European Journal of Social Psychology, Vol. 26, 961–980 (1996)

Probabilistic National Stereotypes

JOACHIM KRUEGER
Department of Psychology, Brown University, U.S.A.

Abstract

A stereotype of a group can be expressed by the estimated percentage of members that possess certain personality attributes (Brigham, 1971). In a multi-group design, the properties of the percentage measure were examined, and three common assumptions about stereotypes were supported. First, there was high consensus *among American and Italian raters regarding the attributes of Americans, Italians, English, and Germans. Second, the perceived typicality of a trait depended largely on* contrasts *with other traits attributed to the same target group. Contrasts between attributions of the same trait to different groups were largely irrelevant. Third, most stereotype judgments revealed* consistency *biases. Compared with a Bayesian model of probability estimation, raters exaggerated the similarities between trait attributions (the percentage measure), social categorizations (percentage of people that belong to a group given they possess the trait), and Likert-scaled typicality ratings. Raters underestimated the effects of the traits' global base rates on the typicality ratings.*

INTRODUCTION

Social stereotypes are beliefs about the attributes of groups of people. In conjunction with beliefs about the attributes of the self and individual others, social stereotypes lie at the heart of social cognition. Since Lippmann's (1922) theoretical analysis and Katz and Braly's (1933) empirical work on national, ethnic, and racial stereotypes,

Addressee for correspondence: Joachim Krueger, Department of Psychology, Box 1853, Brown University, Providence, RI 02912, U.S.A., phone: (401) 863-2503, fax: (401) 863-1300, email: Joachim_Krueger @Brown.Edu.

I am deeply indebted to Anne Maass for advice and assistance with data collection. Nora Paglionico, and Francesca Montalcini translated materials into Italian and collected the data at the Università di Padova. Russ Clement assisted in the collection of the American data. Hilary Broadbent, Tami Bryan, Russ Clement, Robyn Dawes, Kari Edwards, Bill Heindel, Oliver John, Jill Krueger, Myron Rothbart, and Rashmi Sinha provided valuable comments and suggestions on a draft of this manuscript.

CCC 0046–2772/96/060961–20

Received 15 September 1995
Accepted 16 January 1996

several assumptions about stereotypes have been prevalent. According to the assumption of *consensus*, stereotypes are shared among members of a social group (e.g. Devine, 1989; Tajfel, 1969). According to the assumption of *contrast*, the attribution of a particular trait to a particular group derives its stereotypic character in part from attributions of other traits to the same group or from attributions of the same trait to other groups (e.g. Ashmore & Del Boca, 1981; Zawadski, 1948). According to the assumption of *consistency*, stereotypes are biased in that they portray groups in overly simplistic and internally consistent ways (e.g. Judd, Ryan & Park, 1991; Krueger & Clement, 1994).

When Katz and Braly (1933) established the adjective checklist as the empirical paradigm for stereotype measurement, they simultaneously stimulated and constrained research on these assumptions. In the checklist paradigm, subjects nominate the traits they consider characteristic of a target group. For each group, traits are then ranked according to the percentage of raters who nominated them. The fewer traits account for 50 per cent of all choices, the more uniform the stereotype is considered to be.

The checklist method has found numerous applications and it continues to be used today, hence the stimulation. As a method, however, it guarantees the confirmation of the three assumptions, hence the constraint. Katz and Braly (1933), for example, found *consensus* in that Americans considered Italians to be passionate (37 per cent) and Germans to be efficient (16 per cent). Replications in subsequent generations of Princeton students yielded similar results, demonstrating the social or shared nature of the stereotypes (Karlins, Coffman & Walters, 1969). It is not clear, however, what minimum degree of consensus is required to make a stereotype social. Katz and Braly arbitrarily chose to report the 12 most frequently nominated traits for each group. Of the included traits, consensus ranged from 84 per cent (superstitious for Negroes) to 11 per cent (e.g. practical for Germans). The stereotype uniformity index, intended to express the consensuality of a group's entire trait profile, has limited use because it confounds the variability between raters on a given trait with the variability between traits.

Contrasts between traits and between groups emerge if there is any variability in the data. Thus, there are two ways to interpret the Princeton data. On the one hand, the trait passionate may be stereotypically Italian because fewer raters nominated the trait efficient for Italians. This contrast is between traits and within group. On the other hand, to be passionate may be stereotypically Italian because fewer raters nominated the trait passionate for Germans. This contrast is between groups and within trait. The checklist data do not reveal on which type of contrast raters focus. Hence, important psychological processes underlying stereotyping remain ambiguous.

The checklist format also ensures a certain degree of stereotype *consistency*. Stereotypes appear to be oversimplified because all-or-none responses ignore gradations of belief or perceived distributions of group characteristics. One rater may consider Germans efficient because all relevant contacts he or she has had supported this belief, whereas another rater may check the trait after prolonged deliberation weighting confirming and disconfirming instances derived from memory. Checklist responses cannot distinguish these discrepant levels of experience and meaning; yet, they invite the conclusion that both raters hold the belief that all Germans are efficient.

Probabilistic trait attributions

The present research operationalizes the assumptions of consensus, contrasts, and consistency as testable hypotheses. Brigham's (1971) probabilistic stereotype measure allows expressions of degrees of belief—or stereotype strength—on each trait. Each rater estimates the percentage of members of a certain target group who possess a certain trait. With such probabilistic trait attributions, the three assumptions, implicit in much of the early stereotype research can be examined empirically. Do raters belonging to the same group or to different groups significantly agree on the attributes of a target group (*consensus*)? Does the perceived stereotypicality of a trait depend on whether the trait is seen as more or less prevalent than other traits in the target group or does it depend on whether the same trait is seen as more or less prevalent in other target groups (*contrasts*)? Are stereotypes oversimplified in that intergroup differences are exaggerated or in that probabilistic beliefs about a target group are overly homogeneous (*consistency*)?

Consensus

Stereotypes are social to the extent that they are shared by members of the same or different groups (Gardner, 1973; Stangor & Schaller, 1996). If, for example, raters agree that there are more passionate than efficient Italians and that there are more efficient than passionate Germans, these stereotypes are consensual. For an individual trait, the social aspect of a stereotype can be represented statistically by the standard deviation of percentage estimates. Where multiple traits form a stereotype profile, the social aspect can be represented by the average interrater correlation across traits. Stereotypic images of a group may not only be shared by individual raters; they may also be shared across groups. Similarities between auto- and heterostereotypes are important because they may indicate accurate intergroup perceptions (Peabody, 1985). Intergroup consensus would be a most convincing cue for accuracy if raters from different groups agreed on the differences in the traits of multiple target groups. The first hypothesis of this research is that such a pattern of interrater and intergroup consensus will emerge. It should be noted, however, that consensus is an index of reliability, not validity. The accurate perception of group characteristics is only one factor that may create consensus; shared cultural myths is another.

Contrasts

The belief that 80 per cent of Italians are passionate has little meaning if the same percentage of Italians is believed to possess every other trait and if the trait passionate is considered to be equally prevalent in other groups. Which of the two contrasts, between traits or between groups, do raters primarly invoke when deciding how typical a trait is of a group? Perhaps because of the evident relevance of stereotypes for intergroup relations, theorists have favoured between-group

contrasts. Zawadski's (1948) thinking greatly influenced contemporary conceptions. He argued a trait could be stereotypic regardless of the perceived prevalence of a trait in a group. 'The popular conception of a group characteristic seems to be: *a characteristic which is present in the majority of the members of the group.* According to this concept, it is a necessary and sufficient condition for a group characteristic to be present in at least 51 per cent of the members of the group. A moment's reflection, however, suggests another concept. A group characteristic is a characteristic which makes possible a distinction between two groups. Such a distinction is possible, for instance, if the group A possesses the characteristic c in 40 per cent and the group B in 20 per cent' (pp. 135–136, emphasis in the original).

McCauley and Stitt (1978) suggested that their data supported Zawadski's (1948) hypothesis. They argued that traits are stereotypic to the extent that they distinguish a target group from the general population. As an index McCauley and Stitt suggested the *diagnostic ratio* (DR), which is the estimated percentage of group members who possess the trait (trait attribution) divided by the estimated percentage of people in the world who have the trait (base rates). Although simple trait attributions predicted typicality ratings as well as or better than the DR did (see also Biernat & Crandall, 1994; Jonas & Hewstone, 1986; Stephan, Ageyev, Stephan, Abalakina, Stefanenko & Coates-Shrider, 1993), McCauley and Stitt concluded that 'the diagnostic ratio provides a quantitative and individual measure of stereotyping that should enormously facilitate the understanding of stereotypes and their effects' (p. 935). In this spirit, they defined stereotypes exclusively in terms of the DR. 'According to the new definition, the stereotype of any group of people [...] is composed of those attributes for which within-group predictions differ from base-rate predictions' (p. 938).

Despite its intuitive appeal, the DR has serious limitations[1]. Psychologically, it is doubtful that people routinely perform mental division when thinking about the stereotypicality of traits. Statistically, the DR is liable to be redundant relative to simple trait attributions. A ratio tends to covary with its numerator. The higher the trait attribution is, the larger is the DR. In McCauley and Stitt's (1978) study, American raters' trait attributions to Germans were highly correlated with the DR ($r = 0.46$, Study 1, calculated from Table 1, p. 932). In a study of Americans' and Russians' perceptions of each other, this correlation was high as well (mean $r = 0.64$; Stephan *et al.*, 1993). Correlations of this magnitude are to be expected when trait attributions and base-rate estimates are independent. A Monte Carlo simulation with random and independently generated trait attributions and base rates produced a correlation between the DR and trait attributions that was similar to the empirical ones ($r = 0.51$)[2].

Because the DR is a composite and potentially redundant index of stereotyping, the question of its predictive power should be rephrased: Does the DR predict typicality ratings when the effects of probabilistic trait attributions have been controlled? In other words, the correlations between the DR and typicality ratings

[1]Suppose a person believes that 1 per cent of all people are manic, and that 2 per cent of people in group G are manic (DR = 2). For a trait with a base rate of 40 per cent, a group prediction of 80 per cent is needed to obtain the same ratio, and for a trait with a base rate of 60 per cent, it is impossible. The higher the base rate is, the more constricted becomes the range of the DR.

[2]The simulation consisted of two series, A, and B, of 100 casts of a die. A and B were independent ($r = 0.03$) and A was positively correlated with A/B ($p < 0.01$).

need to be computed while trait attributions are partialled out. If these partial correlations are near zero, variations in trait attributions within the target groups are sufficient to explain trait typicality. One can also ask whether trait attributions predict typicality ratings when the DR is partialled out. If they do, one can conclude that, contrary to Zawadski's hypothesis, raters rely primarily on contrasts between traits and within group when judging typicality.

Between-group diagnosticity can also be expressed by the difference between trait attributions and estimated base rates, or by the difference between trait attributions to an outgroup and trait attributions to the ingroup (Diehl & Jonas, 1991; Gardner, Lalonde, Nero & Young, 1988; Martin, 1987; McCauley, Stitt & Segal, 1980; McCauley & Thangavelu, 1991). The limitations of the DR also apply to these difference scores. Therefore, the second hypothesis in this research is that raters judge the trait typicality for a target group primarily through contrasts between traits rather than between groups. The DR and similar indices of intergroup differentiation will not increase the predictability of typicality ratings beyond the effects of probabilistic trait attributions.

It should be noted that the hypothesis that trait typicality primarily depends on contrasts between traits and within groups does not imply that raters do not differentiate between target groups. The opposite is true. The more a rater differentiates between two groups, the lower is the correlation between the trait attributions to the two groups. The more a rater differentiates the Italians from the Germans, for example, the less correlated are the trait attributions. It is important to note that low correlations between trait attributions to different target groups are especially likely to render indices of diagnosticity redundant.

Consistency

The notion of stereotype consistency is akin to Katz and Braly's (1933) notion of uniformity. A consistent stereotype depicts the target group as a well-defined entity, different from other groups and homogeneous within (Campbell, 1958). Three potential biases may contribute to exaggerations of stereotype consistency. The first of these is *differentiation bias.* As noted above, a stereotype about a particular group is differentiated if the perceived attributes of the group differ from the perceived characteristics of other groups or from the perceived characteristics of people in general. Differentiation yields low correlations between target groups and across trait attributions. On average, unbiased correlations of differentiation should be positive, however, because the traits of the world population are the aggregated traits of the constituent groups. For example, if there are more pleasure-seeking than knowledge-seeking people in the world, it is—*a priori*—likely that Americans, Italians, and other specific national groups choose pleasure over knowledge.

Differentiation bias occurs if the correlation between trait attributions and base-rate estimates falls below a standard provided by a normative model (Krueger & Clement, 1996). Bayes's rule provides a normative standard. Analyses require four sets of probability estimates from each rater. *Trait attributions* are the estimated

percentage of members of national group G believed to have trait T, $p(T|G)$[3]. *Social categorizations* are the estimated percentage of people with trait T believed to belong to group G, $p(G|T)$. Trait *base rates* are the estimated percentage of people in the world who have trait T, $p(T)$, and *group sizes* are the estimated percentage of people who belong to group G, $p(G)$. Bayes's rule is the identity of two probability ratios:

$$\frac{p(T|G)}{p(T)} = \frac{p(G|T)}{p(G)} <=> \frac{\text{Trait attribution}}{\text{Base rate}} = \frac{\text{Social categorization}}{\text{Group size}}$$

For example, the probability that an Italian is passionate, divided by the probability that any person is passionate, is equal to the probability that a passionate person is Italian, divided by the probability that a person is Italian. Each of the four probabilities can be derived from the other three. Trait attributions are the product of base rates and social categorizations divided by group size.

$$p(T|G) = \frac{p(T)p(G|T)}{p(G)}$$

Differentiation bias occurs if estimated base rates are less correlated with the rater's expressed trait attributions than with the Bayesian values of those trait attributions. In other words, differentiation bias in stereotyping is a special case of the base-rate fallacy in probability estimation (Kahneman & Tversky, 1972; Koehler, 1996).

The underuse of base rates is a well-documented outcome of representative thinking (Krueger, 1996; Nisbett & Ross, 1980). Reliance on the representativeness heuristic also entails a perceived interchangeability of inverse conditional probabilities. This bias is the second aspect of stereotype consistency, namely *simplification bias*. Stereotypes are oversimplified if trait attributions and social categorizations are more highly correlated than a normative Bayesian analysis would suggest. Raters may consider the question 'What is the percentage of Italians who are passionate?' to be equivalent to the question 'What is the percentage of all the passionate people who are Italians?' In Bayes's formula, trait attributions, $p(T|G)$, and social categorizations, $p(G|T)$, are identical only if the base rate of the trait, $p(T)$, and the size of the group, $p(G)$, are the same. If, however, base rates or group sizes differ, the conditionals differ. The simplification bias can be tested by comparing correlations between trait attributions and social categorizations with their normative standards, that is, the correlations between trait attributions and Bayesian values of the social categorizations (Dawes, Mirels, Gold & Donahue, 1993).

Thus far it has been hypothesized that, among other things, trait attributions will be highly correlated with typicality ratings (contrast hypothesis) and with social categorization (oversimplification hypothesis). It seems thus likely that social

[3]In Bayesian analysis, these estimates quantify probabilities, but in the current context, stereotype estimates are percentage estimates. The two are mathematically equivalent, and in this exposition, both notations will be used.

categorizations will also be highly correlated with typicality ratings. If, compared with a normative model, this correlation is too high, the stereotype implies a *categorization bias*. To illustrate this third aspect of stereotype consistency, suppose the trait 'passionate' is stereotypic of Italians and the trait 'efficient' is counterstereotypic. A passionate person may be more likely to be categorized as Italian than an efficient person. That is, typicality ratings predict categorizations across traits. Categorizations are biased toward stereotype consistency, however, if typicality ratings are more highly correlated with social categorizations than with the rater's own Bayesian values of those social categorizations (i.e. $p(G)Xp(T|G)/P(T)$). Since group size does not vary across traits, the normative standard is the correlation between typicality ratings and the DR (i.e. $p(T|G)/P(T)$). A biased rater includes more people with stereotypic traits in the target group than can be expected from the rater's own trait attributions and base rate estimates (overcategorization). By the same token, categorization bias involves the exclusion of too many people with counterstereotypic traits (undercategorization).

Overview and hypotheses

The properties of the percentage measure of stereotypes were assessed in two samples of American and Italian students. There were three sets of hypotheses:

(1) *Consensus*: national stereotypes are held with considerable intragroup and intergroup consensus. Judgments about a specific target group will be more highly intercorrelated than judgments about different target groups.
(2) *Contrasts*: probabilistic trait attributions are the best predictor of trait typicality ratings because raters rely on contrasts between traits rather than on contrasts between groups. The DR (or similar indices of diagnosticity) will not predict typicality when trait attributions are partialled out. The inverse will not be true.
(3) *Consistency*: compared with a Bayesian model, (A) base rates will be insufficiently correlated with trait attributions (*differentiation bias*), (B) trait attributions will be too highly correlated with social categorizations (*simplification bias*), and (C) typicality ratings will be too highly correlated with social categorizations (*categorization bias*).

METHOD

Raters and materials

The American sample consisted of 169 undergraduate students at Brown University who participated in partial fulfilment of a course requirement or in exchange for $5. After the removal of the data from foreign nationals, data of 140 subjects (59 per cent women) remained for analysis. The Italian sample consisted of 80 students (65 per cent women), enrolled at the Università di Padova, who were approached individually in their dormitories by one of two experimenters and asked to complete

a survey of national characteristics. Data of six students were omitted from analyses because they were incomplete.

Two native speakers of Italian with excellent knowledge of English and a native speaker of German fluent in both English and Italian translated the trait terms and the instructions into Italian. The 16 adjectives were listed alphabetically and the order of the four national groups varied randomly across subjects. In Karlins *et al.*'s (1969) checklist study, the traits artistic, impulsive, passionate, and talkative described the Italians; conservative, practical, reserved, and sophisticated described the English; aggressive, efficient, industrious, and scientifically-minded described the Germans; and ambitious, materialistic, and pleasure-loving described the Americans. These traits were used, and the trait 'individualistic' was added as a (likely) American characteristic[4].

Procedures

Part 1 of the questionnaire was presented as 'a cross-cultural survey in social perception'. The instructions for making *trait attributions*, $p(T/G)$, and *social categorizations*, $p(G|T)$, were as follows:

> This study examines perceptions of how certain personality attributes are distributed in various social groups. While it may rarely be meaningful to attribute a certain trait to all members of a target group (e.g. all Chinese are polite), it seems reasonable to think about the *percentage* of group members that can be described by a certain trait. On the following pages you will find a series of 16 traits. For each trait please make two kinds of judgments with respect to four national groups (American, Italian, English, German). First, what is the percentage of members of each of these groups who have the trait? Second, what is the percentage of the world's people with the trait who are members of the group? Notice that there are no definite correct answers. There are no reliable ways of assessing the 'true' distribution of personality traits in different countries with different cultures and languages. However, we are interested in the overall picture as you see it.

After completing part 1 of the survey, raters worked on unrelated tasks. About 20 minutes later, they received part 2. For each rater, the order of the national groups was the same as in part 1. Ratings of the *typicality* of each trait for each group were made on a scale from 1 (very uncharacteristic) to 9 (very characteristic). Finally, raters estimated the world *base rates*, $p(T)$, of each trait ('For each trait, what is the percentage of the world's population that you think may be described by it') and the *group sizes*, $p(G)$, of the four national groups (e.g. 'What percentage of the world's people are Italians?').

[4]The choice of the term 'individualistic' reflects primarily my own (German) experience with Americans. Interestingly, well-known studies on American self-stereotypes (Karlins *et al.*, 1969) or other nations' stereotypes about Americans (Peabody, 1985) do not list the trait individualistic. Similarly, sociological theory (Riesman, 1950) stresses other-directedness rather than individualism. Recent sociological work, however, delineates mythic, civic, utilitarian, and expressive individualism as a cultural pattern specific to American society (Bellah, Madsen, Sullivan, Swidler & Tipton, 1985).

RESULTS

Preliminary analyses

The mean trait attributions ratings in Table 1 show close correspondence with the national characterizations traditionally obtained with the adjective checklist (Karlins *et al.*, 1969). The trait 'materialistic' was seen as more prevalent than other traits among Americans and more prevalent among Americans than among other groups. Pleasure-loving was the outstanding trait among Italians, conservative among the English, and industrious among the Germans. The similarity of the trait profiles derived from ratings by American and Italian raters lends preliminary support to the consensus hypothesis.

Table 2 shows the mean intercorrelations (obtained from r–Z–r transformations) of the rating variables when averaged across individuals, their national group, and the target group. It is consistent with the contrast and the consistency hypotheses that trait attributions, social categorizations, and typicality ratings were highly intercorrelated, whereas base-rate estimates were only moderately correlated with any of the other ratings.

Consensus

The average pairwise correlations across traits (bottom of Table 1) illustrate the social or collective aspect of national stereotypes. As expected, there was considerable interrater consensus for both groups of raters and all four target groups (all $p<0.001$). There were also unexpected differences. American raters agreed more with one another than did Italian raters, but this difference was limited to trait attributions about the American and the Italian target group. This finding demonstrates that the degree of social consensus in stereotyping depends jointly on the stereotyping and the stereotyped group[5].

To assess intergroup consensus, each rater's trait attributions were correlated with the averages obtained from each group of raters. Table 3 shows the results of the analyses involving estimates about Americans and Italians. There was considerable intergroup consensus (mean $r^a=0.46$). These correlations show how closely ratings by one national group were associated with the averages of the ratings made by the other group. For example, the coefficient $r=0.61$ in the upper right quadrant of Table 3 indicates that trait attributions by individual American raters about Italians predicted the averages of the ratings Italians made about themselves[6].

[5]The average interrater correlation across traits is a measure of reliability, where raters represent scale items. With $n=140$ (American sample), each reliability index is the mean of 9730 interrater correlations ($n(n\text{-}1)/2$). Because of the large number of raters, indices of scale homogeneity (e.g. Cronbach's α) were high, and thus uninformative about differences between the consensuality of different types of stereotype judgments and different target nations. The average standard error of the mean interrater correlation was 0.03, showing that all reliability coefficients were substantially larger than zero. Interrater consensus in typicality ratings (mean $r=0.38$) was similar to the consensus in trait attributions (mean $r=0.35$). The reliability of social categorizations ($p(G|T)$) was lower (mean $r=0.19$).

[6]For the English and the German target groups, the average consensus between the American and the Italian raters was $r=0.47$.

Table 1. Mean trait attributions to the four national groups

Target group	American		Italian		English		German	
Rater group	American	Italian	American	Italian	American	Italian	American	Italian
1. Aggressive	55.34	47.76	56.47	38.09	34.70	44.73	53.05	56.35
2. Ambitious	58.84	60.12	46.35	47.93	47.99	42.46	55.54	47.32
3. Artistic	31.65	31.39	45.62	47.01	31.62	29.38	32.38	26.81
4. Conservative	52.53	44.61	40.16	47.69	64.43	62.77	55.43	43.31
5. Efficient	38.74	52.86	37.18	40.26	51.08	56.20	59.21	64.30
6. Impulsive	49.02	49.41	54.91	55.45	31.24	43.69	32.87	39.53
7. Individualistic	54.09	56.74	43.38	47.93	40.16	54.53	36.37	46.14
8. Industrious	47.54	53.59	41.79	49.50	48.21	47.49	60.33	56.42
9. Materialistic	73.48	62.30	50.59	50.88	54.20	49.46	49.27	52.74
10. Passionate	49.91	42.30	68.11	68.39	36.70	23.27	40.65	28.39
11. Pleasure-loving	73.27	59.05	72.70	72.08	55.29	35.58	54.20	39.51
12. Practical	46.82	61.51	42.50	53.35	59.12	54.47	61.74	56.62
13. Reserved	34.40	33.46	29.85	34.65	61.77	65.14	49.06	55.92
14. Scientifically-minded	34.48	47.85	27.50	32.11	39.74	42.97	49.69	49.82
15. Sophisticated	31.93	30.60	41.19	38.78	52.45	39.31	39.44	28.61
16. Talkative	56.91	57.53	64.60	62.93	46.55	31.97	42.79	29.61
Interrater consensus								
Mean pairwise *r*	0.48	0.24	0.47	0.31	0.35	0.33	0.28	0.31

Table 2. Average correlations among the rating variables

	Trait attributions	Social categorizations	Base rates
Typicality	0.68	0.62	0.25
Trait attributions		0.73	0.37
Social categorizations			0.18

The correlations in the main diagonal represent the degree of consensus within groups (mean $r^b = 0.66$). On the average, a rater's attributions were highly predictive of the average attributions made by his or her group. Not surprisingly, these part–whole correlations were larger than the between-group consensus correlations. Attributions by one group about the two target groups were less correlated, indicating a high degree of intergroup differentiation (mean $r^c = 0.22$). Interestingly, intergroup differentiation was greater among the Italian raters (mean $r = 0$) than among the American raters (mean $r = 0.42$). Finally, correlations between trait attributions made by different groups of raters about different target groups were expectably low (mean $r^d = 0.08$). In sum, there were high levels of intra- and intergroup consensus in trait attributions, and raters of both national groups differentiated between the target groups.

Contrasts

It was hypothesized that raters would primarily rely on contrasts between traits and within groups when rating trait typicality. There was reason to be sceptical of the possibility that contrasts between groups best predict trait typicality. To fully examine the possible contribution of between-group contrasts, three indices of diagnosticity were computed: the DR was the ratio of trait attributions divided by

Table 3. Consensus: mean correlations between trait attributions and group averages

Rater group	Americans		Italians	
Target group	Americans	Italians	Americans	Italians
Individual raters				
Americans				
Judging Americans	0.70[b]	0.41[c]	0.40[a]	0.21[d]
Judging Italians	0.43[c]	0.77[b]	−0.05[d]	0.61[a]
Italians				
Judging Americans	0.30[a]	−0.04[d]	0.52[b]	−0.003[c]
Judging Italians	0.18[d]	0.48[a]	0.002[c]	0.59[b]

[a]Different rater group, same target group. [b]Same rater group, same target group. [c]Same rater group, different target group. [d]Different rater group, different target group.

Table 4. Contrasts: mean correlations between typicality ratings, trait attributions, and three indices of diagnosticity

	Target group			
	Americans	Italians	English	Germans
American raters				
Simple correlations				
Typicality with				
trait attributions $p(T/G)$	0.76	0.74	0.68	0.65
DR	0.26	0.41	0.37	0.33
DS	0.41	0.50	0.44	0.43
DSI	–	0.40	0.45	0.46
Partial correlations				
Typicality with				
DR by $p(T/G)$	−0.10	0.04	0.04	−0.01
DS by $p(T/G)$	−0.11	0.02	0.01	0.01
DSI by $p(T/G)$	–	0.01	0.10	0.08
Italian raters				
Simple correlations				
Typicality with				
Trait attributions $p(T/G)$	0.62	0.64	0.65	0.67
DR	0.25	0.35	0.37	0.44
DS	0.32	0.43	0.46	0.50
DSI	0.40	–	0.59	0.59
Partial correlations				
Typicality with				
DR by $p(T/G)$	−0.10	0.00	−0.06	0.00
DS by $p(T/G)$	−0.14	−0.02	−0.03	0.01
DSI by $p(T/G)$	0.03	–	0.22	0.16

base-rates estimates ($p(T|G)/p(T)$). The DS was the difference between trait attributions and base-rate estimates ($p(T|G)-p(T)$). The DSI was the difference between trait attributions for an outgroup and trait attributions for the ingroup ($p(T|G)-p(T|\text{in-group})$).

The mean correlations shown in Table 4 support the hypothesis that raters focus on contrasts between traits rather than between groups when judging how typical a trait is of a group. Typicality ratings were more highly

Table 5. Differentiation bias: mean correlations between base-rate estimates and actual and Bayesian trait attributions

	Target group			
	Americans	Italians	English	Germans
American raters				
Base-rate estimates with				
Estimated $p(T|G)$	0.60	0.41	0.35	0.35
Bayesian $p(T|G)$	0.77	0.72	0.71	0.71
Italian raters				
Base-rate estimates with				
Estimated $p(T|G)$	0.43	0.35	0.21	0.20
Bayesian $p(T|G)$	0.71	0.64	0.59	0.62

Table 6. Categorization bias: mean correlations between typicality ratings and actual and Bayesian social categorizations

	Target groups			
	Americans	Italians	English	Germans
American raters				
Typicality with				
Estimated $p(G\|T)$	0.66	0.71	0.62	0.66
Bayesian $p(G\|T)$	0.26	0.41	0.36	0.31
Estimated $p(G\|T)$ − Bayesian $p(G\|T)$	0.37	0.40	0.25	0.23
Estimated $p(G\|T)$ with				
Bayesian $p(G\|T)$	0.29	0.45	0.44	0.47
Italian raters				
Typicality with				
Estimated $p(G\|T)$	0.55	0.59	0.57	0.60
Bayesian $p(G\|T)$	0.25	0.35	0.37	0.44
Estimated $p(G\|T)$ − Bayesian $p(G\|T)$	0.10	0.35	0.22	0.29
Estimated $p(G\|T)$ with				
Bayesian $p(G\|T)$	0.41	0.35	0.48	0.53

Note. Underlined correlations indicate categorization bias. With the standard errors of the mean Z scores being about 0.04, even the smallest correlation (mean $r = 0.10$ for Italian subjects and the American target group), was larger than zero, $t(73) = 3.5$, $p < 0.002$, two-tailed.

correlated with trait attributions than with any index of diagnosticity. Also as expected, all indices of diagnosticity were redundant because they were correlated with trait attributions (American data: $r_{\text{averaged across groups}} = 0.54$, 0.69, and 0.60, for DR, DS, and DSI, respectively; Italian data: $r_{\text{averaged across groups}} = 0.62$, 0.78, and 0.74, respectively). Because of this redundancy the partial correlations between the indices of diagnosticity and typicality ratings were near zero when trait attributions were controlled. Twenty-one of the 22 partial correlations were smaller than 0.12, and one was 0.22 (i.e. the correlation between the DSI and typicality ratings of the English target group in the Italian sample). In contrast, when the indices of diagnosticity were partialled out, the correlations between trait attributions and typicality ratings were *not* appreciably reduced (American data: partial $r_{\text{averaged across groups}} = 0.66$, 0.61, and 0.60, for DR, DS, and DSI, respectively; Italian data: $r_{\text{averaged across groups}} = 0.52$, 0.37, and 0.44, respectively).

Consistency

According to the hypothesis of *differentiation bias*, trait attributions would not be similar enough to raters' own base-rate estimates of trait prevalence. In other words, the correlations between trait attributions and base-rate estimates were expected to be too low relative to a Bayesian norm. This hypothesis was supported because in the American sample estimates of $p(T)$ were less highly correlated with estimates of $p(T|G)$ (mean $r = 0.43$) than with the Bayesian values of $p(T|G)$ (mean $r = 0.73$),

$F(1,132) = 144.2$, $p < 0.001$ (see Table 5, top). Similar results emerged in the Italian sample. The correlations between raters' estimates were smaller (mean $r = 0.30$) than the raters' own normative correlations (mean $r = 0.64$), $F(1,68) = 144.7$, $p < 0.001$ (see Table 5, bottom). This differentiation bias was reliable for all target groups (all $p < 0.001$). In other words, raters consistently underweighted their base-rate estimates in making trait attributions. But did they ignore base rates entirely? If they did, the correlations between estimated base rates and trait attributions should be eliminated when the ratios of social categorizations over group size are partialled out. The partial correlations in the American sample were positive, however, (mean partial $r = 0.50$, 0.36, 0.30, and 0.32, for the American, Italian, English, and German target groups, respectively), indicating that base-rate estimates played a significant, albeit insufficient, role in trait attributions. Results in the Italian sample were similar (mean partial $r = 0.34$, 0.26, 0.20, and 0.14, for the American, Italian, English, and German target groups, respectively).

According to the hypothesis of *simplification bias*, trait attributions and social categorizations would be more closely related than a Bayesian analysis would suggest. Indeed, the correlations between estimates of $p(T|G)$ and $p(G|T)$ were larger (mean $r = 0.72$ and 0.74 for the American and the Italian sample, respectively) than the correlations between $p(T|G)$ and the Bayesian values of $p(G|T)$ (Americans: mean $r = 0.54$, $F(1,119) = 40.6$; Italians: mean $r = 0.62$, $F(1,70) = 17.14$, both $p < 0.001$). This overuse of social categorizations in trait attributions is consistent with the underuse of base-rates estimates.

The hypothesis of *categorization bias* stated that estimates of social categorization would be insufficiently regressive[7]. Mean correlations are presented in Table 6. As expected, typicality ratings were more closely related to estimated social categorizations than to the Bayesian values of social categorizations (Americans: $F(1, 121) = 255.1$; Italians: $F(1, 70) = 75.2$, both $p < 0.001$). One way of looking at this bias is to examine the correlations between typicality ratings and the *differences* between estimated and Bayesian social categorizations (see Table 6). The results indicated systematic over- and undercategorization depending on the typicality of the trait. For stereotypic traits, raters overcategorized. They estimated the percentage of people belonging to the national group as higher than their own probabilistic beliefs would allow. For counterstereotypic traits, they undercategorized. Although the correlations between estimated social categorizations and their Bayesian values were reliable (bottom rows in Table 6), they were substantially lower than the correlations between typicality ratings and estimated social categorizations (Americans: $F(1, 120) = 125.8$; Italians: $F(1, 70) = 20.0$, both $p < 0.001$). The perceived typicality of

[7]Mean-level analyses showed that without trait information, American raters believed that 12.67, 6.95, 6.70, and 7.47 per cent of the world population were American, Italian, English, and German, respectively. The Italian data were 29.97, 10.05, 12.78, and 11.80 per cent. The actual percentage are approximately 4.5, 1.0, 0.9, and 1.4, respectively (calculated from data in the *World Almanac 1993*). For the American raters, the means of the social categorization estimates were 16.67, 11.59, 11.28, and 11.37 per cent when averaged across traits. None of these probabilities differed from their Bayesian values (i.e. $p(G|T) = p(G)*p(T|G)/p(T)$), which were 17.84, 9.56, 9.98, and 10.40 per cent (all $p > 0.10$). Similarly, for the Italian raters, the means of the social categorization estimates were 19.61, 16.60, 14.49, and 16.54 per cent. Of these, only the mean social categorization of Americans differed reliably from its Bayesian value (45.63 per cent, $t(72) = 6.2$, $p < 0.001$). The other estimates were 15.35, 16.18, and 17.25 per cent (for Italians, English, and Germans, respectively, all $p > 0.07$). That is, on the average, there was no systematic over- or undercategorization when trait information was present.

traits was a better predictor of social categorizations than were the Bayesian values of these social categorizations.

DISCUSSION

Three assumptions, long held about social stereotypes, found strong empirical support. The social aspect of stereotypes emerged in the considerable *consensus* in estimates of group attributes, both within each of the two groups of raters and between these groups. Both groups of raters also differentiated between target groups. The consensual validation of stereotypes by members of different groups is a necessary condition for stereotype accuracy. Without it, the perceptions of at least one group of raters are incorrect. It should be noted, however, that high intergroup consensus is not a sufficient condition for accuracy. Even if raters agree on the attributes of a target group, they may express internationally shared misconceptions rather than veridical perceptions. Whereas the present findings suggest that the perceptions of Americans and Italians about each other and about the English and the Germans are, at least in part, accurate, the data in no way imply that stereotypes are always accurate. As functional approaches to stereotyping emphasize, the dynamics of intergroup relations may shape and even distort perceptions (Snyder & Miene, 1994). Competition is a powerful cause of assertive intergroup behaviour. Such behaviour may be interpreted and misinterpreted as aggressive and even hostile (Rothbart & Hallmark, 1988). Intergroup relations among the groups in the present study were marked by peaceful, if distant, co-existence. There were no salient conflicts of interest between Americans and Italians, and under these circumstances, the idea that the stereotypes were largely accurate has some credence.

The percentage measure of perceived group attributes was closely related to typicality ratings. When raters decide how typical a trait is of a group they appear to focus primarily on *contrasts* with other traits within the target group rather than contrasts with other groups for that trait. 'The popular conception of group characteristics', challenged by Zawadski (1948, p. 135) seems to capture the process of stereotyping better than the more academic notion of diagnosticity. The three indices of diagnosticity (the DR, DS, and DSI) predicted typicality ratings less well. Biernat and Crandall (1994) reported similarly disappointing results. In their study, the DR also failed to predict contact with and liking for stereotyped groups. What is most important is that the correlations between trait diagnosticity and typicality were spurious, by-products of the correlations between diagnosticity and trait attributions. Hence, they disappeared when trait attributions were statistically controlled[8].

Perhaps the idea of trait diagnosticity gained momentum in stereotype measurement because it is intuitively appealing in single-trait examples. It is easy to imagine how an infrequent trait may be stereotypic if its base rate or its prevalence in other groups is even lower. Similarly, it is easy to imagine how a frequent trait

[8]To assess the incremental predictive power of a composite index, one can ask if the index yields reliable correlations with the criterion variable when all of its individual components are partialled out (Evans, 1991; Stephan, Ageyev, Coates-Shrider, Stephan & Abalakina, 1994). In this study, base-rate estimates (or ingroup attributions) were not partialled out because simply partialling out trait attributions was sufficient to eliminate the contributions of the indices of diagnosticity.

may be counterstereotypic if it is more frequent in a comparison group. Like McCauley and Stitt (1978), Judd and Park (1993) suggested that trait diagnosticity is a necessary and sufficient condition for the stereotypicality of a trait. The stereotypic or counterstereotypic 'attribute is seen as *more or less prevalent in this group relative to other comparison groups, and the stereotype, therefore describes the differences among the groups*' (p. 110, emphasis added). Judd and Park illustrated this definition with an example and added the criterion of intergroup consensus.

> If both Black and White subjects rate Blacks as more athletic than Whites, then this attribute is used as stereotypic of Blacks and counterstereotypic of Whites [...] It may also be that the attribute is seen as relatively prevalent in both groups (e.g. greater than 50 per cent of both target groups are seen as athletic), but again this does not matter. As long as estimates from both subject groups are greater for one target group than for the other, the attribute can be classified for purposes of the analysis as stereotypic of that one target group and counterstereotypic of the other (p. 115).

Two consequences of this definition are noteworthy. First, *any* intergroup difference is evidence for stereotypy in one group and counterstereotypy in the other as long as the two groups agree. If they do not agree, neither is stereotyping. According to this definition, the Nazis did not stereotype the Jews because the Jews did not agree with the way they were described. Second, the link between trait prevalence and stereotypicality is defined away. Defining stereotypes exclusively in terms of intergroup diagnosticity implies an untenable degree of relativism. Whether a trait is stereotypic or counterstereotypic would vary with the group that is invoked for comparison. Consider Judd and Park's (1993) hypothetical example of athleticism attributed to blacks and whites. If both blacks and whites believe that 80 per cent of blacks and 70 per cent of whites are athletic, athleticism is counterstereotypic of whites. Put a third group, Asians, on the scene. If whites and Asians believe that 70 per cent of whites and 60 per cent of Asians are athletic, athleticism is stereotypic of whites. That is, a trait with a fixed perceived prevalence in a group can be stereotypic or counterstereotypic solely as a function of shifting comparison groups. The more groups there are, the more likely is the illogical conclusion that any attribute is both stereotypic and counterstereotypic of any group.

The percentage measure of trait attributions avoids the contradictions of the diagnosticity indices. In this study, traits considered typical of a group were attributed to a majority in the group; traits considered atypical were attributed to a minority. If one accepts traits with average typicality ratings of greater than 6 (on a 9-point scale) as stereotypic of a group, 39 of the 128 traits (16 traits × two groups of raters × four target groups) met this criterion. Only two of these (artistic for Italians as rated by Americans ($M = 6.25$) and by Italians ($M = 6.53$)) were attributed to a minority of Italians (46 per cent and 47 per cent as rated by Americans and Italians, respectively), while the estimated base rates were even smaller (28 per cent and 31 per cent by American and Italian raters, respectively). At the same time, there were two traits with this pattern of percentage estimates that had mean typicality ratings of less than 4 (Italians' ratings of sophisticated for Americans ($M = 3.79$) and Germans ($M = 3.86$); trait attributions = 31 per cent and 29 per cent, respectively, and an

estimated base rate of 28 per cent). If traits with a mean typicality rating of less than 4 are considered counterstereotypic, 10 traits met this criterion. None of these was attributed to a group majority.

To abandon the DR is not to relinquish the concept of intergroup differentiation. In fact, the DR was redundant *because* raters differentiated between the four target groups. The relatively low correlations between trait attributions to different groups showed that raters perceived the four nations as being different (see Table 3). It is also likely that ratings exaggerated actual between-groups differences. Experimental research on contrast effects has shown that people exaggerate mean-level differences between groups (Krueger & Rothbart, 1990). For example, people may estimate that 30 per cent of people in their national ingroup are reserved. If this trait seems typical of a salient outgroup, they may push up their estimates for that group to 60 per cent (Diab, 1963a,b). If this process is consistent across traits, typicality ratings correlate with trait attributions and with the DR. When trait attributions are partialled out, the DR will again fail to contribute independently to the prediction of trait stereotypicality.

Three biases, predicted on the basis of the representativeness heuristic, supported the idea that people hold overly *consistent* national stereotypes. Raters differentiated target groups too much from their own base-rate estimates of trait prevalence (differentiation bias). Similarly, they overestimated the covariation between trait attributions and social categorizations across traits (simplification bias). As Gigerenzer and Murray (1987) pointed out, these two misestimations are each other's mirror image. The underuse of base-rate information has received greater attention in the social psychological literature, whereas the 'confusion of the inverse' (Kardes, 1994) is better known in the literature of decision-making and especially clinical judgment. Dawes (1988) argued that representative thinking creates a symmetry in thought that does not exist in reality. Medical professionals, for example, may erroneously equate the probability that a patient who has disease X displays symptom Y with the probability that a patient who displays symptom Y has disease X. Sherman, McMullen, and Gavanski (1992) attributed the confusion of the inverse to differences in the accessibility of appropriate sample spaces. Physicians may categorize patients primarily according to diseases rather than symptoms. After categorization, they assess whether the patient displayed the symptom in question. Categorization by symptoms may be comparatively difficult (e.g. individual symptoms rarely qualify as medical categories), and the accurate assessment of the conditional probability of the disease may suffer. In social stereotyping, national groups may be 'natural categories' (Rothbart & Taylor, 1992). The group label provides a familiar sample space on which an observer can condition a probabilistic trait attribution, $p(T/G)$. The reverse does not hold. The people of the world who have a certain trait do not constitute a natural sample space on which subjects can condition a probabilistic social categorization, $p(G/T)$. To make social categorizations, raters may resort to the simple but fallible assumption that the inverse conditionals are highly similar.

The confusion of the inverse implied that judgments of social categorization were insufficiently regressive (categorization bias). The correlations between social categorizations and typicality ratings were higher than their Bayesian standards. This finding has important implications for stereotyping, which 'relies upon the execution of some rather rudimentary skills: most notably, the ability to assign people to meaningful social categories' (Macrae, Milne & Bodenhausen, 1994, p. 37).

Given a stereotypic trait, raters overestimated the probability that a person belonged to the stereotyped group. Biased social categorization is likely when there are no salient categorical features defining a person's group membership. For example, the categorization of a person as 'Jewish' or 'gay' is less reliable than the categorization of a person as black or white, male or female (Berger, Hank, Rauzi & Simkins, 1987; Elliott & Wittenberg, 1955). Suppose an observer uses attributes of uncertain cue validity (e.g. a prominent nose or a limp wrist) to categorize someone as Jewish or gay. The present findings suggest that the categorization may not only be incorrect, but even inconsistent with the observer's own probabilistic beliefs. If the observer holds anti-Semitic or homophobic attitudes, pejorative inferences about other attributes or behaviours may follow. Moreover, an erroneous categorization can linger if, as often happens, the observer does not bother to verify it. Categorization bias implies that stereotyping works backwards and thus perpetuates itself. Once a trait has been accepted as stereotypic (i.e. frequent in the target group), the presence of the trait serves to categorize the person when there is no direct knowledge of group membership.

REFERENCES

Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1–36). Hillsdale, NJ: Erlbaum.

Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A., & Tipton, S. M. (1985). *Habits of the heart*. Berkeley, CA: University of California Press.

Berger, G., Hank, L., Rauzi, T., & Simkins, L. (1987). Detection of sexual orientation by heterosexuals and homosexuals. *Journal of Homosexuality*, **13**, 83–100.

Biernat, M., & Crandall, C. S. (1994). Stereotyping and contact with social groups: Measurement and conceptual issues. *European Journal of Social Psychology*, **24**, 659–678.

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

Campbell, D. T. (1958). Common fate, similarity, and other indices of status of aggregates of persons as social entities. *Behavioral Science*, **3**, 14–25.

Dawes, R. M. (1988). *Rational choice in an uncertain world*. New York: Harcourt Brace Janovich Publishers.

Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E.(1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, **4**, 396–400.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, **56**, 5–18.

Diab, L. N. (1963a). Factors affecting studies of national stereotypes. *Journal of Social Psychology*, **59**, 29–40.

Diab, L. N. (1963b). Factors determining group stereotypes. *Journal of Social Psychology*, **61**, 3–10.

Diehl, M., & Jonas, K. (1991). Measures of national stereotypes as predictors of the latencies of inductive versus deductive stereotypic judgements. *European Journal of Social Psychology*, **21**, 317–330.

Elliott, D. N., & Wittenberg, B. H. (1955). Accuracy of identification of Jewish and Non-Jewish photographs. *Journal of Abnormal and Social Psychology*, **51**, 339–341.

Evans, M. G. (1991). The problem of analyzing multiplicative composites. *American Psychologist*, **46**, 6–15.

Gardner, R. C. (1973). Ethnic stereotypes: The traditional approach, a new look. *Canadian Psychologist*, **14**, 133–148.

Gardner, R. C., Lalonde, R. N., Nero, A. M., & Young, M. Y. (1988). Ethnic stereotypes: Implications of measurement strategy. *Social Cognition*, **6**, 40–60.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition* as *intuitive statistics*. Hillsdale, NJ: Erlbaum.
Jonas, K., & Hewstone, M. (1986). The assessment of national stereotypes: A methodological study. *Journal of Social Psychology*, **126**, 745–754.
Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, **100**, 109–128.
Judd, C. M., Ryan, C.S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology*, **61**, 366–379.
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430–454.
Kardes, F. R. (1994). Consumer judgment and decision processes. In R. S. Wyer & T. K. Srull (Eds), *Handbook of social cognition* (Vol. 2, pp. 399–466). Hillsdale, NJ: Erlbaum.
Karlins, M., Coffman, T. L., & Walters, G.(1969). On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology*, **13**, 1–16.
Katz, D., & Braly, K. (1933). Racial stereotypes on one hundred college students. *Journal of Abnormal and Social Psychology*, **28**, 280–290.
Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences*, **19**, 1–53.
Krueger, J. (1996). Studying the use of base rates: Normal science or shifting paradigm? *Behavioral and Brain Sciences*, **19**, 30.
Krueger, J., & Clement, R. W. (1994). Memory-based judgments about multiple categories: A revision and extension of Tajfel's accentuation theory. *Journal of Personality and Social Psychology*, **67**, 35–47.
Krueger, J., & Clement, R. W. (1996). Inferring category characteristics from sample characteristics: Inductive reasoning and social projection. *Journal of Experimental Psychology: General*, **125**, 52–68.
Krueger, J., & Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology*, **59**, 651–663.
Lippmann, W. (1922). *Public opinion*. New York: MacMillan.
Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, **66**, 37–47.
Martin, C. L. (1987). A ratio measure of sex stereotyping. *Journal of Personality and Social Psychology*, **52**, 489–499.
McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology*, **36**, 429–440.
McCauley, C., Stitt, C. L., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, **87**, 195–208.
McCauley, C., & Thangavelu, K. (1991). Individual differences in sex stereotyping of occupations and personality traits. *Social Psychology Quarterly*, **54**, 267–279.
Nisbett, R., & Ross, L. (1980). *Human inference*. Englewood Cliffs, NJ: Prentice-Hall.
Peabody, D. (1985). *National characteristics*. Cambridge, MA: Cambridge University Press.
Riesman, D. (1950). *The lonely crowd*. New Haven, CT: Yale University Press.
Rothbart, M., & Hallmark, W. (1988). In-group-out-group differences in the perceived efficacy of coercion and conciliation in resolving social conflict. *Journal of Personality and Social Psychology*, **55**, 248–257.
Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds? In G. R. Semin & K. Fiedler (Eds), *Language, interaction, and social cognition* (pp. 11–36). London: Sage.
Sherman, S. J., McMullen, M. N., & Gavanski, I. (1992). Natural sample spaces and the inversion of conditional judgments. *Journal of Experimental Social Psychology*, **28**, 401–421.
Snyder, M., & Miene, P. (1994). On the functions of stereotypes and prejudice. In M. P.Zanna, & J. M. Olson (Eds), *The psychology of prejudice: The Ontario symposium* (Vol. 7, pp. 33–54). Hillsdale, NJ: Erlbaum.
Stangor, C., & Schaller, M. (1996). Stereotypes as individual and collective representations. In N. Macrae, C. Stangor, & M. Hewstone (Eds), *Foundations of stereotypes and stereotyping* (pp. 3–37). New York: Guilford Press.

Stephan, W. G., Ageyev, V., Coates-Schrider, L., Stephan, C. W., & Alabakina, M. (1994). On the relationship between stereotypes and prejudice: An international study. *Personality and Social Psychology Bulletin*, **20**, 277–284.

Stephan, W. G., Ageyev, V., Stephan, C. W., Abalakina, M., Stefanenko, T., & Coates-Shrider, L. (1993). Measuring stereotypes: A comparison of methods using Russian and American samples. *Social Psychology Quarterly*, **56**, 54–64.

World Almanac 1993. New York: Pharos, 1992.

Zawadski, B. (1948). Limitations of the scapegoat theory of prejudice. *Journal of Abnormal and Social Psychology*, **43**, 127–141.