, subject pools or conditions of contact (such as equal status or a common goal). For example, one recent study found that prejudice increased when non-Hispanics were exposed to, but did not interact with, randomly assigned confederates speaking Spanish at commuter train stations (Enos, 2014).

Allport did not believe that 'mere contact' would reduce prejudice. Indeed, Allport warned that without moving beyond casual contact into a deeper engagement characterized by the conditions he set forth, "the more contact the more trouble" (1954, p. 263). This prediction stands in direct contrast to Pettigrew and Tropp's conclusion that "Allport's conditions are not essential for intergroup contact to achieve positive outcomes" (2006, p. 766). Given the lack of experiments that systematically test the moderating impact of these conditions on prejudice reduction, we conclude that the literature is not in a place where we can adjudicate between these two positions.

Reinvigorating the study of these moderating conditions means rediscovering innovative experimental designs from decades past. An example of how to experimentally manipulate the conditions of contact comes from Cohen and Roper (1972), who attempted to create an experience of equal status contact between white and black male junior high-school students. In the study, groups of four students, some black and some white, played a strategy game involving cooperation and collective decision-making. The investigators varied study participants' perceptions of outgroup status by preparing them

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

differently in terms of skills and behavioral expectations. While neither this study nor a follow-up replication (Riordan & Ruggiero, 1980) measured prejudice, nor had long-term outcome measures, they make the point that "equal-status contact should not be assumed," but rather experimentally manipulated and tested (Riordan & Ruggiero, 1980, p. 131).

Discovering whether Allport's conditions are important for prejudice reduction is not just a matter of theoretical importance – it is an urgent policy question. Scholars reviewing the contact literature often express skepticism about the feasibility of orchestrating the kinds of high-quality contact that Allport (1954) had prescribed. Dixon *et al.* (2005), for example, lament that contact in "rarefied conditions" may not generalize to "everyday life in divided societies" (p. 697). Amir (1969), meanwhile, writes that

> if most studies have appeared to prove that contact between ethnic groups reduces prejudice, it does not necessarily follow that these results are typical of real social situations. Intergroup contact under the circumstances studied is unfortunately quite rare in actual life, and even when it occurs, it produces only casual interactions rather than intimate acquaintances. (p. 337)

If future research concludes that Allport's conditions are in fact necessary, then policy-makers have a challenging but clear recipe for improving intergroup relations. However, if Allport's conditions are not always necessary, this knowledge could contribute to less expensive interventions that are more readily scalable. Thus, we conclude by renewing Pettigrew and Tropp's call for further investigation of the conditions under which contact reduces prejudice. The contact hypothesis has profound policy implications for the potential benefits of bringing groups together in schools, workplaces and housing. The surge in high-quality research outside the lab and outside the USA brings the policy community closer to answers about the long-term effects of intergroup contact, but important gaps must be addressed before this research can reliably guide future policy decisions.

Acknowledgements

For invaluable research contributions, the authors thank Jason Chin and Kulani Dias. For detailed comments on their meta-analyses, we thank Linda Tropp, Thomas Pettigrew, Gunnar Lemmer and Ulrich Wagner. For helpful comments on an early draft, we thank Ethan Busby, Alex Coppock, Ruth Ditlmann, Jamie Druckman, Al Fang, Nour Kteily, Arnfinn H. Midtbøen, Laura Paler, Alexandra Scacco and Jay Van Bavel. For help assembling and evaluating the contact literature, we thank David Broockman, Lucia Corno, Kristin Davies, Rafaela Dancygier, Greg Duncan, Ryan Enos, Joe Evans, Sarah Gaither, Michael Green, Josh Kalla, David Laitin, Winston Lin, Elizabeth Page-Gould, Gautam

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Rao, Bruce Sacerdote, Anna Schickele, Nicole Shelton, Natalie Shook, James Sidanius, Samuel Sommers, Todd Stinebrickner, Thomas Trail, Colette van Laar, Wolfgang Viechtbauer, Tessa West and Shahar Zaks. All errors are our own. This research was supported by an NSF Grant #1322356 to Elizabeth Levy Paluck.

Supplementary Material

To view supplementary material for this article, please visit https://doi.org/10.1017/bpp.2018.25.

References

Abrams, D. (2010), 'Processes of prejudice: Theory, evidence and intervention', *Human Rights*.

Allport, G. (1954), *The nature of prejudice*, New York, NY: Basic Books.

Amir, Y. (1969), 'Contact hypothesis in ethnic relations', *Psychological Bulletin*, **71**(5): 319.

American Psychological Association (2006), 'Resolution on prejudice, stereotypes, and discrimination', *Council Policy Manual*. Washington, DC: Author.

Barnhardt, S. (2009), 'Near and dear? Evaluating the impact of neighbor diversity on interreligious attitudes', (Unpublished working paper).

Bertrand, M., and E. Duflo, (2017), 'Field experiments on discrimination', In *Handbook of Economic Field Experiments*, Vol. **1** Amsterdam: North Holland Publishing 309–393.

Boisjoly, J., G. J. Duncan, M. Kremer, D. M. Levy and J. Eccles (2006), 'Empathy or antipathy? The impact of diversity', *The American Economic Review*, **96**(5): 1890–1905.

Broockman, D. and J. Kalla (2016), 'Durably reducing transphobia: A field experiment on door-to-door canvassing', *Science*, **352**(6282): 220–224.

Burns, J., L. Corno and E. La Ferrara (2015), 'Interaction, prejudice and performance: Evidence from South Africa', (Working paper).

Camargo, B., R. Stinebrickner and T. Stinebrickner (2010), 'Interracial friendships in college', *Journal of Labor Economics*, **28**(4): 861–892.

Campbell, D. T. and J. C. Stanley (1963), *Experimental and quasi-experimental designs for research*, Chicago, IL: Rand McNally.

Clunies-Ross, G. and K. O'Meara (1989), 'Changing the attitudes of students towards peers with disabilities', *Australian Psychologist*, **24**(2): 273–284.

Cohen, E. G. and S. S. Roper (1972), 'Modification of interracial interaction disability: An application of status characteristic theory', *American Sociological Review*, **37**: 643–657.

Cook, S. W. (1985), 'Experimenting on social issues: The case of school desegregation', *American Psychologist*, **40**(4): 452.

Cooper, H., L. V. Hedges and J. C. Valentine (2009), *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Dessel, A. B. (2008), *Measuring the effects of intergroup dialogues on teachers' attitudes, feelings, and behaviors regarding lesbian, gay, and bisexual students and parents*, (Doctoral thesis). University of Tennessee, Knoxville, TN.

Dessel, A. B. (2010), 'Effects of intergroup dialogue: Public school teachers and sexual orientation prejudice', *Small Group Research*, **41**(5): 556–592.

DiTullio, B. J. (1982), *The effect of employing trainable mentally retarded (TMR) students as workers within the Philadelphia public school system: Attitudes of supervisors and non-handicapped co-workers towards the retarded as a result of contact*, (Doctoral thesis). Temple University, Philadelphia, PA.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Dixon, J., K. Durrheim and C. Tredoux (2005), 'Beyond the optimal contact strategy: A reality check for the contact hypothesis', *American Psychologist*, **60**(7): 697.

Enos, R. D. (2014), 'Causal effect of intergroup contact on exclusionary attitudes', *Proceedings of the National Academy of Sciences*, **111**(10): 3699–3704.

Evans, J. H. (1976), 'Changing attitudes toward disabled persons: An experimental study', *Rehabilitation Counseling Bulletin*, **19**(4): 572–579.

Finseraas, H., A. A. Johnsen, A. Kotsadam and G. Torsvik (2016), 'Exposure to female colleagues breaks the glass ceiling—evidence from a combined vignette and field experiment', *European Economic Review*, **90**: 363–374.

Finseraas, H. and A. Kotsadam (2017). Does personal contact with ethnic minorities affect anti-immigrant sentiments? Evidence from a field experiment. *European Journal of Political Research*, **56**(3), 703–722.

Fiske, S. T. (2011), *Envy up, scorn down: How status divides us*. New York, NY: Russell Sage Foundation.

Ford, W. S. (1986), 'Favorable intergroup contact may not reduce prejudice: Inconclusive journal evidence', *Sociology and Social Research*, **70**(4): 256–258.

Fuegen, K. (2000), 'Comparing paradigms of stereotype change: The case for contact', *Arbor Ciencia Pensamiento Y Cultura*.

Furuto, S. B. and D. M. Furuto (1983), 'The effects of affective and cognitive treatment on attitude change toward ethnic minority groups', *International Journal of Intercultural Relations*, **7** (2): 149–165.

Green, D. P. and J. S. Wong (2009), 'Tolerance and the contact hypothesis: A field experiment', in *The political psychology of democratic citizenship*: 1–23.

Green, S. A. (2014), 'The effects of roommate assignment on racial affect', (Unpublished working paper).

Grutzeck, S. and C. A. Gidycz (1997), 'The effects of a gay and lesbian speaker panel on college students' attitudes and behaviors: The importance of context effects', *Imagination, Cognition and Personality*, **17**(1): 65–81.

Hall, E. P. (1969), *An experimental study of the modification of attitudes toward the mentally retarded*, (Doctoral thesis). University of Alabama, Tuscaloosa, AL.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn and M. D. Jennions (2015), 'The extent and consequences of p-hacking in science', *PLOS Biology*, **13**(3).

Hewstone, M. (2003), 'Intergroup contact: Panacea for prejudice?', *The Psychologist*, **16**(7): 352–353.

Hopkins, N., S. Reicher and M. Levine (1997), 'On the parallels between social cognition and the new racism', *British Journal of Social Psychology*, **36**(3): 305–329.

Hull, W. F. (1972), 'Changes in world-mindedness after a cross-cultural sensitivity group experience', *The Journal of Applied Behavioral Science*, **8**(1): 115–121.

Ioannidis, J. P. (2005), 'Why most published research findings are false', *PLOS Medicine*, **2**(8).

Katz, P. A. and S. R. Zalk (1978)m 'Modification of children's racial attitudes', *Developmental Psychology*, **14**(5): 447–461.

Kelman, H. C. (1998), 'Social-psychological contributions to peacemaking and peacebuilding in the Middle East', *Applied Psychology*, **47**(1): 5–28.

Krahe, B. and C. Altwasser (2006), 'Changing negative attitudes towards persons with physical disabilities: An experimental intervention', *Journal of Community & Applied Social Psychology*, **16**(1): 59–69.

Lazar, A. L., J. T. Gensley and R. E. Orpet (1971), 'Changing attitudes of young mentally gifted children toward handicapped persons', *Exceptional Children*, **37**(8): 600–602.

Lemmer, G. and U. Wagner (2015), 'Can we really reduce ethnic prejudice outside the lab? A meta-analysis of direct and indirect contact interventions', *European Journal of Social Psychology*, **45**(2): 152–168.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Maoz, I. (2010), 'Educating for peace through planned encounters between Jews and Arabs in Israel: A reappraisal of effectiveness', *Handbook on Peace Education*, 303–313.
Markowicz, J. A. (2009), *Intergroup contact experience in dialogues on race groups: Does empathy and an informational identity style help explain prejudice reduction?*, (Doctoral dissertation). The Pennsylvania State University, State College, PA.
Marmaros, D. and B. Sacerdote (2006), 'How do friendships form?', *The Quarterly Journal of Economics*, **121**(1): 79–119.
McClendon, M. J. (1974), 'Interracial contact and the reduction of prejudice', *Sociological Focus*, **7** (4): 47–65.
Meshel, D. S. (1997), *The contact hypothesis and the effects of intergenerational contact on adolescents' attitudes and stereotypes toward older people*, (Doctoral dissertation). Texas Tech University, Lubbock, TX.
Meshel, D. S. and R. P. McGlynn (2004), 'Intergenerational contact, attitudes, and stereotypes of adolsecents and older people', *Educational Gerontology*, **30**(6): 457–479.
Morris, S. B. (2008), 'Estimating effect sizes from pretest-posttest-control group designs', *Organizaitonal Research Methods*, **11**(2): 364–386.
Mussen, P. H. (1950), 'Some personality and social factors related to changes in children's attitudes toward Negroes', *The Journal of Abnormal and Social Psychology*, **45**(3): 423.
Myrdal, G. (1944), *An American dilemma: The Negro problem and modern democracy*, New York, NY: Harper & Bros.
Nosek, B. *et al.* (2015), 'Promoting an open research culture', *Science*, **348**(6242): 1422–1425.
Olken, B. A. (2015), 'Promises and perils of pre-analysis plans', *The Journal of Economic Perspectives*, **29**(3): 61–80.
Page-Gould, E., R. Mendoza-Denton and L. R. Tropp (2008), 'With a little help from my cross-group friend: Reducing anxiety in intergroup contexts through cross-group friendship', *Journal of Personality and Social Psychology*, **95**(5): 1080.
Pagtolun-an, I. G. and J. M. Clair (1986), 'An experimental study of attitudes toward homosexuals', *Deviant Behavior*, **7**(2): 121–135.
Paluck, E. L. and D. P. Green (2009), 'Prejudice reduction: What works? A review and assessment of research and practice', *Annual Review of Psychology*, **60**: 339–367.
Patchen, M. (1999), *Diversity and unity: Relations between racial and ethnic groups*, Chicago, IL: Nelson-Hall Publishers.
Pettigrew, T. F. (1979), 'Tension between the law and social science: An expert witness's view', in *Schools and the courts: Desegregation*, Vol. 1, Eugene, OR: ERIC Clearinghouse for Educational Management, 23–44.
Pettigrew, T. F. (1997), 'Generalized intergroup contact effects on prejudice', *Personality and Social Psychology Bulletin*, **23**(2): 173–185.
Pettigrew, T. F. (2016), 'In pursuit of three theories: Authoritarianism, relative deprivation, and intergroup contact', *Annual Review of Psychology*, **67**: 1–21.
Pettigrew, T. F. and L. R. Tropp (2006), 'A meta-analytic test of intergroup contact theory', *Journal of Personality and Social Psychology*, **90**(5): 751–783.
Pettigrew, T. F., L. R. Tropp, U. Wagner and O. Christ (2011), 'Recent advances in intergroup contact theory', *International Journal of Intercultural Relations*, **35**(3): 271–280.
Riordan, C. (1978), 'Equal-status interracial contact: A review and revision of the concept', *International Journal of Intercultural Relations*, **2**(2): 161–185.
Riordan, C. and J. Ruggiero (1980), 'Producing equal-status interracial interaction: A replication', *Social Psychology Quarterly*, **43**(1): 131–136.
Rosenthal, R. (1979), 'The file drawer problem and tolerance for null results', *Psychological Bulletin*, **86**(3): 638.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25

Sayler, R. I. (1969), *An exploration of race prejudice in college students and interracial contact*, (Doctoral thesis). University of Washington, Seattle, WA.

Scacco, A. and S. S. Warren (2018), 'Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria', *American Political Science Review*, 1–24.

Sheare, J. B. (1974), 'Social acceptance of EMR adolescents in integrated programs', *American Journal of Mental Deficiency*, **78**(6): 678–682.

Simonsohn, U., L. D. Nelson and J. P. Simmons (2014), 'P-curve: A Key to the file-drawer', *Journal of Experimental Psychology: General*, **143**(2): 534.

Smith, F. T. (1943), *An experiment in modifying attitudes toward the Negro*, (Doctoral dissertation). Columbia University, New York, NY.

Smith, S. J., A. M. Axelton and D. A. Saucier (2009), 'The effects of contact on sexual prejudice: A meta-analysis', *Sex Roles*, **61**(3): 178–191.

Sorensen, N. A. (2010), *The road to empathy: Dialogic pathways for engaging diversity and improving intergroup relations*, (Doctoral thesis). University of Michigan, Ann Arbor, MI.

Stephan, W. G. (1987), *The contact hypothesis in intergroup relations*, Thousand Oaks, CA: Sage Publications, Inc.

Tucker, E. W., and M. Potocky-Tripodi (2006), 'Changing heterosexuals' attitudes toward homosexuals: A systematic review of the empirical literature', *Research on Social Work Practice*, **16**(2): 176–190.

van Laar, C., S. Levin, S. Sinclair and J. Sidanius (2005), 'The effect of university roommate contact on ethnic attitudes and behavior', *Journal of Experimental Social Psychology*, **41**(4): 329–345.

Williams, D. (1934), *The effects of an interracial project upon the attitudes of Negro and white girls within the Young Women's Christian Association*, (Unpublished M.A. thesis). Columbia University, New York, NY.

Williams Jr., R. M. (1947), 'The reduction of intergroup tensions: A survey of research on problems of ethnic, racial, and religious group relations', in *Social Science Research Council Bulletin*, Vol. 57, Ann Arbor, MI: University of Michigan.

Yablon, Y. B. (2012), 'Are we preaching to the converted? The role of motivation in understanding the contribution of intergroup encounters', *Journal of Peace Education*, **9**(3): 249–263.

Yuker, H. E. (1994), 'Variables that influence attitudes toward persons with disabilities: Conclusions from the data', *Journal of Social Behavior and Personality*, **9**(5): 3.

Downloaded from https://www.cambridge.org/core. IP address: 71.59.64.243, on 31 Jul 2018 at 14:21:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bpp.2018.25