The Contact Hypothesis Revisited: Supplementary

Materials

Elizabeth Levy Paluck, Seth A Green, and Donald P. Green

May 1, 2018

Contents

1 Overview of Seven Borderline Studies 1

2 Other Related Meta-Analyses 6

3 Heterogeneity by Target and Population 6

4 P-Values Distribution and P-Curve Analysis 8

5 Effect Decay Over Time 9

6 Supplementary Tables 11

7 Supplementary Figures 13

1 Overview of Seven Borderline Studies

We considered hundreds of studies for inclusion before settling on our final sample of 27. A few

in particular were di cult, borderline cases over which we deliberated. Here, we describe seven

interesting, high-quality studies that we decided not to include, detail their results, and explain

our reasons for exclusion.

First, we considered the “railroad studies" by Stuart Cook and coauthors (Cook et al. 1971; Cook

1985; Blanchard et al. 1975). These seminal studies aimed to test the contact hypothesis “under

conditions approximating those of laboratory control" (Cook 1985, p. 453). Subjects, who were

“white Anglo students from the border South. . . [at] two fundamentalist colleges" (p. 454), were

recruited for a “part-time job" in which they operated “an imaginary railroad system with the help

of two co-workers" (p. 453), one white and one black, both confederates. The job took place for 2

hours per day for 20 days, and provided interracial contact in two contexts: a co-working setting,

and also a “30-minute break" in which team members ate lunch and participated in conversations

“guided by the trained confederates" that were intended to introduce the black co-worker and

allow the white co-worker “opportunities to voice egalitarian views regarding race relations" (p.

454). Subjects were tested four weeks after the jobs began and retested “[s]everal months later"

(p. 454). The study was conducted twice, three years apart. Relative to a control group that did

not experience interracial contact, experimental subjects showed significant attitude changes on

measures including whether they would vote “for a black person for Congress" or interview for a

job with a black person (p. 455).

We did not include these studies because random assignment “was not possible" (Cook et al.

1971, p. 46). Instead, Cook and his coworkers “approached the most negative case in each group

of potential subjects. If this person accepted, the next most negative was assigned to the control

group...our subject assignment procedure resulted in slightly more negative subjects than the

controls in the initial experiment" (p. 46).

Second, we considered Enos (2014), whose treatment entailed hiring Mexican nationals living

in the United States to wait on Boston commuter train platforms and speak Spanish to one another.

They were neither given scripts, nor informed of the purpose of the study, nor instructed to interact

with commuters. Following either three or ten days of exposure, commuters in the treatment

condition were significantly more likely to say that immigration levels should be decreased, and to

oppose offering undocumented immigrants without a criminal record a path to citizenship.

Although the study featured exposure to members of an outgroup, following Pettigrew and

Tropp (2006), we “define intergroup contact as actual face-to-face interaction between members of

clearly defined groups" (p. 754), and thus do not include the study. As Hewstone (2015, p. 431)

writes, the study did not have“meaningful cross-group face-to-face interaction or opportunity to

build a relationship and get beyond stereotypes."[[1]](#footnote-0)

Third, Van Laar et al. (2005) study the effects of interracial roommate pairings at UCLA, and

find that such roommate relationships typically “reduce prejudice"; that results do not generalize

to other outgroups besides those to which subjects’ roommates belong; and that whites with Asian

roommates tended to have “more negative" attitudes towards other groups (p. 329). “[E]xposure

to African American roommates had a particularly positive effect on respondents’ intergroup

attitudes" (p. 337).

There is some question, however, about whether roommate assignments can be analyzed as

though they are unconditionally random, rather than random conditional on information students

provide on housing questionnaires. Boisjoly et al. (2006, p. 1891) draw attention to this issue, writing:

we used data from the university housing o ce (instead of from a student survey) containing information on student housing preferences and initial assignment of roommates. This allows us to have reliable information on whether the roommate was randomly assigned, deal with nonresponse bias, and use initial roommate assignment rather than final roommate living arrangement in our estimations. It also allows us to statistically control for housing preferences in our estimations, which is important since roommate assignment is random, conditional on these housing preferences.

We reviewed approximately 20 interracial roommate studies, and found four that either control

for student housing preferences or verify that roommate assignment was unconditionally random.

We included these four.

It is difficult to estimate the extent to which validity is threatened by missing information about

roommate assignment preferences. Consider, however, a hypothetical about smoking, a common

field on roommate questionnaires. In the United States, both race and propensity to smoke are

correlated with socioeconomic status and geographic background, which might lead to systematic

differences in potential outcomes between whites paired with black or white roommates.

Readers who are unconcerned about this or similar hypotheticals should know that interracial

roommate pairings are typically associated with lower levels of prejudice (Van Laar et al. 2005) and

intergroup anxiety (Trail et al. 2009), but that the relationships themselves were often strained and

more likely to dissolve than same-race pairings (Shook and Fazio 2008).

Fourth, we considered Carrell et al. (2015), a study of the effects of racial heterogeneity within

squadrons at the Air Force Academy. The authors note “three sources of exogenous variation": the

placement of female cadets within squadrons without respect to race, the attrition provided by

students failing to matriculate, and the enrollment of students “who suffered injuries or illness

during the previous year’s basic training," which collectively “provide empirical evidence consistent

with random assignment into squadrons with respect to academic ability, athletic ability, and

leadership activity" (p. 7). The authors find that for white cadets, a one standard deviation increase

in the percentage of black squadron-mates has statistically significant effects both on stated racial

attitudes and on likelihood of choosing a black roommate during their sophomore years. We did

not include this study in our main sample, however, because the roommate assignment process

itself is not neutral with respect to race. The Air Force Academy creates squadrons by first assigning

women, then “male ethnic and racial minorities, then white male recruited athletes, then white

males who attended a military preparatory school, and then all remaining white male students" (p.

7). The authors argue that this creates balance “broadly consistent with a random draw from the

USAFA stratified random sorting algorithm" (p. 9), but we decided to limit our sample to studies in

which assignment was either controlled by the experimenters or strictly verified as random.

Fifth, Rao (2013) studied a policy change in India in 2008 that “forced many private schools in

Delhi to meet a quota of poor children in admissions" (p. 1), which, among students from wealthy

families, led to sharp discontinuities in exposure to poor classmates. Rao exploits a number ofeconometric strategies to identify treatment effects. However, not all schools complied immediately,

“either because they expected the policy to be overturned or because they felt the order was issued

too late for them to modify their admission procedures" (p. 7), while a separate subset of schools

was exempt for "historical reasons" (p. 1); these schools functioned as a primary control group.

Within treated schools, study groups are formed by alphabetic order of first name (p. 13), creating

levels of mixing between rich and poor students. Overall, for a group of about 2,000 students, Rao

finds positive, prosocial effects. Treated students are more likely to volunteer for charitable causes

at school, to split money evenly in a dictator game, and to discriminate less against poor children

outside of school in a play setting.

This study features strong treatment effects and admirably unobtrusive outcome measures;

however, we do not include studies whose assignment mechanisms are “plausibly exogenous" (p. 1)

rather than verifiably random. The variation created by study group assignment, moreover, leads

to both treatment and control groups receiving contact in the classroom.

Sixth, Fuegen (2000) studied the effects of interacting with a confederate who either behaved

as a “stereotype-consistent" or “stereotype-disconfirming" feminist (p. 21) in conversations with

female undergraduates. In both conditions, subjects were given a “list of interview questions" to

which the confederate gave "scripted responses" (p. 21). The stereotype-disconfirming confederate

“reported enjoying spending time with her boyfriend, cooking, shopping, and reading beauty

magazines. She also dressed in fashionable clothing, wore makeup and jewelry, and was congenial

with the participant" (p. 21). When speaking to the control group, the confederate conveyed a

“stereotype-consistent" demeanor by “naming History of American Women as her favorite class,

identifying with female role models, expressing interest in a career, and identifying the need to

educate women about inequality as one of the issues she felt most strongly about" (p. 21-22).

Fuegen found positive effects for treatment subjects on attitudes towards feminists immediately

after, while "positive affect decreased" (p. 46) for treatment subjects three to four months later.

However, we consider this study’s interventionto be related to, but distinct from, contact. As was

true of Enos (2014), we do not consider the subjects to have interacted with feminists, in the sense

that the word "interaction" implies a give and take, rather than one party reading strictly from a script.

Finally, Adams et al. (2003) studied the “effects of a Multicultural Relationship Enhancement

Workshop" (p. 281) in which undergraduate education majors participated in eight hours of

workshops, over four sessions, that trained them “in expressive speaking and empathetic listening

based on Relationship Enhancement Therapy" (p. 286). Groups were mainly composed of Latinxs

and Whites; subjects who participated in the workshop scored higher than waitlist control subjects

on measures of empathetic listening and speaking.

Though many studies in our final sample included contact as part of a bundled treatment, the

contact in Adams et al. (2003) does not seem exceptional for the environment from which the

students were drawn. The university “is an hour from the US-Mexico border, and the majority of

the participants were Latino-Americans" (p. 289). The treatment is not theoretically isomorphic

with contact, but rather a series of structured interactions within a setting in which both treatment

and control groups are already accustomed to intergroup contact. We did not consider this to be of

probative value for the contact hypothesis itself.

2 Other Related Meta-Analyses

We were fortunate to have a number of related meta-analyses to build from, many of whose results

we did not incorporate directly but nonetheless influenced our thinking. We appreciate the efforts

of Aboud et al. (2012) on anti-prejudice interventions in early childhood; Davies et al. (2011) on

cross-group friendships; Johnson et al. (1983, 2000) on intergroup cooperation; Lipsey and Wilson

(1993) on psychological interventions generally; Miles and Crisp (2013) on imagined contact;

Bowman (2011) on civic engagement; Gilbert and Ricketts (2008) and Christian et al. (2014) on

ageism; and Smith et al. (2009) on sexual prejudice.

3 Heterogeneity by Target and Population

Tables A1 and A2 present an overview of effect size heterogeneity by target of prejudice and

population from which participants are drawn. Table A1 separates studies by target of prejudice,and ranks the effect sizes from smallest to largest. Table A2 provides a cross-tab of effect sizes by

both target and population.

[Tables A1 and A2 about here]

As Table A1 suggests, there is significant heterogeneity of effect size by target. Four categories

of prejudice—against immigrants and foreign nationals, LGBTQ individuals, the elderly, and

women—come in below what is typically thought of as a small effect in the social sciences ( = .2).

Two others—members of other ethnic or racial groups, or other religious groups—fall within a

small to medium range (.2 < < .5). By contrast, contact’s effects on reducing prejudice towards

individuals with intellectual or physical disabilities are apparently quite large.

What drives this heterogeneity is di cult to say. A consequence, however, is that the long right

tail of the distribution plays a large part in driving average effect sizes. When studies focusing on

reducing prejudices towards individuals with intellectual disabilities are removed from the analysis,

the average effect size drops from 0.394 to 0.225, a reduction of 43%. By contrast, removing studies

aimed at reducing prejudice towards immigrants and foreign nationals raises the effect size to

0.422, an increase of 7%.

We suggest three lessons from this exercise. First, researchers and policy-makers should expect

smaller effects of contact with immigrants and foreign nationals; this accords with results from Enos

(2014), which suggest that incidental exposure to foreigners may increase exclusionary attitudes.

Second, contact appears to be a comparatively effective means at reducing prejudice towards people

with physical and intellectual disabilities, which has important implications for discrimination

in schools and the workplace. Third, researchers and policymakers focusing on racial, ethnic,

religious, gender, and sexual minority prejudice should interpret the pooled result of this and

similar meta-analyses with caution. It appears that not all attitudes are equally amenable to change

following contact, and that the prejudices typically underpinning social conflict are more stable

than a pooled estimate of effect sizes may imply.

4 P-Values Distribution and P-Curve Analysis

Another way to present the effects of contact on prejudice is to examine the distribution of p-values

(Simonsohn et al. 2014). Of the 27 statistical calculations comprising our primary meta-analytic

sample, twelve are statistically significant at p < .05. Two others Boisjoly et al. (2006); Burns et al.

(2015) are significant at the p < .1 level, and the remaining 13 are not statistically significant. We

conducted a p-curve test of the 12 statistically significant results. Figure A1, displays the results of

this test:

[Figure A1 about here]

Overall, according to the criteria laid out by Simonsohn et al. (2014) our collated studies have

high evidentiary value and are consistent with a hypothesis of no publication bias. At the same

time, the majority of the statistical tests we evaluated are not statistically significant.

We caution readers, however, not to place too much weight on this test for a number of

reasons. First, a number of studies in our sample include both experimental and non-experimental

comparisons, and many of the non-experimental comparisons are statistically significant while the

experimental comparisons are not. Hull (1972) and Furuto and Furuto (1983) both include outside

control groups of subjects who are similar to randomly assigned subjects on observables but who

had zero probability of assignment to treatment. Page-Gould et al. (2008), meanwhile, present

a variety of statistically significant analyses partitioned by covariates that were not randomly

assigned, and thus were not included in our primary analysis. For Marmaros and Sacerdote (2006),

roommate contact is just one of many explanatory variables; the main focus of the paper is the

effects of geographic clustering on college students’ social networks. It is thus di cult to comment

on whether significant results are systematically more likely to be published, as most papers in

our sample present a wide variety of outcome measures, at least some of which are statistically

significant, but not all of which are relevant to our meta-analysis.

Second, we note that the relationship between publication and effect size is very noisy. The 27

effect sizes in our primary sample come from 26 separate papers. Nine of those papers were, as

of January 2017, unpublished; five are working papers, and four are unpublished dissertations.

Collapsing unpubished papers into one category, we find a weak relationship between publication

status and effect size ( = .036, p = 0.867) Analyzed separately, published papers have a random

effects estimate of 0.388; working papers an estimated effect size of 0.169; and dissertations an

average estimated effect of 0.824, giving the impression that unpublished dissertations have larger

effects on average.

These effects, however, are in large part driven by one outlier: DiTullio (1982), who found that

contact with intellectually disabled coworkers reduced prejudice for 38 Philadelphia custodians by

2.6 standard deviations. With this one study removed, the average effect size of dissertations drops

to 0.287, and the average estimated effect size of all unpublished studies without DiTullio is 0.195.

However, the difference between published and unpublished studies is still not significant ( =

-.167, p = .231).

5 Effect Decay Over Time

Given our focus on delayed outcome measurements, some readers might wonder if effects in

our sample attenuate over time. To investigate this possibility, we recorded the number of days

separating the beginning of treatment from outcome measurement, paying special attention to

studies that featured multiple post-tests. Overall, we found very little evidence for effects decaying

over time, either across or within studies. A graphical overview of our results is presented in figure

A2.

[ Figure A2 about here ]

Here, we note non-monotonic relationships between time and effect size in two of seven studies

with multiple post-tests (Broockman and Kalla 2016; Camargo et al. 2010), though the samples

within these studies differ over time due to attrition. One study with two treatment arms (Clunies-

Ross and O’Meara 1989) shows increased effect sizes over time, and three show small declines

Krahé and Altwasser (2006); Sorensen (2010); Hull (1972). With all of these outcomes included,

controlling for target and within-study fixed effects, a day of delay is associated with a decline in

effect size of -0.0004. Results are substantially the same when comparing effect sizes only within

studies with multiple post-tests.

We also note that estimating how much time has elapsed in between treatment and outcome

measurements involved some guesswork. Many studies record only approximate timelines, or

measured outcomes within a range of a few weeks. We adjudicated such cases by taking the

approximate midpoint within the ranges given. for instance, Scacco and Warren (2018) measured

outcomes “four to six weeks" (p. 14) after a 16-week program, which we recorded as taking place

21 weeks, or 147, days after treatment.

References

Aboud, F. E., Tredoux, C., Tropp, L. R., Brown, C. S., Niens, U., and Noor, N. M. (2012). Interventions

to reduce prejudice and enhance inclusion and respect for ethnic differences in early childhood:

A systematic review. Developmental Review, 32(4):307–336.

Adams, E. M., Waldo, M., Steiner, R., Mayfield, R., Ackerlind, S. J., and Castellanos, L. P. (2003).

Creating peace by confronting prejudice: Examining the effects of a multicultural communication

*skills group intervention. International Journal for the Advancement of Counselling, 25(4):281–291.*

Blanchard, F. A., Adelman, L., and Cook, S. W. (1975). Effect of group success and failure upon

interpersonal attraction in cooperating interracial groups. Journal of Personality and Social

Psychology, 31(6):1020.

Boisjoly, J., Duncan, G. J., Kremer, M., Levy, D. M., and Eccles, J. (2006). Empathy or antipathy? the

impact of diversity. The American Economic Review, 96(5):1890–1905.

Bowman, N. A. (2011). Promoting participation in a diverse democracy: A meta-analysis of college

diversity experiences and civic engagement. Review of Educational Research, 81(1):29–68.

Broockman, D. and Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-

door canvassing. Science, 352(6282):220–224.

Burns, J., Corno, L., and La Ferrara, E. (2015). Interaction, prejudice and performance: Evidence

from south africa. Technical report, Working Paper. forthcoming.

Camargo, B., Stinebrickner, R., and Stinebrickner, T. (2010). Interracial friendships in college.

*Journal of Labor Economics, 28(4):861–892.*

Carrell, S. E., Hoekstra, M., and West, J. E. (2015). The impact of intergroup contact on racial

attitudes and revealed preferences. Technical report, National Bureau of Economic Research.

working paper.

15

Christian, J., Turner, R., Holt, N., Larkin, M., and Cotler, J. H. (2014). Does intergenerational

contact reduce ageism: When and how contact interventions actually work? Journal of Arts and

Humanities, 3(1):1–15.

Clunies-Ross, G. and O’Meara, K. (1989). Changing the attitudes of students towards peers with

disabilities. Australian Psychologist, 24(2):273–284.

Cook, S. W. (1985). Experimenting on social issues: The case of school desegregation. American

*Psychologist, 40(4):452.*

Cook, S. W., Wrightsman, L. S., and Wrightsman, S. (1971). The Effect of Unintended Interracial

*Contact Upon Racial Interaction and Attitude Change. Educational resources in information center,*

US Department of Health, Education & Welfare.

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., and Wright, S. C. (2011). Cross-group friendships

and intergroup attitudes: A meta-analytic review. Personality and Social Psychology Review,

15:332–351.

*DiTullio, B. J. (1982). The effect of employing trainable mentally retarded (TMR) students as workers*

*within the Philadelphia public school system: Attitudes of supervisors and non handicapped co-workers*

*towards the retarded as a result of contact. PhD thesis.*

Enos, R. D. (2014). Causal effect of intergroup contact on exclusionary attitudes. Proceedings of the

*National Academy of Sciences, 111(10):3699–3704.*

Fuegen, K. (2000). Comparing paradigms of stereotype change: The case for contact. Arbor Ciencia

*Pensamiento Y Cultura.*

Furuto, S. B. and Furuto, D. M. (1983). The effects of affective and cognitive treatment on attitude

*change toward ethnic minority groups. International Journal of Intercultural Relations, 7(2):149–*

165.

Gilbert, C. N. and Ricketts, K. G. (2008). Children’s attitudes toward older adults and aging: A

synthesis of research. Educational Gerontology, 34(7):570–586.

Hewstone, M. (2015). Consequences of diversity for social cohesion and prejudice: The missing

dimension of intergroup contact. Journal of Social Issues, 71(2):417–438.

Hull, W. F. (1972). Changes in world-mindedness after a cross-cultural sensitivity group experience.

*The Journal of Applied Behavioral Science, 8(1):115–121.*

Johnson, C., Schaller, M., and Mullen, B. (2000). Social categorization and stereotyping: ’you mean

*i’m one of ’them’?’. British Journal of Social Psychology, 39(1):1–25.*

Johnson, D. W., Johnson, R. T., and Maruyama, G. (1983). Interdependence and interpersonal

attraction among heterogenous and homogeneous individuals: A theoretical formulation and a

meta-analysis of the research. Review of Educational Research, 53(1):5–54.

Krahé, B. and Altwasser, C. (2006). Changing negative attitudes towards persons with physical

*disabilities: An experimental intervention. Journal of Community &amp; Applied Social Psychology,*

16(1):59–69.

Lipsey, M. W. and Wilson, D. B. (1993). The e cacy of psychological, educational, and behavioral

treatment. Confirmation from meta-analysis. The American psychologist, 48(12):1181–1209.

Marmaros, D. and Sacerdote, B. (2006). How do friendships form? The Quarterly Journal of

Economics, pages 79–119.

Miles, E. and Crisp, R. J. (2013). A meta-analytic test of the imagined contact hypothesis. Group

*Processes & Intergroup Relations, 17(1):3–26.*

Page-Gould, E., Mendoza-Denton, R., and Tropp, L. R. (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.*

Pettigrew, T. F. and Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. Journal of

*Personality and Social Psychology, 90(5):751–783.*

Rao, G. (2013). Familiarity Does Not Breed Contempt: Diversity, Discrimination and Generosity in

Delhi Schools.

Scacco, A. and Warren, S. S. (2018). Can social contact reduce prejudice and discrimination?

evidence from a field experiment in nigeria. American Political Science Review, pages 1–24.

Shook, N. J. and Fazio, R. H. (2008). Roommate relationships: A comparison of interracial and

same-race living situations. Group Processes & Intergroup Relations, 11(4):425–437.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: a key to the file-drawer. Journal of

*Experimental Psychology: General, 143(2):534.*

Smith, S. J., Axelton, A. M., and Saucier, D. A. (2009). The effects of contact on sexual prejudice: A

meta-analysis. Sex Roles, 61(3-4):178–191.

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

intergroup relations. PhD thesis, The University of Michigan.

Trail, T. E., Shelton, J. N., and West, T. V. (2009). Interracial roommate relationships: Negotiating

*daily interactions. Personality and Social Psychology Bulletin, 35(6):671–684.*

Van Laar, C., Levin, S., Sinclair, S., and Sidanius, J. (2005). The effect of university roommate

contact on ethnic attitudes and behavior. Journal of Experimental Social Psychology, 41(4):329–345.

1. One could argue, however, that Enos’s study provides a test of what Pettigrew and Tropp (2006, p. 753) call “the mere exposure perspective [which] suggests that, all things being equal, greater contact and familiarity with members of other groups should enhance liking for those groups." In this light, the study could be taken as evidence against the perspective that brief, incidental contact will produce changes and suggests instead that Allport’s structuring conditions are crucial for changing minds. [↑](#footnote-ref-0)