Supporting Information

Christ et al. 10.1073/pnas.1320901111

SI Text

Computation of Contextual Effect. We followed Raudenbush and Bryk (1) when estimating the contextual effect of intergroup contact by specifying the following relation:

$$\text{Level 1}: \text{prejudice}_{ij} = \beta_{0j} + \beta_1 \text{direct intergroup contact}_{CWC} + r_{ij}, \quad \textbf{[S1]}$$

where prejudice_{ij} is the outcome for respondent i in context j modeled as a function of the intercept β_{0j} of context j, the slope β_{1j} for direct intergroup contact in context j, and an error r_{ij}. The predictor, direct intergroup contact, is centered at the group mean (centering within cluster), removing all between-context variation in direct intergroup contact and yielding a pooled-within (i.e., level 1) estimate (β_W) for the relation of direct intergroup contact with prejudice. The level 1 coefficients β_{0j} and β_{1j} are then modeled at level 2. Level 2 coefficients are typically notated as γ.

$$\text{Level 2}: \begin{matrix} \beta_{0j} = \gamma_{00} + \gamma_{01} \text{direct intergroup contact}_{\textit{Group Mean}} + u_{0j} \\ \beta_{1j} = \gamma_{10,} \end{matrix} \quad \textbf{[S2]}$$

where γ_{00} and γ_{10} are the level 1 intercepts, and γ_{01} is the slope relating the group mean of direct intergroup contact to the intercepts from the level 1 equation. The slope γ_{01} captures the between-context relation between direct intergroup contact (mean of intergroup contact within contexts) and prejudice. It is important to note that we model a random-intercept model (1) here; therefore, only the level 1 intercepts have a level 2 residual u_{0j}. The effect of direct intergroup contact on prejudice at level 1 is fixed, allowing for no variation in this effect between contexts. A contextual effect is present if γ_{01} is significantly larger or smaller than γ_{10}, meaning that the relationship at the aggregated level is stronger or weaker than the relationship at the individual level.

As Eq. **S2** shows, at level 2 the group mean is used as an estimate for the level 2 effect of intergroup contact. However, sampling error in the group mean can cause biased and less efficient estimates of the true population effect (here γ_{01}) with a consequential biased estimate for the contextual effect, therefore a multilevel latent covariate (MLC) approach that corrects for the unreliability in the level 2 construct has been recommended, resulting in unbiased estimates of the level 2 effect (2).* We therefore applied the MLC approach to estimate the contextual effect of intergroup contact in all studies. We implemented MLC using the maximum-likelihood procedure in Mplus 6.1 (3). Due to "missingness," we used full-information maximum-likelihood estimates with robust SEs. Missing values in no case exceeded 2.2%.

To assess the effect size of the contextual effect of contact we used an effect size measure (ES2) proposed by Marsh et al. (4). The effect size is calculated with the following formula:

$$ES2 = \frac{2 * B * SD_{\textit{intergroup contact}}}{\sigma^2_{\textit{intergroup contact}}}, \quad \textbf{[S3]}$$

where B is the unstandardized regression coefficient of the contextual effect in the multilevel model, $SD_{\textit{intergroup contact}}$ is the SD of intergroup contact at the social context level, and σ^2 is the total variance of intergroup contact at the individual level. The resulting effect size describes the difference in the dependent variable between two level 2 groups that differ by two SDs on the predictor variable. This effect size is comparable with Cohen's d (5). An effect can be defined as small with d (or ES2) = 0.2, medium with d (or ES2) = 0.5, and large with d (or ES2) = 0.8.

A prerequisite for estimating the contextual effect of intergroup contact is sufficient between-level variance in all relevant measures. In studies 1a to 1e, the intraclass correlations (ICC) for all indicators and the composite measures of direct intergroup contact, ingroup norms, and prejudice were small to large in size (M = 0.17, SD = 0.08, minimum = 0.07, maximum = 0.30), showing that there was substantial between-level variance. Likewise, ICCs in studies 2a and 2b were small to medium for intergroup contact, social norms, and prejudice at time 1 and 2 (M = 0.08, SD = 0.06, minimum = 0.02, maximum = 0.18), indicating sufficient between-level variability in these measures.

SI Materials and Methods: Sampling Information and Measures in Study 1 and Study 2

Study 1a. ***Sampling information.*** Data came from the first round of the European Social Survey (ESS) (6). This representative cross-national survey was conducted from September 2002 to October 2003 and covered 22 countries (21 European countries and Israel). In total, 42,359 face-to-face interviews were achieved. We dropped all respondents without national citizenship, with place of birth outside the country of data collection, or who classified themselves as belonging to a minority ethnic group in their country, resulting in a reduced sample size for all analyses of n = 36,334 respondents (for four respondents, regional codes were not available).

As the regional level measure, we used the country-specific indicator variables available in the ESS to group respondents into regional units corresponding to the nomenclature of statistical units classification scheme (NUTS) (7). The NUTS classifies European regions according to socioeconomic, cultural, and historical characteristics (7). Conceptually, the NUTS comprises three different regional subdivisions that divide each country into large-scale (NUTS 1), medium-scale (NUTS 2) and small-scale (NUTS 3) regions. However, the NUTS levels provided by the ESS differed somewhat between countries. Whereas for the majority of countries, respondents are grouped into NUTS 2 regions, for four countries, respondents were grouped according to NUTS 1. To base our analyses on regions of comparable size, we reclassified the NUTS 2 codes into NUTS 1 codes and recalculated all analyses using only NUTS 1 regions (n = 91 NUTS 1 regions). The pattern of results was almost identical compared with the analyses using the mix of NUTS 1 and NUTS 2 regions. We therefore report only the results using the original regional codes provided with the ESS data (n = 248 NUTS 1/NUTS 2 regions). The mean number of

*As Lüdtke et al. (2) have recently shown, the estimation of the contextual effect cannot only be biased due to sampling error, but also due to measurement error. The authors (2) distinguished between different approaches to correct for sources of error in estimating contextual effects, proposing a 2 × 2 taxonomy of multilevel contextual models correcting for no error source, for only one source of error, or for all error sources. Lüdtke et al. (2) showed in a simulation study that depending on specific data circumstances, the uncorrected and the partial correction approaches can result in biased estimates of the contextual effect. However, when the data provides only limited information on the level 2 constructs (i.e., small number of groups, low intraclass correlations), partial correction approaches outperform the doubly latent approach. The authors therefore suggest that researchers juxtapose the different approaches (where possible) and use the estimates from the different approaches as bounds for the true parameter. We were able to implement these different approaches in study 1b because multiple items for intergroup contact and prejudice were available and therefore latent variables on both levels could be specified. In all approaches, a significant estimate for the contextual effect emerged, ranging from −0.142 to −0.331.

respondents per NUTS 1/ NUTS 2 region was $M = 146.51$. In all analyses, we controlled for between-country differences, using country as a level 3 unit in the analyses.

Measures. Direct intergroup contact was measured with a single indicator: "Do you have any friends who have come to live in [country] from another country?" (1 = no, none at all, 2 = yes, a few, 3 = yes, several).[†]

Prejudice toward foreigners was assessed with four items (Cronbach's $\alpha = 0.72$): "Average wages and salaries are generally brought down by people coming to live and work here"; "People who come to live and work here generally harm the economic prospects of the poor more than the rich"; "If people who have come to live and work here are unemployed for a long period, they should be made to leave"; "If people who have come to live here commit any crime, they should be made to leave" (1 = disagree strongly to 5 = agree strongly).

Social norms were measured with two items ($r = 0.61$, $P < 0.001$): "Would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?" (1 = cultural life undermined to 10 = cultural life enriched); "Is [country] made a worse or a better place to live by people coming to live here from other countries?" (1 = worse place to live to 10 = better place to live).[‡]

Control variables were age, sex, and education on the individual level. There were no controls available on the social context level.

Study 1b. ***Sampling information.*** Data were obtained from a probability survey of the German adult population (16 y of age and older) conducted in May/June 2002, excluding those with a migration background. Respondents were randomly selected from a two-stage probability sample, resulting in a representative sample of the German adult population. A total of $n = 2,722$ respondents were interviewed by a survey company using computer-assisted telephone interviews (CATI). These data contain district codes that indicate the place of residence of each respondent interviewed. A district is a state organizational unit usually composed of a big city or a number of smaller cities, towns, or rural areas. Sizes of districts vary between 35,700 and 3,382,200 inhabitants. Altogether, Germany is divided into 440 districts, of which 418 districts were sampled. The mean number of observations per district was $M = 6.50$ respondents.

Measures. Direct intergroup contact was measured with three items ("How many of your friends are foreigners living in Germany?"; "How often have you had an interesting conversation with a foreigner?"; "How often have you been helped by a foreigner?"; Cronbach's $\alpha = 0.75$). All items were answered using four-point rating scales ranging from 1 = none/never, 2 = few/sometimes, 3 = fairly many/often, to 4 = very many/very often.

Prejudice toward foreigners was assessed with three items ("If jobs become scarce, foreigners should be sent back to their home countries?"; "There are too many foreigners in Germany"; "Foreigners are a burden for our social security system"; Cronbach's $\alpha = 0.82$). Each item was answered on a four-point rating-scale (1 = fully disagree to 4 = fully agree).

Control variables were age, sex, and education on the individual level and an index of regional deprivation (gross domestic product, unemployment rate, rate of people receiving social welfare) on the social context level.

Study 1c. ***Sampling information.*** The data were collected from mid-May to mid-July 2005 as part of the U.S. Citizenship, Involvement, Democracy Survey conducted by the Center for Democracy and Civil Society at Georgetown University (8). This national survey is comprised of 1,001 face-to-face interviews of adults throughout the United States. We restricted our analyses to White respondents ($n = 725$) because sample sizes for Blacks, Hispanics, and Asians were too small for analysis. As the regional level measure, we used information at the level of census tracts ($n = 174$). Census tracts generally have a population size between 1,200 and 8,000 people. The mean number of observations per district was $M = 4.08$ respondents.

Measures. Direct intergroup contact was measured with one item: "Now I want to ask you some questions about people you are really close to, that is, people you feel at ease with and can talk to about whatever's on your mind, or call on for help. Though this may include family members, in the questions that follow I will refer to these people as your close friends. How many of your close friends are of a different race from yours? By race I mean such groups as Asians, Blacks, Hispanics, and Whites" (1 = none to 9 = all).

Prejudice toward Blacks, Hispanics, and Asians was assessed with the difference between an indicator for liking Whites (the ingroup) and a composite based on three indicators for liking Blacks, Hispanics, and Asians, respectively ("How do you feel about the following groups, in general?"; 1 = dislike a great deal to 11 = like a great deal).

Control variables were age, sex, education, and income on the individual level, and educational level and income on the social context level. Moreover, on the individual level, we also controlled for the quantity of close friends (whether ingroup or outgroup, to control for being more outgoing or sociable, which may be related to contact and/or prejudice).

Study 1d. ***Sampling information.*** The data came from a survey in England (2009–2010) with $n = 868$ White British respondents (level 1) from $n = 217$ neighborhoods (level 2; mean number of observations per neighborhood was $M = 4.00$). Neighborhoods constituted so-called middle-layer super output areas in England, which are small geographical units with an average size of 7,200 residents. Data collection was subcontracted to a professional survey organization that used computer-assisted personal interviewing by trained social survey interviewers, involving face-to-face interviews in respondents' own homes. Random location quota

[†]An anonymous reviewer of an earlier version of this paper queried whether respondents all understand the distinction between "yes, a few" (coded 2) and "yes, several" (coded 3) in the same manner. As an authority on survey methodology has noted, though there is random and systematic variation in the meanings of some verbal labels to respondents, "many labels do appear to have sufficiently universal meanings to be used in attitude measurement in this manner" (9, p. 151). If there were such random and/or systematic variation in understanding this distinction, this would likely cause an unsystematic relation between contact and prejudice scores at these two levels of the contact measure. A trend test using regression analysis showed, however, a strong and significant linear trend between the contact and the prejudice measure ($\beta = -0.66$, $t = 16.75$, $P < 0.001$) over the whole scale of contact and prejudice, and a significant but smaller quadratic trend ($\beta = 0.09$, $t = 8.61$, $P < 0.001$). Although the linear trend is less steep on higher ratings in the friendship measure (i.e., between scale point 2 and scale point 3), there is still a strong relationship between contact and prejudice on the two higher scale points, as corroborated by a significant difference in prejudice scores between these two scale points ($M = 3.05$, SD = 0.85 for respondents with score 2 on the contact measure; $M = 2.84$, SD = 0.85 for respondents with score 3 on the contact measure; $t = 14.02$, df = 16, 022, $P < 0.001$). These results do not support the assumption that respondents vary systematically in their understanding of the distinction between scale points 2 and 3. Finally, given that the size of the contextual effect in this study is comparable to those found in the other studies, we are confident in the robustness and validity of our findings in study 1a.

[‡]Exploratory factor analysis of the data from study 1a confirmed that our measures for norms and prejudice are related, but separable, constructs. In study 1a, we were able to use multilevel exploratory factor analysis (ML-EFA) as implemented in Mplus (3) to examine the factorial structure for both the norms and prejudice measure on the individual and social context levels simultaneously. In Mplus, ML-EFA is based on maximum-likelihood estimates, allowing us to compare different factorial solutions by means of fit statistics known from a structural equation modeling framework. The two within- and two between-factor solutions (involving separate factors of norms and prejudice on both levels, respectively) showed the best fit to the data compared with all other possible combinations ($\chi^2 = 2,771.42$, df = 8, $P < 0.001$; Comparative Fit Index = 0.894; Root Mean Square Error of Approximation = 0.098; Standardized Root Mean Square Residual $(SRMR)_{within} = 0.041$; $SRMR_{between} = 0.043$). These findings are consistent both with prior theoretical work on the construct of diversity beliefs (10) and prior empirical work distinguishing diversity beliefs and prejudice (11). We could not test the factorial structure of norms and prejudice in studies 1d, 1e, and 2b, due to single-item measures of prejudice in studies 1d and 1e, and small sample size on the social context level in study 2b. Full results of the EFA are available from the first author.

sampling ensured that the profile of respondents interviewed in each neighborhood reflected the profile of the neighborhood with regard to key demographics (age, sex, working status, and ethnicity). Data collection took place from October 2009 to February 2010.

Measures. Direct intergroup contact was measured with a single item: "What proportion of your close friends are from ethnic minorities?" (1 = none or very few, 2 = a few, 3 = about half, 4 = a lot, 5 = almost all or all).

Prejudice toward ethnic minorities was assessed with an ingroup bias measure computed by subtracting the outgroup rating from the ingroup rating: "How warm or cold do you feel about White British people?" and "How warm or cold do you feel about ethnic minorities?" (feeling thermometer ranged from 0 = cold to 100 = warm).

Social norms were measured with two items ($r = 0.40$, $P < 0.001$): "The mix of different ethnic groups in my neighborhood enriches local life" and "The mix of different ethnic groups in my neighborhood creates social disorder" (reverse coded; 1 = definitely disagree, 2 = tend to disagree, 3 = neither agree nor disagree, 4 = tend to agree, 5 = definitely agree).

Control variables were age, sex, and education on the individual level and an index of multiple deprivation score on the social context level.

Study 1e. ***Sampling information.*** Data came from a city survey in Cape Town, South Africa, conducted in September and November 2011 with $n = 897$ respondents (level 1) from $n = 97$ neighborhoods (level 2; mean number of observations per neighborhood was $M = 9.25$ respondents). The survey included information from Black ($n = 438$) and Colored respondents ($n = 459$). Data collection was subcontracted to a professional survey organization that used trained interviewers to undertake face-to-face interviews in respondents' own homes.

Measures. Direct intergroup contact with White South Africans was measured with one item: "What proportion of your close friends are White people?" [1 = none to 5 = (almost) all].

Prejudice toward White South Africans was measured with one item asking how much the respondents felt that White South Africans can be trusted (1 = cannot trust them at all to 5 = all can be trusted). This item was reverse coded for the analyses, so that high scores denoted prejudice.

Social norms were measured with four items (Cronbach's $\alpha = 0.86$): "How important do you think it is to have people from different racial backgrounds in your workplace, or place of study?"; "How important do you think it is to have people from different racial backgrounds among your friends?"; "How important do you think it is to have people from different racial backgrounds in your neighborhood?"; and "How important do you think it is to have people from different racial backgrounds in South Africa?" (1 = not at all important to 5 = very important).

Control variables were age, sex, and education on the individual level. There were no controls available on the social context level.

Study 2a. ***Sampling information.*** Data were collected as waves 1 and 4 of a multiwave panel study representative of the German adult population (16 y and above) with no migration background. A total of 1,024 respondents were interviewed via a survey company using CATI in both waves in 2002 (time 1) and 2006 (time 2). These respondents form a part of the sample used in study 1b. As such, the level 2 units were the same districts as in study 1b. We sampled a random subsample of respondents from study 1b who agreed to be recontacted for the panel survey ($n = 2,363$; response rate = 43%). Therefore, the number of districts was reduced to $n = 345$; the average number of respondents within districts was $n = 2.97$. Missing data were negligible (<1%). Systematic panel mortality was negligible.

Measures. The panel included the same indicators for direct intergroup contact (Cronbach's $\text{alpha}_{\text{time 1}} = 0.73$; Cronbach's $\text{alpha}_{\text{time 2}} = 0.75$) and prejudice (Cronbach's $\text{alpha}_{\text{time 1}} = 0.81$; Cronbach's $\text{alpha}_{\text{time 2}} = 0.82$) as used in study 1b.

Control variables were age, sex, and education on the individual level, and an index of regional deprivation (gross domestic product, unemployment rate, rate of people receiving social welfare) on the social context level.

Study 2b. ***Sampling information.*** Data were collected in Germany by a professional survey organization, using trained social survey interviewers and CATI (www.mmg.mpg.de/en/publications/working-papers/2012/wp-12-21/). Respondents were purposefully sampled from neighborhoods varying in their proportional share of foreign residents, resulting in a two-level hierarchical data structure with respondents nested in neighborhoods. Fifty neighborhoods (so-called "Wohnviertel"; minimum $n = 2,800$ residents, average $n = 7,500$ residents) from 16 different cities in Germany were randomly sampled. Data collection took place from May to July 2010 for wave 1 ($n = 1,976$) and May to July 2011 for wave 2 ($n = 1,056$; response rate: 53.44%). The final sample size for this study was $n = 1,056$ respondents (level 1) from $n = 50$ neighborhoods (level 2) who took part at both time points. The average number of respondents within districts was $n = 21.12$. Missing data were negligible (<1%), as was systematic panel mortality.

Measures. Direct intergroup contact was measured with an index based on the product of two items assessing the frequency and quality of contact with foreigners within the neighborhood of the respondents: "In your neighborhood, how often do you talk to people who are themselves not native Germans or whose parents are not from Germany?" and "How do you perceive the conversations with immigrants in your neighborhood?". Both items were answered on five-point rating scales (1 = never to 5 = daily; 1 = very unpleasant to 5 = very pleasant).

Prejudice toward foreigners was assessed with four items (Cronbach's $\text{alpha}_{\text{time 1}} = 0.72$; Cronbach's $\text{alpha}_{\text{time2}} = 0.73$): "Foreigners in Germany threaten the German way of life"; "The values of foreigners living in Germany are incompatible with the values of Germans"; "Foreigners living in Germany make it more difficult for Germans to find jobs"; and "Foreigners living in Germany are a burden on the social welfare system". The items were answered on a five-point rating scale (1 = fully disagree to 5 = fully agree).

Social norms were measured with two items ($r_{\text{time 1}} = 0.48$, $P < 0.001$; $r_{\text{time2}} = 0.48$, $P < 0.001$): "It is enriching for a city when people come from different backgrounds and cultures" and "Muslims living in Germany should have the right to build mosques, including in your own neighborhood". The items were answered on a five-point rating scale (1 = fully disagree to 5 = fully agree).

Control variables were age, sex, and education on the individual level, and unemployment rate on the social context level.

1. Raudenbush SW, Bryk AS (2002) *Hierarchical Linear Models* (Sage, Newbury Park, CA), 2nd Ed.
2. Lüdtke O, Marsh HW, Robitzsch A, Trautwein U (2011) A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychol Methods* 16(4):444–467.
3. Muthén LK, Muthén BO (1998–2010) *Mplus User's Guide* (Muthén & Muthén, Los Angeles), 6th Ed.
4. Marsh HM, et al. (2009) Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behav Res* 44(6):764–802.
5. Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (Erlbaum, Mahwah, NJ).
6. Norwegian Social Science Data Services (2002) *European Social Survey Round 1 Data*. Data file edition 6.2. Available at www.europeansocialsurvey.org/data/download.html?r=1.
7. Eurostat (2003) *European Regional Statistics Reference Guide* (Office for Official Publication of the European Communities, Luxemburg).
8. Howard MH, Gibson JL, Stolle D (2005) *The U.S. Citizenship, Involvement, Democracy Survey* (Center for Democracy and Civil Society, Georgetown University, Washington, DC).
9. Krosnick JA, Fabrigar LR (1997) *Survey Measurement and Process Quality*, eds Lyberg L, et al. (Wiley-Interscience, New York), pp 141–164.

10. Tropp LR, Bianchi RA (2006) Valuing diversity and interest in intergroup contact. *J Soc Issues* 62(3):533–551.

11. Adesokan AA, Ullrich J, van Dick R, Tropp LR (2011) Diversity beliefs as moderator of the contact-prejudice relationship. *Soc Psychol* 42(4):271–278.

Table S1. Unstandardized estimates (SE in brackets) for model 2 in study 2b

	Study 2b	
Model 2	β (SE)	*P*
Level 1		
$contact_{time1} \rightarrow contact_{time2}$	0.636 (0.039)	<0.001
$prejudice_{time1} \rightarrow prejudice_{time2}$	0.569 (0.031)	<0.001
$contact_{time1} \rightarrow prejudice_{time2}$	−0.007 (0.004)	0.089
$prejudice_{time1} \rightarrow contact_{time2}$	−0.239 (0.244)	0.327
$norms_{time1} \rightarrow norms_{time2}$	0.056 (0.035)	0.108
$contact_{time1} \rightarrow norms_{time\ 2}$	0.051 (0.006)	<0.001
$norms_{time1} \rightarrow prejudice_{time2}$	0.004 (007)	0.615
Level 2		
$contact_{time1} \rightarrow contact_{time2}$	0.989 (0.160)	<0.001
$prejudice_{time1} \rightarrow prejudice_{time2}$	0.467 (0.221)	0.034
$contact_{time1} \rightarrow prejudice_{time2}$	−0.043 (0.015)	0.006
$prejudice_{time1} \rightarrow contact_{time2}$	4.028 (2.492)	0.106
$norms_{time1} \rightarrow norms_{time2}$	0.391 (0.146)	0.007
$contact_{time1} \rightarrow norms_{time\ 2}$	0.084 (0.030)	0.005
$norms_{time1} \rightarrow prejudice_{time2}$	−0.146 (0.057)	0.011