Pre-Analysis Plan:

Does Exposure to Ethnic Minorities Affect Support for Welfare Dualism? Evidence From a Field Experiment.

Henning Finseraas[1] *and Andreas Kotsadam*[2]

Introduction

An influential argument on how ethnic diversity might undermine public goods production, cooperation and collective action rests on assumptions from sociological/social-psychological theories of group conflict. The core assumptions of this perspective are that people will develop social group identifications where ethnic similarities typically function as group boundaries. Negative views on out-group members are caused by real or perceived competition between your in-group and out-groups over scarce resources such as rights and social status (see e.g. Bobo 1999; Semyonov, Raijman, and Gorodzeisky 2006).

The core assumptions and implications of conflict theory stand in stark contrast to those of the competing inter-group contact theory (see e.g. Pettigrew 1998). According to this perspective, prejudice and negative stereotyping of minorities might decline with contact with out-group members. Since the frequency of contact will increase with ethnic diversity, any negative effects of immigration on welfare state support caused by increased competition over resources might be off-set by the positive consequences of increased contact with members of the minority group. While inter-group contact theory is frequently used to debunk the bleak perspective of conflict theory, it is often forgotten that

[1]Institute for Social Research, P.box 3233 Elisenberg, 0208 Oslo, Phone: +47 95169855, Norway e-mail: henning.finseraas@samfunnsforskning.no

[2]Department of Economics, University of Oslo, P.box 1095 Blindern, 0317 Oslo, Norway, Phone: +47 40338176, e-mail: andreas.kotsadam@econ.uio.no.

contact theory proposes quite restrictive conditions for in what contexts contact will reduce majority-minority conflict: Contact will reduce tensions only if those in contact have equal status in the particular context, if they share common goals, if they are in a cooperative context, and if the contact takes place under some form of authority (see Pettigrew 1998).

The existing empirical literature on the consequences of ethnic diversity tends to overlook how important these different underlying assumptions are, and simply regress e.g. some indicator of views on diversity on an indicator of ethnic diversity (e.g. Senik, Stichnoth, and Van der Straeten 2009). The discrepancy between the theoretical and empirical model implies that the empirical estimates are not very informative about the importance of minority-majority contact. We take the underlying assumptions of contact theory more seriously than in the previous literature and test it in a setting where theory suggests it is most likely to hold. We also set up a research design with a random allocation to give our estimates a causal interpretation.

Specifically, we propose an explicit test of contact theory in a context where the strict assumptions of the theory is plausible, namely in the military. In fact, the initial inspiration and empirical support for contact theory is from a study of integration of Black soldiers into the US Army (see Pettigrew 1998). Soldiers of private rank have equal social status within the army, they share the common goals of the unit, they need to cooperate to solve their tasks, and contact takes place in a context with an explicit, enforcing authority. Moreover, the army is a promising venue to study social interaction since the soldiers cannot determine who they want to share rooms with and who they want to serve with. Thus, biases due to self-selection into social interactions based on own preferences (such as prejudice) are reduced, and we have exogenous exposure to contact with out-group members.

In this plan we describe the analytic decisions that will be made in the analysis of the data in the project. That is, we describe the hypotheses to be tested and how they will be tested. The description includes how the variables will be created, how we deal with attrition and missing values, and how the estimation equation will look like. We also conduct a power analysis based on a pilot study. All deviations from the plan will be highlighted in the final paper.

Description of sample

Our sample is the whole population of soldiers going to the North Brigade of the Norwegian Armed Forces (NAF), August 2014. The population was around 2000 in the first round of data collection carried out in August, at the first day of military service where medical examinations and other tests are also conducted. At this stage the soldiers have not been allocated to rooms yet. The population will decrease from first to second round of data collection since many are discharged during the first two months. We expect a final sample of between 1200 and 1600 individuals based on interviews with experienced staff in the NAF. We randomize roommates by providing the NAF with a excel sheet which allocates soldiers randomly into rooms. Men and women share rooms, but room assignment follows a rule where mixed rooms consist of at least two women (if possible). The share of Norwegians sharing room with ethnic minorities is thought to be around 25 percent (it was 26 percent in the pilot, see below).

We will examine randomization balance according to sibling composition, parental employment and education, educational aspirations, parents divorced/separated, and own IQ. See Appendix 1 for operationalizations. We consistently condition on platoon ("tropp") fixed effects.[3] We will examine attrition on these

[3]In fact, not all battalions divide soldiers into platoons during the first 8 weeks and we will create quasi-troops where real troops do not exist. These quasi-troops will be created by combining battalion with room building.

same covariates.

Key data sources

The data sources are survey data collected at the first day of military service (Baseline/t_1) and a similar data collection at the end of the recruitment period, i.e. after about two months service (t_2). The survey data will be supplemented with test score data from the mandatory pre-registration interview ("sesjon"). The main survey questions are described in the text below and added in the Appendix. Depending on resources we will conduct a follow-up at the end of the military service.

Hypotheses and construction of dependent variables

Following the classic studies of contact theory (e.g Allport 1954), we expect majority member soldiers who are randomly allocated a roommate of ethnic minority background to develop more positive attitudes toward ethnic minorities, and we expect support for welfare dualism to decrease among those with a roommate with minority background. We particularly expect this result to hold in our setting as it is a setting of cooperation and as laboratory experiments in cooperative settings often find relations across groups to improve while the opposite holds in competitive settings (see Boisjoly, Duncan, Kremer, Levy, and Eccles 2006, for an overview). Nonetheless, we are open to the possibility that there may also be negative effects on attitudes of exposure to second generation immigrants, and we will therefore conduct two-sided tests of our hypotheses.

Our main outcome of interest is "immigrants same rights", which measures welfare universalism versus welfare dualism. The variable is a categorical vari-

able based on the following question asked to the soldiers at day one and in week 8:

"Do you agree or disagree with the following statements:

Refugees and immigrants should not have the same rights to social assistance as Norwegians.

1= Strongly agree

2= Agree

3= Neither agree nor disagree

4= Disagree

5= Strongly disagree".

Recode into "immigrants same rights": 4 and 5=1, 1 to 3=0

We will also explore two possible mechanisms for this results. Is the result driven by changed views of immigrants work ethic or is it a broader positive impression of immigrants that is affected? First we will thus test whether the respondents think the work ethic of immigrants is more similar if exposed to immigrants. This is measured with the variable "immigrants similar work ethic", which is based on the following question:

"In general, immigrants have poorer work ethics than Norwegians.

1=Strongly agree

2= Agree

3= Neither agree nor disagree

4= Disagree

5= Strongly disagree".

Recode into "immigrants similar work ethic": 4 and 5=1, 1 to 3=0

We will further explore whether there is an effect on attitudes towards immigrants in general ("better country"). The specific question asked is:
"Is Norway made a worse or better place to live in by people coming to live here from other countries?
1= Worse place to live
7= Better place to live".

We further expect there will be a more positive effect of minority roommate if the minority roommate has a higher relative IQ score (see how we measure relative IQ below). We expect roommate IQ to matter in so far as negative views on minorities reflects statistical discrimination which will be more strongly updated if one has contact with a high-IQ minority person (Carrell, Hoekstra, and West 2013).

Independent variables of main interest

The main independent variable is a dummy variable which equals one if there is at least one person with at least one parent born in a non-Western country. This variable is based on the answers to two different questions about the mother and father respectively. The question asked is:
"In what country is your mother/father born?
1=Norway
2=Other Nordic country
3=Other European country
4=A country in North America
5=A country in South America
6=A country in Asia

7=A country in Oceania

8=A country in Africa."

We code the person as having a non-Western parent if he or she answers 5 to 8.[4] We choose to have non-Western ethnic background rather than simply foreign background as we think the effect is larger for this group. In fact, all are Norwegian citizens and having a parent from e.g. another Nordic country is not as likely to even be visible and hence not noticed by the other peers.

In addition to using a dummy variable we will also use the share of second generation non-Western immigrants among the others in the room (jackknifed). This is further discussed in the section on identification of peer effects. IQ of ethnic minority roommates is measured as a dummy equal to 1 if the ethnic minority roommate has an IQ score above the median of the minority soldiers in the battalion.

Identification of peer effects

Peer effects interests social scientists across a range of disciplines. The notion that people are affected by other people is commonly held yet it turns out to be difficult to prove. The by far most commonly estimated model of peer effects (Sacerdote 2011) is some version of the following equation:

$$Y_i = a + \beta_1 \overline{Y} - i + \gamma_1 X_i + \gamma_2 \overline{X} - i + \epsilon_i \quad (1)$$

Where Y_i is the outcome of interest for individual i which is thought to

[4] As Oceania can arguably be coded as both Western or non-Western we do a robustness test and code them as Western and we also test to drop these observations. We do not expect that many respondents have parents born in Oceania, however.

be a function of the average outcomes of the peers ($\overline{Y} - i$), the individuals own characteristics (X_i), and the characteristics of the peers ($\overline{X} - i$). Being interested in welfare dualism, one could imagine a test of attitudes towards welfare dualism as a function of the peers attitudes toward dualism and the individuals own and the peers background (including e.g. ethnicity). Without random (or at least plausibly exogenous) allocation of individuals to peers, identification of equation 1 will most likely be subject to severe selection bias. It is common that people choose or become allocated to people similar to themselves. More concretely, if one were to test the contact hypothesis using observational data on e.g. friends it is very likely that there would be a positive bias in the estimation of the peer effect as individuals with negative attitudes toward immigrants are more likely to support welfare dualism and less likely to be friends with people of other ethnic groups.

The selection problem is not the only problem facing researchers interested in identifying equation 1. Following Manski (1993) it is common to distinguish between three types of effect:

1) Endogenous effects whereby the individual is affected by the behavior of the other individuals. People try to estimate this effect by looking at β_1.

2) Exogenous effects whereby individuals are affected by the characteristics of the peers. The hope of the researcher is to identify this by looking at γ_1.

3) Correlated effects whereby there is a correlation between individuals and their peers because they face similar environments or because of selection.

The problem of correlated effects is a major one as stated above but it can be solved by randomly allocating peers to individuals (see Sacerdote 2011, for a review of the literature). In estimating endogenous effects the problem is that if peers affect the outcomes of each other it is difficult to separate the effects of

the peers' on individual i's outcome from the effect of individual i on the peers' outcomes. Manski (1993) labels this the reflection problem. Another practical problem is that, even with random assignment of peer groups, separate identification of β_1 and γ_1 is often difficult, in particular since peer characteristics affect peer outcomes, and most papers therefore rely on the reduced form effect which is a combination of the two (Sacerdote 2011). Angrist (fortcoming), however, notes that identification of β_1 is often more problematic than this as it is driven by a common variance in outcomes and he strongly cautions against using outcome-on-outcome estimations.

Angrist (fortcoming) is also skeptical to studies where individuals whose background characteristics are thought to be important are also included in the sample thought to be affected by other individuals. He instead argues that the most compelling evidence comes from studies whereby there is a clear separation of the individuals thought to be affected and the peers thought to provide the mechanisms for the peer effects. This type of design is applied in Kling, Liebman, and Katz (2007) who analyze the effects of neighborhoods on individuals randomly assigned to receive housing vouchers in the Moving to Opportunity program. The neighborhood effects are only estimated by using characteristics of the neighbors but the neighbors themselves do not otherwise play any role and no effects on these old neighbors is estimated. Similarly, Angrist and Lang (2004) investigate the effects of low-income peers in the classroom by estimating the effects on individuals where low-income individuals were bussed in as part of the Metco program. Again, the low income students themselves were not included in the regression but were only used to calculate peer characteristics.

The treatment effect equation

Based on the discussion above, we will limit the sample to only individuals whom are themselves western and the non-western second generation immigrants (hereafter immigrants for short) will only be used to define the room characteristics. The following regression models will be estimated:

$$Y_{t2} = \alpha_J + \beta_1 Treatment + \beta_2 Y_{t1} + \beta_n X + \epsilon \tag{2}$$

Where α_J refers to the platoon fixed effects, β_2 is the outcome measured at baseline, while β_n is the vector of coefficients of the covariates. The vector X will be restricted to only include control variables for which treatment and control differ and results with and without these controls will be presented. Standard errors will be clustered on rooms as treatment is at this level. The baseline variables Y_{t1} are included for reasons of power (see below).

In this specification, β_1 is the intention to treat effect. As such it will capture the causal effect of being assigned to a room with an immigrant. We will also estimate the equation:

$$Y_{t2} = \alpha_J + \beta_1 Share_{t2} + \beta_2 Y_{t1} + \beta_n X + \epsilon \tag{3}$$

Where $Share_{t2}$ is the actual share of immigrants in the room. The effects of the share of the immigrants are interesting for discussing magnitudes. To the extent that the initial allocation is found not to be completely followed so that some individuals assigned to treated rooms become untreated and some assigned to control rooms become treated we will also instrument the share by the assigned treatment (T). β_1 in that case measures the treatment effect on the treated.

The treatment heterogeneity across minority IQ will be estimated in the following models:

$$Y_{t2} = \alpha_J + \beta_1 HighIQmin + \beta_2 LowIQmin + \beta_3 Y_{t1} + \beta_n X \tag{4}$$

$$+\beta_n X \times HighIQmin + \beta_n X \times LowIQmin + \epsilon$$

Where HighIQmin is a dummy representing high IQ-score minority roommate, LowIQmin is a dummy representing a low IQ-score minority roommate. The reference category is having no minority roommate (these three categories are mutually exclusive). β_1 and β_2 test whether High IQ and Low IQ groups differ from the control group. We are also interested in the difference between β_1 and β_2 and will rely on F-tests to examine whether they are statistically significant from each other. To the extent that control variables are included (i.e. if they differ by treatment and control), these will also be interacted with high and low IQ (and so will the platoon fixed effects).

Interpretation and robustness tests

The two Liekert scale dependent variables, "immigrants same rights" and "immigrants similar work ethic", will also be transformed to dummy variables whereby people answering "Disagree" and "Strongly disagree" will be coded as 1 and all other responses will be coded as zero. This is done in order to ease interpretation of the magnitude of the effects and to see the share of people affected. The variable "better country" ranges from 1 to 7 and will be treated as linear.

In interpreting the effects as running via the ethnicity of the peers we might be worried that we pick something correlated with ethnicity of peers, in particular low educated parents. To address this concern we examine the effect of

share of low educated parents on our outcomes, and compare the size to that of ethnic minority peers.

In order to check the robustness of the results we will also carry out placebo tests using two variables for which we expect less of an effect. These questions are linked to gender equality. While one may of course imagine circumstances whereby attitudes toward gender equality are affected by sharing room with someone from an ethnic minority we propose that the effect on these variables should be smaller. The tests will be carried out using standardized coefficients in order to be able to compare the effects. The two questions used and their recodings are:

"It is important that men and women share household work equally.

Original:

1= Strongly agree

2= Agree

3= Neither agree nor disagree

4= Disagree

5= Strongly disagree."

Recode into "household equality not important": 4 and 5=1, 1 to 3=0

"Which sex do you think is the best in leading a troop?

Original:

1=Men

2=Equally good

3=Women"

Recode into "sex of leader not important": 2=1, 1 and 3=0.

For the placebo tests with gender equality as outcome we will exclude all rooms containing women. This is so since no women are foreign (at least not in the pilot) and as sharing room with women is likely to affect attitudes toward gender equality.

Power calculation based on pilot data

In order to test the survey, the logistics, and to be able to do power calculations we conducted a longitudinal pilot study based on a similar but much smaller set of soldiers going to the North Brigade of NAF in January 2014. We followed 297 soldiers and collected baseline data in January and follow up data 8 weeks later. These soldiers attend a similar year of military training as the group of soldiers enrolling in August. we can conduct a power analysis with quite few assumptions as we have longitudinal pilot data of a similar duration as our final data and which includes all our outcomes of interest (see Ard and Edland 2012, for a discussion of the importance of these factors).

We use the following formula for power calculations with grouped errors and longitudinal data:

$$MDE = \frac{t_{1-\kappa} + t_{\alpha/2}}{\sqrt{P(1-P)J}} \sqrt{\rho + \frac{1-\rho}{n}} \sqrt{\sigma_u^2 + \sigma_e^2}$$

Where $t_{1-\kappa} + t_{\alpha/2}$ is 2.8 as a result of a power of 80 percent and a statistical significance level of 5 percent and at a large enough sample , P is the share treated, n is the total number of observations, J is the number of rooms, σ_u^2 is the group variance (room) and σ_e^2 is the individual variance, and ρ is the

intracluster correlation (σ_u^2/ (σ_u^2+ σ_e^2)).

There are between 4 and 8 people in the rooms. The room size is likely to be roughly 6 on average. We choose to be conservative in our baseline estimation of the MDE and set N to be 1000 and J=N/6. The share of treated individuals in the pilot data was 26 percent. The main power calculation is based on our main independent variable and our MDE without controls for anything but baseline outcome is .30 which is quite low. This corresponds to standardized beta coefficients and 0.2 is usually considered low, 0.5 medium, and 0.8 high (Duflo, Glennerster, and Kremer 2007).

Adding controls should not affect the expected value of the outcome as treatment is randomly assigned. Nonetheless it may reduce the variance but in some cases also increase the variance. Adding control variables always reduce the degrees of freedom so they should be restrictively added. In particular, controlling for variables that have a large correlation with the outcome can help reduce standard errors. We will therefore only control for baseline outcomes as they are highly predictive. As stated above, we will only add control variables if there are statistically significant differences between the treatment and control groups.

Multiple outcomes

One of the most critical features of the pre-analysis plan is to specify as exactly as possible the outcomes to be tested. This is so since the researcher can then not select the most publishable results ex post at the expense of accuracy (Miguel 2014). This would be of little use, however, without an understanding of the limits to power of testing multiple hypotheses. If we were to test, say 50 hypotheses we would end up with several being statistically significant by chance alone. It is no problem to generate many hypotheses about the peer

effects. For instance, in the educational realm it is often suggested that the linear-in-means model is restrictive as there are likely to be different effects of e.g. high ability peers on low and high ability students. One could imagine similar effects in our case whereby there are different effects of different types of immigrant peers depending on all possible characteristics of soldier i's type and i's roommate's type. Such an open ended investigation should in our view be considered as a generation of hypotheses rather than as a test of them. Taking into account the limits to power by testing multiple hypotheses, on the other hand, forces us to be very restrictive in the number of hypotheses tested. To the extent that we test other outcomes we will be clear about the deviations from the pre-analysis plan and the results will be seen as suggestive or even as new hypotheses to be tested in later research.

In order to deal with the problem of multiple comparisons we therefore restrict the number of outcomes and we also impose pre-specified decision rules (Rosenblum and van der Laan 2011). We restrict ourselves to one major hypothesis and we test two different mechanisms. The test of the mechanisms therefore involves multiple hypotheses. In addition, we test treatment heterogeneity on IQ.

To account for having four different outcomes (one main outcome, two mechanism, and one heterogeneity test) we follow the recommendations of Fink, McConnell, and Vollmer (2014) and use a method developed by Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) to minimize the false non-discovery rate (see also Almeida 2012). The idea is that tests of multiple hypotheses may lead to false discoveries while at the same time some correction procedures are too restrictive. For instance, the classical Bonferroni method (which corrects the critical level for rejection of an hypothesis to be the desired level a divided by the number of hypotheses, $0.05/4$ in our case) is too restrictive

when there are reasons to believe that the hypotheses are correlated. The false discovery rate method developed by Benjamini and Hochberg (1995) implies that the m p-values of the i hypotheses are ordered from low to high and that the test is then $p(i) <= a \times i/m$

To take a concrete example, assume our 4 hypotheses have the following ordered p-values:

$p(1) = 0.002, p(2) = 0.0126, p(3) = 0.040, p(4) = 0.051.$

Without any correction for multiple testing, the first 3 hypotheses would be rejected. With Bonferroni correction the critical p-value would be 0.0125 and hence only one hypothesis would be rejected. With the FDR method the decision rule implies that 2 hypotheses would be rejected as:

$p(1) = 0.002 < 0.05 * 1/4, p(2) = 0.0126 < 0.05 * 2/4, p(3) = 0.040 > 0.05 * 3/4, p(4) = 0.051 > 0.05 * 4/4$

This is an effective way of limiting the risk of false discoveries while still retaining some statistical power. The main advantage is that the method only adjusts the critical values based on other true hypotheses.

Survey attrition and missing observations

We have two sources of attrition. One source is due to people leaving the population because they are discharged from the military. We will use these observations to calculate room characteristics, but they will otherwise be discarded. The second is due to missing data.

The first test we will do is to see whether attrition is related to treatment status. We will thus estimate the following regression:

$$Attrition_i = \alpha_J + \beta_1 Treatment + \beta_n X + \epsilon \quad (5)$$

The hope is that $\beta_1 = 0$. If β_1 is different from 0 at the 5 percent level of significance we will conduct an extreme bounds analysis, where all missing observations are assigned extreme values, to examine how the treatment effect differs between the extreme scenarios.

In addition we also run the following regression:

$$Attrition_i = alpha_J + \beta_1 Y_{t1} + \beta_n X + \epsilon \quad (6)$$

For observations where there are missing values on our key variables we will conduct a list wise deletion and extreme bounds analyzes.

Limited variation

In order to limit noise caused by variables with limited variation, questions for which 95 percent of the observations have the same value within the relevant sample will be omitted from the analysis.

Archive

The pre-analysis plan is archived before the second wave of data is collected. We archive it at the registry for randomized controlled trials in economics held by The American Economic Association: `https://www.socialscienceregistry.org/`

Appendix: Question wordings and recoding of survey items for tests of balance

How many brothers/half-brothers/stepbrothers do you have?

Original: Continuous

Recode: We recode into a binary variable of no brother (=0) versus any number of sisters (1/x=1).

How many sisters/half-sisters/stepsisters do you have?

Original: Continuous

Recode: We recode into a binary variable of no sister (=0) versus any number of sisters (1/x=1).

Are your parents divorced/separated?

Original: 1=Yes, 2=Don't know, 3=No

Recode: 2 to missing and 3=1.

Are your parents in work?

Original: 1= Yes, both, 2=My mother is in work, my father is not, 3=My father is in work, my mother is not, 4=No, neither of them is in work

Recode: We recode into two variables:Mother in work (1/2=1, 3/4 = 0) and Father in work (1 and 3=1, 2 and 4=0)

Do your parents have higher education (university/college)?

Original: 1= Yes, both have higher education, 2=My mother has higher education, my father has not, 3= My father has higher education, my mother has not, 4=No, neither of them have higher education

Recode: We recode into two variables: Mother with high education (1/2=1, 3/4 = 0) and Father with high education (1 and 3=1, 2 and 4=0)

Do you plan to take higher education?

Original: 1=Yes, 2=Don't know, 3=No

Recode: 2/3=0.

The IQ measure is a composite score from three speeded IQ tests: arithmetic, word similarities, and figures (see Sundet, Jon Martin, Dag G. Barlaug, and Tore M. Torjussen, "The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Test Scores of Norwegian Conscripts During Half a Century", Intelligence , XXXII (2004), 349-362). The composite IQ test score is an unweighted mean of the three sub-tests.

Original: The IQ score is reported in stanine (Standard Nine) units, a method of standardizing raw scores into a nine point standard scale with a normal distribution, a mean of 5, and a standard deviation of 2.

Recode: We rely on the original coding.

References

Allport, G. W. (1954): *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.

Almeida, R. e. a. (2012): "The Impact of Vocational Training for the Unemployed in Turkey: Pre-Analysis Plan," Pre-analysis plan posted at poverty-actionlab.org.

Angrist, J. D. (fortcoming): "The perils of peer effects," *Labour Economics*.

Angrist, J. D., and K. Lang (2004): "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program," *American Economic Review*, pp. 1613–1634.

Ard, M. C., and S. D. Edland (2012): "Power Calculations for Two-Wave, Change from Baseline to Follow-Up Study Designs," *International Journal of Statistics in Medical Research*, 1(1), 45–50.

Benjamini, Y., and Y. Hochberg (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

Benjamini, Y., and D. Yekutieli (2001): "The Control of the False Discovery Rate in Multiple Testing under Dependency," *Annals of Statistics*, pp. 1165–1188.

Bobo, L. D. (1999): "Prejudice As Group Position: Microfoundations of a Sociological Approach to Racism and Race Relations," *Journal of Social Issues*, 55(3), 445–472.

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.

Carrell, S. E., M. Hoekstra, and J. E. West (2013): "Racial Preference Formation," Paper presented at Department of Economics, University of Oslo.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using Randomization in Development Economics Research: A Toolkit," CEPR Discussion Paper No. 6059.

Fink, G., M. McConnell, and S. Vollmer (2014): "Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures," *Journal of Development Effectiveness*, 6(1), 44–57.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75(1), 83–119.

Manski, C. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60(3), 531–542.

Miguel, E. e. a. (2014): "Promoting Transparency in Social Science Research," *Science*, 343(6166), 30–31.

Pettigrew, T. F. (1998): "Intergroup Contact Theory," *Annual Review of Psychology*, 49(1), 65–85.

Rosenblum, M., and M. J. van der Laan (2011): "Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment," *Biometrika*, 98(4), 845–860.

Sacerdote, B. (2011): "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?," in *Handbook of the Economics of Education*, ed. by E. A. Hanushek, S. J. Machin, and L. Woessmann. Elsevier.

Semyonov, M., R. Raijman, and A. Gorodzeisky (2006): "The Rise of Anti-Foreigner Sentiment in European Societies, 1988-2000," *American Sociological Review*, 71(3), 426–449.

Senik, C., H. Stichnoth, and K. Van der Straeten (2009): "Immigration and Natives' Attitudes towards the Welfare State: Evidence from the European Social Survey," *Social Indicators Research*, 91(3), 345–370.