# Effective reduction of prejudice and discrimination: Methodological considerations and three…

**Article** *in* Revue Internationale de Psychologie Sociale · January 2010

CITATIONS
6

READS
30

**3 authors**, including:

Abdelatif Er-rafiy
Université de Poitiers
**13** PUBLICATIONS **105** CITATIONS
SEE PROFILE

Markus Brauer
University of Wisconsin–Ma…
**95** PUBLICATIONS **1,777** CITATIONS
SEE PROFILE

# EFFECTIVE REDUCTION OF PREJUDICE AND DISCRIMINATION: METHODOLOGICAL CONSIDERATIONS AND THREE FIELD EXPERIMENTS

**Abdelatif Er-rafiy, Markus Brauer, Serban C. Musca**

**Presses univ. de Grenoble** | « Revue internationale de psychologie sociale »

2010/2 Tome 23 | pages 57 à 95

ISSN 0992-986X
ISBN 9782706116438

RIPS / IRSP, 23 (2/3), 57-95 © 2010, Presses universitaires de Grenoble

# Effective reduction of prejudice and discrimination: Methodological considerations and three field experiments

*Réduire efficacement les préjugés et la discrimination : considérations méthodologiques et trois exemples d'études sur le terrain*

*Abdelatif Er-rafiy**
*Markus Brauer***
*Serban C. Musca***

**Abstract**

The social psychology literature on prejudice reduction reports only very few field studies from which researchers can draw reliable causal conclusions. The present paper has two goals. First, we encourage social psychologists and public policy decision-makers to carry out more randomized field experiments on the reduction of prejudice and discrimination. To this end we begin with a discussion of methodological considerations and guidelines for designing and conducting such experiments. Our second goal is to contribute to the growing literature on effective interventions aimed at reducing prejudice. In the second part of the article, we report three field experiments that evaluated the effectiveness of different interventions. Experiment 1 examined the

**Résumé**

La littérature en psychologie sociale scientifique sur la lutte contre les préjugés est caractérisée par peu d'études sur le terrain qui permettent de tirer des conclusions causales fiables. L'objectif de cet article est d'inciter les psychologues sociaux et les acteurs associatifs et politiques à réaliser davantage d'études de terrain randomisées. Dans la première moitié de l'article, nous proposons des considérations méthodologiques et des conseils utiles sur la mise en place de ce genre d'études. Dans la deuxième moitié, nous présentons trois exemples d'études ayant permis d'évaluer l'efficacité de deux interventions différentes, dont l'une est un «atelier de diversité» (Étude 1) et l'autre une affiche mettant en avant les différences au sein d'un groupe

**Key-words**
Perceived variability, diversity training, randomized field experiments, prejudice reduction

**Mots-clés**
Variabilité perçue, atelier de diversité, études de terrain randomisées, réduction des préjugés

We wish to thank Elizabeth Paluck and Paula Niedenthal for extremely helpful feedback on earlier drafts of this article.
*Clermont Université, LAPSCO/CNRS, Av. Carnot 34, F-63037 Clermont-Ferrand.
E-Mail: abdelatif.er.rafiy@gmail.com
**Centre National de la Recherche Scientifique & Clermont Université.
E-mail : markus.brauer@univ-bpclermont.fr, serbancmusca@gmail.com

beneficial effect of "diversity training" whereas in Experiments 2 and 3, we tested the effectiveness of a poster highlighting differences among members of a minority group. In all three experiments, half of the participants were subjected to the intervention, and the other half were not. Both types of interventions were shown to be effective. The results are discussed in light of other methods of prejudice reduction.

minoritaire (Études 2 et 3). Les deux interventions se sont révélées efficaces.

Attempts to reduce prejudice and discrimination are characterized by a paradox. On the one hand, governments and companies spend billions of euros on interventions aimed at reducing prejudice and discrimination, without evaluating the effectiveness of these interventions. For example, in 2008 the French government agency HALDE (acronym for "High Authority against Discrimination and for Equality") sponsored an intervention campaign entitled "You are against discrimination – write it out loud and clear". The campaign encouraged students to write about their opposition to discrimination on the internet. The price of this campaign was 600.000 euros, and the potential beneficial outcomes were never evaluated empirically. On the other hand, social psychologists offer very little practical advice on how to design and carry out effective interventions. In their recent paper entitled "Prejudice reduction: What works?", Paluck and Green (2009) review the empirical work on prejudice reduction since 1958. This analysis revealed that only 13 publications in which researchers examined the effectiveness of a given intervention through a randomized field experiment with a non student population. According to the authors, "the literature provides little empirical guidance to policymakers seeking to intervene with populations living in conflict or postconflict environments" (p. 352).

In our opinion, a false but common belief underlies the paradox described in the previous paragraph. The belief is that it is impos-

sible to conduct reliable field studies on prejudice reduction. In our professional contacts with decision-makers it becomes clear that many of them believe that the effectiveness of an intervention cannot be evaluated because it is impossible to reliably measure abstract concepts such as prejudice and discrimination. And given the predominance of laboratory experiments in prejudice research, many social psychologists seem to think that field studies do not satisfy their high standards of scientific rigor. In the present paper, we argue against these ideas. We want to show that it is possible – and necessary – to conduct randomized field experiments under conditions that are sufficiently controlled to warrant valid conclusions.

Prejudice-related field studies have generally pursued one of the following two goals: to develop interventions that can subsequently be used by decision-makers who want to fight prejudice, or to evaluate the effectiveness of an intervention that has been carried out by a government agency. Those who develop interventions often have to examine the effectiveness of preliminary versions of these interventions in the field in order to be able to improve them. Thus, the two goals are closely related in practice. We argue here that studies on prejudice interventions ought to demonstrate that the observed reduction in prejudice is caused by the intervention (i.e., internal validity) and that the beneficial effects of the intervention can be generalized to other situations (i.e., external validity). In the first part of this article we discuss some methodological issues and offer advice that should prove useful in applied research on prejudice reduction. In the second part we report the results of three field experiments in which we evaluated the effectiveness of different prejudice reduction interventions.

## Contribution of scientific social psychology

For nearly a century the topic of prejudice has generated one of the richest and most prolific literatures in social psychology. An abbreviated list of the theoretical approaches aimed at explaining the origins of prejudice includes authoritarian personality theory (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950), scapegoat theory (Hovland & Sears, 1940), social dominance theory

(Sidanius & Pratto, 1999), social identity theory (Tajfel, 1978), and realistic conflict theory (Sherif & Sherif, 1953). Another line of research focused on the consequences of prejudice. This research revealed, for example, that being the target of prejudice or discrimination induces a high level of depression, anxiety and stress (Landrine & Klonoff, 1996), aggressiveness (Dion, 1986), a low level of self-esteem and life satisfaction (Broman, 1997), and low performance (e.g., stereotype threat: Steele & Aronson, 1995). Other studies have shown that members of minority groups are treated in an unfair manner in employment, housing and judicial settings (Blank, Dabady, & Citro, 2004). The repertoire of prejudice measures includes self-evaluation scales (e.g., Osgood, Suci, & Tannenbaum, 1957), the observation of nonverbal behavior (Crosby, Bromley, & Saxe, 1980), measures of reaction times (Brauer, Wasel, & Niedenthal, 2000) and measures of cerebral activity through brain imaging techniques (Phelps & Thomas, 2003).

Most social scientists studying prejudice are in part motivated by the hope that their research will directly or indirectly contribute to the reduction of prejudice. Yet, only a small part of the research in social psychology has investigated the ways to reduce prejudice (Oskamp, 2000). Based on Allport's (1954) *contact hypothesis*, it has been shown that face to face contact between members of two groups in conflict promotes positive relationships and therefore leads to prejudice reduction. In the 1980s, some researchers examined if it is possible to reduce prejudice by modifying people's representations. Brewer and Miller (1984), for example, proposed that *decategorization* decreases the prominence of the categories "us" and "them" and makes people interact with each other as individuals rather than as members of different groups. Gaertner and Dovidio (2000) suggested that the *recategorization* of the in-group and the out-group into a unique, inclusive category reduces prejudice. Other methods were based on the empathy (Batson, 1991), education (Stephan & Stephan, 2001), and accountability (Dobbs & Crano, 2001). Recently, new promising approaches to prejudice and discrimination reduction have been proposed, such as diversity training (Paluck, 2006) and the modification of perceived variability (Brauer & Er-rafiy, 2009; Er-rafiy & Brauer, 2010). These latter and more recent approaches are the focus of this paper.

Most of the research on the reduction of prejudice and discrimination was carried out in laboratory settings (Paluck & Green, 2009). Laboratory studies have many advantages because they are characterized by (a) the test of a wide range of prejudice reduction theories with a high degree of creativity and precision (b) an environment and experimental methods that lead to internally valid conclusions about the causal impact of the intervention, and (c) an understanding of the basic mechanisms that intervene in intergroup relations. Laboratory studies yield promising results, but they suffer from shortcomings that limit the impact of their findings. First, laboratory studies often involve short-term interventions. For instance, in studies conducted with the "minimal group paradigm" prejudice can be created, modified, and examined in less than one hour. The duration of the beneficial effect of an intervention is rarely examined (Hill & Augustinos, 2001). Second, most laboratory studies involve artificial groups that are created during the experiment and that have no real life counterpart. Often, participants have never interacted with a member of the artificial out-group and there is no real conflict of interest between in-group and out-group. It is thus difficult to know if an intervention that has been tested with artificial groups is equally effective with real world groups (Bourhis & Leyens, 1999). Third, most researchers use psychology students as participants, which raises the question of the generalization of the results to other populations (Henry, 2008). Finally, certain interventions that work in a laboratory setting are impossible to implement in the field. Indeed, while it may be possible to convince participants to give up their social identity when it comes to an artificial group created in the laboratory (e.g., decategorization approach: Brewer & Miller, 1984), such an objective is virtually impossible to attain outside the laboratory (e.g., when it comes to ethnic group identity).

The little research in which social scientists conducted both laboratory and field studies revealed that it is not unusual for an intervention to be effective in the laboratory but to be ineffective in the field (e.g., cross categorization approach, Rich, Kedem, & Shlesinger, 1995). In some cases, it is even the case that an intervention had positive effects in the laboratory but negative effects in the field. To cite just an example, consider the recent study of

Paluck (in press). Based on the observation that discussion is an effective method to reduce conflicts in the laboratory (Allport, 1954; Mutz & Martin, 2001), Paluck created a field intervention through a radio broadcast that encouraged people to discuss discrimination-related topics with each other. She then tested the effectiveness of this field intervention. The study was carried out in the Democratic Republic of Congo and the radio broadcast was aired in three "test" regions and was not aired in three "control" regions. After one year of broadcasting, attitudinal and behavioral measures were collected from 850 individuals in the six regions. Results showed that at the end of the study individuals in the "test" regions were more prejudiced and less willing to help out-group members than those from the "control" regions. This study illustrates how important it is to carry out field studies before one can claim that a particular prejudice reduction method is efficient.

Field studies are very informative for the creation and evaluation of prejudice reduction interventions. Paluck and Green (2009) reviewed all publications in which the authors examined the reduction of prejudice and discrimination. Out of a total of 985 studies, sixty percent (i.e., 591) were nonexperimental and thus do not allow us to draw reliable causal conclusions. Of the remaining studies, 287 were conducted in the laboratory, and only 107 articles reported randomized field experiments. Of these 107 randomized field experiments, 36 were concerned with cooperative learning ("jigsaw classroom") and the remaining 71 examined other prejudice reduction methods. Of the 107 randomized field experiments only 13 were carried out with a non student population. Importantly, only a minority of the prejudice reduction methods developed in scientific laboratories were systematically tested in the field. The effectiveness of the most prominent scientific approaches—such as decategorization, recategorization, and cognitive training—has rarely been examined in field studies. Thus, the contribution of scientific social psychology to the reduction of prejudice and discrimination may be qualified as modest at best.

Social psychologists in France are more often than not absent from the main decision-making bodies concerned with prejudice

reduction such as the HALDE, the "National Agency for Social Cohesion and Equality of Chances", and the Ministry of Housing and Urban Planning. Social psychologists rarely take part in public debates on prejudice in the mass media (Brauer, Martinot, & Ginet, 2004). This lack of participation is not surprising if one considers the modest contribution of scientific social psychology to the creation and evaluation of prejudice reduction programs. The methodological considerations in the next sections of this article should prove useful to any researcher who wants to reduce this lack of impact.

## Carrying out randomized experiments in the field

Imagine that a social psychologist wants to create a prejudice reduction intervention in the field and a decision-maker who wants to evaluate the efficiency of an intervention s/he just conducted. Both probably ask themselves the same questions: Is it necessary to include a control group? How should participants be assigned to the experimental conditions? What is the most appropriate experimental design? What are the best dependent measures? In the following we will attempt to provide some answers to these questions.

Just as in laboratory experiments, a control group is necessary to be able to draw reliable conclusions in field experiments. At least two groups are required, one with participants who are exposed to the intervention (the "test group") and one with participants who are not (the "control group"). To test the intervention as a whole, control group participants should not be exposed to the intervention at any moment. If the concern is a specific "ingredient" of the intervention, then control group participants should be exposed to the entire intervention except the specific part that one wants to test. It should be emphasized that a simple pretest-posttest design, in which participants are measured once before and once after the intervention and in which there is no control group, is not satisfactory. In such a design, the researchers may observe a reduction in prejudice between the pretest and the posttest but this reduction may well be due to causes other than the intervention: maturation, exposure to events reported by the mass media, seasonal mood change, becoming familiar with the

questionnaire items, etc. Only an experimental design with a control group makes it possible to rule out such alternative explanations (Judd & Kenny, 1981).

Randomization is a second key element. As in laboratory experiments, randomization is like a lottery whereby "units" (generally participants) are assigned either to the test group or the control group. Ideally, randomization is achieved by a coin toss or by generating random numbers on a computer. Randomization ensures that the assignment of participants to the different experimental conditions is unbiased. The larger the sample, the more probable it is that participants with different characteristics (i.e., motivation, intelligence, self-esteem) are evenly distributed between the different experimental conditions. In addition, by including appropriate measures the investigator may also estimate the extent to which the groups are equivalent across conditions. By randomly assigning participants to experimental groups, the investigator can determine whether prejudice reduction is due to chance or to the intervention (Gerber, Green, & Kaplan, 2004).

Although it is desirable to randomly assign participants to experimental conditions, it is not always possible to do so in the field. For example, if one attempts to reduce prejudice with a poster that praises diversity, it is impossible to randomly assign individuals to levels of exposure to the poster in the field. Most of the time, the investigator will thus assign higher-order units to the experimental conditions. For example, s/he may choose two schools and assign one to each experimental condition. In such a case, however, it is crucial to ensure that the higher order units (i.e., the two schools in our example) are similar on all relevant dimensions (e.g., school size, rural vs. urban location, pupils'socioeconomic status, percent of foreign pupils, etc.). From a methodological viewpoint it is recommended to include in the experiment the largest possible number of higher-order units and to assign them randomly to the experimental conditions. For example, one may choose thirty schools and assign fifteen of them to each experimental condition. The method can be further improved by using a matched randomized design. In such a design, the investigator forms pairs of higher-order units

such that the two members of the same pair share a maximum of common characteristics, and then randomly assigns, for each pair separately, one member to the test group and one member to the control group. When participants are part of higher-order units (e.g., schools), specific statistical procedures have to be used. These procedures will be presented in later sections of this article.

Including a pretest is a double-edged sword. On the one hand, a pretest is useful because it may be used to test whether the units assigned to the different experimental conditions are indeed equivalent before the intervention. A pretest is highly recommended if one only has two higher order units (e.g., two schools) and each of them is assigned to one experimental condition (test *vs.* control). Investigators should be aware, on the other hand, that a pretest may bias the results, especially when one deals with the measurement of prejudice[1]. The mere fact of answering a scale measuring prejudice towards an out-group and of participating in a study on prejudice may reduce participants'prejudice level, even in the control condition (Brown, 1995). Also, some prejudice measures do not lend themselves well to being administered twice in a short time period. For instance, it is not easy to imagine how the "lost letter" technique (see below) could be used multiple times with the same participants. In our research (Brauer & Er-rafiy, 2009; Er-rafiy & Brauer, 2010) we only used a pretest in the following cases: the pretest and the posttest were sufficiently distant in time, there were only two higher order units available, it was not clear whether participants in the test and control groups had the same prejudice level before the intervention, and if there were good reasons to think that including the pretest would not influence the prejudice level at the time of the posttest.

Researchers should conduct field studies with different populations and in different contexts. More precisely, in addition to pupils and students, investigators can resort to other populations, such as employees, neighborhood residents, medical

1. In studies that include a pretest, it is possible to measure participants' awareness through the "Solomon 4-group design" (half of the participants in each experimental condition are pretested and the other half are not).

practice patients, car wash clients, public transportation users, etc. It is also possible to run field studies in countries that are characterized by highly conflictual intergroup relations, such as Northern Ireland (e.g., Hewstone & Cairns, 2001) or Israel (e.g., Salomon, 2004). Also, as mentioned above, Paluck (in press) tested the efficiency of a prejudice reduction method in the Democratic Republic of Congo, one of the world's most conflict-prone regions.

Investigators who conduct field studies should bear in mind the objective of diversifying the dependent measures. In order to avoid conceptual ambiguity, it is important to be aware of the distinction between the three following concepts: stereotype, prejudice, and discrimination. A *stereotype* is defined as beliefs about the attributes of outgroup members (Leyens, Yzerbyt, & Schadron, 1994). A *prejudice* is a generalized negative affect toward members of an out-group (Allport, 1954). *Discrimination* is a negative behavior towards the members of an out-group, such as refusing to hire or refusing to rent an apartment to a member of a minority group (Dovidio & Gaertner, 1986). Stereotype, prejudice, and discrimination correspond respectively to the cognitive, affective, and behavioral components of an intergroup attitude (see Brauer, 2005a, 2005b). It should be noted that stereotype and prejudice are by no means interchangeable constructs: one may associate some traits with an out-group without having a negative affective reaction towards that group. In support of this view, Dovidio, Brigham, Johnson, and Gaertner (1996) found a correlation of .16 between stereotypes and prejudice. Among the Caucasian majority participants studied by Judd and his colleagues (Judd, Park, Ryan, Brauer, & Kraus, 1995) the stereotype-prejudice correlation was small and nonsignificant. In conclusion, the relevant dependent measure should specifically assess the aspect(s) of an intergroup attitude one is interested in measuring.

There are different ways of measuring an aspect of an intergroup attitude in the field. One such possibility is through a questionnaire that surveys participants' stereotypes or level of prejudice. Questionnaires are of two types, direct and indirect. Direct questionnaires, as their name indicates, measure stereotypes and

prejudice in such a way that the objective of the questionnaire is transparent to participants. The social distance scale of Bogardus (1925), which measures prejudice, is an example of such a questionnaire. It uses a set of assertions describing situations that involve the presence of a member of a given social group at varying social distances, and for each situation respondents state whether they would accept or not that situation (e.g., whether they would accept a Mexican as their neighbour, whether they would approve of a Mexican marrying into their family, etc.). Brigham (1971) suggested measuring stereotypes by asking respondents to estimate the percentage of members from a given group that possess a given trait. The limitation of direct questionnaires, as researchers have acknowledged for a long time, is that answers are biased by social desirability or politically correct norms (Brown, 1995). Indirect questionnaires do not eliminate these biases, but they reduce them because the questions make the investigator's goal less obvious. The indirectness is achieved by the use of more abstract formulations, i.e., formulations that are less directly related to disliking the target group under consideration (e.g., the *Modern Racism Scale* of McConahay, 1986; the *Modern Sexism Scale* of Swim, Aikin, Hall, & Hunter, 1995).

Behavioral measures are another way of measuring intergroup attitudes. This kind of measures is one of the strength of field studies. Behavioral measures make sense in the field because it is more probable that discriminatory behaviors occur in the field, especially when participants feel unobserved. Among the behavioral prejudice measures that are easily applied in the field are measures such as the percentage of people who help an outgroup member (Gaertner & Dovidio, 1977), the amount of a precious resource that participants give to a member of a disliked outgroup (Paluck, in press), choosing or not a member of a minority group to be part of a game team (Johnson & Johnson, 1989), donating to an association that fights for the defense of ethnic minority rights (Rokeach, 1971), the physical distance separating a participant from a member of the outgroup (Katz & Zalk, 1978), mailing a letter that a member of a minority group forgot in a public place (Milgram, Mann, & Harter, 1965), the willingness of a child to take part in a picnic with children who are members of an outgroup

(Weiner & Wright, 1973), and the number of votes cast for a member of an ethnic minority (Altemeyer, 1994).

The most effective interventions are those that induce a long-lasting change in attitudes and behaviors (e.g., Hill & Augustinos, 2001). To test for long-term effects, longitudinal measures are required, i.e., one needs to measure participants' intergroup attitudes multiple times after the end of the intervention (Aboud & Fenwick, 1999).

In sum, we advise researchers and public policy decision-makers willing to conduct field experiments to take into account the following points. It is necessary to include a control group and to randomly assign participants (or groups of participants) to the test group and the control group. If higher-order units (e.g., schools, neighborhoods, workplaces) are used, we advise the use of a matched randomized design, in which similar units are paired and then randomly assigned to the experimental conditions. We also advocate the use of a variety of measures, from questionnaires to behavioral indices. Ideally, these measures should be administered a reasonable interval after the intervention. Finally, investigators should obtain data from a variety of populations and not only from students.

## How to analyze data from randomized field experiments

In the following paragraphs, we discuss two data analysis procedures, one that is relatively easy to comprehend and one that requires more elaboration. The first data analysis procedure is called regression adjustment. When analyzing data from a study including both a pretest and a posttest, it is preferable to enter the pretest as a covariate in an analysis in which the dependent variable is the posttest, rather than analyzing the data as if pretest and posttest were the two levels of a within-subject variable (i.e., repeated measures, see Judd & Kenny, 1981). The inclusion of the pretest as a covariate is statistically more powerful, because the repeated-measures approach makes the (probably incorrect) assumption that there is a one-to-one correspondence between pretest and posttest (for further details see Brauer & Judd, 2007).

The regression adjustment approach can be used with a variety of experimental designs in which a pretest and a posttest are considered in the presence of one or more (continuous or categorical) between-participants variables.

The second data analysis procedure concerns experimental designs in which participants are part of higher-order units (e.g., schools, workplaces, neighborhood houses, etc.). In this case, the statistical approach needs to take into account the fact that the data are non-independent. On average, two pupils from the same school will give responses that are more similar to each other that those of two pupils from different schools. The way to analyze non-independent data depends on the number of higher-order units. Data from studies that included more than 20 higher-level units can be analyzed with a family of statistical techniques that have numerous desirable features. If there are less than 20 higher-order units, researchers are obliged to use different, less convincing analytic approaches.

If the randomized field experiment contains more than 20 higher-level units, one may aggregate the data within each higher-order unit and run the analyses with the higher-order unit as the level of analysis. For example, if an investigator collected data from 30 schools, s/he would average the responses from all the pupils in the same school and run analyses with school as the level of analysis. These analyses would then be based on an $n$ of 30 observations. If the experimental condition (test group *vs.* control group) were the only independent variable of interest, the investigators would run an independent samples t-test with $n_{higher\ order\ units} - 2$ degrees of freedom. In the example above, they would end up with 28 degrees of freedom. Although the drastic reduction in sample size due to the utilization of the higher-order unit as the unit of analysis may be seen as a disadvantage, the loss of statistical power is often negligible. Aggregating within each higher-order unit diminishes random error, which compensates for the loss of statistical power.

If the investigator disposes of more than 20 higher-level units, an even more desirable approach is the use of multilevel (or hierarchical) modeling (see HLM software, the "proc mixed" procedure in SAS, or the "linear mixed models" module in SPSS). Although

multilevel modeling does essentially the same thing – it takes into account the non-independence of participants' responses by averaging within higher-order units – it has two advantages over the approach described in the previous paragraph (Bryk & Raudenbush, 1992). First, multilevel models offer results based on both ordinary standard errors and robust standard errors. The latter, as the name indicates, converge to the true standard error value even if the assumptions about the distribution and covariance structure of the random effect are incorrect (Raudenbush & Bryk, 2002). Moreover, a comparison of ordinary and robust standard errors can be used to check for model misspecification: a large discrepancy means the model can be improved (by incorporating slope heterogeneity into the model for example). Second, multilevel modeling offers more flexibility in the questions that may be asked from the data. For example, it allows researchers to examine, individually or jointly, the influence of independent variables related to the participants (e.g., gender, age, number of family members from other cultures) and independent variables related to the higher-order units (e.g., school size, percent of minority pupils in school, director's endorsement of anti-discrimination measures).

If there are fewer than 20 higher-level units (e.g., fewer than 20 schools), the researcher cannot use any of the statistical techniques described above. The main reason is insufficient statistical power. Given that those techniques average across higher-order units, the statistical analyses are based on as many observations as there are higher-order units. And it simply does not make sense to run a t-test or an ANOVA with fewer than 20 observations. With such a test, the researcher has a very low chance of detecting an effect even if this effect were to exist (Murphy & Myors, 2004). The second reason concerns multilevel modeling only. This statistical technique estimates parameters by iterative algorithms that yield unbiased estimates only when there are at least 20 higher order units (Raudenbush & Bryk, 2002)[2].

With fewer than 20 higher-order units, the researcher is obliged to use statistical techniques that deal with the non-independence

2. Some experts even suggest a minimum of 30 higher-order units (Gelman & Hill, 2007).

of responses in a different way. We advise the use of a correction procedure developed by Kish (1965, 1987). This approach consists of running a standard *t*-test with participants as the unit of analysis and diminishing the *t*-value depending upon the degree of non-independence caused by higher-order units. Remember that in a standard independent-samples *t*-test, the *t*-value is computed by dividing the difference of the two group means ($\overline{X}_1 - \overline{X}_2$) by the standard error (SE). Kish's correction procedure increases the standard error, and thus decreases the *t*-value and reduces the probability of rejecting the null hypothesis. The extent to which the standard error is increased depends upon the number of higher order units and on the size of the intraclass correlation (i.e., the degree of non-independence). The Kish-corrected standard error is defined as

$$SE_{corrected} = SE * (1 + (n_{b\,o\,units} - 1)\,\rho) \quad (1),$$

where $n_{b\,o\,units}$ is the number of higher order units (e.g., number of schools) and $\rho$ is the intraclass correlation coefficient. The latter is defined as

$$\rho = \sigma_{u0}^2/(\sigma_{u0}^2 + \sigma_e^2) \quad (2),$$

where $\sigma_{u0}^2$ is the between-group variance and $\sigma_e^2$ is the within-group variance[3]. To sum up, the Kish correction uses the corrected standard error from Equation 1 as a denominator and results in a corrected *t*-value that has N-2 degrees of freedom (N being the total number of participants):

$$t_{corrected} = \frac{\overline{X}_1 - \overline{X}_2}{SE_{corrected}} = \frac{\overline{X}_1 - \overline{X}_2}{SE * [1 + (n_{b\,o\,units} - 1)\rho]}$$

$$= \frac{\overline{X}_1 - \overline{X}_2}{SE * \left[1 + (n_{b\,o\,units} - 1)\left(\frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_e^2}\right)\right]} \quad (3)$$

3. Most statistics packages allow researchers to easily obtain between-group and within-group variances. In SPSS 17.0, for example, one has to run an ANOVA with the higher order units as an independent variable (e.g., twelve schools coded as 1, 2, 3, ... , 12). Then, in "Options", one selects "Fixed and random effects". In the output, the between-group variance is called "Between-component variance" (in the "Descriptives" table), whereas the within-group variance is reported in the cell that is at the intersection of the column labeled "Mean Square" and the row labeled "Within Groups" (in the "ANOVA" table). With Statistica 6.1, one has to choose the "General Linear Models" module, use the higher order units as the independent variable, and check the "Random Factors" option under "Options". In the output, one clicks on "Variance Components" under "Summary". The between-group variance can be found in the line with the name that was given to the variable that contained the higher order units. The within-group variance is reported in the line labeled "Error".

The corrected *t*-value is then compared to the critical *t*-test value for the corresponding number of degrees of freedom (available from *t*-test tables or several spreadsheet programs) in order to determine whether it is statistically significant. For a concrete example of how to compute a Kish-corrected *t*-value, see Experiment 3 below.

Note that if there are several dependent variables, the intraclass correlation coefficient will of course vary from one dependent variable to another, so the Kish correction has to be computed separately for each dependent variable.

In the standard General Linear Model, it is possible to test for the effect of the experimental manipulation while controlling for the effect of other variables that may covary with the dependent variable (e.g., participants' age or gender). A similar approach can be used when the data are non-independent. In such a case, the researcher would regress the dependent variable on the experimental condition and on the variables to be controlled for. This regression yields a regression coefficient for the experimental condition and the standard error of this regression coefficient. Instead of simply dividing the regression coefficient by its standard error, which is how the *t*-value is computed in standard multiple regression, the researcher should divide the regression coefficient by the Kish-corrected standard error, which can be computed in the same way as described above.

The only statistical approach we strongly advise against is the use of a standard *t*-test (or regression analysis) that ignores the non-independence problem. Numerous statisticians have pointed out that analyzing non-independent data as if they were independent leads to increased type-I error rates (i.e., a significant result although the prejudice intervention had absolutely no beneficial effect). Recent work by Musca et al. (2010) demonstrates that the type-I error rates can be extremely high in these types of analyses, even when the intraclass correlation coefficient is quite low. For example, with 10 higher-order units, 100 participants per higher-order unit, and an intraclass correlation of .2, the type-I error rate is 68%, which is well above the conventional level of 5%. Musca and colleagues also show that the Kish correction ameliorates the problem but does not eliminate it under all circumstances. When

the experiment includes few higher-order units and a large number of participants, the type-I error rates are elevated even if Kish's correction procedure is used. For the case described above ($n_{b\ o\ units}$ = 10, N = 1000, and $\rho$ =.2), the type-I error rate after application of Kish's correction is 26%. Interestingly, the type-I error rate drops to acceptable levels as the number of participants per higher-order unit decreases or as the number of higher-order units increases. This finding leads to the unintuitive suggestion that it is preferable to use an intermediate rather than a large number of participants per higher-order unit when researchers dispose of relatively few higher-order units. This is because the type-I error rate rises to an unacceptable level when the number of participants per higher-order unit becomes too large.

To summarize, we suggest that researchers make every effort to include as many higher-order units as possible. If they have more than 20 higher-order units, they can either average within higher-order units or use multilevel modeling. If they have fewer than 20 higher-order units, they should use a correction procedure that adjusts for the non-independence of participants' responses. Note that the application of the Kish correction is preferable to ignoring the non-independence problem, but that Kish-corrected results still have to be interpreted with caution because under certain circumstances, the type-I error rate is higher than the acceptable level.

## Experiment 1

We now report three field experiments on prejudice reduction in order to achieve two goals. First, the experiments allow us to illustrate the points we raised in the previous sections of this article. Second, they contribute to the small but growing literature on effective prejudice interventions. If Paluck and Green's (2009) list is correct, the present article is the 14th in the literature on prejudice interventions in which a randomized field experiment was conducted with a non-student population (see Experiment 3). In the first experiment described below, we evaluated the effectiveness of a "diversity training" intervention. The two other experiments examined the effectiveness of an intervention consisting of a poster that highlights within-group differences in a minority

group (Experiments 2 and 3). All three experiments addressed prejudice against Arabs, one of the largest minority groups living in France. The three experiments do not comply with all the suggestions made above, but they allow us to illustrate our main purpose.

The focus of Experiment 1 was "diversity training". This type of training generally has two objectives. First, it attempts to make participants aware of ethnic and cultural differences and to teach them to value these differences (Hollister, Day, & Jesaitis, 1993). A second objective is to promote acceptance of, and empathy, compassion and solidarity with members of disadvantaged groups (Paluck, 2006). The training sessions are usually led by a "diversity consultant" who proposes different activities such as watching movies, role playing, reading short texts, and group discussions. In most cases, these activities also include a presentation on how to define prejudice and discrimination, and the best ways to reduce them (for further details, see Paluck, 2006). Stephan and Stephan (2001) object that most of these training programs are not based on theory or empirical data, and their effectiveness is never empirically evaluated. The aim of Experiment 1 was to illustrate how one can design a field experiment that evaluates a particular aspect of diversity training. More precisely, we examined the influence of empathy on prejudice reduction. Indeed, Stephan and Finlay (1999) showed that empathy originating from reading a report of discriminatory acts perpetrated against Blacks causes readers to evaluate Blacks more positively.

Experiment 1 included a control group in addition to the test group. First, we measured pupils' attitudes towards Arabs in two different school classes (pretest). Then the pupils of one class (i.e., the diversity training group) viewed a documentary showing Arab individuals who recounted a situation in which they had been discriminated against in a hiring situation. The pupils in the other class (i.e., the control group) were not exposed to any particular information about Arabs. Finally, attitudes towards Arabs were measured again (posttest). Our hypothesis was that pupils in the diversity training group would have more favorable attitudes toward Arabs than pupils in the control group. Specifically, we expected to find (a) less in-group bias, (b) less

prejudice, (c) a lesser social distance, and (d) more positive contacts with Arab individuals in the diversity training group than in the control group.

## Method

### Participants

Thirty-six high-school seniors of French nationality from two different classes were included in the experiment. There were 25 females (*mean age* = 17.56, *SD* = .96) and 11 males (*mean age* = 17.45, *SD* = .82). The intervention took place in the classroom. The two classes were randomly assigned to one of the two experimental conditions "diversity training" and "control".

### Material

All participants filled out a questionnaire that measured their attitudes towards Arabs. They completed the questionnaire twice, once at the beginning and once at the end of the experiment. Participants responded to each item by drawing a vertical line on a continuous scale ranging from "I disagree entirely" to "I agree entirely". The questionnaire included the following measures.

The first scale measured *in-group bias* (Aboud, 2003). Participants were asked to rate the extent to which French people and Arab people possess several attributes (e.g., selfish, aggressive, hard-working, warm; see Dambrun & Guimond, 2004, for the choice of the attributes). Question order was counterbalanced so that half of the participants evaluated first Arabs and then the French, whereas the other participants proceeded in the inverse order.

To measure pupils' level of *prejudice* against Arabs, we used an adapted version of the Modern Racism Scale (McConahay, 1986). The original scale has 15 items but we selected the 8 items that correlated most highly with the two main factors of the original scale, nationalism and intolerance (see Dambrun & Guimond, 2001, for the choice of the items). Four items express positive affect (e.g., "One may easily understand the anger that Arabs living in France feel") and four items express negative affect (e.g.,

"The high unemployment rate in France is due to the Arabs, who take away the jobs from the French people"). Given that the Cronbach's alpha values were .92 and .96 for the pretest and the posttest respectively, this scale seems reliable.

Another scale measured *social distance* (Green & Wong, 2001). It has 3 items, one positive ("If a person of a different religion was put in charge of me, I would not mind taking advice and direction from him or her") and 2 negative ("I would probably feel a little self-conscious dancing with an individual of Arab origin in a public place", and "I wouldn't want to be around a teenager who is of Arab origin"). The Cronbach's alpha values were .95 and .72 for the pretest and the posttest, respectively.

Finally, the participants evaluated the quality of their contact with Arabs and with French people by answering the question "To what extent do you feel at ease with Arabs [the French]?" (Berry & Kalin, 1995). Question order was counterbalanced in the same way as for the in-group bias measure.

Participants in the diversity training group viewed a 15-minute excerpt from a national TV documentary that showed people of Arab origin who had been victims of discrimination at the work-place. Participants in the control group did not see the documentary.

### *Procedure*

Pupils participated in the experiment in their own classroom. The experiment was run in several phases. During the first week, a female experimenter of European (i.e., Caucasian) appearance entered the classroom and asked the pupils to take part in a brief study conducted by researchers from the Department of Psychology. The week after, participants completed the questionnaire described above (pretest). Three weeks later, the pupils from the diversity training class viewed the documentary during a one of their "social and economic sciences" classes. In order to make the link between the questionnaires and the documentary less obvious, the questionnaires were handed out by the experimenter, while the documentary was presented by one of the pupils' usual teachers. During the fifth week, participants completed again the

questionnaire measuring their attitudes towards Arabs. The experimenter explained that the first questionnaire had some spelling errors and could not be used for data analysis. At the end of the experiment, the participants were debriefed and thanked.

### *Results and Discussion*

We recoded participants'responses by attributing each of them a score between 1 and 28. Preliminary analyses showed that gender, age, and question order had no influence on the results, so these variables were not included in the analyses reported below. An analysis of covariance (ANCOVA) was conducted for each dependent variable, with the posttest as dependent variable, the experimental condition (diversity training *vs.* control) as independent variable, and the pretest as covariate.

The *in-group bias* was computed as follows. A liking score was computed for each target group by subtracting the average rating of the negative traits from the average rating of the positive traits. Then, we obtained an in-group bias score by subtracting the liking score for Arabs from the liking score for the French (i.e., higher scores mean greater in-group bias). The results revealed an effect of the pretest, $F$ (1, 33) = 6.21, $p$ <.02. In addition, there was less in-group bias in the diversity training group (*adjusted mean* = -2.11, *SD* = 1.78) than in the control group (*adjusted mean* = 1.77, *SD* = 1.78), but this difference did not reach conventional levels of statistical significance, $F$ (1, 33) = 2.38, $p$ = .13.

The *prejudice* score was computed as the mean of the eight corresponding items (taking into account the reverse-coded items). The pretest effect was significant, $F$ (1, 33) = 32.30, $p$ <.001, as well as the effect of the experimental condition, $F$ (1, 33) = 7.51, $p$ =.01. Participants from the diversity training condition exhibited less prejudice towards Arabs than participants from the control condition (see Figure 1, top panel).

The *social distance* score was computed as the mean of the three corresponding items (taking into account the reverse-coded item). A pretest effect was observed, $F$ (1, 33) = 12.56, $p$ <.001, as well as an effect of experimental condition, $F$ (1, 33)

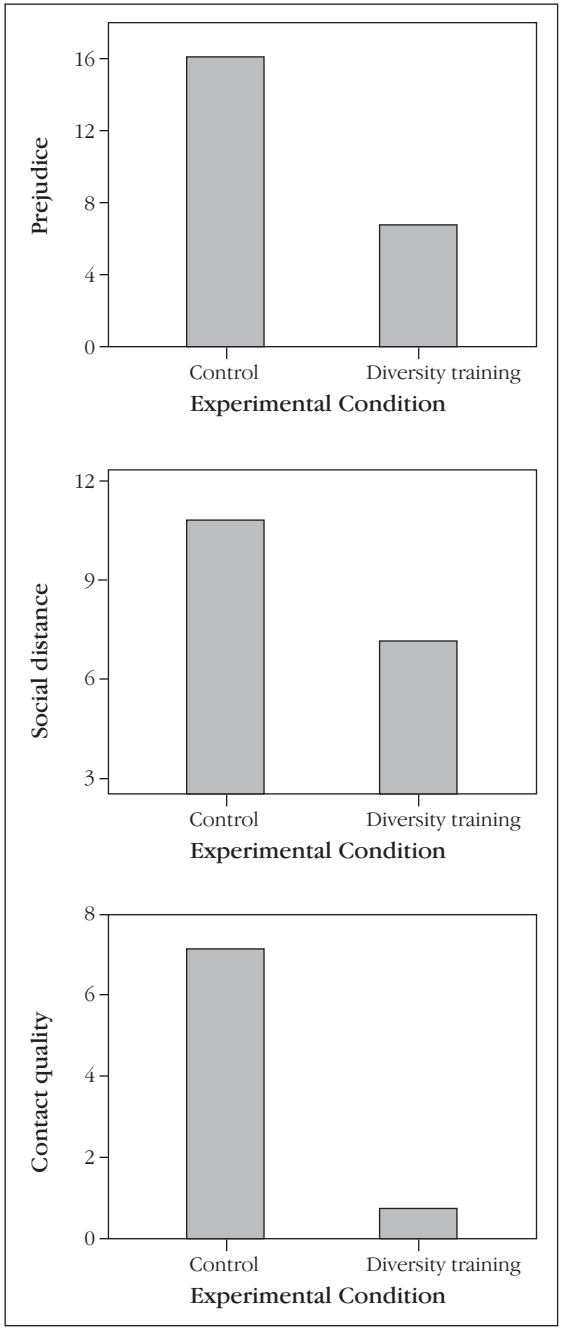


FIGURE 1: Prejudice, social distance and contact quality at posttest as a function of experimental condition in Experiment 1. The displayed values are adjusted means.

$= 5.95, p < .02$. As can be seen in Figure 1 (see middle panel), the score of the participants in the diversity training condition shows a lesser social distance to Arabs in this condition, as compared to the control condition.

To determine whether the intervention influenced the *contact quality*, a difference score was computed between the estimation of the contact quality with French and that with Arabs. The higher the scores, the more a participant feels more comfortable around French than around Arabs. The analyses revealed a pretest effect, $F\ (1,\ 33) = 42.57,\ p < .001$, as well as an effect of the experimental condition, $F\ (1,\ 33) = 20.95,\ p < .001$. At posttest, participants in the diversity training condition felt equally comfortable with members of both target groups, whereas participants in the control condition felt more comfortable around French people than around Arab people (see Figure 1, bottom panel).

Taken together, these results show that it is possible to influence positively individuals' attitudes towards a minority group by making them aware of the discrimination that members of the minority group (here, Arabs) experience. Participants who viewed a documentary on people of Arab origin who were discriminated against at their workplace (a) were less prejudiced toward Arabs, (b) felt socially less distant to Arabs, and (c) felt more at ease with Arabs than participants in the control condition. No significant difference in ingroup bias was found between the diversity training group and the control group. The results of this first experiment also revealed an effect of the pretest on each of the dependent variables. The fact that participants are consistent in their responses is an indicator of the good reliability of our measures (i.e., high test-retest reliability). Experiment 1 illustrates how one can design a field experiment that empirically tests the effectiveness of diversity training. We resorted to an approach in which we compared participants who were subjected to an intervention aimed at reducing prejudice ("test group") to participants were not exposed to the intervention ("control group").

While various dependent variables were used here, we did not measure behavior. To exemplify how measures of behavior can be included in a field experiment we report the following two exper-

iments. These experiments examine the impact of modifying perceived variability on helping behavior (Experiment 2) and on social distance (Experiment 3).

## Experiment 2

*Perceived variability* refers to the fact that two individuals, while associating the same characteristic with a given group, may have a different perception of the extent to which the members of that group share the characteristic under consideration. For example, one of the two individuals may perceive the members of an outgroup as very similar to each other on a given characteristic (homogeneous perception), while the other individual perceives the members of the same out-group as rather different from each other on that same characteristic (heterogeneous perception; Park & Rothbart, 1982). Laboratory research has shown that increasing the perceived variability of an outgroup has a positive effect on the stereotyping, prejudice, and discrimination. More precisely, the more heterogeneous a group is perceived, (a) the less likely it is that a group stereotype is applied to a given member of the group (Ryan, Judd, & Park, 1996), and (b) the less likely it is that a characteristic of a given member is generalized to the entire group s/he belongs to (Park & Hastie, 1987).

We recently conducted an experiment that provides experimental field evidence for the idea that perceived variability is positively related to the reduction of discrimination. This experiment, which is submitted elsewhere (Brauer & Er-rafiy, 2009), illustrates well our point. In the experiment, participants were asked to take a seat in a waiting room near the laboratory (*participants'mean age* = 19.47, *SD* = 1.75). While they were waiting, participants were exposed (test group) or not (control group) to information that drew their attention to differences among Arabs. There were six posters on the walls of the waiting room. Five posters, identical in the two experimental conditions, promoted changes in favor of desirable behaviors (e.g., quitting smoking, not driving when under the influence of alcohol, recycle more). The sixth poster varied between the two experimental conditions. In the "heterogeneity condition", the poster showed the pictures of 12 Arab individuals. Eight pictures were accompanied by a box

containing the name, the age and one characteristic of the individual (e.g., "Fatima, 49 years old, lawyer", "Btissam, 24 years old, unemployed for one year"). At the bottom of the poster a slogan in very large print read "Notre point commun: la diversité" ("What makes us the same – is that we are all different"; see Figure 2). In the "control condition", a poster with a similar layout encouraged people to eat more fruits and vegetables. After allowing the participants wait for a few minutes, an experimenter took them to an experimental room and seated them in front of a computer, where they completed a distractor task for fifteen


FIGURE 2:
The poster used to modify participants' perceived variability of Arabs in Experiments 2 and 3.

minutes. Upon completion of the task, participants were asked to fill out a questionnaire. The questionnaire measured their perception of the variability of Arab and French people[4]. When participants completed the questionnaire, they were told that the experiment was over, debriefed about the distractor task, and thanked. The experimenter instructed the participants to go to another building to get their experimental credit validated. In the other building, a female confederate of Arab origin went past the participant and dropped a large bag such that the bag's contents spilled out. We then measured whether the participant spontaneously offered to help the confederate. In the literature, helping behavior is considered an indirect measure of discrimination (Crosby et al., 1980). Note that the confederate was unaware of the condition the participant was in (test or control).

The results showed that participants in the heterogeneity condition perceived Arabs as more different from each other than participants in the control condition. The behavioral measure showed that participants in the heterogeneity condition were more likely to help the Arab confederate than participants in the control condition (see Figure 3). Mediation analyses suggested that this latter effect was due to the fact that the poster effectively modified participants' perception of variability of Arabs. While one may question whether this experiment should be considered a laboratory or a field experiment, one should take into account that the dependent variable was measured outside the laboratory and that participants were unaware of the fact that their helping behavior was assessed.

## Experiment 3

In the previous experiments, participants were not part of higher order units and all observations were independent. In Experiment 3, participants were recruited in the offices of six different physical therapists. This experiment thus allows us to illustrate the analyses of non-independent data with a concrete example.

4. The participants indicated the extent to which they thought Arabs [the French] were different from each other by drawing a single vertical line on a continuous scale ranging from "not at all different" to "very different". The scales were later divided into 28 intervals of equal size, and each response was a number from 1 to 28.

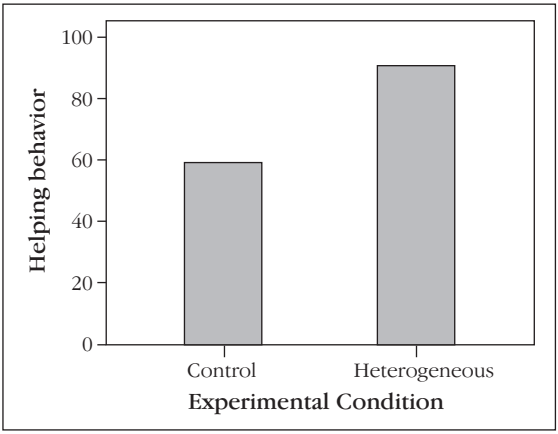


FIGURE 3:
Percent of participants who helped the Arab confederate in Experiment 2.

## *Method*

### *Participants*

Three hundred and thirty-one men (154) and women (177) from a medium-sized French city participated in the experiment. Both the experimental manipulation and the behavioral measure took place in the waiting rooms of various physical therapists.

### *Material and procedure*

First, three pairs of physical therapist offices were created in such a way as to pair offices that shared a maximum of common characteristics (e.g., socio-economic status of the neighborhood, size of the seats in the waiting room, percent of foreigners among patients). Then, within each pair, one physical therapist office was randomly assigned to the heterogeneity condition and the other to the control condition.

The experiment ran over a course of four weeks. During the first phase participants' perception of variability of Arabs was manipulated: a poster highlighting the differences among Arabs (see Experiment 2) was displayed during two weeks in the waiting rooms of the three physical therapists in the heterogeneity condition, while no poster was displayed in the waiting room of the three physical therapists in the control condition. During the fourth week of the experiment (i.e., seven to ten days after the poster had been removed), we evaluated participants' attitudes

towards Arabs through a measure of social distance (Word, Zanna, & Cooper, 1974). In order to do that, the six waiting rooms were set up so that all the seats were lined up. A male confederate of Arab origin took place in the waiting room, always on the seat that was closest to the door. Once a patient entered the waiting room and sat down, the confederate inconspicuously wrote down the number of seats that separated him from the patient. The seat immediately next to the confederate was coded 1, the seat second next to the confederate was coded 2, and so forth. The confederate also noted the participant's gender, an estimation of participant's age, and evaluated the participant's appearance (on a Likert scale ranging from 1, very cheap clothes, to 7, very expensive clothes). Only situations in which the confederate was alone in the waiting room when a new patient arrived were counted as experimental trials.

### *Results*

Results show that the participants from the heterogeneity condition sat down closer to the Arab confederate (*mean* = 1.93, *SD* = .76) than participants in the control condition (*mean* = 2.57, *SD* = .84). Remember that participants were part of higher-order units (physical therapist offices) and that their behaviors are not independent of each other. We therefore analyzed our data by following our own suggestions about the analysis of non-independent data (see first part of this article). We first conducted a one-way ANOVA with physical therapists as the independent variable (numbered from 1 to 6) and with the distance of the chosen seat as the dependant variable. The unit of analysis was the participant (N = 331). This analysis yielded a between-group variance ($\sigma_{u0}^2$) of .19 and a within-group variance ($\sigma_e^2$) of .58. Equation (2) allowed us to compute the intraclass correlation: $\rho = \sigma_{u0}^2/(\sigma_{u0}^2 + \sigma_e^2) = .19 / (.19 + .58) = .24$. Note that an intraclass correlation of .24 is far from negligible, which suggests that a relatively high degree of non-independence was present in our data. We then conducted an independent samples *t*-test with the participants as the unit of analysis, experimental condition as the independent variable, and distance of the chosen seat as the dependant variable. This *t*-test indicated that the difference between the heterogeneity group and the control group ($\overline{X}_1 - \overline{X}_2$) was .64 chairs, and that the (non corrected) standard error (SE) was .09. We used

equation (1) to compute the Kish-corrected standard error: $SE_{corrected} = SE * [1 + (n_{h\,o\,units} - 1)\,\rho] = .09\,[1 + (6 - 1).24] = .09 * 2.2 = .19$. Note that the Kish-corrected standard error is more than twice as large as the uncorrected standard error. The Kish-corrected *t*-value is then $t\,(329) = .64/.19 = 3.37$, which corresponds to $p < .001$. This result reveals that the difference between the two experimental conditions is statistically significant, even with the non-independence of the data taken into account. A poster highlighting the variability among Arabs had a beneficial effect on participants' behavior in the presence of an Arab individual.

One may object that with 6 higher-order units, more than 50 participants per higher-order unit, and an intraclass correlation of .24, our Kish-corrected t-test has an inflated type-I error rate (see Musca et al., 2010). It is therefore appropriate to adopt a more conservative alpha level. With 329 degrees of freedom, the critical t-value is 2.61 for $p = .01$ and 3.35 for $p = .001$. Our Kish-corrected t-value of 3.37 exceeds these values, and it is thus highly unlikely that the observed difference between the experimental conditions is due to chance.

Given the differences between patients in different physical therapist offices, it is necessary to show that the effect of the experimental manipulation remains significant even if one statistically controls for participants' personal characteristics. As before, we examined this question by following the advice given in the first part of this article. We first ran a regression analysis with participants as the unit of analysis. We regressed the distance of the chosen seat on experimental condition (coded as -1 and 1), gender, estimated age, and appearance. The unstandardized regression coefficient associated with the experimental condition had a value of .32, and its associated standard error was .04. The Kish-corrected standard error is $SE_{corrected} = .04 * [1 + (6 - 1).24] = .10$, so the Kish-corrected t-test value testing the effect of experimental condition is $t\,(321) = .32/.10 = 3.19$, which corresponds to $p < .002$. Regardless of whether one chooses to adopt a more conservative alpha-level or to statistically control for participant characteristics (or both), the effect of the experimental condition is significant. Individuals who were exposed to a poster highlighting the differences among Arabs later sat closer

to an Arab individual than individuals who were not exposed to the poster.

## General Discussion

Three studies were reported to illustrate how to carry out field experiments that examine the effectiveness of prejudice interventions. In Experiment 1 participants viewed a documentary with Arabs who were the target of discrimination. Results showed that the intervention influenced intergroup attitudes. Participants who saw the documentary were less prejudiced, were less socially distant, and felt more at ease while interacting with Arabs than participants who did not see the documentary.

Experiment 2 and Experiment 3 are examples of field studies with a behavioral measure as a dependent variable. The manipulation of perceived variability of Arabs was achieved through a poster that highlighted the differences among Arabs. The results of Experiment 2 revealed that a female confederate of Arab origin received more help from participants who had been exposed to the poster than from participants in the control condition. The results also showed that the aforementioned poster caused participants to perceive Arabs as a more heterogeneous group. In Experiment 3, in which a manipulation identical to that of Experiment 2 was used, participants in the heterogeneity condition kept less distance to a male Arab confederate than participants in the control condition.

The methodology used in the three experiments reported here illustrates that it is relatively easy to conduct field experiments that examine the effectiveness of a given prejudice reduction method. The utility of the studies can be taken further by considering the strong and weak points of each. Experiment 1 has two main strengths. First, a control group was included in a repeated measures experimental design, that is, a design in which the attitudes of participants in both test and control group were assessed twice, once at the beginning and once at the end of the experiment. Such a pretest/posttest design including a control group allows researchers to draw reliable conclusions about the causal influence of one variable (here, the diversity training) on another

variable (here, prejudice; see Shadish, Cook, & Campbell, 2002). Second, many different dependent variables were used. More precisely, attitudes towards Arabs were measured in form of in-group bias, prejudice, social distance, and contact quality. Experiment 2 also has several strong features. First, we created a situation in which participants were induced to think the experiment was over, and their helping behavior (the dependent measure) was subsequently measured in a way that ensured that participants did not relate it to the experimental manipulation. Second, Experiment 2 included a behavioral measure of discrimination towards a member of a minority group, in this case a helping behavior. The strengths of Experiment 3 are the following. First, an unobtrusive behavioral measure of social distance was used as a measure of prejudice. Second, this measure was not taken immediately after the experimental manipulation, but more than 7 days later. Actually, our choice of running this experiment in physical therapist offices rather than in other offices was based on the observation that clients of physical therapists come back regularly. We could thus conduct a medium-term experimental manipulation and examine the effectiveness of the intervention after an extended period of time.

In spite of the strong points, the three experiments presented here could be improved in different ways. For instance, we tested only pupils and students in Experiments 1 and 2. It would have been preferable to include other populations as well. Also, we tested only the pupils of two classes in Experiment 1. It would have been preferable to obtain permission to run the experiment in many classes, to make pairs of classes, and randomly assign the classes in each pair to the experimental conditions. Finally, we measured only the short-term effect of the diversity training on participants' attitudes in Experiment 1. The inclusion of a delayed measurement, for example six months later, would have been a beneficial addition to the experiment. In Experiment 3, there were only 6 higher-order units and a relatively large number of participants per higher-order units. In order to avoid problems with inflated type-I error rates and to be able to conduct multilevel analyses, it would have been better to measure participants' behaviors in the waiting room of 20 or more physical therapists.

To summarize, the methodological suggestions made here are intended to be used by social psychologists and public policy decision makers who wish to conduct field experiments, whether their objective is to develop a new prejudice reduction method or to assess the effectiveness of a prejudice reduction intervention. We recommend the inclusion of a control group, in addition to the test group. We also recommend a random assignment to the test and control conditions, either at the participant level (e.g., pupils) or at the group level (e.g., classes of pupils). We encourage the use of indirect questionnaires and of behavioral measures, in order to eliminate response biases. These measures should be extended in the long-term, as much as many months after the intervention. Finally, researchers interested in field studies should consider many different populations, such as neighborhood residents, baker's store costumers, public transportation users, etc.

Though seldom used, field experiments are crucial to applied research in social psychology. First, they allow researchers to test in the field an intervention that was initially developed in the laboratory. Second, they make it possible to examine the long-term effects of an intervention in the field, that is, in a real-world situation characterized by political and economical influences, by social pressure and by distraction from features of the environment. Finally, field research helps generate new predictions and theories that improve our knowledge of social phenomena. In the future, field experiments on prejudice reduction should not only target the "What works?" question, but also other questions, such as "Why does it work?" and "Under what conditions does it work?" For example, many questions related to the diversity training technique remain unanswered. Is it desirable that members of a minority group are physically present during the diversity training? Should participants be encouraged to express blatantly their stereotypes and prejudice during the diversity training? If the diversity training has positive effects, how exactly does it reduce prejudice? Is the prejudice reduction due to empathy or to other factors?

All in all, there are many open questions in the field of prejudice and discrimination. The public image of social psychology

depends in part on our ability to answer these questions. The funding of applied or basic research in social psychology by public policy decision-makers may well hinge on social psychologists'ability to give practical and effective advice on how to reduce prejudice, backed up with solid empirical evidence. Thus, the future of social psychology depends in part on researchers'willingness to pursue applied research and to conduct randomized experiments in the field.

## References


Aboud, F. E. (2003). The formation of in-group favoritism and out-group prejudice in young children: Are they distinct attitudes? *Developmental Psychology, 39*, 48-60.

Aboud, F. E., & Fenwick, V. (1999). Exploring and evaluating school-based interventions to reduce prejudice in preadolescents. *Journal of Social Issues, 55*, 767-785.

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality.* New York: Harper.

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

Altemeyer, B. (1994). Reducing prejudice in right-wing authoritarians. In M. Zanna & J. Olson (Eds.), *The psychology of prejudice: The Ontario symposium* (p. 131-148). Hillsdale, NJ: Erlbaum.

Batson, C. D. (1991). *The altruism question: Toward a social-psychological answer*. Hillsdale, NJ: Erlbaum.

Berry, J. W., & Kalin, R. (1995). Multicultural and ethnic attitudes in Canada: An overview of the 1991 national survey. *Canadian Journal of Behavioural Science, 27*, 301-320.

Blank, R. M., Dabady, M. C., & Citro, C. F. (2004). *Measuring racial discrimination*. Washington, DC: National Academies Press.

Bogardus, E. S. (1925). Measuring social distances. *Journal of Applied Sociology, 9*, 299-308.

Bourhis, R. Y., & Leyens, J. P. (1999). *Stéréotypes, discrimination et relations intergroupes*. Sprimont, Belgium: Mardaga.

Brauer, M. (2005a). Préjugés et stéréotypes. In M. Borlandi, R. Boudon, M. Cherkaoui, & B. Valade (Eds.), *Dictionnaire de la Pensée Sociologique*. (pp. 571-572). Paris: Presses Universitaires de France.

Brauer, M. (2005b). Stéréotypes et rapports de domination. *Cerveau & Psycho*, *13*, 24-28.

Brauer, M., & Er-rafiy, A. (2009). *Modifying perceived variability in order to improve intergroup relations: A new way to reduce prejudice and discrimination?* Manuscript submitted for publication.

Brauer, M., & Judd, C. M. (2007). *L'ajustement par régression: Une méthode plus puissante pour analyser des plans expérimentaux avec prétest et posttest.* Unpublished manuscript. Clermont Université.

Brauer, M., Martinot, D., & Ginet, M. (2004). Current tendencies and future challenges for social psychologists. *Current Psychology of Cognition, 22*, 537-558.

Brauer, M., Wasel, W., & Niedenthal, P. M. (2000). Implicit and explicit components of prejudice. *Review of General Psychology, 4,* 79-101.

Brewer, M. B., & Miller, N. (1984). Beyond the contact hypothesis: Theoretical perspectives on desegregation. In N. Miller & M. Brewer (Eds.), *Groups in contact: The psychology of desegregation*. New York: Academic Press.

Brigham, J. C. (1971). Ethnic stereotypes. *Psychological Bulletin, 76*, 15-38.

Broman, C. L. (1997). Race-related factors and life satisfaction among African Americans. *Journal of Black Psychology, 23*, 36-49.

Brown, R. (1995). *Prejudice: Its social psychology*. Oxford: Blackwell.

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of black and white discrimination and prejudice: A literature review. *Psychological Bulletin, 87*, 546-563.

Dambrun, M., & Guimond, S. (2001). La théorie de la privation relative et l'hostilité envers les Nords-Africains. *Revue Internationale de Psychologie Sociale, 14*, 57-89.

Dambrun, M., & Guimond, S. (2004). Implicit and explicit measures of prejudice and stereotyping: Do they assess the same underlying knowledge structure? *European Journal of Social Psychology, 34*, 663-676.

Dion, K. L. (1986). Responses to perceived discrimination and relative deprivation. In J. M. Olson, C. P. Herman, & M. P. Zanna (Eds.), *Relative deprivation and social comparison: The Ontario Symposium* (pp. 159-179). Hillsdale, NJ: Erlbaum.

Dobbs, M., & Crano, W. D. (2001). Out-group accountability in the minimal group paradigm: Implications for aversive discrimination and social identity theory. *Personality and Social Psychology Bulletin, 27*, 355-364.

Dovidio, J. F., Brigham, J. C., Johnson, B. T., & Gaertner, S. L. (1996). Stereotyping, prejudice and discrimination: Another look. In N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Foundations of stereotypes and stereotyping* (pp. 276-319). New York: Guilford.

Dovidio, J. F., & Gaertner, S. L. (1986). Prejudice, discrimination and racism: Historical trends and contemporary approaches. In J. E. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination and racism*. New York: Academic Press.

Er-rafiy, A., & Brauer, M. (2010). L'effet bénéfique de l'augmentation de la variabilité perçue sur le niveau de préjugés et la discrimination, *L'Année Psychologique*, *110*, 103-125.

Gaertner, S. L., & Dovidio, J. F. (1977). The subtlety of white racism, arousal and helping behavior. *Journal of Personality and Social Psychology, 35*, 691-707.

Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing intergroup bias: The common in-group identity model*. Philadelphia, PA: Psychology Press.

Gelman, A., & Hill, J. (2007*). Data analysis using regression and multilevel/hierarchical models*. Cambridge, NY: Cambridge University Press.

Gerber, A. S., Green, D. P., & Kaplan, E. H. (2004). The illusion of learning from observational research. In I. Shapiro, R. Smith, & T. Massoud (Eds.), *Problems and methods in the study of politics* (pp. 251-73). New York: Cambridge University Press.

Green, D. P., & Wong, J. S. (2001). *Tolerance and the contact hypothesis: A field experiment.* Unpublished manuscript, Yale University, CT.

Henry, P. J. (2008). Student sampling as a theoretical problem. *Psychological Inquiry, 19*, 114-125.

Hewstone, M., & Cairns, E. (2001). Social psychology and intergroup conflict. In D. Chirot & M. P. E. Seligman (Eds.), *Ethnopolitical warfare: Causes, consequences and possible solutions*. Washington, DC: American Psychological Association.

Hill, M. E., & Augustinos, M. (2001). Stereotype change and prejudice reduction: Short- and long-term evaluation of a cross-cultural awareness program. *Journal of Community and Applied Social Psychology, 11*, 243-262.

Hollister, L., Day, N. E., & Jesaitis, P. T. (1993). Diversity programs: Key to competitiveness or just another fad? *Organization Development Journal, 11*, 49-59.

Hovland, C. I., & Sears, R. R. (1940). Minor studies of aggression: Correlations of lynchings with economic indices. *Journal of Psychology, 9*, 301-310.

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Assessing mediation in treatment evaluations. *Evaluation Review, 5,* 602-619.

Judd, C. M., Park, B., Ryan, C. S., Brauer, M., & Kraus, S. L. (1995). Stereotypes and ethnocentrism: Interethnic perceptions of African American and White American college samples. *Journal of Personality and Social Psychology, 69*, 460-481.

Katz, P. A., & Zalk, S. R. (1978). Modification of children's racial attitudes. *Developmental Psychology, 14,* 447-461.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1987). *Statistical design for research*. Wiley, New York.

Landrine, H., & Klonoff, E. A. (1996). The schedule of racist events: A measure of racial discrimination and a study of its negative physical and mental health consequences. *Journal of Black Psychology, 22,* 144-168.

Leyens, J. P., Yzerbyt, V. Y., & Schadron, G. (1994). *Stereotypes and social cognition*. London: Sage.

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-125). San Diego: Academic Press.

Milgram, S., Mann, L., & Harter, S. (1965). The lost-letter technique. *Public Opinion Quarterly, 29*, 437-438.

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Hillsdale, NJ: Erlbaum.

Musca, S. C., Kamiejski, R., Er-rafiy, A., Méot, A., Brauer, M., & Nugier, A. (2010). *Data with Hierarchical structure: what one ought to know and do*. Manuscript submitted for publication.

Mutz, D., & Martin, P. (2001). Facilitating communication across lines of political difference: The role of mass media. *American Political Science Review, 95,* 97-114.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Oskamp, S. (2000). *Reducing prejudice and discrimination*. Mahwah, NJ: Lawrence Erlbaum Associates.

Paluck, E. L. (2006). Diversity training and intergroup contact: A call to action research. *Journal of Social Issues, 62*, 439-451.

Paluck, E. L. (in press). Is it better not to talk? Group polarization, extended contact, and perspective-taking in eastern Democratic Republic of Congo. *Personality and Social Psychology Bulletin*.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339-369.

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance- versus abstraction-based stereotypes. *Journal of Personality and Social Psychology, 53*, 621-635.

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology, 42*, 1051-1068.

Phelps, E. A., & Thomas, L. A. (2003). Race, behavior and the brain: The role of neuroimaging in understanding complex human behaviors. *Political Psychology, 24*, 747-758.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Rich, Y., Kedem, P., & Shlesinger, A. (1995). Enhancing intergroup relations among children: A field test of the Miller-Brewer model. *International Journal of Intercultural Relations, 19,* 539-553.

Rokeach, M. (1971). Long-range experimental modification of values, attitudes, and behavior. *American Psychologist, 26,* 453-459.

Ryan, C. S., Judd, C. M., & Park, B. (1996). Effects of racial stereotypes on judgments of individuals: The moderating role of perceived group variability. *Journal of Experimental Social Psychology, 32*, 71-103.

Salomon, G. (2004). *Does peace education make a difference in the context of an intractable conflict?* Unpublished manuscript. University of Haifa, Center for Research on Peace Education, Haifa, Israel.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Sherif, M., & Sherif, C. W. (1953). *Groups in harmony and tension: an integration of studies on intergroup relations*. New York: Harper.

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. New York: Cambridge University Press.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797-811.

Stephan, W. G., & Finlay, K. (1999). The role of empathy in improving intergroup relations. *Journal of Social Issues, 4*, 729-743.

Stephan, W. G., & Stephan, C. W. (2001). *Improving intergroup relations*. Thousand Oaks, CA: Sage Publications.

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology, 68,* 199 – 214.

Tajfel, H. (1978). *Differentiation between social groups.* London: Academic Press.

Weiner, M. J., & Wright, F. E. (1973). Effects of undergoing arbitrary discrimination upon subsequent attitude toward a minority group. *Journal of Applied Social Psychology, 3*, 94-102.

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology, 10*, 109-120.