Discussion Paper No. 1319

ISSN (Print) 0473-453X
ISSN (Online) 2435-0982

REDUCING PREJUDICE: A META-ANALYSIS AND EXPERIMENT ON THE CONTACT HYPOTHESIS

Gwen-Jirō Clochard

August 2026

The Institute of Social and Economic Research
The University of Osaka
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Reducing Prejudice: A Meta-Analysis and Experiment on the Contact Hypothesis

Gwen-Jirō Clochard[1,2,*]

[1]Institute of Social and Economic Research, Osaka University

[2]Joint Initiative for Latin American Experimental Economics

[*]gjclochard@iser.osaka-u.ac.jp

This version: August 2026

Abstract

For decades, intergroup contact has been considered one of the most important tools for reducing prejudice and improving intergroup relations. This paper meta-analyses the experimental literature on the contact hypothesis. The analysis is based on 209 outcomes from 61 papers, involving more than 30,000 people. The main finding is that contact interventions *are* effective in reducing prejudice, with an average effect of approximately 0.3 standard deviation and few instances of backlash. However, I identify several worrying signs of publication bias and p-hacking, although these do not impact the significance of the main result. I also find substantial heterogeneity in contact effects, but the characteristics of contact interventions considered in the literature have limited power to explain this heterogeneity. In particular, the common-goals condition, often emphasized as beneficial, is negatively correlated with effect size, although this relationship is descriptive. An analysis using external participants and Large Language Models also reveals considerable diversity in how contact interventions satisfy the conditions emphasized in the literature. Overall, the findings suggest that while contact interventions can reduce prejudice, the evidence currently provides limited guidance on which types of contact are most effective, highlighting the need for experiments that directly manipulate contact characteristics and for a better theoretical understanding of the mechanisms underlying contact effects.

Keywords: contact hypothesis, meta-analysis, prejudice reduction, field experiments

JEL Codes: C93, C12, C83

1 Introduction

Because of its consequences on social inequality (Durante and Fiske, 2017), xenophobia (Kumar et al., 2011) or reduced economic output (Hjort, 2014), solutions to reduce prejudice have been studied in psychology, sociology and economics for decades. Among candidate solutions, *contact* interventions have received the most attention (Bertrand and Duflo, 2017).

The contact hypothesis posits that interpersonal contact between members of different groups can, under certain conditions, lead to reduced prejudice and improved intergroup relations. Proposed by Gordon Allport in 1954, this theory suggests that direct interaction between individuals from diverse backgrounds can challenge stereotypes, increase understanding, and foster empathy (Allport, 1954). Since Allport, interventions that promote intergroup contact have been seen as one of the main tools for reducing prejudice (Bertrand and Duflo, 2017; Paluck et al., 2021).

The literature examining the contact hypothesis developed rapidly after Allport's book, with the landmark meta-analysis by Pettigrew and Tropp (2006) identifying no fewer than 515 studies, involving more than 250,000 people from 38 countries, over a period spanning from the 1940s to the year 2000. Their overwhelming conclusion was that contact is effective in reducing prejudice, noting that "results from the meta-analysis conclusively show that intergroup contact can promote reductions in intergroup prejudice" (p.751).

However, only a small fraction of the 515 studies analyzed in the meta-analysis used an experimental design, with a randomized assignment to treatment and control groups. After removing non-experimental protocols and adding more recent experimental studies, Paluck et al. (2019) analyze 27 studies investigating the effects of contact. Of these, the largest proportion (33%) investigated the effect of contact on racial and ethnic prejudice among university students or young adults (18-25 years old), and only two were conducted in developing countries - Scacco and Warren (2018) in Nigeria and Corno et al. (2022)

in South Africa.

The purpose of this paper is threefold. First, I deepen the work by Paluck et al. (2019) and Lowe (2025) by investigating which characteristics of interventions are associated with larger effects, by examining whether contexts of interventions (prejudice targeted, type of population, region of the world, etc.) correlate with effect size. I also use multiple outcomes per paper, enabling a better understanding of the effects. For example, I distinguish, within paper, the target of the outcome–does the outcome about the entire outgroup or the specific individuals met, sometimes called the generalizability of contact (Lowe, 2025)–or the timing of measurement–right after the contact or several months to years later. This method allows to control for statistical dependence within paper (following Imai et al. (2021)).

Second, I analyze which contact conditions are important for the effect of contact. Allport (1954) identified three conditions of effectiveness of contact interventions - equal status between groups, common goals, and the support of authorities. Subsequent work (Pettigrew and Tropp, 2006; Lemmer and Wagner, 2015) also identified positive contact and friendship potential of the encounter as a potential moderator of the effect of contact. However, with the exception of Lowe (2021), no experimental test of the relevance of these hypotheses has been conducted (Paluck et al., 2019).[1] Because deciding whether papers meet the conditions is highly subjective, I supplement my own coding of the corresponding variables with two additional approaches. First, I use an online experiment in which participants are incentivized to report second order beliefs about whether they think that papers meet the conditions, when given the design of the experiment. Second, I use a Large Language Model, prompting with the same text as the online experiment (ChatGPT-5.3).

Third, this paper expands the analysis of Paluck et al. (2019) by incorporating a significant number of more recent papers, in particular with more robust designs (pre-registration, larger sample size). Indeed,

[1] Greene et al. (2025) also provides evidence of the role of equal status. The experiment was not included in the present meta-analysis because the interaction was online, though. Moreover, it can be argued that randomizing *only* one characteristic at a time is almost impossible, as some other characteristics would be changed too.

the number of papers on the topic of contact has more than doubled since this meta-analysis.

The analysis sample consists of 209 outcome measures from 61 experiments (both registered and non-registered) conducted in almost all regions of the world (no experiment in Latin America could be found) and involving more than 30,000 people. There exists a large heterogeneity in the prejudices studied - the modal prejudice being racial/ethnic/caste but also covering e.g., age, immigration or religion, contact type - discussions, sports, army, etc.- or length of intervention (from one day to 2.5 years). There is also a great deal of heterogeneity in the magnitude of the effects.

Following Imai et al. (2021), I start the analysis with the estimation of the meta-analytic average, using a random-effect model, with robust standard errors clustered at the paper level (Hedges et al., 2010). I then follow the three-level meta-analysis method (Van den Noortgate et al., 2013) to account for potentially statistically dependent estimates within paper.

After this analysis, I evaluate the potential threat of publication bias, whereby papers that report stronger treatment effects might be more likely to be written-up and subsequently published (Andrews and Kasy, 2019; Brodeur et al., 2016, 2020; Chopra et al., 2024) and p-hacking, i.e., the tendency for papers to selectively report models based on p-values. To study publication bias, I use the Limit Meta-Analysis method (Rücker et al., 2011), which models the dependence between effect size and its precision. To study p-hacking, I use the Right Truncated Meta-Analysis method (Mathur, 2024).

I then study drivers of heterogeneity, first by estimating if each paper characteristic (including contact conditions) moderate the magnitude of the treatment effect. Finally, I perform a lasso estimation on all the characteristics of the experiments to identify the main correlates of effect size.

The meta-analysis has five main findings. The first finding is positive, in that, overall, contact interventions *are* effective in reducing prejudice. The point estimate of the meta-analytic average for the three-level model is 0.292 standard deviations (95% CI: $[0.214, 0.371]$), and highly statistically signifi-

cant ($p < 0.001$).[2] In a standard trust game, this corresponds to an increase of approximately one token out of 10. Importantly, very few papers find significant negative effects of contact. I also find that contact is found to be effective both when the outcome is related to the individuals specifically met and when it is related to the entire outgroup.

The second finding, however, is a cautionary one, as I find several indicators of potential publication bias and p-hacking in the contact literature. Indeed, using Rücker et al. (2011)'s Limit Meta-Analysis method, I find that studies with more precise estimates tend to have smaller treatment effects. This finding is negative, but when correcting for potential publication bias, the net effect of contact interventions remains statistically significant, although the magnitude is marginally smaller. Similarly to Lowe (2025), I also find that pre-registered experiments find significantly smaller effects. I further find evidence of bunching of z-score just on the right of the 1.96 threshold, indicating potential p-hacking (Simonsohn et al., 2014). Using Mathur (2024)'s Right-Truncated Meta-Analysis method, I find again that the corrected estimate, although smaller in magnitude, is still significantly positive. Thus, despite threats of publication bias and p-hacking, contact interventions cause a reduction in prejudice.

The third finding is related to the online experiment that was conducted to determine whether the papers meet the contact conditions set forth in the literature. First, I find that no paper fully satisfies any condition, and that there is substantial variation for each condition across papers. Second, I find that the results from the experiment are very consistent, in that participants are quite good at predicting what others are reporting.

The fourth finding is that some of the contact intervention conditions identified by Allport (1954) and Pettigrew and Tropp (2006) are correlated with the magnitude of the effect size. Consistent with expectations from the literature, having a more positive encounter and stronger support of authorities are associated with more positive effect sizes, while sharing an equal status and friendship potential are not

[2]Restricting to registered studies, as done by Lowe (2025), yields an estimate of 0.152 [0.082, 0.222].

significantly correlated with effect sizes. However, I find that the common goals condition is *negatively* correlated with effect size. This finding is new to the literature and requires further analysis.

The fifth and final finding is that the analysis of heterogeneity has a limited predictive power. Indeed, while the negative coefficient for the common goals condition is one of the only five variables selected by a lasso penalized regression of the contact effects, the results of the post-lasso estimation indicate that the best prediction captures only a limited fraction of the variance ($R^2 < 0.30$). This result suggests that future work is still needed to better understand the determinants of heterogeneity of contact interventions.

This paper contributes to the contact hypothesis literature in three ways. First, I update the meta-analysis of Paluck et al. (2019) and include a large number of papers published since then. Furthermore, I allow for the inclusion of multiple outcomes per paper, increasing the number of observations and the possibility to study heterogeneity (some papers measure several outcomes at the same time, others evaluate the lasting effects of contact, etc.). Similar to their main finding, I find that contact interventions *are* effective at reducing prejudice and improving intergroup relations. I also find substantial heterogeneity in the effect of contact. I also further the recent meta-analysis by Lowe (2025) by studying both registered and non-registered experiments, in the context of in-person contact. Importantly, while Lowe (2025) focuses on the magnitude of the average effect of contact in the most credible subset of the experimental literature (by studying 41 pre-registered experiments and their pre-registered primary outcomes, but including online contacts), this paper asks why contact effects differ across interventions, contexts and measurements. In this respect, the paper complements Lowe's analysis by using the heterogeneity in the broader literature to investigate the characteristics associated with larger or smaller contact effects. Finally, consistent with the comparison between Paluck et al. (2019) and Lowe (2025), I find that registered experiments tend to report smaller effects than non-registered experiments, although the estimated effect remains positive and statistically significant.

Second, I present an in-depth analysis of the determinants of the heterogeneity of contact effects. As previously highlighted in the literature, Lowe (2021) is the only paper to experimentally vary the assignment of one condition (in this case, common goals)[3], and an in-depth analysis of the drivers of the effects of contact is missing (Paluck et al., 2019), particularly for the evaluation of the conditions identified in Allport (1954). This paper contributes by extensively studying whether differences in context, nature of the interventions or nature of the measurement explain part of the heterogeneity in the treatment effects of contact.

Third, I propose a methodology for evaluating, with the help of external reviewers, whether the experimental design used in papers satisfies Allport (1954)'s conditions. In the literature, most papers claim that their experimental design satisfies the conditions, usually without much discussion to support this claim. The results of the online experiment I conducted, which are internally consistent, actually show that no paper fully satisfies the contact conditions. The exercise of correlating the conditions and effect size is descriptive, and no causal claim can be made. For instance, a possible explanation for the observed negative correlation between effect size and common goals could be that the only papers with limited common goals that end up being written up are those with stronger effects.

Finally, this paper contributes to the literature on meta-analyses in the social sciences (see for example Brodeur et al. (2020); Imai et al. (2021); Brown et al. (2024); Chopra et al. (2024); Irsova et al. (2025); Opatrny et al. (2026)) by using external observers to reduce subjectivity in the definition of variables. To the best of my knowledge, the only paper who used such outside observers in the context of meta-analyses is the "red team" of Schaerer et al. (2023), composed of independent researchers and laypeople, recruited to forecast the evolution of gender bias in hiring decisions. In this paper, I use outside observers not to predict the outcomes of the experiments, but to define the variables of interest of the

[3]Greene et al. (2023) also randomizes equal status among participants. This paper is not included in the meta-analysis though, as the contact takes place online.

interventions themselves, to help in the analysis of the determinants of heterogeneity. I find that external evaluators' measures are strongly correlated with my own and LLMs', indicating that, although these conditions are difficult to comprehend (Pettigrew and Tropp, 2006), a ground truth can be found using this method. Interestingly, using several evaluators could reduce noise in evaluations, thereby reducing attenuation bias (Hausman, 2001), which might explain why metrics from the experiment correlate more with the effect size than my assessments or ChatGPT's.

The overall effect of contact interventions is positive, but the implications for the implementability of contact interventions remain speculative. First, the wide variety of protocols identified as "contact" makes clear predictions about the potential effectiveness of any particular intervention - say, playing sports with immigrants - relatively uncertain. Second, there is a general lack of discussion on the external validity of the papers,[4] in particular due to the lack of representativeness of the subject pool or the outgroup they encounter (List, 2022). Third, and potentially very important for policy, none of the papers identified in this meta-analysis perform a welfare analysis of the interventions, either in terms of their monetary values (the wealth created by increased interactions) or in terms of the return in prejudice reduction per dollar invested in the intervention.

After presenting the results, I discuss in detail the potential limitations that need to be addressed before contact interventions can be reliably used as a policy tool to reduce prejudice.

The remainder of the paper is organized as follows. Section 2 presents the method of selection of the papers in this analysis, describes the variables of interest and the experimental design to determine whether the contact conditions were fulfilled, as well as the empirical strategy. Section 3 presents the descriptive statistics. Section 4 presents the results of the analysis of whether contact is effective at reducing prejudice and evaluates publication bias and p-hacking. Section 5 analyses the drivers of heterogeneity

[4] Elwert et al. (2023), for example, state that "if the positive effects of interethnic contact on interethnic relations hinged on coresidence, the policy potential of contact theory would be limited."

in the effect of contact on prejudice. Lastly, Section 6 discusses the implications of the findings and concludes on future for research on the contact hypothesis.

2 Data

In this section I present the paper selection for the meta-analysis as well as the coding of variables of interests for papers.

2.1 Paper Search and Selection

To select papers, I systematically searched Google Scholar and Web of Science in January 2026, using the software Publish or Perish (Harzing, 2007). The search was performed using the terms "Experiment" AND ("Contact hypothesis" OR "Prejudice"). This systematic search yielded 980 papers, after removing duplicates. I also included papers that were included in two previous meta-analyses (Lemmer and Wagner, 2015; Paluck et al., 2019), as well as papers that cited these meta-analyses. I added papers that cited Allport (1954) since 2015 (previous papers would have been included in the previous meta-analyses). Lastly, I included papers that cited the working-paper version of the present paper.

The paper search yielded a total of 1,048 papers. After removing obviously irrelevant papers (e.g., "Oxytocin promotes human ethnocentrism") and papers that were not about contact (approximately 500 papers), I applied the following criteria for inclusion.

The first criterion was to look for empirical papers, which excluded purely theoretical papers, as well as reviews. The second criterion was to select only original research papers, which removed book chapters, policy briefs, and theses. Third, I only focus on real groups, as opposed to minimal groups (Turner, 1978), to avoid the issue of high levels of discrimination in artificial groups (Lane, 2016) and thus limit the potential for experimenter demand (Zizzo, 2010; Lenz and Mittlaender, 2022).[5] Furthermore, using

[5]This distinction only removed one paper that passed the other criteria (Whitt et al., 2021).

real groups gives more policy relevance to study contact as a policy tool to reduce "real" prejudice.

Fourth, I only included papers that used an experimental (i.e., randomized) creation of contact. The experimental criterion was one of the main reasons for excluding papers from the analysis (approximately 200 papers). I include experiments that use either contact-only effects (e.g., White vs Black roommates) and bundled contact effects (where the comparison is, e.g., heterogeneous group meeting vs no meeting). This criterion implies that studies involving quasi-experimental variations (Vertier and Viskanic, 2018; Rao, 2019; Steinmayr, 2021; Bursztyn et al., 2024) were not included in the analysis. Studies with no random assignment at all (Alesina et al., 2003; Danckert et al., 2017) were also excluded from the analysis. This inclusion criterion improves comparability across papers.

Fifth, I only included papers that induced in-person contact. This criterion excluded a growing number of studies (particularly since the COVID-19 pandemic) involving online encounters, such as Lenz and Mittlaender (2022).[6] Two other studies, Gaikwad et al. (2025) and Green and Hyman-Metzger (2025), were removed because the treatment involved increasing the likelihood of being in contact with another group (increasing the likelihood of being recruited for international work in Gaikwad et al. (2025) and likelihood of interacting with diverse squadrons in Green and Hyman-Metzger (2025)) but not directly inducing contact. To ensure measurement consistency (effect of contact rather than intent-to-treat of facilitation of contact), I decided not to include these papers.

Lastly, I excluded three papers because the authors used a randomized method to allocate treatment, but did not follow the control group.

After the application of the inclusion criteria (see summary in Figure A.1), I was left with 61 papers, published between 1972 and 2026, spanning nearly all continents and covering, in total, more than 30,000 individuals. The full list of papers is presented in Table A.1.[7]

[6]For a meta-analysis of contact in online contexts, see Imperato et al. (2021), who find a positive effect of online contact with outgroup members.

[7]The overlap with previous analyses by Paluck et al. (2019) and Lowe (2025) is presented in Table A.2.

Below, I detail how each paper and outcomes were coded. Descriptive statistics are presented in the following Section.

2.2 Paper coding

The first set of variables for each paper are objective metrics from the papers, detailed below.

Variables on papers For the variables on papers, I define four variables of interest, which I categorize as follows.

Publication year: I split the sample in four categories: before 2000, between 2001 and 2010, between 2011 and 2020, and after 2021. For correlations, I also use the linear and quadratic years of publication.

Publication status: I coded a dummy variable for whether the paper is published, as of February 15, 2026.

Pre-registration status: I coded a dummy variable for whether I could find a pre-registration for the paper. To code this variable, I searched for the information on four sources: the papers themselves, the AEA RCT registry, the OSF registry and the authors' personal websites. Including both pre-registered and non-pre-registered papers, contrary to Lowe (2025), enables to evaluate if registered papers find smaller effects.

Variables on contexts For the variables on contexts, I define five variables of interest.

Sample size: I used four categories: [0,50], [51,100], [101,500] and 501+. This definition is arbitrary but is used to illustrate the general trend between sample size and effect size. For the correlation analyses, I directly use the number of observations as a linear predictor.

Mean age: The variable uses the average age of participants provided by the paper. Three main categories were identified, 0-18 years old, 18-25 and 25+. These categories can be broadly thought of as

corresponding to "Children to high-school students", "University students or young adults" and "Adult population". For the correlation analyses, I additionally use the average age and its quadratic term.

Region: I divided papers according to geographical areas. This category includes (alphabetically) East Asia and Oceania, Europe, Middle East and North Africa, North America, South Asia and Sub-Saharan Africa. No experiment in Latin America could be found.

Type of population: This categorical variable is used to describe the targeted population, with a division into five categories: University students and young adults, children and adolescents, army recruits, specific adults (such as shop keepers in Gu et al. (2019) or recent university graduates in Okunogbe (2024)), and general adults.

Type of prejudice: This categorical variable captures the dimension of prejudice targeted by the contact intervention in the paper. The variable is divided in eight categories: age, disabilities, gender, immigrants, LGBTQ+, politics, race / ethnicity / caste, religion and others.

Variables on interventions For variables on the interventions, I define three variables of interest.

Type of intervention: The variable is the type of contact intervention used, divided in height categories: discussions, cooperative games, sports, classmates, neighbors, roommates, army soldiers and others.

Number of encounters: Divided into three categories: 1, 2-10 and 11+. Contrary to other variables, this category cannot be precisely (numerically) identified. Indeed, while for some protocols the number of interactions is clear (e.g., in Page-Gould et al. (2008) the pairs met three times), it is much less the case in experiments involving neighbors, army squadrons or roommates, for example. Furthermore, it is likely that the number of encounters might differ even within a single paper for such interventions -some neighbors might meet every day, others less so, even within the same experiment. For these reasons, I only use this categorical variable for the analysis.

Duration of contact: The variable is defined as the time duration between the start and the end of the contact (in days) and is categorized as 1 (typically one shot intervention), 2-30 and 31+. For correlation analyses, I use the linear and quadratic number of days of the interaction.

Variables on outcomes For all selected papers, I included in the data all the main outcome variables included in the articles, as well as a different measure for different treatments (that involve contact). This departs from collapsing multiple point estimates into a single effect size index for each paper, as was done by Paluck et al. (2019) and Lowe (2025), for example.[8] This distinction was made because some papers have outcomes that measure conceptually different parameters: for instance, in Clochard (2025), I examine the effect of contact separately on trust in the specific police officers met, but also on trust in the police in general; other papers report similar measures at different points in time. For all papers, I prioritized the use of mean differences between the treatment and control groups (as opposed to linear effects of meeting additional outgroup members). This left a total of 209 measures from the 61 papers.

I additionally define three variables of interest for the type of outcome measure.

Type of outcome: This variable defines the type of outcome used in the paper. This variable is divided into four categories. The first category is explicit survey measures, usually in the form of a series of Likert scales. This typically involves participants to declare whether they agree with a pre-defined set of statements explicitly about the other group, e.g. "Affirmative action in college admissions should be abolished" (Boisjoly et al., 2006), "Disabled people are often grumpy and moan about everything" (Krahe and Altwasser, 2006). The second category involves experimental games, such as trust or dictator games (Finseraas et al., 2019; Clochard et al., 2026). The third category is actions measured in the field, such as marrying an outgroup member (Okunogbe, 2024), trading goods (Lowe, 2021) or choosing a roommate for future years in college (Camargo et al., 2010). The fourth and final category is implicit

[8] As a robustness check, I also perform an analysis by collapsing outcomes at the paper level, to give similar weights to all papers. Results are unchanged.

behavior, measured using the Implicit Association Test (Greenwald et al., 1998), such as in Barnhardt (2009), Page-Gould et al. (2008) and Corno et al. (2022).

Entire group: This variable is a dummy for whether the outcome is directed at the outgroup in general (e.g., "Disabled people are often grumpy and moan about everything" in Krahe and Altwasser (2006)) or toward the people the participants met specifically (e.g., playing a trust game with the person met in Clochard et al. (2026) or electing the best players among their peers in Mousa (2020)).

Time since contact: This variable is coded as the distance (in days) between the end of the contact and the measurement. It is categorized as 0 (immediately after the intervention, or while the intervention is still taking place), 1-30 days and 31+ days. For correlation analyses, I use the linear and quadratic terms of the variable (not the categorical).

Effect size For all outcomes in the analysis, the effect size was normalized using Cohen (1969)'s d statistic ($d = \frac{Effect\ Size}{Standard\ Deviation}$). If the statistic was directly available (with its associated standard error), I used it as is. If it was not available, I used the treatment effect along with descriptive statistics to create the normalized effect size. Finally, if the treatment effect was not provided, I used descriptive statistics of the control and treatment groups to create the d-statistic. The standard error was coded in the same way.

The variable was coded so that the effect is positive if contact improves intergroup perceptions (increased trust, more outgroup friends, etc).

2.3 Measurement of contact conditions

In his seminal book, Gordon Allport identified conditions that would be linked to stronger effects of contact, stating 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." (Allport, 1954, p. 281). The literature typically divides these conditions in three categories: common goals, equal status, and support of authorities.[9]

In their groundbreaking review, Pettigrew and Tropp (2006) find that these conditions do not appear to be either necessary nor sufficient for contact to be effective. The literature further suggests that "positive contact" and "friendship potential" should be added as conditions to explain effectiveness of contact interventions.

However, very little experimental evidence exists on whether these scope conditions are actually relevant (Paluck et al., 2019). One noTable exception is Lowe (2021) who randomizes whether participants engage in collaborative or adversarial contact by having them play in different cricket teams. And even in this example, it is very hard to argue that one is manipulating *only* one condition - for instance, one could say that having adversarial contact also reduces the positivity and the friendship potential of the contact. Furthermore, the extent to which each paper fulfills each condition is very subjective.

In this paper, I contribute by investigating the relative importance of each of these conditions in explaining the magnitude of the effect of contact. The extent to which papers fulfill the conditions is subjective (Pettigrew and Tropp, 2006), and using my own evaluation alone might be biased, because of the knowledge of the results of the experiments and the theoretical predictions. For these reasons, I used three methods to code variables related to how much experimental protocols fulfill Allport (1954)'s conditions.

First, I designed an online experiment on Prolific in which I asked participants to rate how much they felt the experimental designs fulfilled the conditions. The experiment was conducted as follows.

[9]The literature sometimes differentiates the "common goals" and "intergroup cooperation". In this review, I combine the two notions.

Before the experiment, I selected extracts from the "Experimental Design / Methods" section of all the papers that describe how participants interacted (see examples in Appendix B). The experiment was then composed of two stages. First, a group of Prolific users were invited to answer how much they felt the papers fulfilled the condition, on a scale from 0 to 100. Each participant evaluated 10 papers. Each participant received a participation-fee of US$4 and three participants were selected at random to earn a bonus of US$ 25.

For the second stage of the experiment, another group of Prolific users were asked to guess what the average of responses was for all first-stage participants. Their participation was incentivized with a performance-based lottery (Evans et al., 1988; Camilleri et al., 2023) to earn a US$ 25 bonus if they were close to the true value. Specifically, for a paper to evaluate, the number of tokens earned by the participant was the maximum of $\{100 - 10 \times \text{Difference}, 0\}$, where Difference is the difference between the average from the first experiment and the participant's guess. One paper was randomly selected per participant. At the end of the experiment, the number of tokens is converted into a lottery ticket, and one ticket was selected for every 10 participants. This scheme is incentive compatible, as the best response for participants is to report their true belief. A screenshot of the experiment is displayed in Figure B.1.

Second, I coded the variables myself. Third, I used ChatGPT 5.3 to code the variables, using the same questions as for the Prolific experiment (see prompts in Table B.1).[10]

For the analysis, I therefore defined five variables of interest at the paper level,[11] *common goals*, *Equal status*, *Positive contact*, *Support of authorities* and *Friendship potential* as the average response across participants in the second stage of the experiment.

[10] Another solution would have been to use a survey of experts to assess the validity. However, as for my evaluations, reports from authors of these papers could be biased. As an illustration, nearly all papers in the database claimed to satisfy Allport's conditions, meaning that the measurement would likely be heavily biased.

[11] The personal contact in all the experiments is by definition fulfilled, as the papers were selected to include in person contact.

3 Descriptive statistics

This Section presents the descriptive statistics of all papers. The complete data sets, variable dictionary, and code are accessible on the Harvard Dataverse, accessible here.

3.1 Descriptive statistics: Papers and contexts

The first clear trend is that the number of experiments investigating the effects of contact has increased substantially over the period. While a few papers were published before 2000, a cumulative distribution of the publication date (Figure C.1) shows that the bulk of the experimental contact literature is recent, with more than 70% of papers published after 2010. More than 80% of the papers (51/61) in the sample are published, with the majority of unpublished papers being recent (thus, likely to be published in the future).

Second, most experiments found on the contact hypothesis were not pre-registered (less than 40% of the sample), although a clear shift has occurred in recent years, with more than 70% of experiments post-2020 being pre-registered (Table C.1). This later trend is a clear departure from previous literature, as Paluck et al. (2019) could only identify two pre-registered papers for their analysis.

Table 1 shows descriptive statistics on the contexts and the interventions. A clear characteristic of the literature, previously identified by Lemmer and Wagner (2015); Paluck et al. (2019); Lowe (2025), is the geographical concentration, with a large fraction of the experiments conducted in North America (especially the US), and no papers found in Latin America. Regarding population and age, a very large proportion of the interventions were conducted on young samples, from children to university students, although more recent papers focused more on general adult populations. For the type of prejudice studied, the mode by far is race / ethnicity / caste, with a significant number of studies on interreligious prejudice and prejudice against immigrants.

Table 1: Descriptive statistics of experiments

Region	*N*	*Type of population*	*N*
East Asia / Oceania	3	University students / Young adults	22
Europe	12	Children / adolescents	13
Middle East and North Africa	8	Army recruits	5
North America	24	Specific adults	6
South Asia	5	General adults	15
Sub-Saharan Africa	9		
Mean age		*Sample size*	
0-17	14	[0, 50]	5
18-25	30	[51, 100]	10
26+	17	[101, 500]	23
		501+	23
Type of prejudice		*Type of intervention*	
Age	2	Army	5
Disabilities	2	Classmates	10
Gender	2	Cooperative games	8
Immigrants	9	Discussions	19
LGBTQ+	3	Neighbors	3
Politics	2	Roommates	7
Race / Ethnicity / Caste	25	Sports	2
Religion	8	Others	7
Others	8		
Number of encounters		*Duration of contact (days)*	
1	18	1	18
2-10	13	2-30	14
11+	30	31+	29

Regarding the type of intervention, I find substantial heterogeneity. The most common intervention typically involves short discussions (Broockman and Kalla, 2016; Clochard, 2025). The army, sports teams and roommates also provide types of interventions which have been studied by several papers. There is also variety of duration and number of encounters, with some contacts lasting for a long time - e.g. roommates sharing a room for the entire first year of university (Boisjoly et al., 2006; Corno et al., 2022) - and some being very short (Page-Gould et al., 2008; Mae Boag and Wilson, 2014).

Lastly, contrary to what was found in the broader prejudice-reduction literature (Paluck et al., 2021), the samples for the experiments on contact are relatively large, with more than a third of papers having a sample size larger than 500 subjects.

Table 2: Descriptive statistics of outcome measures

Type of outcome	
Survey measure	152
Experimental Game	21
Action (in the field)	25
Implicit Association Test	11
Entire group or specific individual	
Entire outgroup	180
Specifically met	29
Time since end of contact intervention (in days)	
0	117
1-30	48
31+	44

Table 2 shows the characteristics of the outcomes. The modal metric used (70% of the outcomes) is a survey to measure explicit attitudes. Interestingly, most measures focus on the entire outgroup, rather than the individuals specifically concerned by the contact (team members, roommates, etc.). Finally, while a large fraction of the literature investigated the immediate effects of contact (56%), there is a large number of studies investigating the lasting effect of contact across months and even years (Camargo et al., 2010; Okunogbe, 2024).

3.2 Descriptive statistics: Contact conditions

Table 3: Descriptive of contact conditions from experimental sample

Variable	N	Experiment		Author		ChatGPT	
		Mean	SD	Mean	SD	Mean	SD
		(1)	(2)	(3)	(4)	(5)	(6)
Common goal	61	68.39	(10.56)	72.79	(13.89)	67.05	(20.11)
Equal status	61	59.66	(10.12)	76.97	(16.23)	65.98	(14.91)
Positive encounter	61	67.68	(8.92)	73.85	(9.06)	70.90	(14.19)
Support of authorities	61	69.21	(8.34)	78.11	(13.91)	78.36	(15.65)
Friendship potential	61	62.01	(12.23)	71.39	(17.61)	60.25	(16.79)

Note: This table presents mean and standard deviation (between papers) of the degree of fulfilment of Allport's conditions. All variables are defined on scales 0-100, 100 meaning complete fulfillment of the condition. Columns 1 and 2 presents the averages from the experiment conducted on Prolific. Columns 3 and 4 presents the values based on the author's own calculations. Columns 5 and 6 presents the values from ChatGPT.

Table 3 shows the average response for each of the contact condition put forward by Allport (1954). While most papers claim that they fulfill all these conditions,[12] it is interesting to see that no paper has an average guess above 90 percent, meaning that no paper completely fulfills any of the conditions, according to the Prolific participants, ChatGPT or my evaluation. Nevertheless, on average Prolific participants believe that papers fulfill approximately two thirds of the contact conditions. There is also substantial variation between papers (Figures D.1a to D.1e). It is also noteworthy that conditions appear to be very correlated with each other for a given paper (Table D.1). Importantly, the three different types of measures of fulfillment (from the experiment, from the author and the LLM) are highly consistent, in that they are highly significantly correlated (Tables D.2 to D.4).

4 Is Contact Effective?

In this section are presented the main results from the meta-analysis, i.e. whether contact interventions are effective at reducing prejudice. I also investigate the potential for publication bias and

[12] A word search in papers reveals that all but 8 papers claim to fulfill at least one condition.

p-hacking.

4.1 Meta-analytic average

To study the meta-analytic average effect of contact interventions in the context of multiple, potentially statistically dependent measures per paper, I follow the methodology by Imai et al. (2021). I start with estimating a random-effects model (DerSimonian and Laird, 1986):

$$d_i = \overline{d}_i + \varepsilon_i = \bar{d} + \mu_i + \varepsilon_i, \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, se_i^2)$ is a sampling variation of d_i as an estimate of $\overline{d}_i$, and the observation-specific "true" treatment effect $\overline{d}_i$ is decomposed into $\bar{d}$ (the overall mean) and the variation μ_i. It is further assumed that $\mu_i \sim \mathcal{N}(0, \tau^2)$, where τ^2 is the heterogeneity in treatment effects. I weigh the estimates with the weights given by $w_i = 1/(se_i^2 + \hat{\tau}^2)$, with cluster-robust variance estimation, which helps dealing with within-study dependence (Hedges et al., 2010), and residual maximum likelihood (REML) to account for small-sample bias (Piepho et al., 2024).

I then apply the three-level meta-analysis method (Van den Noortgate et al., 2013; Imai et al., 2021) to handle statistically dependent estimates. Denoting $d_{i,j}$ the j-th treatment effect from paper i, this method decomposes the heterogeneity into three components: a sampling error $\epsilon_{i,j} (\sim \mathcal{N}(0, se_{i,j}^2))$, a within-paper variation $\mu_{i,j}^{within} (\sim \mathcal{N}(0, \tau_{within}^2))$ and a between-paper variation $\mu_i^{between} (\sim \mathcal{N}(0, \tau_{between}^2))$. The three levels (sampling, within-paper and between-paper) are then combined as Equation 2, where the meta-analytic average is $\bar{d}$.

$$d_{i,j} = \bar{d} + \mu_{i,j}^{within} + \mu_i^{between} + \varepsilon_{i,j} \tag{2}$$

Results of the random-effects model (Equation 1) and the multi-level model (Equation 2) are presented in Table 4.[13] To help visualization with multiple outcomes per paper, I follow Popova et al. (2025) to display the distribution of the treatment effects in Figure 1.[14]

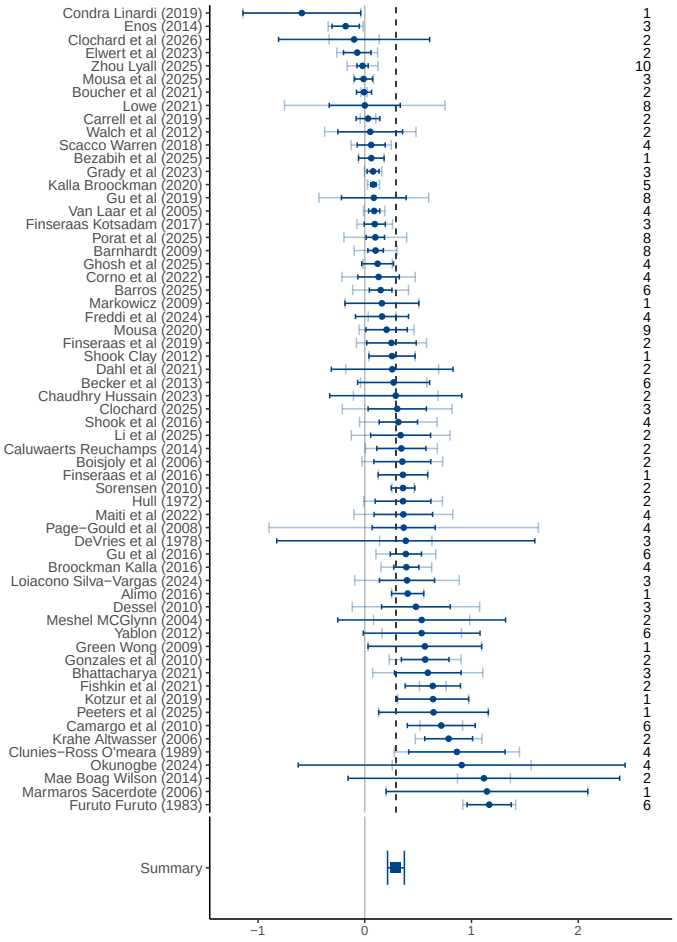


Figure 1: Modified forest plot of paper-level contact estimates

Note: In the top panel, each dot represents the pooled estimate from a paper (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). Numbers on the right denote the number of outcomes in each paper. The bottom panel shows the average effect of contact (and 95% CI) estimated using a multi-level model. Visualization follows Popova et al. (2025).

The main result from this paper is that contact *is* effective at reducing prejudice. The analysis of the random-effect model (Column 1) reveals that the average effect of contact is positive, with a magnitude

[13] As a robustness check, I also collapse the multiple outcomes for a given paper into a single index, as in Lowe (2025). The results show no difference with the main estimates (Table E.1).

[14] A raw forest plot is also displayed in Figure E.1.

Table 4: Meta-analytic average effect of contact

	Random Effect (1)	Multi-Level (2)
$\bar{d}$	0.275***	0.292***
SE	(0.047)	(0.039)
95% CI	[0.181, 0.369]	[0.214, 0.371]
τ^2	0.132	
I^2	96.512	
τ^2_{within}		0.062
$\tau^2_{between}$		0.059
I^2_{within}		49.434
$I^2_{between}$		46.784
Num. clusters	61	61
Num. obs	209	209

Note: The dependent variable is the standardized effect size. Column 1 presents the results of the random-effect model (Equation 1). Column 2 presents the results of the multi-level model (Equation 2). * p < 0.05, ** p < 0.01, *** p < 0.005.

of $\bar{d} = 0.275$ (95% CI: $[0.181, 0.369]$). This measure is highly significant ($p < 0.001$). The average effect corresponds to approximately 1 extra token (out of 10) to be sent in a trust game, using Banerjee et al. (2021)'s benchmark of a standard deviation of 3.

To study heterogeneity, I use Higgins and Thompson (2002)'s I^2 statistic, which is defined as $I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2+s^2} \times 100$, where $\hat{\tau}^2$ is the estimated value of τ^2 and $s^2 = \frac{(n-1)\sum_{i=1}^{n} w_i}{(\sum_{i=1}^{n} w_i)^2+\sum_{i=1}^{n} w_i^2}$ is the "typical" sampling variance of the observed effect sizes, where $w_i = 1/se_i^2$ and n is the number of observations. I find that more than 95% of the total variability in estimates is due to between-observation heterogeneity.

The multi-level model (Column 2) yields an average effect of contact of 0.292 ($p < 0.001$, 95% CI: $[0.214, 0.371]$), which is slightly larger than the random-effect model, but not statistically significantly different (t-test, $p = 0.78$), meaning that the dependence nature of the sample does not seem to influence the main result. The I^2 measure, adapted to the multi-level specification shows that heterogeneity comes as much from within-paper variation as from between paper variation. The remainder ($\simeq 4\%$) is due to sampling variation.

The meta-analytic average is slightly smaller that that found in Paluck et al. (2019) (0.39), largely due to addition of more recent papers with smaller effect sizes (possibly due to larger sample sizes, see the following Section).

It is also noteworthy that very few studies report significantly negative effects of contact (8 measures out of 209), including the adversarial contact measure of Lowe (2021).

4.2 Evidence of Publication Bias and p-hacking

Taken at face value, the previous result would indicate that contact interventions could be applied in a wide variety of contexts. However, recent meta-analyses in economics and other social sciences have highlighted the very serious risk of publication bias, meaning that the published evidence may not represent the full universe of conducted studies due to selective reporting or publication of findings. Publication bias can manifest in various forms. The typical form of bias inflates effect sizes by favoring statistically significant or positive estimates that reject the null hypothesis (Andrews and Kasy, 2019; Brodeur et al., 2016, 2020; Chopra et al., 2024). In the context of the contact literature, this concern is particularly salient because there seems to be a strong consensus among social scientists that contact *should* be effective (Doucouliagos and Stanley, 2013; Opatrny et al., 2026). Consequently, studies finding weak or insignificant effects of contact may be less likely to be written, and then appear in published outlets, potentially distorting the empirical distribution used in the meta-analytic estimation.

The initial analysis of publication bias here is to correlate the effect size with several characteristics of the papers. Results, displayed in Table F.1, show that the treatment effect is positively correlated with standard error, indicating that more precise papers tend to report smaller effects (see also the asymmetric funnel plot in Figure F.2, often used as a graphical sign of publication bias (Egger et al., 1997; Stanley and Doucouliagos, 2010)). Papers with larger sample sizes also report smaller effects. Published and non-pre-registered papers also tend to report stronger effects. All these regressions point to a potential

publication bias.

To formally test the presence of publication bias, and to test whether the average effect of contact is only positive because of such publication bias, I use the limit meta-analysis (LMA) method (Rücker et al., 2011), a method designed to correct for publication bias while explicitly modeling between-study heterogeneity within a random-effects framework (Brown et al., 2024; Clochard et al., 2025). The LMA models the dependence between a study's effect size and its precision by regressing effect sizes on a transformation of the study-specific standard error that incorporates the estimated between-study variance (τ^2):

$$d_i = \beta + \sqrt{se_i^2 + \tau^2}\,(\alpha + \epsilon_i)\,, \tag{3}$$

where d_i is the treatment effect estimate, se_i is its standard error, and ϵ_i is a standard-normal error term. This model yields two primary parameters: a mean effect (β) and a publication bias term (α), the latter representing the degree to which less precise studies deviate from the estimates of large studies. Specifically, a positive value of α would indicate publication bias.

The method then computes a bias-adjusted pooled effect by extrapolating the fitted model to the hypothetical situation of infinite precision (i.e., as the standard error goes to zero). The resulting limit estimate $\bar{d}_{lim} = \beta + \alpha\tau$ represents the effect if there were no publication bias.[15]

Results, presented in Table 5, indicate that there is evidence of some publication bias. Indeed, the parameter α is statistically significantly positive ($p < 0.05$), indicating that studies with a larger treatment effect (in absolute terms) tend to have a lower precision. Because of this, correcting for publication bias pushes the meta-analytic average toward 0, with an LMA estimate for the average of 0.237 (95% CI: [0.173, 0.302]). However, the magnitude of the correction is not dramatic, as the LMA estimate is still

[15] While the limit meta-analysis offers advantages (e.g., it explicitly models between-study variance), some caveats remain. Most importantly, it assumes a specific functional relationship between effect size and precision (via the term $\sqrt{se_i^2 + \tau^2}$). If the true mechanism of publication bias is more complex (e.g., selection dependent on p-values or effect size in non-linear ways), the model may mis-specify the bias. For additional contemporary approaches to correcting publication bias, see Irsova et al. (2025).

highly significantly different than 0 ($p < 0.001$) and as the two previous, non-bias-adjusted estimates of the meta-analytic average (Table 4) fall within the confidence interval of the bias-adjusted estimate.

Table 5: Meta-analytic average effect of contact, adjusting for publication bias

	Effect of contact (1)
$\overline{d}_{lim}$	0.237***
SE	(0.033)
95% CI	[0.173, 0.302]
Publication bias (α)	1.060*
SE(α)	(0.493)
Num. obs.	209

Note: Limit meta-analysis (Equation (3)), when the dependent variable is the standardized effect size. * p < 0.05, ** p < 0.01, *** p < 0.005.

Another potential problem identified in past meta-analyses in social sciences (Brodeur et al., 2020; Irsova et al., 2025; Opatrny et al., 2026) is p-hacking, or the tendency to selectively report models, within papers, based on the p-values of the test. Stefan and Schönbrodt (2023), for example, identify selectively choosing which variables to include as controls or which outliers to exclude as potential methods used to reduce p-values. One diagnostic test for the presence of p-hacking is to look at the distribution of the Z-score ($\frac{d_i}{SE(d_i)}$), and to check whether we observe bunching on the right of the 1.96 threshold, equivalent to a p-value of 0.05 in a two-sided t-test (Simonsohn et al., 2014). Figure F.3 indicates that such bias is present in the contact literature.

To test if the overall positive meta-analytic average of contact interventions is entirely due to p-hacking, I use the Right-Truncated Meta-Analysis method (RTMA), proposed by Mathur (2024). The RTMA assumes that tests that reject the null in the expected direction (so-called "affirmative results", in this case statistically significant positive treatment effects) are expected to be p-hacked, while the non-affirmative (null or negative results) are not. The method consists in reconstructing the distribution of outcomes based on non-affirmative results, and re-run a random-effects model with the reconstructed

data, which gives a conservative estimate (biased toward 0) of the meta-analytic average in the case without p-hacking. Performing the test using Mathur (2024)-designed tool,[16] I find that a conservative estimate of the meta-analytic average is $\bar{d} = 0.17$ (95% CI: [0.076, 0.43]). While this coefficient is again lower than the baseline models, it is still highly statistically significant, and the non-corrected estimates still fall within the confidence interval.

To summarize the main result, the meta-analytic average effect of contact is positive, and although there is evidence of publication bias and p-hacking, correcting for such biases reduces the average, but the effect is still statistically significantly positive.

5 Drivers of heterogeneity of effects

In this Section, I examine potential moderators of the effect of contact. I first conduct a separate analysis for each characteristic. Second, I examine the relationship between effect size and contact conditions, as determined by the experiment. Third, I perform a lasso penalized regression to identify variables that correlate the most with effect size.

5.1 What characteristics matter for contact?

In this section, I present the results of the analysis of the effects of contact disaggregated by all characteristics presented in Section 2. In Appendix G are presented all the corresponding figures and tables. For each category, I first present the meta-analytic average for the category, and then perform a meta-analytic regression to test for differences between categories.

[16] https://metabias.io/phacking/.

5.1.1 Paper characteristics

Meta-analytic averages for each of the variables representing the characteristics of the papers are presented in Figure 2. Details per variable are presented in Tables G.1 to G.3, and Figures G.1 to G.3.

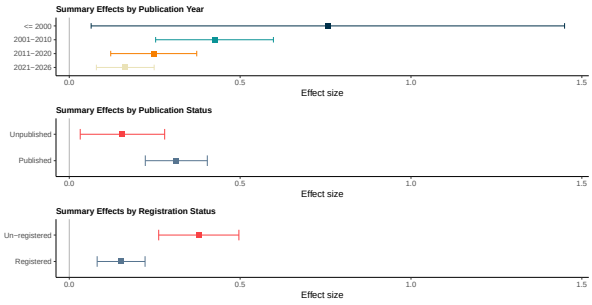


Figure 2: Meta-analytic averages for each of the characteristics of the papers.

Note: The panels show the average effect of contact (and 95% CI) estimated using a multi-level model, for each subsample. The top panel shows the meta-analytic average per publication decade. The second panel shows the average per publication status. The third panel shows the average per registration status.

Publication year: A clear downward trend in the effect size with time is observed. It is important to note, however, that the meta-analysis point estimate is positive and statistically significant for each period.

Publication status: As discussed in the previous Section, published papers have, on average, a larger effect size, although the difference is not statistically significant. Still, the average effect for unpublished papers is still positive and statistically different from 0 ($p = 0.02$).

Registration status: Again as discussed in the previous Section, registered experiments have a meta-analytic average effect size that is smaller than un-registered experiments. That difference is statistically significant, but the meta-analytic average for registered studies is still statistically significantly positive ($p < 0.005$).

5.1.2 Context characteristics

Meta-analytic averages for each of the variables representing the characteristics of the contexts are presented in Figure 3. Details per variable are presented in Tables G.4 to G.8, and Figures G.4 to G.8.

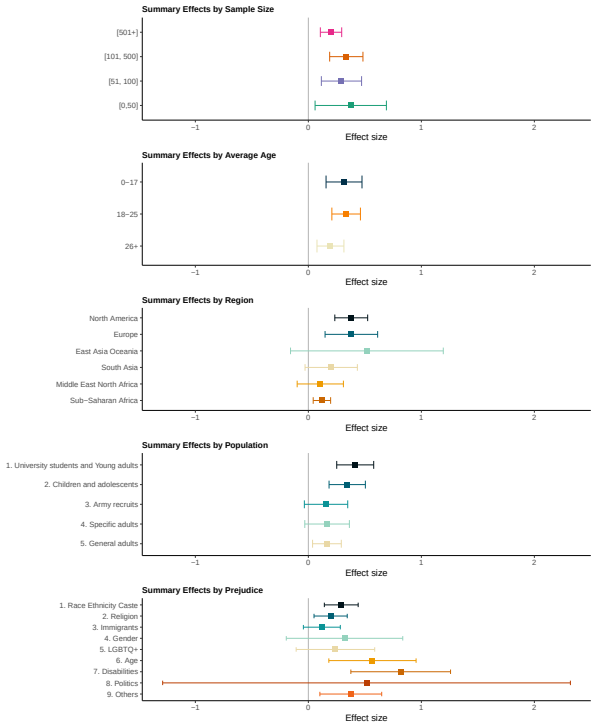


Figure 3: Meta-analytic averages for each of the characteristics of the contexts.

Note: The panels show the average effect of contact (and 95% CI) estimated using a multi-level model, for each subsample. The top panel shows the meta-analytic average per sample size. The second panel shows the average per average age in the sample. The third panel shows the average per region where the experiment was conducted. The fourth panel shows the average per population studied. The fifth panel shows the average per prejudice studied.

Sample: As discussed in the previous Section for signs of potential publication bias, interventions with a larger sample size tend to report smaller effect sizes. However, the meta-analytic average for each group (when grouped by sample size) are statistically significantly different from 0.

Age: It appears that studies involving the general population (age 25+) tend to have a lower average effect, and with a more precise meta-analytic estimate. This result could be the mere consequence of the fact that studies involving the general adult population tend to be better powered, and therefore provide

more accurate estimates, but could also indicate that contact intervention are less effective among adults.

Region: No clear pattern emerge from the separation by region, except that contact appears slightly more effective in North America. Part of the explanation could be that the United States concentrate most of the experiments performed on university students, who tend to have stronger effects (see below).

Population: Results indicate that university students and children tend to be more receptive to the effects of contact, while all the other groups have smaller meta-analytic averages. The difference with other groups is not statistically significant, though.

Type of prejudice: No clear pattern emerges from the analysis of the link between the type of prejudice and the effectiveness of contact. No difference between groups is statistically significant.

5.1.3 Intervention characteristics

Meta-analytic averages for each of the variables representing the characteristics of the measures are presented in Figure 4. Details per variable are presented in Tables G.9 to G.11, and Figures G.9 to G.11.

Type of intervention: As for the type of prejudice, no clear pattern emerges from the comparison of the type of contact intervention, except that discussions tend to be more effective than other interventions, particularly interventions based on classmates and the army.

Number of encounters: No clear patter seems to emerge from the number of encounters. Regardless of the number of times participants met outgroup members, the effect is statistically significantly positive.

Duration of contact: No clear pattern seems to emerge from the analysis of the duration of the contact. Regardless of the duration of the contact, the effects of contact are statistically significantly positive, and the effect of duration is not.

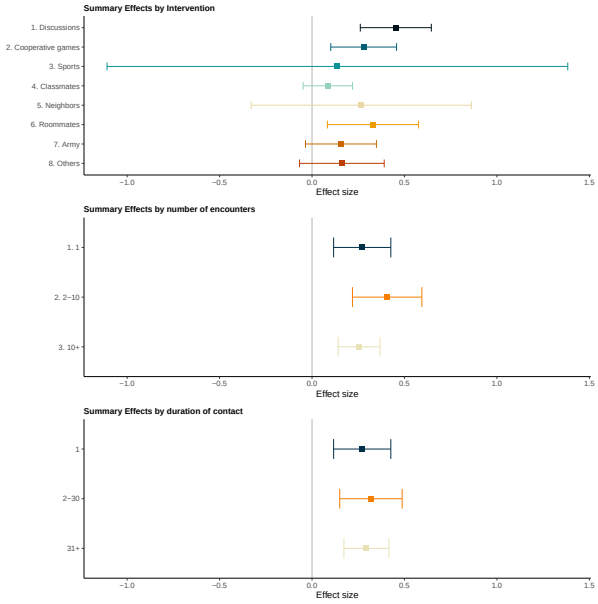


Figure 4: Meta-analytic averages for each of the characteristics of the interventions.

Note: The panels show the average effect of contact (and 95% CI) estimated using a multi-level model, for each subsample. The top panel shows the meta-analytic average per type of intervention. The second panel shows the average per pool of number of encounters. The third panel shows the average per pool of duration of the intervention.

5.1.4 Measurement characteristics

Meta-analytic averages for each of the variables representing the characteristics of the measures are presented in Figure 5. Details per variable are presented in Tables G.12 to G.14, and Figures G.12 to G.14.

Type of outcome: Measures of prejudice from behavior (when pooling experimental games, actions in the field and IATs) tend to be slightly more responsive to contact than explicit measures of prejudice using surveys. This could indicate a form of desirability bias in explicitly asking participants about their bias, making the effects of contact less noticeable (although still statistically significant). Interestingly, the effect is smaller and insignificant for Implicit Association Tests (although the sample size is limited to 11 observations).

Measure for the entire group: The effect of contact on the perceptions of the specific individuals

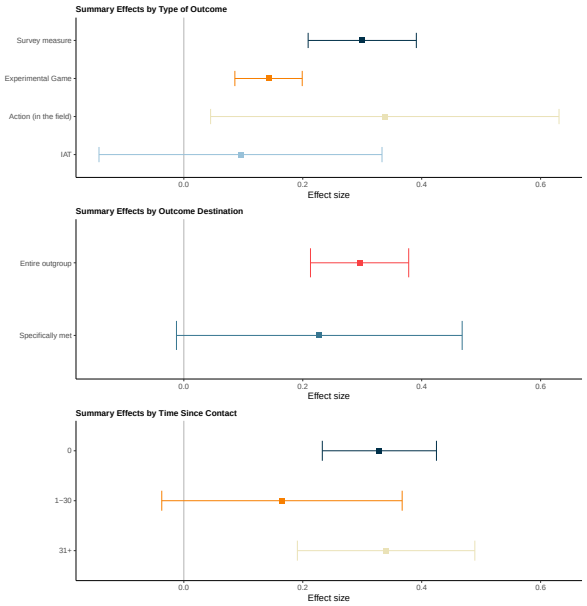


Figure 5: Meta-analytic averages for each of the characteristics of the measures.

Note: The panels show the average effect of contact (and 95% CI) estimated using a multi-level model, for each subsample. The top panel shows the meta-analytic average per type of outcome. The second panel shows the average per the destination of the outcome (the individuals specifically met or the entire outgroup). The third panel shows the average per pool of times between the intervention and measurement.

met during the intervention seems to be slightly stronger than the effects on the perception of the entire outgroup, although the average effect on the latter is still positive and strongly significant. This result could be rationalized by a simple model of incomplete information and a stronger signal for the individual than for the outgroup.

Time between end of intervention and measurement: It appears that the effect of contact does not fade with time. It is possible, however, that there might be some selective measurement and/or reporting of effect in this regard (researchers only measuring the effect of contact in a longer run if they observed significant results right at the end of the contact).

5.1.5 Contact conditions and contact effects

In Table G.15 I present meta-regression results where the outcome variable is the effect of contact, and the regressors are the conditions for contact effectiveness, as defined from the experiment, the Large

Language Model, and my own evaluation. From the experimental variables, it appears that positive encounters are associated with slightly larger effect sizes, and that the common goals condition might actually be negatively associated with the treatment effect of contact. The negative correlation with common goals condition is in contradiction to what was experimentally found by Lowe (2021). Potential explanations are presented in the next section. Contrary to what was posited by Allport (1954), equal status or the support of authorities are not found to be correlated with effect sizes. Similarly and contrary to what was found in the meta-analysis from (Pettigrew and Tropp, 2006), friendship potential does not seem to be significantly linked with the effect of contact.

The correlation between my evaluation of conditions (Column 2) or that of ChatGPT (Column 3) yield noisier and less consistent results.

5.2 Lasso penalized regression

After analyzing characteristics of the interventions one by one, I performed a lasso penalized regression to determine which of these characteristics matter most for the effectiveness of contact. I used the standardized effect size as the outcome, and all the variables described above as regressors. For the selected variables, I then performed an OLS estimation and a meta-analytic regression to obtain the unbiased correlation coefficients for these regressors. For the OLS estimation, I use the inverse of standard errors as analytical weights.[17]

Results are presented in Table 6. They indicate that the treatment effect is smaller for interventions based on neighbors, for papers published later and (counter-intuitively) stronger for outcomes measured longer after the end of the contact. Importantly, I also find that the publication year is strongly and negatively correlated with effect size, and this effect dominates the effect of sample size (which is also included but not selected by the lasso). As in the previous section, the common goals condition from the

[17] Removing weights does not change the results.

experiment is significantly negatively correlated with the results. Interestingly, no other conditions put forth by Allport (1954) are features selected by the lasso penalization.

The negative correlation between the effect of contact and the pursuit of common goals in the protocol is surprising, given that this is the only condition that was experimentally tested by (Lowe, 2021), which found that contact with common goals was more effective than adversarial contact. Several hypotheses could be put forth to make sense of this result. The first could be selection bias, with the only papers that have low common goals that are being written-up being the ones that find stronger effects. This explanation does not seem to hold, as controlling for standard error (one method to control for publication bias) changes neither the significance nor the magnitude of the correlation (Table H.1).[18] An alternative explanation is that the link between common goals and effect size is true over the support for the common goals variable in the present meta-analysis - from 43 to 86% - but the effect is positive if the support was from 0 to 100%, as is assumed in Lowe (2021).

It is also noteworthy that the only five variables selected only explain a relatively low fraction of the variance ($R^2 < 0.30$), meaning that the heterogeneity in the effects of contact is not mainly due to characteristics of the protocol - or rather, not the ones identified in this paper.

6 Discussion

In this paper, I conduct a meta-analysis of the literature on the contact hypothesis. While the sample of the original meta-analysis by Pettigrew and Tropp (2006) consisted almost entirely of descriptive, non-experimental evidence on the effect of contact, the number of experiments using contact is growing rapidly, with the added bonus of broadening the geographic origins of the samples. While the mode of research distribution is still the United States (representing 38% of the papers), there is now a growing

[18] If anything, the correction also makes the friendship potential, identified by Pettigrew and Tropp (2006), to be also negatively correlated with effect size.

Table 6: Post-Lasso OLS and meta-analytic regression on Lasso-selected variables

	OLS	META
	(1)	(2)
Intervention type: Neighbors	-0.246*	-0.127
	(0.107)	(0.252)
Common goal experiment	-0.008***	-0.007*
	(0.003)	(0.003)
Friendship potential experiment	-0.004	-0.003
	(0.003)	(0.004)
Time since contact	0.000***	0.000
	(0.000)	(0.000)
Year	-0.019***	-0.015*
	(0.004)	(0.005)
Constant	39.172***	31.974**
	(8.853)	(9.333)
Num. clusters	61	61
Num. obs	209	209

Note: The outcome variable is the standardized effect. In Column 1 are displayed the coefficients from the post-Lasso OLS regression (weighted by the inverse of the standard error), with standard errors clustered at the paper level. In Column 2 are presented the coefficients from a meta-regression. Explanatory variables for both columns are the ones that have been selected by a Lasso penalized regression. * p<0.05, ** p<0.01, *** p<0.005. Standard errors in parentheses.

number of studies from other parts of the world, notably in developing countries.

Moreover, the experimental literature on contact is also becoming more credible, with the use of relatively large sample sizes. Almost all papers published after 2010 in this analysis would be considered a "large study", using the taxonomy of Paluck et al. (2021), with an average contact group of more than 100 individuals.[19] The vast majority of papers published after 2015 also used a pre-analysis plan, which strengthens the credibility of the results.

The main finding of this paper is that contact is effective in reducing prejudice, and that very few papers find a backlash effect of contact interventions. The magnitude of the effect does not seem to depend much on the characteristics considered relevant, and the overall effect is a significant reduction in prejudice. However, this meta-analysis has found worrying signs of publication bias and p-hacking,

[19]The only exceptions would be Yablon (2012) and Gu et al. (2019).

with larger samples and pre-registration being associated with smaller effect sizes, and a bunching of z-scores above the 1.96 threshold. Despite these worrying signs, the bias-corrected estimates of the contact effect are still statistically significantly positive, and are only marginally smaller than the non-corrected estimate. The overall result from this meta-analysis is therefore positive.

An important caveat regarding the analysis of heterogeneity is multiple hypothesis testing (MHT). Throughout the paper, I examine heterogeneity across many dimensions using a relatively small set of data (209 observations). Although the lasso estimation is intended to limit the risk of overfitting, the large number of potential sources of heterogeneity (particularly those examined in Section 5.1) means that individual findings should be interpreted with caution. Accounting for MHT would likely reduce the statistical significance of several of the heterogeneity results. More importantly, even without such a correction, the analysis does not reveal a clear and consistent pattern of heterogeneity in the effects of contact, with the possible exception of the downward trend in treatment effects. This absence of a systematic pattern reinforces one of the central conclusions of the paper: the variables currently used in the literature provide limited explanatory power for the heterogeneity in the effects of contact.

However, the present meta-analytic exercise has highlighted four major limitations of the current contact literature. The first major limitation of the contact literature is the lack of discussion of pre-experimental prejudice. In the typical literature, the absolute level of prejudice is calculated from a baseline (pre-treatment) survey, but there is no discussion of the meaning of the observed levels. In addition, the counterfactual of a comparison group is typically absent. For example, white participants are asked whether they agree with the statement "Do you think African Americans can be trusted?" but never "Do you think whites can be trusted?" This lack of a comparison group prohibits much of the literature from identifying prejudicial behavior at baseline.

Furthermore, when papers do identify pre-experimental levels of discrimination, such as Finseraas

et al. (2016), they rarely identify the underlying source of prejudice (i.e., taste-based vs. statistical). There is also often a lack of explicit discussion of the absolute levels of prejudice. Identifying the source of prejudice can be critical to analyzing its consequences and identifying methods to combat it (Ahmadi et al., 2024), and at this stage, it is unclear whether contact interventions reduce taste-based, statistical or other forms of discrimination.

The second major limitation of the literature is the lack of consistency in the use of the term "contact," as described in Section 2. The wide variety of protocols that fall under the umbrella of contact - from sports leagues to youth camps to interactions between classmates or army recruits - makes it difficult to compare different interventions. In my view, future experiments should place more emphasis on the precise content of contact interventions, not just the context in which they occur. Now that several meta-analytic works have been carried out, all of which highlight the benefits of contact interventions, the focus of the literature should shift from the question of *whether* contact interventions are effective to the question of *how* they work. In terms of understanding the mechanisms by which contact can affect prejudice, more effort should be devoted to randomizing contact conditions within an experimental sample - *à la* Lowe (2021) or Greene et al. (2025). This would allow the field to understand what features of protocols are effective in improving intergroup relations. This exercise could help confirm or refute the descriptive results from the lasso analysis above.

The third limitation of the literature is the lack of a theoretical framework to explain the effects of contact. Of all the papers analyzed in this paper, only two (Lowe, 2021; Clochard, 2025) make attempts to develop a theoretical framework to explain the effects of contact on prejudice. In case the source of prejudice is statistical discrimination, the contact literature could, for example, work with the literature on belief updating. One possible solution could be to integrate a form of updating into the stereotyping literature (Bordalo et al., 2020). In this regard, it may be useful to analyze a clear distinction between the

effects of contact on the specific individuals encountered and the effects of contact on the entire outgroup. One reason why identifying the mechanisms behind the effects of contact has proven difficult may be that the literature has rarely performed heterogeneity analyses to identify the most affected participants, which could shed light on the underlying mechanisms.

The fourth and potentially most important limitation relates to policy applicability. As discussed in the introduction, the papers lack discussion of the external validity of their findings in other contexts. Furthermore, none of the papers conduct a welfare analysis of contact interventions. Before asserting that contact interventions are desirable policy tools, the literature needs to better understand the returns to each dollar invested. The first dimension could be to use a metric of X% reduction in prejudice per thousand dollars invested -which would require to measure and report the initial level of discrimination, see the first limitation. A second dimension could be in economic terms. It has been shown that prejudice can reduce economic output (e.g., Hjort (2014); Hedegaard and Tyran (2018)) and that contact interventions can reduce prejudice. Now we would need to show that contact interventions can increase economic output and estimate the monetary returns to these interventions.

In conclusion, because of the limitations highlighted above, I can only echo the call of Pettigrew and Tropp (2006) and Paluck et al. (2019) to fill the gaps in the literature before contact interventions can be used effectively as a policy tool.

Acknowledgements

I am grateful to Lori Beaman, Colin Camerer, Patrick DeJarnette, Carlos Gomez-Gonzalez, Taisuke Imai, Nobuyuki Hanaki, Tomáš Havránek, Guillaume Hollard, Yoshio Kamijo, Andreas Kotsadam, John List, Martina Lušková, Pauline Rossi, and Ferdinand Vieider for their helpful comments. I am also grateful for comments from participants at seminars in Waseda University and the Virtual East Asia Experi-

mental and Behavioral Economics Seminar, the International Workshop in Theoretical and Experimental Economics and at internal seminars. I gratefully acknowledge the financial support from Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and the Japan Society for the Promotion of Science (KAKENHI No. JP25K16598). I want to thank Ignacio Camps, Marco DiGiacomo, Maria del Mar Megy, Leandro Monzon, Bruno Pelaccini, Clara Pinasco, and Tomas Villar for excellent research assistance. Artificial Intelligence (ChatGPT 5.3) was used to assess the fulfillment of contact conditions, as described in Section 2, and to assist in the writing of the code. All the code and text was thoroughly reviewed. All errors are my own.

Competing interests

The author declares no competing interests.

Ethics

The online experiment was approved by the Institutional Review Board of the Institute of Social and Economic Research at the University of Osaka (#20240103) on January 29, 2024. No specific ethical issues were identified.

References

Ahmadi, M., G.-J. Clochard, J. Lachman, and J. A. List (2024). Toward an understanding of discrimination: The case of parsing multiple sources. Technical report, Working paper.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic growth 8*(2), 155–194.

Alimo, C. J. (2016). From dialogue to action: The impact of cross-race intergroup dialogue on the development of white college students as racial allies. In *Intergroup Dialogue*, pp. 47–70. Routledge.

Allport, G. W. (1954). *The nature of prejudice*. Addison-wesley Reading, MA.

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–2794.

Banerjee, S., M. M. Galizzi, and R. Hortala-Vallve (2021). Trusting the trust game: An external validity analysis with a uk representative sample. *Games 12*(3), 66.

Barnhardt, S. (2009). Near and dear? evaluating the impact of neighbor diversity on inter-religious attitudes. Technical report, Unpublished working paper.

Barros, H. P. (2025). The power of dialogue: Forced displacement and social integration amid an islamist insurgency in mozambique. *Journal of Development Economics 174*, 103457.

Becker, J. C., S. C. Wright, M. E. Lubensky, and S. Zhou (2013). Friend or ally: Whether cross-group contact undermines collective action depends on what advantaged group members say (or don't say). *Personality and Social Psychology Bulletin 39*(4), 442–455.

Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. In *Handbook of economic field experiments*, Volume 1, pp. 309–393. Elsevier.

Bezabih, M., S. Bezu, T. Getahun, I. Kolstad, P. Lujala, and A. Wiig (2021). Inter-group interaction and attitudes to migrants. Technical report, CMI Working Paper.

Bhattacharya, S. (2021). Intergroup contact and its effects on discriminatory attitudes: Evidence from india. Technical report, WIDER Working Paper.

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

Bordalo, P., N. Gennaioli, and A. Shleifer (2020). Memory, attention, and choice. *The Quarterly Journal of Economics 135*(3), 1399–1442.

Boucher, V., S. Tumen, M. Vlassopoulos, J. Wahba, and Y. Zenou (2021). Ethnic mixing in early childhood: Evidence from a randomized field experiment and a structural model. Technical report, IZA Working Paper.

Brodeur, A., N. Cook, and A. Heyes (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review 110*(11), 3634–3660.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

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

Brown, A. L., T. Imai, F. M. Vieider, and C. F. Camerer (2024). Meta-analysis of empirical estimates of loss aversion. *Journal of Economic Literature 62*(2), 485–516.

Bursztyn, L., T. Chaney, T. A. Hassan, and A. Rao (2024). The immigrant next door. *American Economic Review 114*(2), 348–384.

Caluwaerts, D. and M. Reuchamps (2014). Does inter-group deliberation foster inter-group appreciation? evidence from two experiments in belgium. *Politics 34*(2), 101–115.

Camargo, B., R. Stinebrickner, and T. Stinebrickner (2010). Interracial friendships in college. *Journal of Labor Economics 28*(4), 861–892.

Camilleri, A. R., K. Dankova, J. M. Ortiz, and A. Neelim (2023). Increasing worker motivation using a reward scheme with probabilistic elements. *Organizational Behavior and Human Decision Processes 177*, 104256.

Carrell, S. E., M. Hoekstra, and J. E. West (2019). The impact of college diversity on behavior toward minorities. *American Economic Journal: Economic Policy 11*(4), 159–182.

Chaudhry, Z. and K. Hussain (2023). The economic effects of inter-sectarian contact. Technical report, IGC Working Paper.

Chopra, F., I. Haaland, C. Roth, and A. Stegmann (2024). The null result penalty. *Economic Journal 134*(657), 193–219.

Clochard, G.-J. (2025). Using a brief contact to improve trust in the police by the youth. *Journal of Behavioral and Experimental Economics 117*, 102376.

Clochard, G.-J., S. Dey, S. Sasaki, and T. Imai (2025). Revisiting the price elasticity of charitable giving: Meta-analysis of tax incentives and matching donations. Technical report, CESifo Working Paper.

Clochard, G.-J., G. Hollard, and O. Sene (2026). Bringing contact interventions to the lab: Effects of brief bilateral discussions on interethnic trust in senegal. *World Development 199*, 107247.

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

Cohen, J. (1969). Statistical power analysis for tibe behavioral sciences.

Condra, L. N. and S. Linardi (2019). Casual contact and ethnic bias: Experimental evidence from afghanistan. *The Journal of Politics 81*(3), 1028–1042.

Corno, L., E. La Ferrara, and J. Burns (2022). Interaction, stereotypes, and performance: Evidence from south africa. *American Economic Review 112*(12), 3848–3875.

Dahl, G. B., A. Kotsadam, and D.-O. Rooth (2021). Does integration change gender attitudes? the effect of randomly assigning women to traditionally male teams. *The Quarterly Journal of Economics 136*(2), 987–1030.

Danckert, B., P. T. Dinesen, R. Klemmensen, A. S. Nørgaard, D. Stolle, and K. M. Sønderskov (2017). With an open mind: Openness to experience moderates the effect of interethnic encounters on support for immigration. *European Sociological Review 33*(5), 721–733.

DerSimonian, R. and N. Laird (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials 7*(3), 177–188.

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

DeVries, D. L., K. J. Edwards, and R. E. Slavin (1978). Biracial learning teams and race relations in the classroom: four field experiments using teams-games-tournament. *Journal of Educational Psychology 70*(3), 356.

Doucouliagos, C. and T. D. Stanley (2013). Are all economic facts greatly exaggerated? theory competition and selectivity. *Journal of Economic surveys 27*(2), 316–339.

Durante, F. and S. T. Fiske (2017). How social-class stereotypes maintain inequality. *Current opinion in psychology 18*, 43–48.

Egger, M., G. D. Smith, M. Schneider, and C. Minder (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ 315*(7109), 629–634.

Elwert, F., T. Keller, and A. Kotsadam (2023). Rearranging the desk chairs: A large randomized field experiment on the effects of close contact on interethnic relations. *American Journal of Sociology 128*(6), 1809–1840.

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

Evans, K. M., P. Kienast, and T. R. Mitchell (1988). The effects of lottery incentive programs on performance. *Journal of Organizational Behavior Management 9*(2), 113–135.

Fernández-Castilla, B., L. Declercq, L. Jamshidi, N. Beretvas, P. Onghena, and W. Van den Noortgate (2020). Visual representations of meta-analyses of multiple outcomes: Extensions to forest plots, funnel plots, and caterpillar plots. *Methodology 16*(4), 299–315.

Finseraas, H., T. Hanson, Å. A. Johnsen, A. Kotsadam, and G. Torsvik (2019). Trust, ethnic diversity, and personal contact: A field experiment. *Journal of Public Economics 173*, 72–84.

Finseraas, H., Å. 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.

Fishkin, J., A. Siu, L. Diamond, and N. Bradburn (2021). Is deliberation an antidote to extreme partisan polarization? reflections on "america in one room". *American Political Science Review 115*(4), 1464–1481.

Freddi, E., J. Potters, and S. Suetens (2024). The effect of brief cooperative contact with ethnic minorities on discrimination. *Journal of Economic Behavior & Organization 222*, 64–76.

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.

Gaikwad, N., K. Hanson, and A. Tóth (2025). Bridging the gulf: How migration fosters tolerance, cosmopolitanism, and support for globalization. *American Journal of Political Science 69*(3), 813–830.

Ghosh, A., P. Kundu, M. Lowe, and G. Nellis (2026). Creating cohesive communities: a youth camp experiment in india. *Review of Economic Studies 93*(1), 438–475.

Gonzales, E., N. Morrow-Howell, and P. Gilbert (2010). Changing medical students' attitudes toward older adults. *Gerontology & Geriatrics Education 31*(3), 220–234.

Grady, C., R. Wolfe, D. Dawop, and L. Inks (2023). How contact can promote societal change amid conflict: An intergroup contact field experiment in nigeria. *Proceedings of the National Academy of Sciences 120*(43), e2304882120.

Green, D. P. and O. Hyman-Metzger (2025). The vietnam draft lottery and whites' racial attitudes: Evidence from the general social survey. *American Political Science Review 119*, 2011–2018.

Green, D. P. and J. S. Wong (2009). Tolerance and the contact hypothesis: A field experiment. In *The political psychology of democratic citizenship*, pp. 1–23. Oxford University Press New York, NY.

Greene, K., E. Rossiter, E. Seira, and A. Simpser (2023). Interacting as equals: How contact can promote tolerance among opposing partisans. Technical report, Available at SSRN 4456223.

Greene, K. F., E. L. Rossiter, E. Seira, and A. Simpser (2025). Interacting as equals reduces partisan polarization in mexico. *Nature Human Behaviour 9*(1), 147–155.

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology 74*(6), 1464.

Gu, J., A. Mueller, I. Nielsen, J. Shachat, and R. Smyth (2019). Improving intergroup relations through actual and imagined contact: field experiments with malawian shopkeepers and chinese migrants. *Economic Development and Cultural Change 68*(1), 273–303.

Gu, J., I. Nielsen, J. Shachat, R. Smyth, and Y. Peng (2016). An experimental study of the effect of intergroup contact on attitudes in urban china. *Urban Studies 53*(14), 2991–3006.

Harzing, A.-W. (2007). Publish or perish. https://harzing.com/resources/publish-or-perish. Software available online; accessed 5 March 2026.

Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic perspectives 15*(4), 57–67.

Hedegaard, M. S. and J.-R. Tyran (2018). The price of prejudice. *American Economic Journal: Applied Economics 10*(1), 40–63.

Hedges, L. V., E. Tipton, and M. C. Johnson (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods 1*(1), 39–65.

Higgins, J. P. T. and S. G. Thompson (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine 21*(11), 1539–1558.

Hjort, J. (2014). Ethnic divisions and production in firms. *The Quarterly Journal of Economics 129*(4), 1899–1946.

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

Imai, T., T. A. Rutter, and C. F. Camerer (2021). Meta-analysis of present-bias estimation using convex time budgets. *The Economic Journal 131*(636), 1788–1814.

Imperato, C., B. H. Schneider, L. Caricati, Y. Amichai-Hamburger, and T. Mancini (2021). Allport meets internet: A meta-analytical investigation of online intergroup contact and prejudice reduction. *International Journal of Intercultural Relations 81*, 131–141.

Irsova, Z., P. R. Bom, T. Havranek, and H. Rachinger (2025). Spurious precision in meta-analysis of observational research. *Nature Communications 16*(1), 8454.

Kalla, J. L. and D. E. Broockman (2020). Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review 114*(2), 410–425.

Kotzur, P. F., S. J. Schäfer, and U. Wagner (2019). Meeting a nice asylum seeker: Intergroup contact changes stereotype content perceptions and associated emotional prejudices, and encourages solidarity-based collective action intentions. *British Journal of Social Psychology 58*(3), 668–690.

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.

Kumar, R., N. Seay, and S. Karabenick (2011). Shades of white: Identity status, stereotypes, prejudice, and xenophobia. *Educational studies 47*(4), 347–378.

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review 90*, 375–402.

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.

Lenz, L. and S. Mittlaender (2022). The effect of inter-group contact on discrimination. *Journal of Economic Psychology 89*, 102483.

Li, Z., H. Liao, S. Ma, J. Qiu, and S. Xue (2025). Inter-group contact and national identity: Evidence from hong kong and macau students in the mainland of china. *Journal of Development Economics 179*, 103625.

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Crown Currency.

Loiacono, F. and M. Silva-Vargas (2024). *Matching with the right attitude: The effect of matching firms with refugee workers*. European Bank for Reconstruction and Development.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review 111*(6), 1807–1844.

Lowe, M. (2025). Has intergroup contact delivered? *Annual Review of Economics 17*, 321–344.

Mae Boag, E. and D. Wilson (2014). Inside experience: Engagement empathy and prejudice towards prisoners. *Journal of Criminal Psychology 4*(1), 33–43.

Maiti, S. N., D. Pakrashi, S. Saha, and R. Smyth (2022). Don't judge a book by its cover: The role of intergroup contact in reducing prejudice in conflict settings. *Journal of Economic Behavior & Organization 202*, 533–548.

Markowicz, J. A. (2009). *Intergroup contact experience in dialogues on race groups: Does empathy and an informational identity style help explain prejudice reduction?* The Pennsylvania State University.

Marmaros, D. and B. Sacerdote (2006). How do friendships form? *The Quarterly Journal of Economics 121*(1), 79–119.

Mathur, M. B. (2024). P-hacking in meta-analyses: A formalization and new meta-analytic methods. *Research synthesis methods 15*(3), 483–499.

Meshel, D. S. and R. P. MCGlynn (2004). Intergenerational contact, attitudes, and stereotypes of adolescents and older people. *Educational Gerontology 30*(6), 457–479.

Mousa, S. (2020). Building social cohesion between christians and muslims through soccer in post-isis iraq. *Science 369*(6505), 866–870.

Mousa, S., L. Naumann, and A. Scacco (2025). Intergroup contact, empathy education, and refugee-native integration: Evidence from a field experiment in lebanon. Technical report, Unpublished Working Paper.

Okunogbe, O. (2024). Does exposure to other ethnic regions promote national integration? evidence from nigeria. *American Economic Journal: Applied Economics 16*(1), 157–192.

Opatrny, M., T. Havranek, Z. Irsova, and M. Scasny (2026). Publication bias and model uncertainty in measuring the effect of class size on achievement. *Journal of Labor Economics*.

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.

Paluck, E. L., S. A. Green, and D. P. Green (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy 3*(2), 129–158.

Paluck, E. L., R. Porat, C. S. Clark, and D. P. Green (2021). Prejudice reduction: Progress and challenges. *Annual review of psychology 72*, 533–560.

Peeters, A., G. Ouvrein, A. Dhoest, and C. De Backer (2025). Meat & greet: How inter-and intragroup dyadic interactions influence perceptions of meat eaters and meat avoiders. *Appetite 214*, 108169.

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.

Piepho, H.-P., J. Forkman, and W. A. Malik (2024). A reml method for the evidence-splitting model in network meta-analysis. *Research Synthesis Methods 15*(2), 198–212.

Popova, A., T. Imai, and S. Toussaert (2025). Time to untie odysseus from the mast? evidence review on commitment demand. Technical report, Working Paper.

Porat, R., S. Larry, J.-H. Pezzuto, M. Poupko, and D. Manekin (2025). Collaboration across group lines: Evidence from two field experiments in jerusalem. Technical report, The Hebrew University Working Paper.

Rao, G. (2019). Familiarity does not breed contempt: Generosity, discrimination, and diversity in delhi schools. *American Economic Review 109*(3), 774–809.

Rücker, G., G. Schwarzer, J. R. Carpenter, H. Binder, and M. Schumacher (2011). Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics 12*(1), 122–142.

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 112*(3), 654–677.

Schaerer, M., C. Du Plessis, M. H. B. Nguyen, R. C. Van Aert, L. Tiokhin, D. Lakens, E. G. Clemente, T. Pfeiffer, A. Dreber, M. Johannesson, et al. (2023). On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions. *Organizational Behavior and Human Decision Processes 179*, 104280.

Shook, N. J. and R. Clay (2012). Interracial roommate relationships: A mechanism for promoting sense of belonging at university and academic performance. *Journal of Experimental Social Psychology 48*(5), 1168–1172.

Shook, N. J., P. D. Hopkins, and J. M. Koech (2016). The effect of intergroup contact on secondary group attitudes and social dominance orientation. *Group Processes & Intergroup Relations 19*(3), 328–342.

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.

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

Stanley, T. D. and H. Doucouliagos (2010). Picture this: A simple graph that reveals much ado about research. *Journal of Economic Surveys 24*(1), 170–191.

Stefan, A. M. and F. D. Schönbrodt (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science 10*(2), 1–30.

Steinmayr, A. (2021). Contact versus exposure: Refugee presence and voting for the far right. *Review of Economics and Statistics 103*(2), 310–327.

Turner, J. C. (1978). Social categorization and social discrimination in the minimal group paradigm. *Differentiation between social groups: Studies in the social psychology of intergroup relations 101*, 140.

Van den Noortgate, W., J. A. López-López, F. Marín-Martínez, and J. Sánchez-Meca (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods 45*(2), 576–594.

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.

Vertier, P. and M. Viskanic (2018). Dismantling the 'jungle': Migrant relocation and extreme voting in france. Technical report, Center for Economic Studies and ifo Institute.

Walch, S. E., K. A. Sinkkanen, E. M. Swain, J. Francisco, C. A. Breaux, and M. D. Sjoberg (2012). Using intergroup contact theory to reduce stigma against transgender individuals: Impact of a transgender speaker panel presentation. *Journal of Applied Social Psychology 42*(10), 2583–2605.

Whitt, S., R. K. Wilson, and V. Mironova (2021). Inter-group contact and out-group altruism after violence. *Journal of Economic Psychology 86*, 102420.

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.

Zhou, Y.-Y. and J. Lyall (2025). Prolonged contact does not reshape locals' attitudes toward migrants in wartime settings. *American Journal of Political Science 69*(1), 210–222.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics 13*(1), 75–98.

Appendices

Appendix A Paper selection

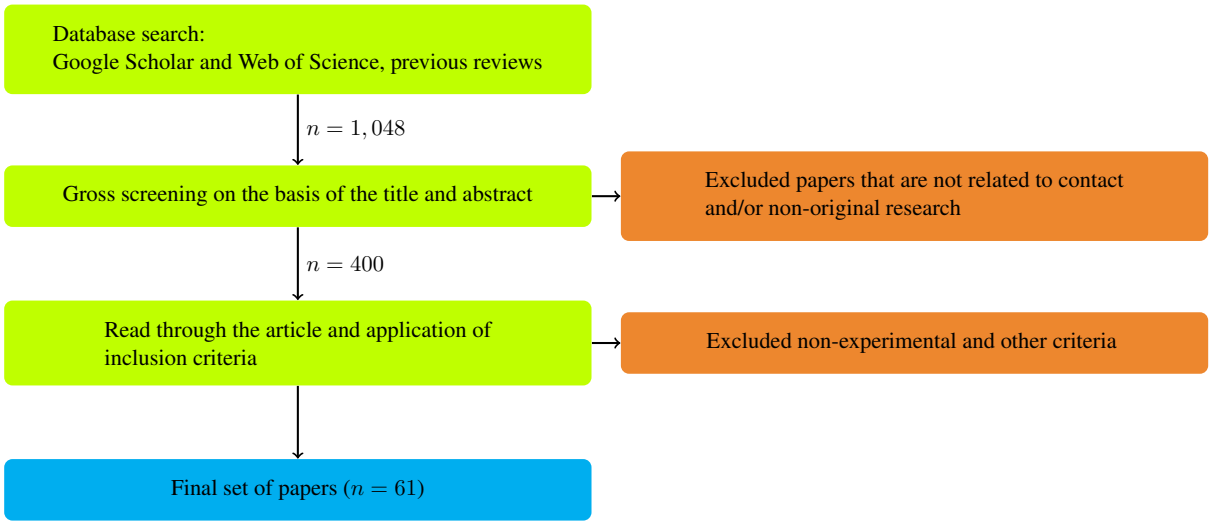


Figure A.1: Literature search and screening

Table A.1: List of papers included in the analysis

Paper	Paper	Paper	Paper
Alimo (2016)	Barnhardt (2009)	Barros (2025)	Becker et al. (2013)
Bezabih et al. (2021)	Bhattacharya (2021)	Boisjoly et al. (2006)	Boucher et al. (2021)
Broockman and Kalla (2016)	Caluwaerts and Reuchamps (2014)	Camargo et al. (2010)	Carrell et al. (2019)
Chaudhry and Hussain (2023)	Clochard (2025)	Clochard et al. (2026)	Clunies-Ross and O'meara (1989)
Condra and Linardi (2019)	Corno et al. (2022)	Dahl et al. (2021)	Dessel (2010)
DeVries et al. (1978)	Elwert et al. (2023)	Enos (2014)	Finseraas et al. (2016)
Finseraas and Kotsadam (2017)	Finseraas et al. (2019)	Fishkin et al. (2021)	Freddi et al. (2024)
Furuto and Furuto (1983)	Ghosh et al. (2026)	Gonzales et al. (2010)	Grady et al. (2023)
Green and Wong (2009)	Gu et al. (2016)	Gu et al. (2019)	Hull IV (1972)
Kalla and Broockman (2020)	Kotzur et al. (2019)	Krahe and Altwasser (2006)	Li et al. (2025)
Loiacono and Silva-Vargas (2024)	Lowe (2021)	Mae Boag and Wilson (2014)	Maiti et al. (2022)
Markowicz (2009)	Marmaros and Sacerdote (2006)	Meshel and MCGlynn (2004)	Mousa (2020)
Mousa et al. (2025)	Okunogbe (2024)	Page-Gould et al. (2008)	Peeters et al. (2025)
Porat et al. (2025)	Scacco and Warren (2018)	Shook and Clay (2012)	Shook et al. (2016)
Sorensen (2010)	Van Laar et al. (2005)	Walch et al. (2012)	Yablon (2012)
Zhou and Lyall (2025)			

Table A.2: Overlap with other meta-analyses on contact

Paper	N overlap	N paper	Share overlap
Paluck et al. (2019)	23	27	85.2%
Lowe (2025)	18	41	43.9%

Appendix B Experimental design

Appendix B.1 Examples of extracts from experimental design / methds sections

Example from Mae Boag and Wilson (2014): Using a field experimental design, the experimental group (n = 87) were randomly selected from a wider participation pool of undergraduate criminology students (n = 143) to take part in a day long carceral tour at an [British] Category B prison. The prisoners (n = 235) were all convicted of serious offences against the person. [...] The carceral tour involved meeting the prison staff, visiting one of five prison wings, taking part in a debate with prisoners and engaging with (conversing, dining and questioning) individuals and groups of prisoners who were present throughout the day. The entire experimental manipulation lasted approximately five hours.

Example from Elwert et al. (2023): The field experiment manipulated interethnic contact between Roma and non-Roma students by randomizing the seating chart within third- through eighth-grade classrooms in rural Hungarian schools for the three core subjects Hungarian reading, Hungarian literature, and mathematics. Depending on grade level, these subjects account for approximately seven to 10 hours per week, or between 25% and 45% of the school week. Students were randomly allocated to freestanding front-facing desks seating two students each, on the basis of class lists provided by the schools (using a random number generator). The intervention commenced at the beginning of the 2017 fall semester, and we encouraged adherence to the seating chart until the end of the semester in January 2018.

Example from Gu et al. (2019): "In this paper, we report the results of three framed field experiments that test whether [...] contact is effective in reducing prejudice between two groups, Malawian entrepreneurs and Chinese migrant entrepreneurs, in Malawi.[...] In each session, the Malawian and Chinese participants were randomly paired to form dyads. [...]

Each dyad was asked to complete two jigsaw puzzles. The Chinese and the Malawian participants

were both unfamiliar with jigsaw puzzles. We believe that this created a level playing field between both groups of participants. We used the following procedure to manipulate the amount of contact between the treatment and the control groups. We gave two comparable jigsaw puzzles per dyad. In the treatment group, each dyad was asked to jointly solve the puzzles one after the other, with the order randomly assigned. In the control group, each dyad was asked to solve both puzzles as well; however, in this case each of the members was randomly assigned one of the puzzles to solve individually. All the participants were given a maximum of 1 hour to complete the jigsaw puzzle task."

Appendix B.2 Screenshot of experimental screen

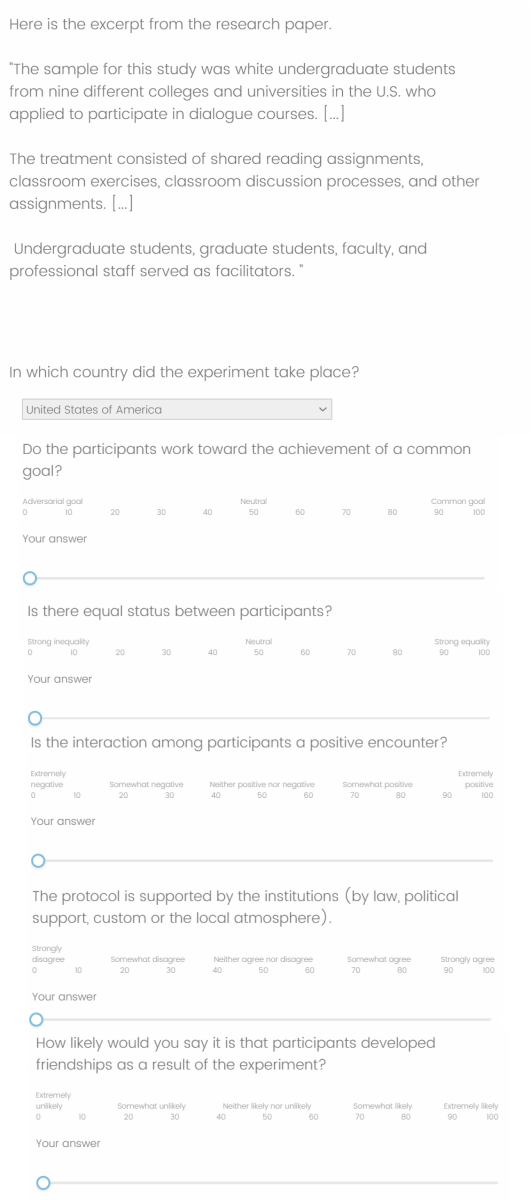


Figure B.1: Screenshot of experimental design (extract from Alimo (2016))

Table B.1: Prompts used for ChatGPT

Common goal	Can you give me on a scale of 0 to 100 whether participants in the following experiment work toward the achievement of a common goal?
Equal status	Can you give me on a scale of 0 to 100 whether participants have an equal status?
Positive encounter	On a scale of 0 to 100, is the interaction among participants a positive encounter?
Support of authorities	On a scale of 0 to 100, how much would you agree that the protocol is supported by the institutions (by law, political support, custom or the local atmosphere)?
Friendship potential	On a scale of 0 to 100, how likely would you say it is that participants developed friendships as a result of the experiment?

Appendix C Trends in publication and registration

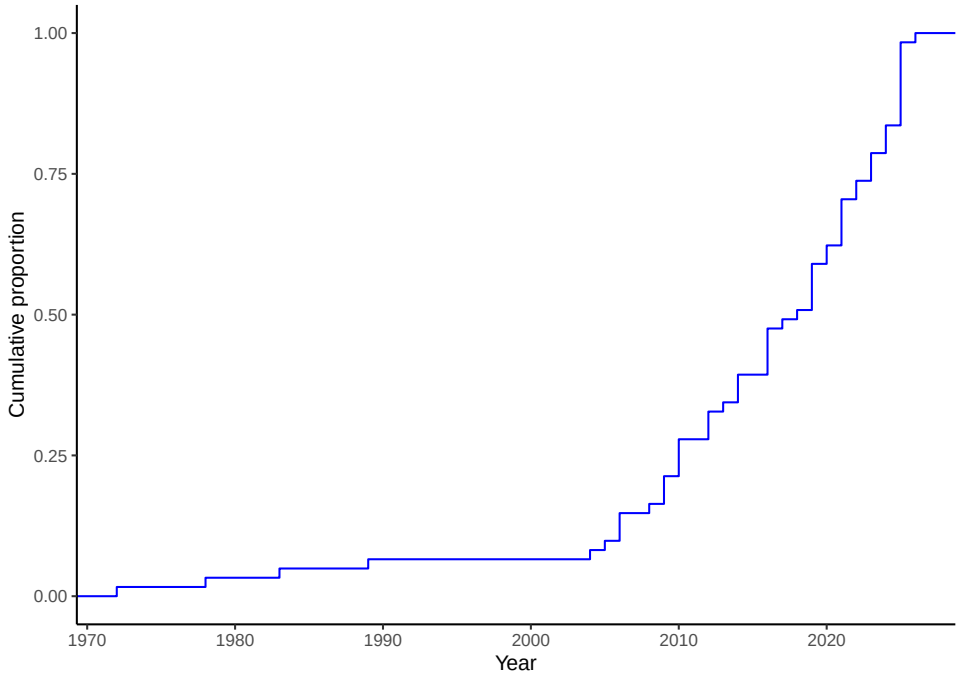
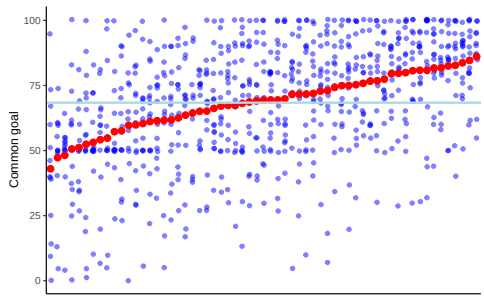


Figure C.1: Cumulative distribution of papers related to the contact hypothesis by year

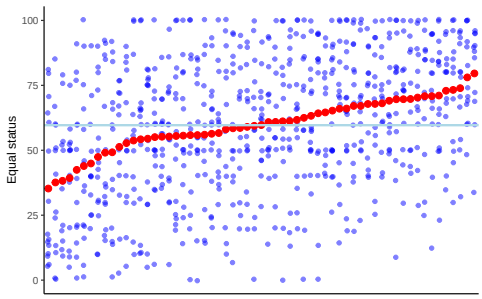
Table C.1: Trends pre-registration

Year of Publication	Percentage of pre-registered papers	N
$\leq$ 2000	25.00	4
2001-2010	0.00	13
2011-2020	28.60	21
2021-2026	73.90	23

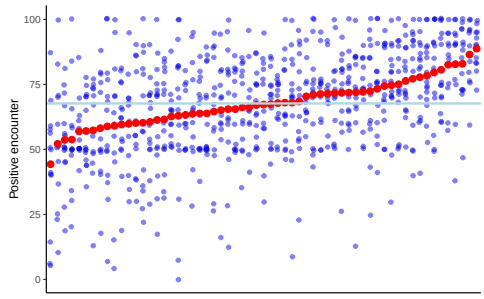
Appendix D Contact conditions



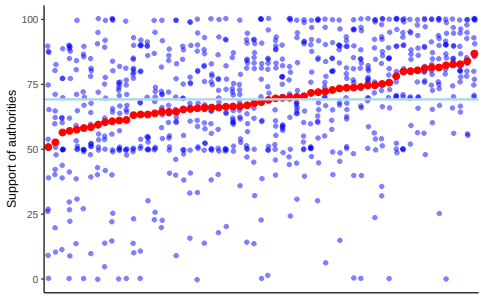
(a) common goals



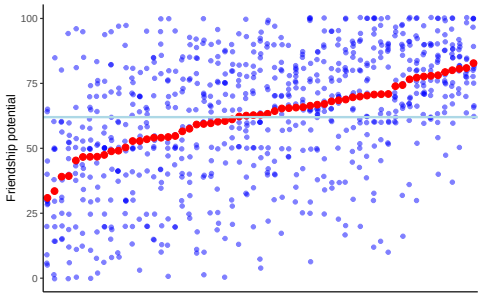
(b) Equal status



(c) Positive encounter



(d) Support of authorities



(e) Friendship potential

Figure D.1: Distributions of responses in the second stage of the Prolific experiment

Note: The Figure displays the scatter plot of responses from the Prolific experiment (incentivized second-order belief, in blue, see description in Section 2.3), the average per paper (in red) and the average for all papers (in light blue). For each figure, papers are sorted by the average value of the variable across experimental participant.

Table D.1: Pairwise correlation of outcomes in the second stage of the experiment

	Common goals (1)	Equal status (2)	Positive encounter (3)	Support of authorities (4)	Friendship potential (5)
Common goals	1.000***				
Equal status	0.304*	1.000***			
Positive encounter	0.622***	0.422***	1.000***		
Support of authorities	0.206	0.411***	0.341**	1.000***	
Friendship potential	0.368***	0.277*	0.565***	0.149	1.000***

Note: The table displays pairwise correlation coefficients for the fulfillment of Allport's conditions. Each varialble corresponds to the average response in the experiment. * p < 0.05,** p < 0.01,*** p < 0.005

Table D.2: Consistency of outcomes in the two experiments

	Common goals 1 (1)	Equal status 1 (2)	Positive encounter 1 (3)	Support of authorities 1 (4)	Friendship potential 1 (5)
Common goals 2	0.709***				
	(0.119)				
Equal status 2		0.397*			
		(0.150)			
Positive encounter 2			0.500***		
			(0.128)		
Support of authorities 2				0.138	
				(0.147)	
Friendship potential 2					0.819***
					(0.113)
Constant	25.103***	38.823***	36.967***	61.352***	17.902*
	(8.240)	(9.098)	(8.708)	(10.221)	(7.155)
Observations	61	61	61	61	61
R^2	0.375	0.106	0.207	0.015	0.470

Note: The dependent variable in each column is the average response of participants in the first Prolific experiment (first-order belief of fulfillment of each condition). The independent variable in each regression is the average response of participants in the second Prolific experiment (incentivized second-order belief). See Section 2.3 for details. * p<0.10, ** p<0.05, *** p<0.01. Standard errors in parentheses.

Table D.3: Consistency of outcomes between the experiments and author's own evaluation

	Common goals 2 (1)	Equal status 2 (2)	Positive encounter 2 (3)	Support of authorities 2 (4)	Friendship potential 2 (5)
Common goals author	0.372***				
	(0.086)				
Equal status author		0.015			
		(0.081)			
Positive encounter author			0.242		
			(0.124)		
Support of authorities author				0.136	
				(0.076)	
Friendship potential author					0.396***
					(0.074)
Constant	41.345***	58.517***	49.783***	58.556***	33.709***
	(6.397)	(6.379)	(9.247)	(6.031)	(5.456)
Observations	61	61	61	61	61
R^2	0.239	0.001	0.061	0.052	0.326

Note: The dependent variable in each column is the average response of participants in the second Prolific experiment (incentivized second-order belief of fulfillment of each condition). The independent variables are the author's own evaluations of whether the protocol satisfies the conditions. See Section 2.3 for details. * p<0.10, ** p<0.05, *** p<0.01. Standard errors in parentheses.

Table D.4: Consistency of outcomes between the experiments and ChatGPT

	Common goals 2 (1)	Equal status 2 (2)	Positive encounter 2 (3)	Support of authorities 2 (4)	Friendship potential 2 (5)
Common goals GPT	0.315***				
	(0.055)				
Equal status GPT		0.115			
		(0.087)			
Positive encounter GPT			0.299***		
			(0.072)		
Support of authorities GPT				0.209***	
				(0.064)	
Friendship potential GPT					0.447***
					(0.075)
Constant	47.294***	52.089***	46.453***	52.830***	35.077***
	(3.828)	(5.888)	(5.203)	(5.101)	(4.677)
Observations	61	61	61	61	61
R^2	0.359	0.029	0.227	0.154	0.377

Note: The dependent variable in each column is the average response of participants in the second Prolific experiment (incentivized second-order belief of fulfillment of each condition). The independent variables are the evaluations of whether the protocol satisfies the conditions by ChatGPT. See Section 2.3 for details. * p<0.10, ** p<0.05, *** p<0.01. Standard errors in parentheses.

Appendix E Additional meta-results

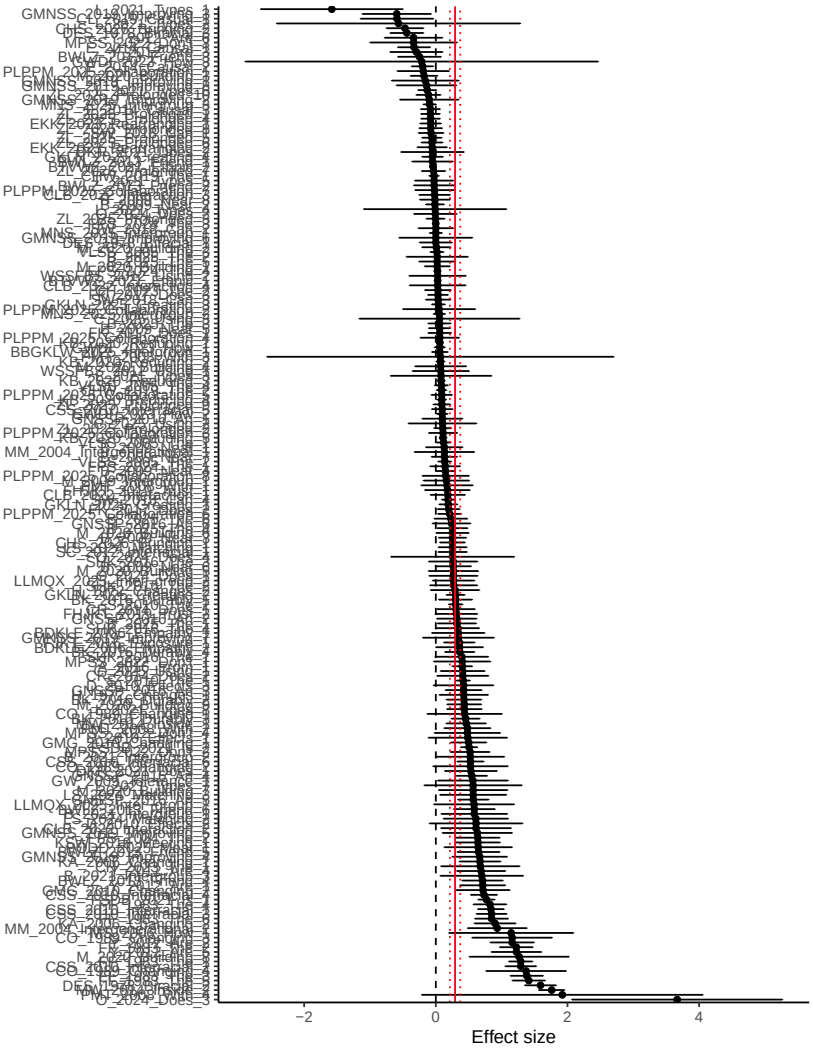


Figure E.1: Raw forest plot, not accounting for statistical dependence between papers. Each paper is represented by a dot with 95% confidence intervals. The red vertical line represents the meta-analytic average of the multi-level analysis, with red dotted lines representing the 95% confidence intervals of the meta-analytic average.

Table E.1: Meta-analytic average effect of contact

	Collapsed effect (1)
$\bar{d}$	0.278***
SE	(0.044)
95% CI	[0.191, 0.365]
τ^2	0.056
I^2	76.827
Num. clusters	61
Num. obs	61

Note: The dependent variable is the average standardized effect size across outcomes throughout the paper. * p < 0.05, ** p < 0.01, *** p < 0.005.

Appendix F Publication bias

Table F.1: Worrying signs of publication bias and/or file drawer problem

	Standardized effect size			
	(1)	(2)	(3)	(4)
Standard error	1.187***			
	(0.264)			
Sample size		−0.0001***		
		(0.00003)		
Published			0.126*	
			(0.060)	
Registered				−0.211***
				(0.048)
Constant	0.080*	0.290***	0.099	0.311***
	(0.035)	(0.033)	(0.054)	(0.035)
Observations	209	209	209	209
R^2	0.089	0.075	0.021	0.086

Note: The dependent variable is the standardized effect size. * p<0.05, ** p<0.01, *** p<0.005. Standard errors in parentheses. Papers are weighed by the inverse of the standard error.

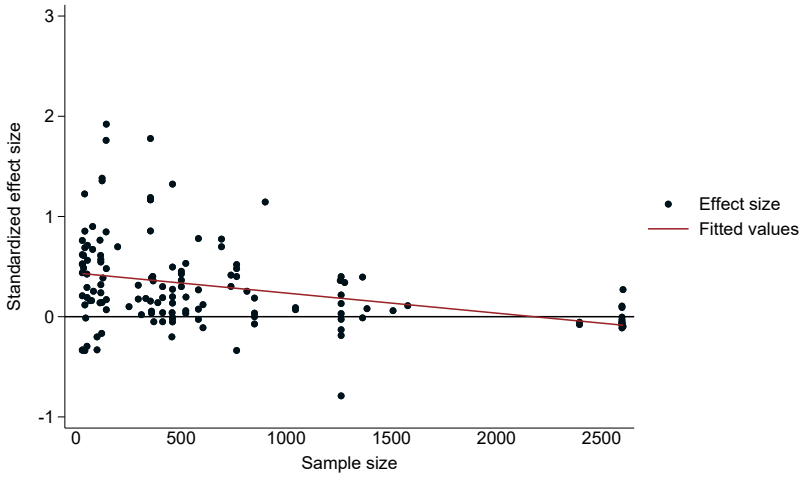


Figure F.1: Effect size as a function of the sample size

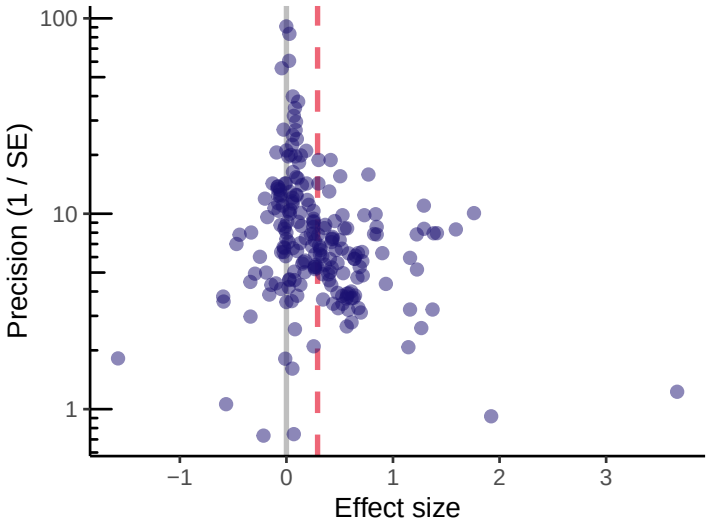


Figure F.2: Funnel plot (relationship between treatment effects and their corresponding precision $(1/SE)$

Notes: The y-axis is shown on a log scale to improve visual clarity. The vertical red line represents the multi-level estimate reported in Column 2 in Table 4.

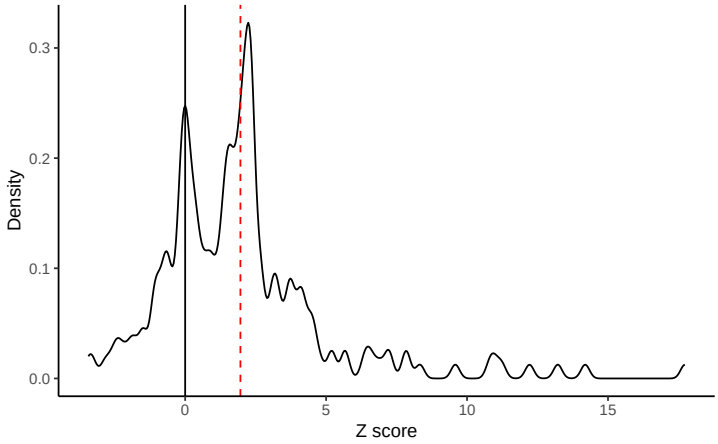


Figure F.3: Kernel distribution of the z-score.

Notes: The vertical red line corresponds to a z-score of 1.96, which yields a p-value of 0.05 for a two-sided t-test. There is suspicion of p-hacking when we observe bunching on the right side of the 1.96 threshold (Simonsohn et al., 2014; Mathur, 2024).

Appendix G Effect size as a function of characteristics

Appendix G.1 Publication year

Table G.1: Average effect size per decade of publication and meta-analytic regression of effect on year

	Panel A. Meta-analytic averages for each decade			
	$\leq$ 2000	2001-2010	2011-2020	2021-2026
	(1)	(2)	(3)	(4)
$\bar{d}$	0.757*	0.425***	0.247***	0.164***
SE	0.211	0.078	0.06	0.041
95% CI	[0.064, 1.449]	[0.253, 0.597]	[0.121, 0.373]	[0.079, 0.248]
τ^2_{within}	0.274	0.037	0.088	0.013
$\tau^2_{between}$	0.083	0.043	0.036	0.025
I^2_{within}	72.491	41.245	67.202	31.258
$I^2_{between}$	21.908	48.104	27.629	61.116
Num. clusters	4	13	21	23
Num. obs	15	38	73	83
	Panel B. Meta-analytic regression			
Year	-0.013*			
	(0.005)			
Constant	27.026*			
	(9.804)			
Num. clusters	61			
Num. Obs.	209			

Note: Panel A represents the meta-analytic averages of a multi-level model (with each paper as cluster), separated by decade. Panel B displays the results of a meta-analytic regression. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

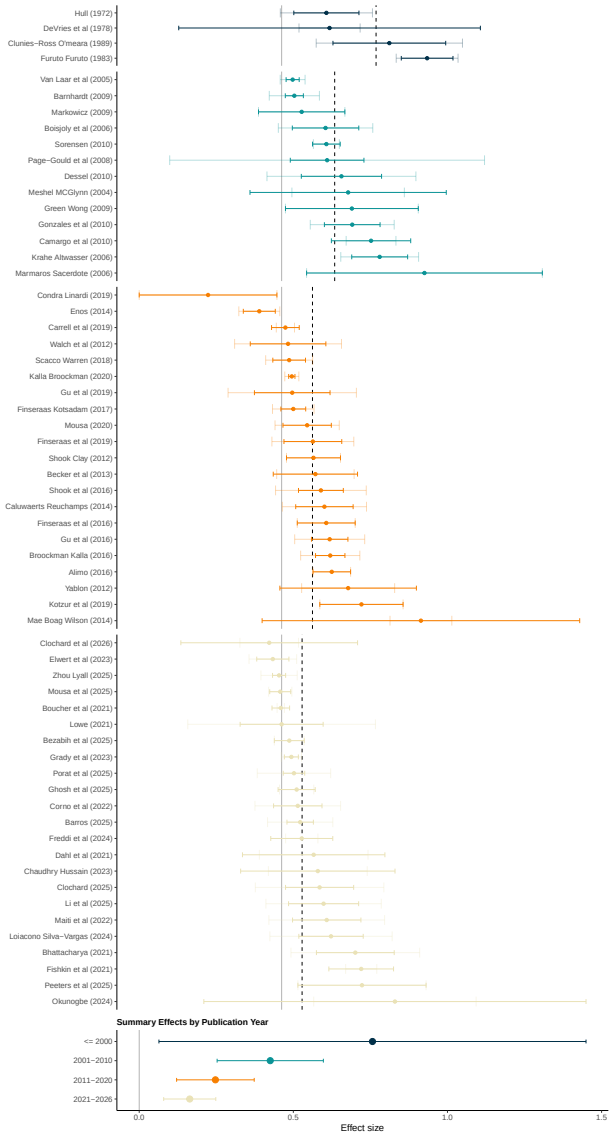


Figure G.1: Effect size as a function of the publication year

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by decade. Visualization follows Popova et al. (2025).

Appendix G.2 Publication status

Table G.2: Average effect size by publication status and meta-analytic regression

Panel A. Meta-analytic averages by publication status		
	Unpublished	Published
	(1)	(2)
$\bar{d}$	0.156*	0.313***
SE	0.054	0.045
95% CI	[0.032, 0.279]	[0.222, 0.404]
τ^2_{within}	0.002	0.078
$\tau^2_{between}$	0.022	0.062
I^2_{within}	7.968	53.401
$I^2_{between}$	80.462	42.599
Num. clusters	10	51
Num. obs	33	176
Panel B. Meta-analytic regression		
Published	0.123	
	(0.078)	
Constant	0.190*	
	(0.064)	
Num. clusters	61	
Num. Obs.	209	

Note: Panel A reports multi-level meta-analytic averages (clustered at the study level) by publication status. Panel B reports a meta-analytic regression of effect size on a Published indicator. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

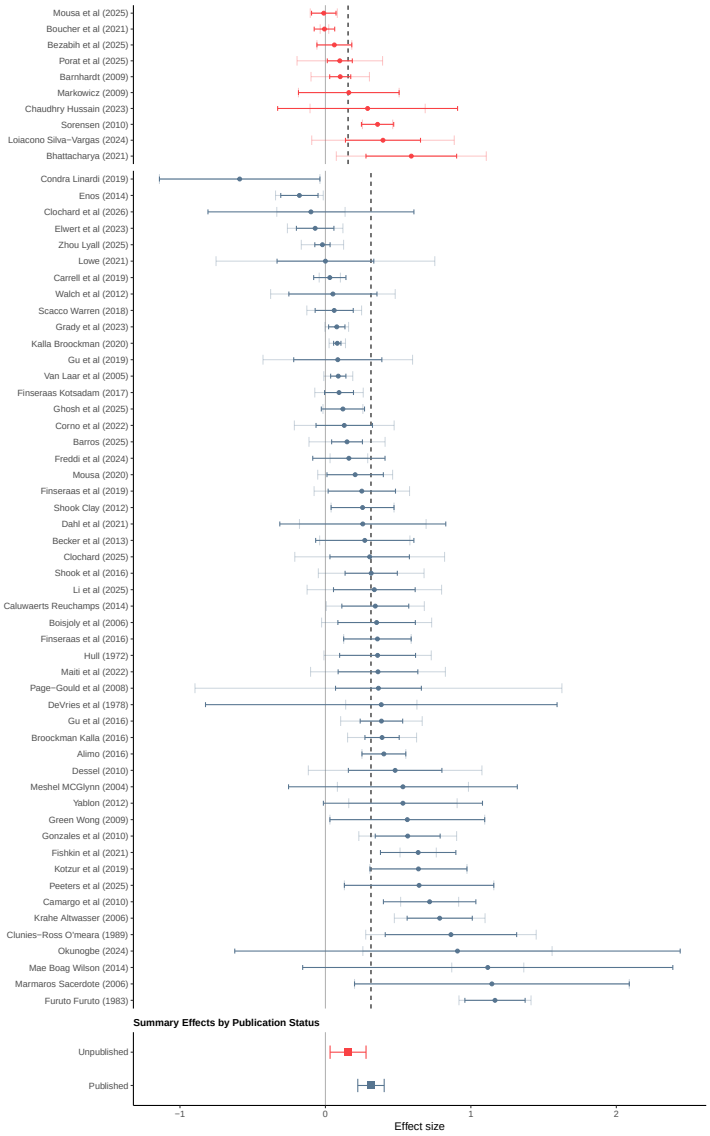


Figure G.2: Effect size as a function of whether the paper is published

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by publication status. Visualization follows Popova et al. (2025).

Appendix G.3 Registration status

Table G.3: Meta-analytic averages by study registration and meta-regression

Panel A. Meta-analytic averages by study registration		
	Un-registered	Registered
	(1)	(2)
$\bar{d}$	0.379***	0.152***
SE	0.058	0.034
95% CI	[0.262, 0.496]	[0.082, 0.222]
τ^2_{within}	0.108	0.010
$\tau^2_{between}$	0.063	0.018
I^2_{within}	60.715	30.466
$I^2_{between}$	35.513	56.804
Num. clusters	37	24
Num. obs	109	100
Panel B. Meta-analytic regression (baseline: Un-registered)		
Registered	-0.199**	
	(0.070)	
Constant	0.377***	
	(0.058)	
Num. clusters	61	
Num. Obs.	209	

Note: Panel A reports subgroup meta-analytic averages from a multilevel model (clustered at the paper level), separated by study registration status. Panel B reports a meta-analytic regression where the omitted category is unregistered studies. Coefficients represent differences relative to this baseline category. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

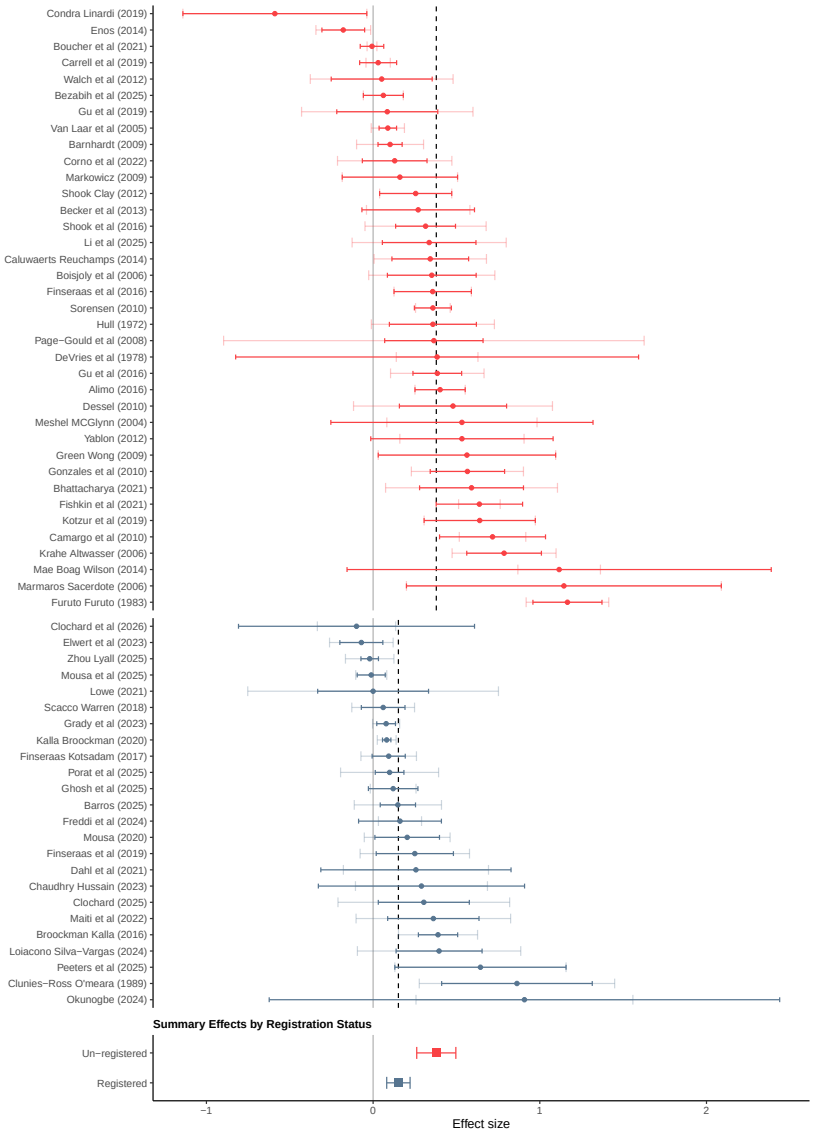


Figure G.3: Effect size as a function of whether the experiment was pre-registered

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by registration status. Visualization follows Popova et al. (2025).

Appendix G.4 Sample size

Table G.4: Meta-analytic averages by sample size and meta-regression

	Panel A. Meta-analytic averages by sample size			
	[0,50]	[51,100]	[101,500]	[501+]
	(1)	(2)	(3)	(4)
$\bar{d}$	0.375*	0.293***	0.336***	0.201***
SE	0.127	0.080	0.071	0.046
95% CI	[0.059, 0.691]	[0.115, 0.471]	[0.189, 0.483]	[0.106, 0.295]
τ^2_{within}	0.147	0.080	0.107	0.002
$\tau^2_{between}$	0.043	0.030	0.077	0.041
I^2_{within}	61.272	63.512	56.269	4.436
$I^2_{between}$	18.069	23.733	40.752	88.581
Num. clusters	7	13	28	24
Num. obs	24	30	79	76
	Panel B. Meta-analytic regression			
N	-0.000***			
	(0.000)			
Constant	0.379***			
	(0.052)			
Num. clusters	61			
Num. Obs.	209			

Note: Panel A reports subgroup meta-analytic averages from a multilevel model (clustered at the paper level) by study sample size. Panel B reports a meta-analytic regression where the moderator is the study sample size. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.005.

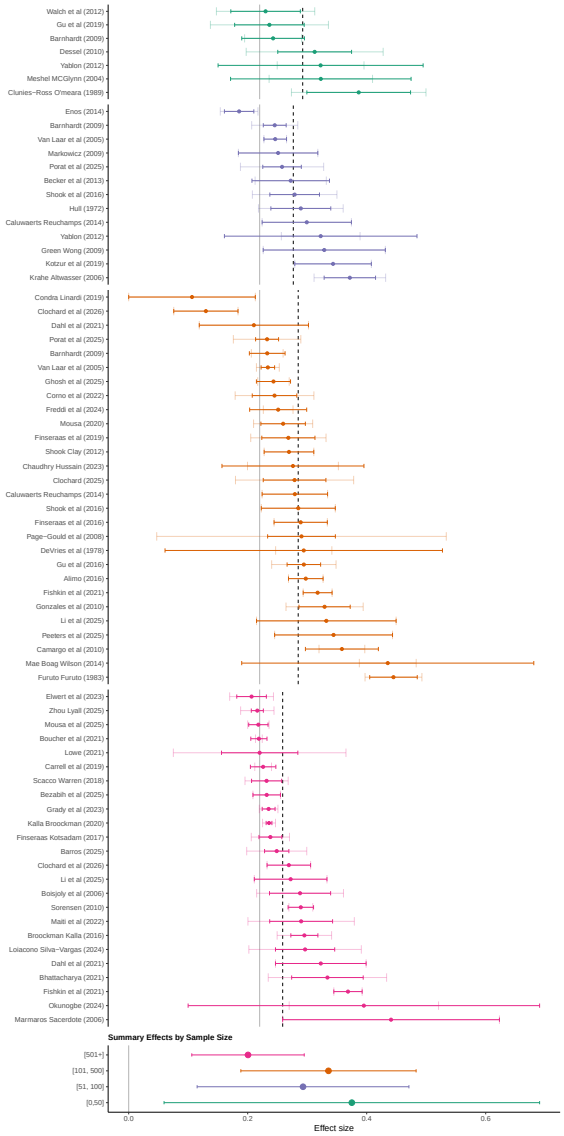


Figure G.4: Effect size as a function of the sample size category

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by sample size category. Visualization follows Popova et al. (2025).

Appendix G.5 Average age

Table G.5: Meta-analytic averages by age group and meta-regression on mean age

	0–17	18–25	26+
	(1)	(2)	(3)
Panel A. Meta-analytic averages by age group			
$\bar{d}$	0.316***	0.334***	0.196***
SE	0.072	0.062	0.056
95% CI	[0.157, 0.475]	[0.208, 0.461]	[0.076, 0.315]
τ^2_{within}	0.161	0.045	0.007
$\tau^2_{between}$	0.014	0.081	0.045
I^2_{within}	90.622	33.177	11.773
$I^2_{between}$	8.137	59.456	77.803
Num. clusters	14	30	18
Num. obs	46	103	60
Panel B. Meta-analytic regression			
Mean age	0.019		
	(0.021)		
Mean age^2	-0.000		
	(0.000)		
Constant	0.106		
	(0.255)		
Num. clusters	61		
Num. Obs.	209		

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by age group. Panel B reports a meta-analytic regression of effect size on mean age and its quadratic term. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

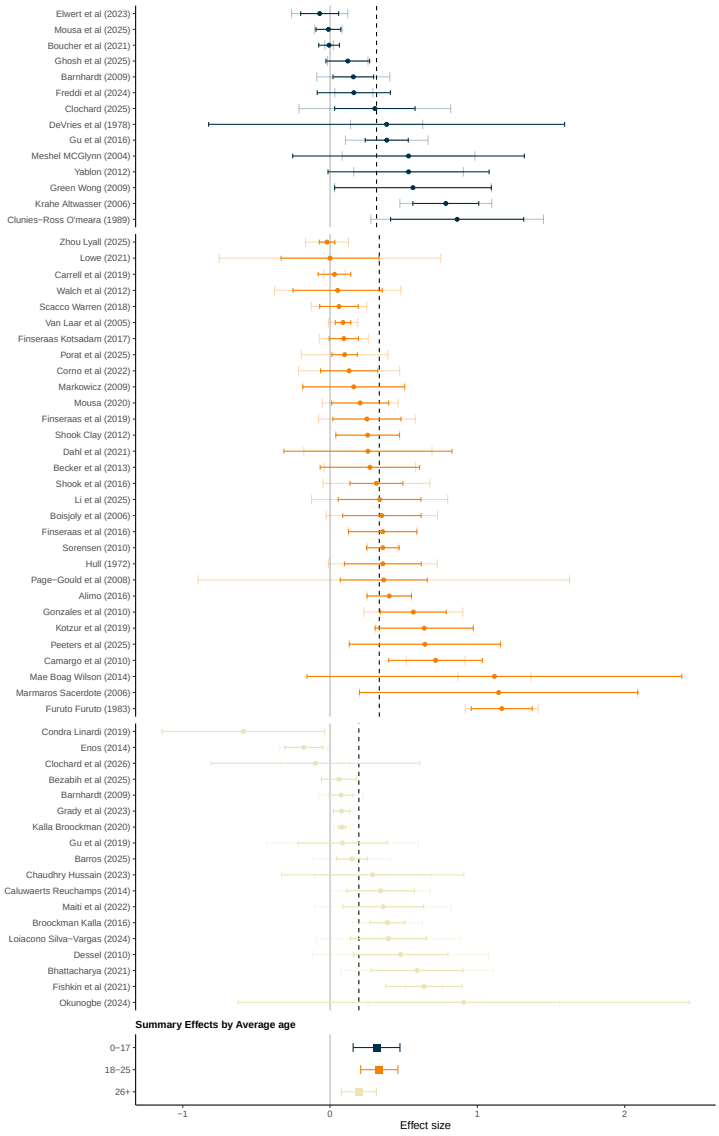


Figure G.5: Effect size as a function of the average age of the sample

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by average age of the sample. Visualization follows Popova et al. (2025).

Appendix G.6 Region

Table G.6: Meta-analytic averages by world region and meta-regression

	Panel A. Meta-analytic averages by world region					
	North America	Europe	East Asia & Oceania	South Asia	MENA	Sub-Saharan Africa
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{d}$	0.380***	0.380***	0.518	0.202	0.105	0.120**
SE	0.070	0.105	0.153	0.081	0.085	0.032
95% CI	[0.234, 0.525]	[0.148, 0.613]	[-0.158, 1.194]	[-0.029, 0.434]	[-0.099, 0.310]	[0.043, 0.197]
τ^2_{within}	0.087	0.091	0.020	0.017	0.045	0.021
$\tau^2_{between}$	0.067	0.069	0.050	0.024	0.039	0.000
I^2_{within}	54.514	55.680	19.994	31.879	51.182	64.377
$I^2_{between}$	42.028	42.254	50.249	45.073	44.413	0.000
Num. clusters	24	12	3	5	8	9
Num. obs	69	25	12	27	41	35
	Panel B. Meta-analytic regression (baseline: East Asia & Oceania)					
Europe	-0.147					
	(0.192)					
Middle East & North Africa	-0.421					
	(0.181)					
North America	-0.151					
	(0.174)					
South Asia	-0.317					
	(0.182)					
Sub-Saharan Africa	-0.388					
	(0.166)					
Constant	0.527					
	(0.159)					
Num. clusters	61					
Num. Obs.	209					

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by world region. Panel B reports a meta-analytic regression where the omitted category is East Asia & Oceania. Coefficients represent differences relative to this baseline category. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

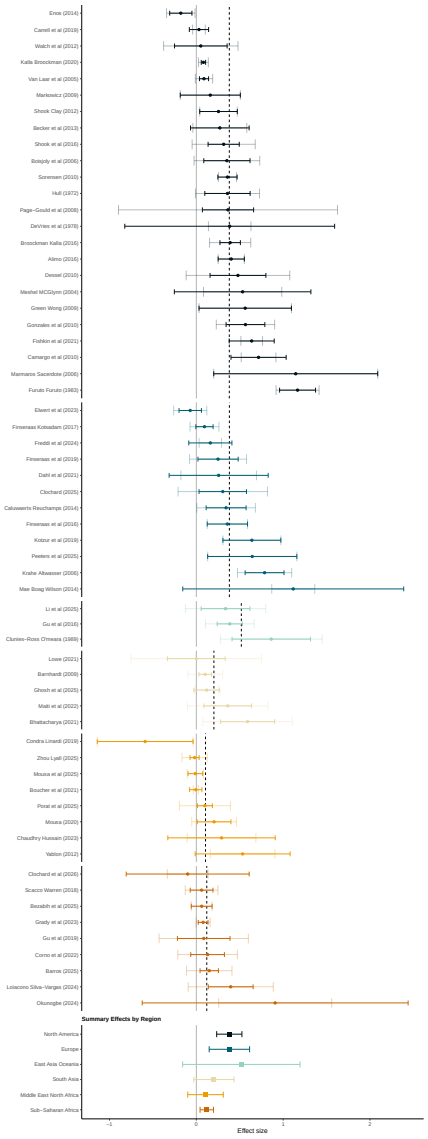


Figure G.6: Effect size as a function of the region

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by region. Visualization follows Popova et al. (2025).

Appendix G.7 Type of population

Table G.7: Meta-analytic averages by population group and meta-regression

	Univ. students & young adults	Children & adolescents	Army recruits	Specific adults	General adults
	(1)	(2)	(3)	(4)	(5)
Panel A. Meta-analytic averages by population group					
$\bar{d}$	0.415***	0.344***	0.156	0.166	0.164*
SE	0.078	0.072	0.064	0.075	0.059
95% CI	[0.251, 0.578]	[0.183, 0.504]	[-0.035, 0.348]	[-0.032, 0.363]	[0.037, 0.291]
τ^2_{within}	0.062	0.176	0.004	0.005	0.016
$\tau^2_{between}$	0.091	0.005	0.010	0.024	0.043
I^2_{within}	37.971	96.044	22.065	12.609	24.708
$I^2_{between}$	55.687	2.734	48.887	57.746	66.663
Num. clusters	22	13	5	6	16
Num. obs	65	43	10	33	58
Panel B. Meta-analytic regression (baseline: Univ. students & young adults)					
Children & adolescents	-0.076				
	(0.109)				
Army recruits	-0.233				
	(0.101)				
Specific adults	-0.235				
	(0.116)				
General adults	-0.228*				
	(0.100)				
Constant	0.412***				
	(0.082)				
Num. clusters	61				
Num. Obs.	209				

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by population group. Panel B reports a meta-analytic regression where the omitted category is university students and young adults. Coefficients represent differences relative to this baseline. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

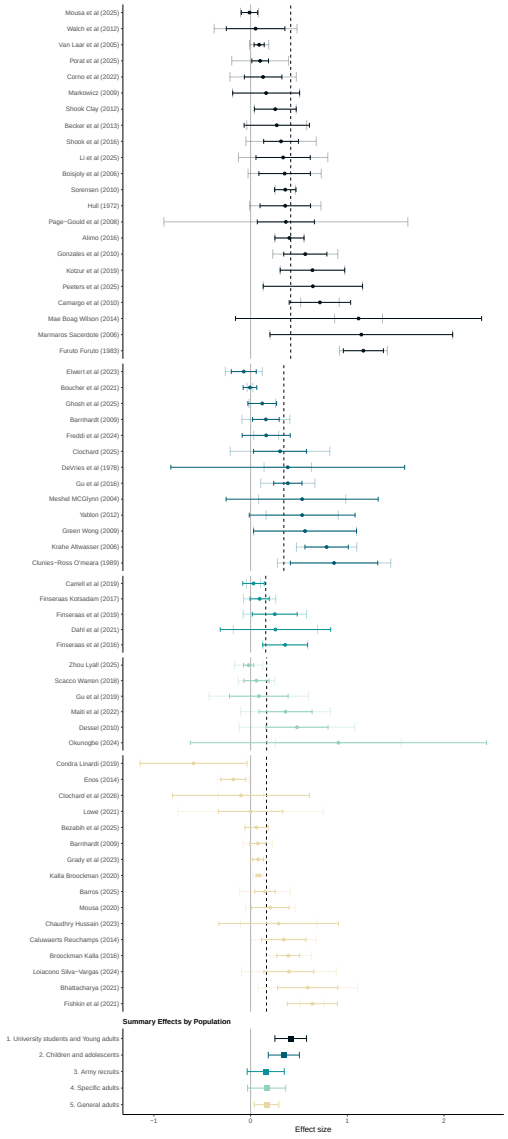


Figure G.7: Effect size as a function of the population

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by population. Visualization follows Popova et al. (2025).

Appendix G.8 Type of prejudice

Table G.8: Meta-analytic averages by target group and meta-regression

	Panel A. Meta-analytic averages by target group								
	Ethnicity / Race / Caste	Religion	Immigrants	Gender	LGBTQ+	Age	Disabilities	Politics	Others
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\bar{d}$	0.291***	0.197*	0.119	0.320	0.240	0.567*	0.817*	0.515	0.375*
SE	0.072	0.061	0.071	0.041	0.104	0.030	0.035	0.142	0.113
95% CI	[0.141, 0.441]	[0.050, 0.344]	[-0.044, 0.283]	[-0.195, 0.835]	[-0.108, 0.587]	[0.181, 0.953]	[0.376, 1.258]	[-1.289, 2.319]	[0.102, 0.649]
τ^2_{within}	0.101	0.053	0.001	0.018	0.000	0.053	0.024	0.021	0.094
$\tau^2_{between}$	0.072	0.015	0.038	0.000	0.033	0.000	0.000	0.023	0.060
I^2_{within}	57.286	65.980	2.679	34.587	0.000	58.331	31.303	39.788	56.125
$I^2_{between}$	41.018	19.370	87.154	0.000	82.389	0.000	0.000	43.715	35.810
Num. clusters	25	8	9	2	4	2	2	2	8
Num. obs	73	45	34	3	11	4	6	4	29
	Panel B. Meta-analytic regression (baseline: Race/Ethnicity/Caste)								
Religion	-0.081								
	(0.095)								
Immigrants	-0.153								
	(0.102)								
Gender	0.022								
	(0.087)								
LGBTQ+	-0.043								
	(0.125)								
Age	0.281								
	(0.078)								
Disabilities	0.538								
	(0.081)								
Politics	0.220								
	(0.158)								
Others	0.094								
	(0.138)								
Constant	0.285***								
	(0.072)								
Num. clusters	61								
Num. Obs.	209								

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by target group. Panel B reports a meta-analytic regression where the omitted category is Race/Ethnicity/Caste. Coefficients represent differences relative to this baseline category. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

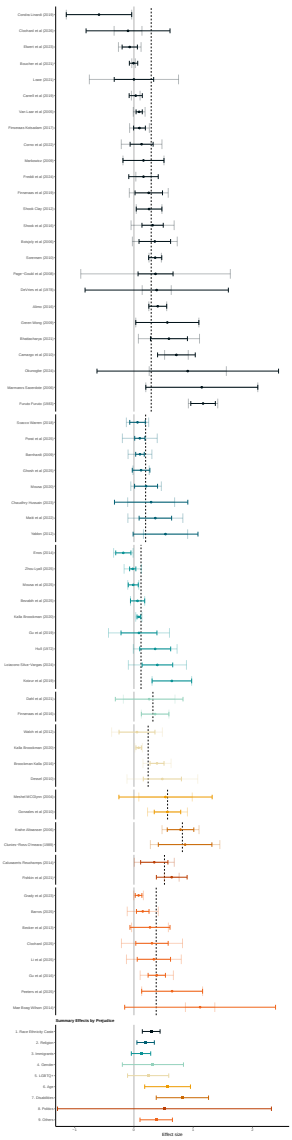


Figure G.8: Effect size as a function of the type of prejudice

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by prejudice. Visualization follows Popova et al. (2025).

Appendix G.9 Type of intervention

Table G.9: Meta-analytic averages by contact type and meta-regression

	Discussions	Coop. games	Sports	Classmates	Neighbors	Roommates	Army	Others
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Meta-analytic averages by contact type								
$\bar{d}$	0.453***	0.279**	0.137	0.085	0.266	0.330*	0.156	0.161
SE	0.091	0.073	0.099	0.054	0.134	0.099	0.064	0.092
95% CI	[0.261, 0.645]	[0.101, 0.457]	[-1.109, 1.383]	[-0.048, 0.219]	[-0.329, 0.862]	[0.083, 0.576]	[-0.035, 0.348]	[-0.068, 0.390]
τ^2_{within}	0.097	0.076	0.062	0.082	0.002	0.050	0.004	0.012
$\tau^2_{between}$	0.100	0.015	0.002	0.002	0.045	0.039	0.010	0.038
I^2_{within}	47.586	80.361	71.069	93.991	3.655	50.881	22.065	20.483
$I^2_{between}$	49.113	15.827	2.805	2.187	71.547	39.438	48.887	61.854
Num. clusters	19	8	2	10	3	7	5	7
Num. obs	57	32	17	38	15	23	10	17
Panel B. Meta-analytic regression (baseline: Discussions)								
Cooperative games	-0.170							
	(0.124)							
Sports	-0.338							
	(0.145)							
Classmates	-0.328**							
	(0.113)							
Neighbors	-0.153							
	(0.172)							
Roommates	-0.128							
	(0.135)							
Army	-0.281*							
	(0.110)							
Others	-0.277							
	(0.136)							
Constant	0.459***							
	(0.093)							
Num. clusters	61							
Num. Obs.	209							

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by contact type. Panel B reports a meta-analytic regression where the omitted category is Discussions. Coefficients represent differences relative to this baseline category. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$. Standard errors in parentheses.

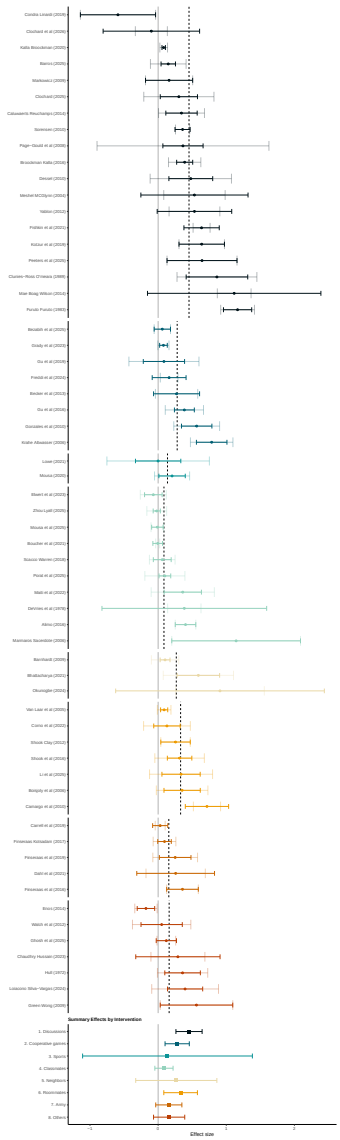


Figure G.9: Effect size as a function of the type of intervention

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by type of intervention. Visualization follows Popova et al. (2025).

Appendix G.10 Number of encounters

Table G.10: Meta-analytic averages by number of encounters and meta-regression

	1	2–10	10+
	(1)	(2)	(3)
Panel A. Meta-analytic averages by number of encounters			
$\bar{d}$	0.271***	0.406***	0.254***
SE	0.073	0.086	0.055
95% CI	[0.117, 0.426]	[0.219, 0.594]	[0.141, 0.368]
τ^2_{within}	0.071	0.077	0.055
$\tau^2_{between}$	0.057	0.051	0.060
I^2_{within}	54.299	51.128	45.551
$I^2_{between}$	43.154	34.349	50.195
Num. clusters	18	13	30
Num. obs	59	44	106
Panel B. Meta-analytic regression (baseline: 1 encounter)			
2–10 encounters	0.136		
	(0.112)		
10+ encounters	-0.017		
	(0.093)		
Constant	0.273***		
	(0.074)		
Num. clusters	61		
Num. Obs.	209		

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by number of encounters. Panel B reports a meta-analytic regression where the omitted category is one encounter. Coefficients represent differences relative to this baseline category. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

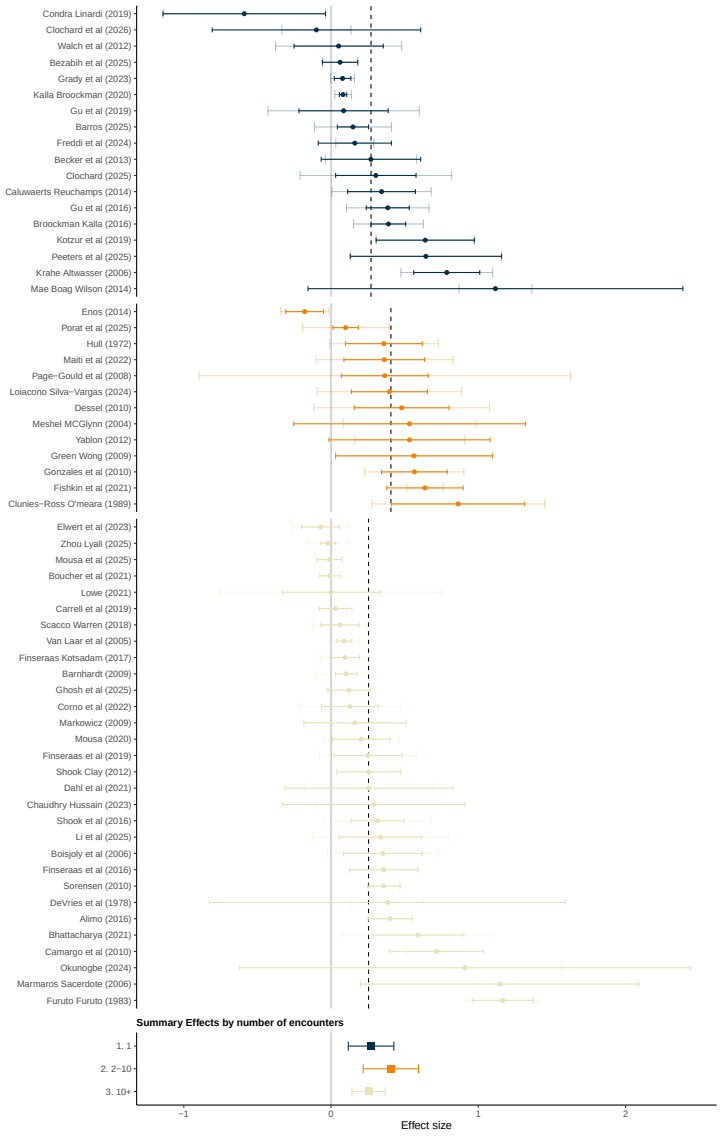


Figure G.10: Effect size as a function of the number of encounters

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by number of encounters. Visualization follows Popova et al. (2025).

Appendix G.11 Duration of intervention

Table G.11: Meta-analytic averages by contact duration and meta-regression

	1	2–30	31+
	(1)	(2)	(3)
Panel A. Meta-analytic averages by contact duration			
$\bar{d}$	0.271***	0.319***	0.294***
SE	0.073	0.078	0.059
95% CI	[0.117, 0.426]	[0.150, 0.487]	[0.172, 0.416]
τ^2_{within}	0.071	0.018	0.075
$\tau^2_{between}$	0.057	0.065	0.063
I^2_{within}	54.299	18.105	52.683
$I^2_{between}$	43.154	65.898	43.713
Num. clusters	18	14	29
Num. obs	59	50	100
Panel B. Meta-analytic regression			
Duration	-0.000		
	(0.001)		
Duration^2	0.000		
	(0.000)		
Constant	0.291***		
	(0.050)		
Num. clusters	61		
Num. Obs.	209		

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by duration category. Panel B reports a meta-analytic regression of effect size on continuous duration (in days). * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

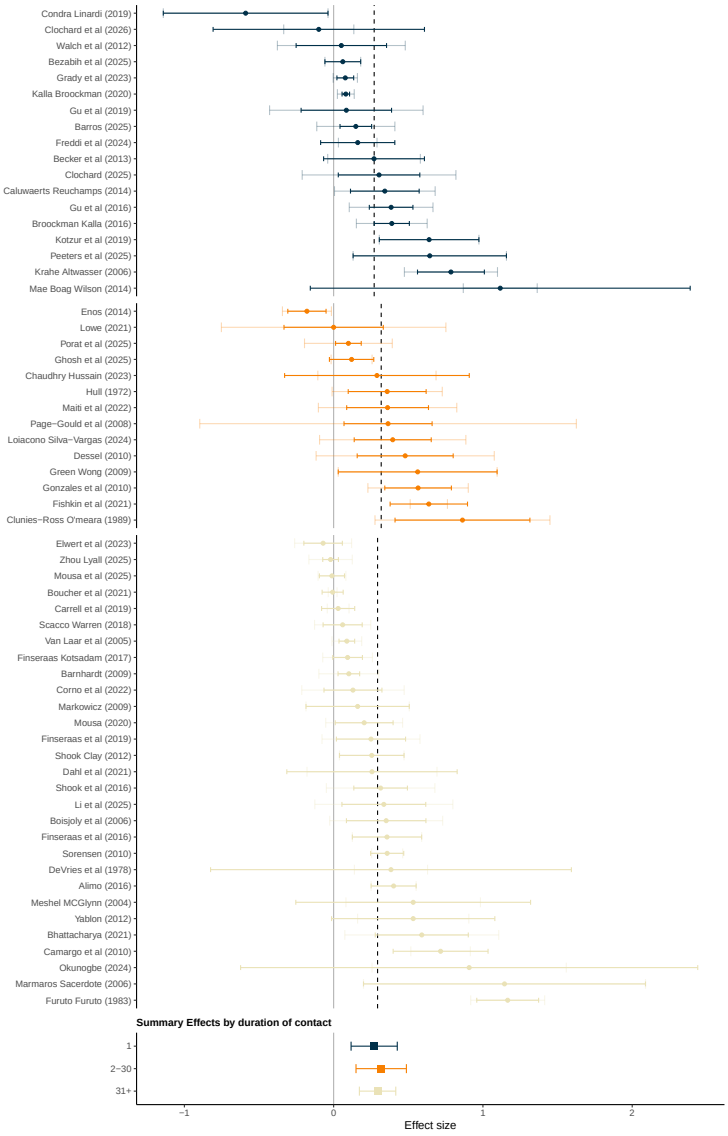


Figure G.11: Effect size as a function of the duration of the contact

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by duration of contact (days). Visualization follows Popova et al. (2025).

Appendix G.12 Type of outcome

Table G.12: Meta-analytic averages by outcome type and meta-regression

	Survey measure	Experimental game	Action (in the field)	IAT
	(1)	(2)	(3)	(4)
Panel A. Meta-analytic averages by outcome type				
$\bar{d}$	0.300***	0.142***	0.338*	0.095
SE	0.045	0.023	0.123	0.021
95% CI	[0.209, 0.391]	[0.086, 0.199]	[0.045, 0.631]	[-0.143, 0.333]
τ^2_{within}	0.068	0.016	0.107	0.005
$\tau^2_{between}$	0.065	0.000	0.062	0.000
I^2_{within}	48.869	86.628	60.736	20.555
$I^2_{between}$	46.960	0.000	34.894	0.000
Num. clusters	52	12	9	4
Num. obs	152	21	25	11
Panel B. Meta-analytic regression (baseline: Survey measure)				
Action (in the field)	0.054			
	(0.094)			
Experimental game	-0.062			
	(0.064)			
IAT	-0.087			
	(0.041)			
Constant	0.298***			
	(0.045)			
Num. clusters	61			
Num. Obs.	209			

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by outcome type. Panel B reports a meta-analytic regression where the omitted category is Survey measure. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

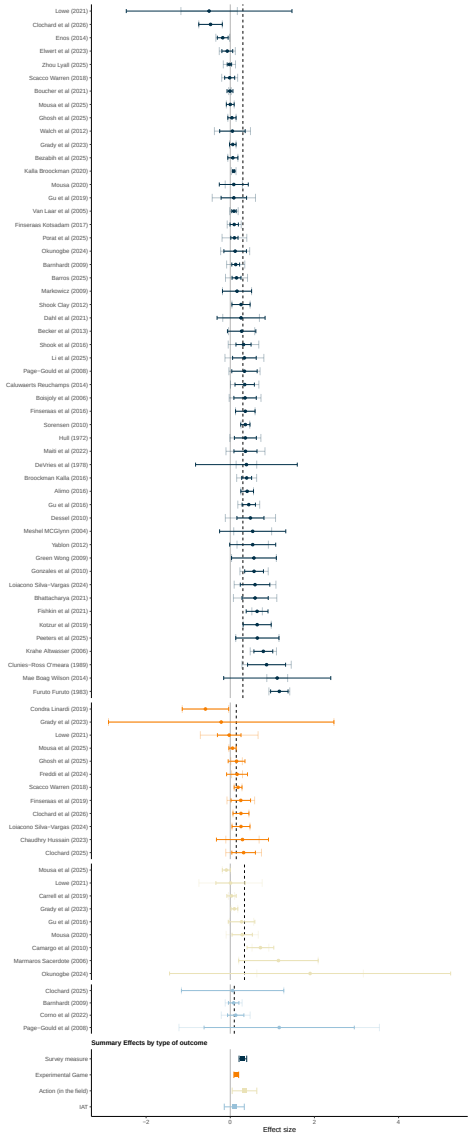


Figure G.12: Effect size as a function of the type of outcome

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by type of outcome. Visualization follows Popova et al. (2025).

Appendix G.13 Entire group or individuals met

Table G.13: Meta-analytic averages by target scope and meta-regression

	Entire outgroup	Specifically met
	(1)	(2)
Panel A. Meta-analytic averages by target scope		
$\bar{d}$	0.296***	0.228
SE	0.041	0.107
95% CI	[0.213, 0.378]	[-0.012, 0.468]
τ^2_{within}	0.057	0.028
$\tau^2_{between}$	0.063	0.095
I^2_{within}	45.119	21.968
$I^2_{between}$	50.419	75.060
Num. clusters	58	11
Num. obs	180	29
Panel B. Meta-analytic regression (baseline: Specifically met)		
Entire outgroup	-0.023	
	(0.137)	
Constant	0.313*	
	(0.131)	
Num. clusters	61	
Num. Obs.	209	

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by target scope. Panel B reports a meta-analytic regression where the omitted category is Specifically met. Coefficients represent differences relative to this baseline category. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

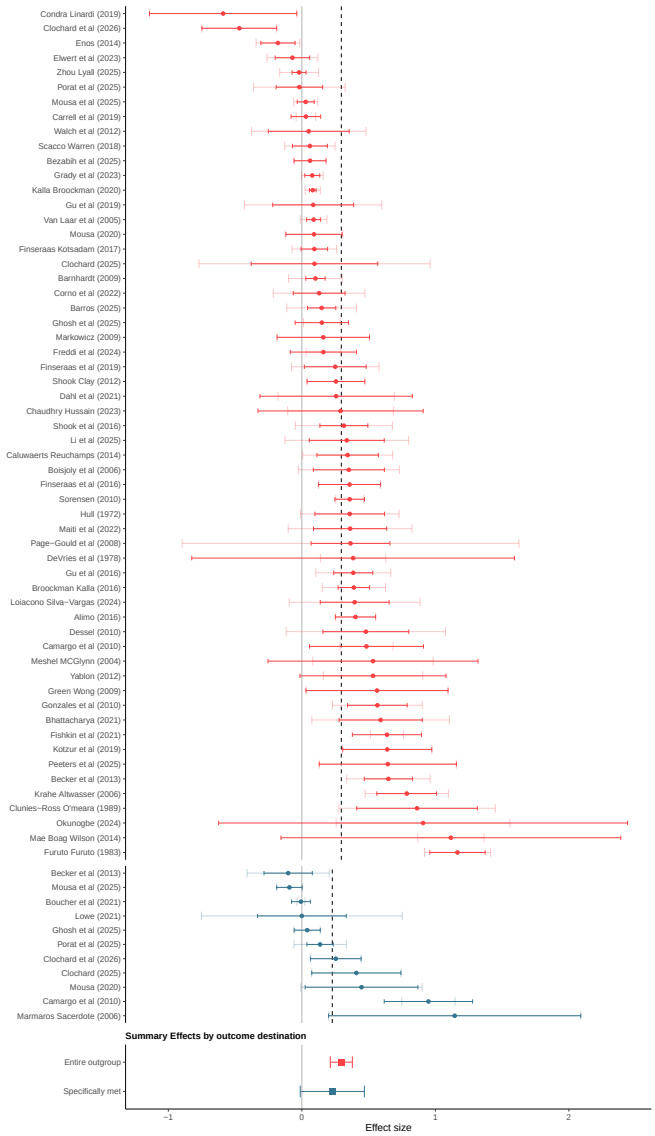


Figure G.13: Effect size as a function of whether the outcome measures prejudice against specific individuals met or the entire outgroup

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by who the outcome is about. Visualization follows Popova et al. (2025).

Appendix G.14 Time since contact

Table G.14: Meta-analytic averages by time since contact and meta-regression

Panel A. Meta-analytic averages by time since contact			
	0	1–30	31+
	(1)	(2)	(3)
$\bar{d}$	0.329***	0.165	0.340***
SE	0.047	0.093	0.071
95% CI	[0.233, 0.425]	[-0.037, 0.367]	[0.191, 0.489]
τ^2_{within}	0.077	0.049	0.017
$\tau^2_{between}$	0.048	0.094	0.068
I^2_{within}	58.774	33.514	18.520
$I^2_{between}$	36.208	64.535	72.150
Num. clusters	44	14	19
Num. obs	117	48	44
Panel B. Meta-analytic regression			
Time since contact	0.000		
	(0.000)		
Time since contact^2	-0.000		
	(0.000)		
Constant	0.287***		
	(0.041)		
Num. clusters	61		
Num. Obs.	209		

Note: Panel A reports subgroup meta-analytic averages from a multi-level model (clustered at the paper level), separated by time since contact category. Panel B reports a meta-analytic regression of effect size on time since contact (in days) and its quadratic term. * p < 0.05, ** p < 0.01, *** p < 0.005. Standard errors in parentheses.

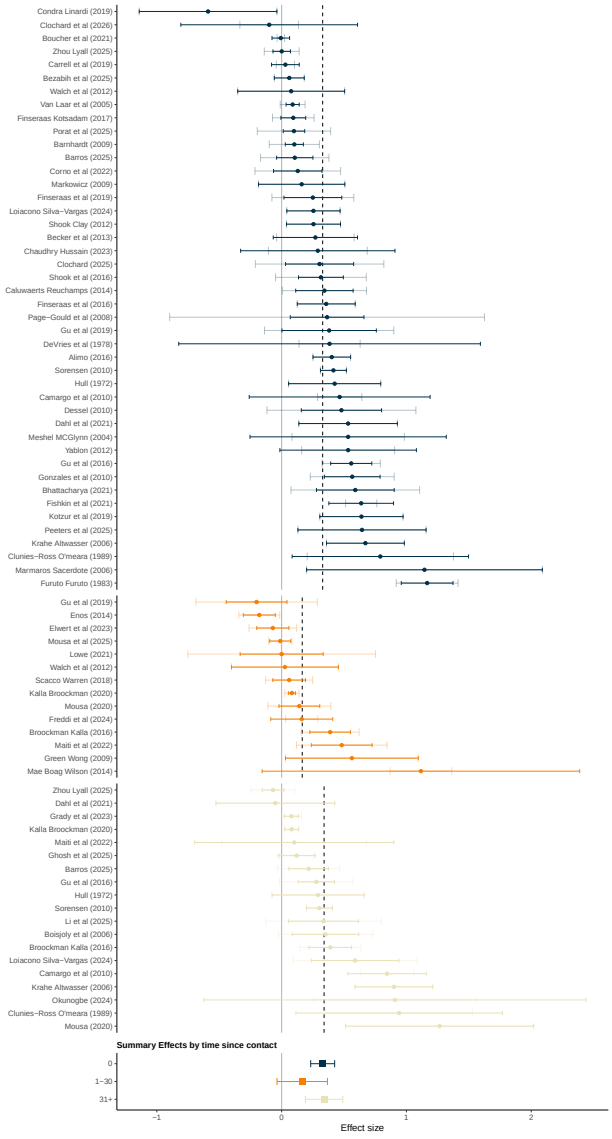


Figure G.14: Effect size as a function of the time between the end of the intervention and the measure

Note: In the top panels, each dot represents the pooled estimate from a study (using a random-effects model), with the dark horizontal lines indicating the corresponding 95% CIs. Light horizontal lines show 95% CIs constructed using the median standard error within each study (Fernández-Castilla et al., 2020). The bottom panel shows the average contact effect (and 95% CI) estimated using a multi-level model by the time since contact. Visualization follows Popova et al. (2025).

Appendix G.15 Allport's conditions: Meta-Analytic Regressions

Table G.15: Meta-analytic regression with by Allport's conditions

	Dependent variable: Effect size		
	Experiment	Author's evaluation	ChatGPT
	(1)	(2)	(3)
Common goals	-0.010*	-0.001	-0.003
	(0.004)	(0.003)	(0.003)
Equal status	-0.010	-0.004	-0.006
	(0.005)	(0.005)	(0.004)
Positive encounter	0.015*	0.009	0.003
	(0.006)	(0.008)	(0.005)
Support of authorities	0.004	0.011*	0.001
	(0.005)	(0.005)	(0.003)
Friendship potential	-0.001	-0.007	0.006
	(0.004)	(0.006)	(0.004)
Constant	0.364	-0.414	0.249
	(0.334)	(0.421)	(0.279)
Num. clusters	61	61	61
Num. Obs.	209	209	209

Note: This table reports the results of a meta-analytic regression of effect size on how much protocols fulfill Allport (1954)'s conditions for contact effectiveness. In Column 1, the variables for conditions are defined by the experiment. In Column 2, the variables are defined by the author. In Column 3, the variables are defined by ChatGPT. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$. Standard errors in parentheses.

Appendix H Publication bias and common goals condition

Table H.1: Publication bias for the negative link between common goal and effect size

	Standardized effect
	(1)
Intervention type: Neighbors	0.168
	(0.204)
Time since contact	0.0003***
	(0.0001)
Common goals experiment	−0.007*
	(0.003)
Friendship potential experiment	−0.007**
	(0.003)
Year	−0.021***
	(0.004)
SE effect	0.110
	(0.445)
Constant	43.361***
	(7.558)
Observations	209
R^2	0.276

The dependent variable is the standardized effect size. * p<0.05, ** p<0.01, *** p<0.005. Standard errors, clustered by paper, are in parentheses.