European Journal of Social Psychology
Eur. J. Soc. Psychol. **32**, 223–245 (2002)
DOI:10.1002/ejsp.68

Perceiving minority members as individuals: the effects of relative group size in social categorization

KARL CHRISTOPH KLAUER*, INGO WEGENER, AND KATJA EHRENBERG
Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

Abstract

According to the account by comparative distinctiveness, minorities draw attention by virtue of their relative size, leading to more individuation and more stereotyping of their members. Using the 'Who said what?' paradigm by Taylor, Fiske, Etcoff, and Ruderman (1978) in Klauer and Wegener's (1998) modified version, relative group size of gender categories and age categories was varied in a pilot study and a main experiment, respectively. In the pilot study, memory for discussion statements and in both studies, memory for individuating information increased as subgroup size decreased. Rating measures obtained in the main experiment revealed most stereotyping of minority members. The findings thereby support major predictions of the account by comparative distinctiveness, but demonstrate dissociations between different modes of category-based processing, i.e. category memory, reconstructive category use, and stereotyping. Copyright © 2002 John Wiley & Sons, Ltd.

INTRODUCTION

In current models of impression formation (e.g. Brewer & Harasty Feinstein, 1999; Fiske, Lin, & Neuberg, 1999), a distinction is drawn between stereotyping or category-based impression formation on the one hand and individuation or person-based impression formation on the other hand. Stereotyping relies on the perceived person's social group memberships (e.g. on the target's race or gender) and the attributes stereotypically associated with such groups. Individuation is guided by the target's personal and unique constellation of attributes. Person-based and category-based processing are distinguished as qualitatively different processing modes, and a host of cognitive, motivational, and situational factors is assumed to determine their relative impacts. An important situational factor is relative group size.

*Correspondence to: Karl Christoph Klauer, Psychologisches Institut, Rheinische Friedrich-Wilhelms-Universität Bonn, Römerstr. 164, 53117 Bonn, Germany. E-mail: christoph.klauer@uni-bonn.de

Received 16 October 2000
Accepted 16 May 2001

Relative group size affects a variety of processes at the interface between social cognition and group processes. Mullen (1991) presents a meta-analytic review of a large portion of this research, in which he shows that consensus in a given group is increasingly overestimated, that the magnitude of ingroup bias increases and that the group is perceived as increasingly more homogeneous, as its relative size decreases.

A prominent account of the effects of relative group size is based on what has later been termed *comparative distinctiveness*. Taylor (1981) argues that 'any person, whether male or female or black or white, is distinctive in inverse proportion to the number of other members of his or her race or sex present in the social context' (p. 94). In this account, small relative group size leads to distinctiveness of the members of the small group and thereby to the allocation of a disproportionate amount of attention to the distinct persons. According to Taylor and coworkers, this has two consequences:

(a) More information of any kind is taken in about distinctive persons, enabling the perceiver to make 'more discriminations within a subgroup, the fewer the members of the subgroup are' (Taylor *et al.*, 1978, p. 780).
(b) The distinct persons' category membership and associated expectancies will furthermore be activated more strongly, leading to increased stereotyping of these persons. Minority status should thus simultaneously increase the extent of person-based and category-based processing.

In a recent theoretical analysis, Brewer and Harasty Feinstein (1999) called impression formation in which an individuated impression is developed under a category-based processing mode *category-based individuation*.

Alternative theoretical accounts of the role of relative group size have been proposed by Mullen (1991) and Oakes and Turner (1986; cf. Oakes, 1987, 1994; Oakes, Haslam, & Turner, 1994). Mullen's model is also based on distinctiveness, but differs from the account by comparative distinctiveness with respect to the predictions for person-based processing. Mullen postulates that minorities draw attention by virtue of their relative size. As a consequence, information about minority groups and their members is argued to be processed in terms of prototype representations; information about majority groups and their members in terms of exemplar representations. Considering the processing characteristics, the distinction between prototype and exemplar representation as conceived by Mullen (1991) roughly corresponds to that between category-based and person-based processing. The model therefore implies increasing category-based processing and *decreasing* person-based processing as subgroup size decreases. Similar predictions have been derived from social identity theory (Simon & Brown, 1987; cf. Brewer, 1993).

A third account has been presented by Oakes and Turner (1986; cf. Oakes, 1987, 1994; Oakes *et al.*, 1994). According to their functional analysis, individuals will tend to perceive themselves and others as members of given social categories to the extent that it is adequate and useful with regard to the characteristics of the situation and the perceivers' goals and motives within that situation. According to Oakes *et al.* (1994), those categories are most appropriate that maximize differences between members of different categories and minimize differences between members of the same category. Other things being equal, category salience and hence category-based processing are argued to be maximal if category membership divides the group into two subgroups of equal size (Oakes & Turner, 1986, p. 330). Conversely, category salience should decrease as the group composition becomes more unbalanced. In other words, an inverted U-shaped relationship of relative subgroup size and category-based processing is predicted with a maximum of category-based processing for balanced groups. Since Oakes and Turner underwrite to the principle of self-categorization theory (Turner, 1987) that there is a functional antagonism between person-based and category-based processing, the opposite prediction follows for person-based processing. Person-based processing should be minimized in balanced groups and should increase as group composition becomes unbalanced.

The account by comparative distinctiveness postulates that the effects of relative group size are mediated by attentional effects in perceiving the group members. Van Twuyver and van Knippenberg (1999) pointed out that these attentional effects can be tested most directly through the memory measures of Taylor *et al.*'s (1978) 'Who said what?' paradigm. From a different perspective, Oakes (1994) argued that this paradigm's measures, by assessing relatively directly the degree of exchangeability of category members, permit especially strong tests between competing accounts of the effects of relative group size. In the 'Who said what?' paradigm, perceivers observe a discussion between members of two or more categories. In a recognition test, they are again shown the discussion statements and are then asked to assign each statement to its speaker. Participants' assignment errors can be classified into two kinds. Within-category errors occur if a statement is wrongly assigned to a person who is a member of the same category as the speaker. Between-categories errors arise if a statement is assigned to a person from the other category. If there are more within-category errors than between-categories errors (after a suitable chance correction), this is interpreted as evidence for category memory. The difference of the two kinds of error is taken as a measure of category salience in perceiving and interpreting the group discussion.

In the remainder of this section, the results of six pertinent studies are briefly reviewed. It will be argued that limitations of the original 'Who said what?' paradigm restrict the interpretability of some of these findings. A pilot study and an experiment are then presented that use an improved version of the paradigm to test predictions of the account by comparative distinctiveness.

Studies Realizing Numerical Minorities

Biernat and Vescio (1993) compared perception of and memory for (a) racially balanced groups and (b) groups with solo black members. The six group members discussed attitudes toward the White Student Union, and three of them argued in favor of the union, whereas the remaining three argued against the union. There were four experimental conditions with different compositions of discussion groups: a balanced group in which all three black members were opposed to the White Student Union, and all three white members in favor of it; a balanced group in which one black member was also in favor of the White Student Union, and one white member opposed to it; a solo black group, in which the black person was opposed to the union; a solo black group, in which the black person was in favor of the union. In two studies, the recognition data showed that the error-difference measure for race was largest in the first balanced group in which race and attitude position correlated perfectly in the stereotypically expected direction, that is, in which so-called comparative as well as normative fit was maximal in terms of Oakes' (1987) analysis of category salience (cf. Biernat & Vescio, 1993; Oakes, 1994).

The interpretation of the error-difference measure as a measure of category salience is, however, problematic in this situation. Participants make use of expectancies in assigning statements when uncertain about the speaker and the speaker's category membership (Klauer & Wegener, 1998, Experiment 4; van Knippenberg, van Twuyver, & Pepels, 1994). In the present case, normative or stereotypical expectancies suggest to assign statements in favor of the White Student Union to white group members and statements in opposition to the White Student Union to black group members. The expectancy-based guessing strategy leads to more within-category than between-categories errors to the extent to which normative fit was actually realized in the discussion. Thus, the results may merely reflect the differential effectiveness, governed by the extent of normative fit, of the expectancy-based guessing strategy rather than any differences between the experimental conditions in terms of category salience. Although expectancy-based guessing clearly presupposes some extent of category salience, even constant amounts of category salience and expectancy-based guessing produce increases in the error-difference measure as normative fit increases.

Another problem is a structural one. In the conditions with a solo member, category memory for the solo member's statements necessarily leads to correct assignments, because category membership uniquely identifies the solo member. Thus, category memory for the solo member cannot contribute to the error-difference measure even if category memory is very high. A third problem relates to the manner in which the different kinds of errors are corrected for chance. It is discussed below.

In van Twuyver and van Knippenberg's (1999) study the contents of the statements were not correlated with category membership. The discussion group of the 'Who said what?' paradigm was actually partitioned into four subgroups along two orthogonal dimensions, gender and academic status (student versus teacher). For each dimension, group composition was varied in three steps. For example, there were groups with four men and twelve women, with eight men and eight women, and those with twelve men and four women. For a given group, the sex ratio was the same within each status subgroup, and the status ratio was kept constant within each gender subgroup. The numerical group compositions with respect to gender and academic status were manipulated orthogonally so that the composition of discussion groups followed a 3×3 design.

Van Twuyver and van Knippenberg (1999) computed the error-difference measure separately for each dimension of categorization along with the number of correct assignments, which was interpreted as a measure of person-based processing. The error-difference measure indicated the highest category salience for balanced groups when the gender categories were considered. In addition, the error difference was larger for minorities than for majorities both in the analysis along the gender dimension as well as in terms of academic status. Simultaneously, more correct assignments occurred for speakers from these minority categories.

In the study by van Twuyver and van Knippenberg (1999), each kind of error was corrected for the total number of possible confusions of that kind. For example, consider the group with four men and twelve women. The probability of a within-women error in random assignments is 11/16. Similarly, the probability of a within-men error is 3/16. The relative frequencies of the different kinds of errors were divided by a factor proportionate to the above probabilities.

This is an adequate correction, if the assignments rely on totally blind guessing. However, when there is partial information such as memory for the speaker's category membership, other correction factors are appropriate: Assume that you correctly remember that a given statement was made by a woman but you do not remember which. In this case, the probability of picking of the wrong woman by chance, and thus to commit a within-women error, is 11/12 and similarly, the chance probability of a within-men error, given category memory, is $3/4 = 9/12$. To the extent to which category membership is actually salient in memory, the appropriate correcting factors are therefore much less disparate for the minority and the majority category (12/9 and 12/11, respectively) than when there is no category memory (16/3 and 16/11, respectively). Note in particular that within-minority errors are overweighted leading to seemingly stronger minority categorization to the extent to which there is in fact category memory. In principle, there are similar problems with the second variable, number of correct assignments, but van Twuyver and van Knippenberg (1999) made an explicit effort to control for partial knowledge in the form of category memory through an analysis of covariance with the categorization index as covariate. The superior *correct* statement assignment to minority members is therefore much more likely to reflect a genuine effect.

Studies Realizing Nominal Minorities

In a related line of research, discussion groups are split into two subgroups of equal sizes, and participants are led to believe that the members of one of these subgroups stem from a small minority and members of the other sub-group from a large majority. In a study by Brewer, Weber, and Carini (1995; Experiment 3) the minority and majority categories were actually artificial categories of overestimators and underestimators as determined ostensibly on the basis of a perceptual test. The contents of the discussion statements were perfectly correlated with estimator category. That is, underestimators made statements that identified them with science-oriented academic studies, whereas all overestimators were associated with arts and humanities. At the outset of the experiment, participants underwent the perceptual estimation task and were assigned to one of the estimator categories, ostensibly on the basis of their test performance, or to a control condition in which they did not receive feedback about their category membership.

It was found that minority members made fewer within-minority errors than within-majority errors, whereas majority members made about as many within-minority as within-majority errors. Participants in the control condition made more within-minority errors than within-majority errors. There were no effects of the participants' group membership on between-categories errors, nor on a free recall measure of discussion statements obtained for about half of the participants. An analysis of recall order revealed evidence for clustering around speakers as well as categories in the minority members' data.

Simon, Aufderheide, and Hastedt (2000) performed a similarly designed study in which categories with low *a-priori* accessibility were used (preference for rural life versus preference for urban life), which were claimed to divide the target population into a small minority and a large majority. An additional independent variable was whether or not the individual self concept had been made accessible through a self-description task. When the individual self concept had not been primed, minority members made more within-ingroup errors than majority members. Given high accessibility of the individual self on the other hand, there were no significant differences between minority and majority members in the frequency of within-ingroup errors. Using basically the same procedures, extended by an attractiveness manipulation of the ingroup, Simon and Hastedt (1997, Study 2) found an interesting pattern of results for the within-ingroup errors, which was in part mirrored in a complementary pattern in the between-categories errors. Given low accessibility of the individual self, there were more within-ingroup errors for minority members than for majority members when the ingroup was attractive, and vice versa, when the ingroup was unattractive.

Summary

Two lines of research have employed the 'Who said what?' paradigm to explore person-based and category-based perception of minority and majority members. A number of methodological problems of the 'Who said what?' paradigm render all of these studies difficult to evaluate. In the studies with numerical minorities, problems with correcting factors may have caused biased estimates of minority and majority categorization as explained earlier. Furthermore, in the studies by Biernat and Vescio (1993), Brewer *et al*. (1995), Simon and Hastedt (1997, Study 2), and Simon *et al*. (2000), the contents of the statements covaried with category membership. Intelligent guessing strategies can exploit this covariation when statements are assigned. For this reason, the observed effects may reflect differences in the use or effectiveness of these guessing strategies rather than differences in the manner in which the group discussion was perceived and represented in memory. Simon and Hastedt (1997, Experiment 2) performed separate analyses on the statements for which there was normative fit and on neutral statements which did not covary with category membership in a transparent manner. They found similar patterns of effects for both sets of statements rendering this alternative interpretation less likely for this data set.

Thus, two different modes of category-based processing may be confounded in these studies: category-based memory for the particulars of the group discussion and later use of generic category-based expectancies in reconstructive guessing.

THE MODIFIED 'WHO SAID WHAT?' PARADIGM BY KLAUER AND WEGENER (1998)

More generally, the pattern of assignment errors is in fact a complex function of several processes. Expectancy-congruent guessing can cause shifts in within-category errors and between-categories errors in the absence of differences in category salience in encoding and interpreting the discussion statements. Similarly, shifts in item memory, person memory, and category memory all affect the frequencies of the different kinds of errors and the error-difference measure.

Klauer and Wegener (1998) proposed a small modification of the paradigm that allows one to disentangle these different processes. Apart from the statements that are shown in the discussion, new statements or *distractors* are presented in the assignment phase. For each statement, the participant is first asked whether or not the statement occurred in the discussion. If the participant judges the statement old, he or she is required to assign the statement to a speaker in a second step. If the statement is judged new, an assignment to a speaker is not required.

Based on this modification, Klauer and Wegener (1998) validated a multi-nomial model (Batchelder & Riefer, 1999) of the processes involved in the 'Who said what?' paradigm. The model parameters assess the separate contributions of different psychological processes:

(1) *Item memory*: Is there recognition of the statement that is to be assigned?

- Strength of item memory is measured by the model parameters D_A and D_B, according to the speaker's social category (A versus B), and D_N, where N denotes new items. D_N refers to the likelihood with which distractors are detected. Note that remembering that an old item had been presented may be psychologically quite distinct from detecting a new item as new. The model therefore provides different parameters for both processes.

(2) *Person memory*: Given item memory, is the speaker of the statement recognized?

- Strength of person memory is measured by the model parameters c_A and c_B for speakers from Category A and B, respectively.

(3) *Category memory*: Given item memory, but no person memory for an individual statement, is the speakers's category membership still available and used in assigning the statement?

- Strength of category memory is measured by the model parameters d_A and d_B for speakers from Category A and B, respectively.

(4) *Guessing the category membership*: If the information retrieved for a given statement is not sufficient for determining the speakers's category membership, category membership is guessed based on whatever information is available at retrieval and the statement is then assigned to a member of the guessed category.

- The strength of the tendency to guess Category A rather than B is measured by the model parameter a. $a = 0.5$ means that A and B are guessed equally frequently, whereas $a > 0.5$

means that there is a bias in favour of Category A. Since there are only two categories, the likelihood to guess Category B is always given by $1-a$.

(5) *Guessing the item status*: If the information retrieved is not sufficient for determining whether the item was presented at all or not, the item status is guessed.

- The strength of the tendency to guess old rather than new is measured by the model parameter b.

(6) *Guessing the speaker within the correct category*: If the information retrieved is sufficient for determining the speakers's category, but not his or her identity, the correct speaker is guessed with probability of success given by $1/n$, where n is the size of the category.

In the Appendix, the model is described in more detail, and the methods of parameter estimation, goodness-of-fit tests and hypotheses tests are explained.

The modified 'Who said what?' paradigm allows one to assess the contributions of a number of different processes that are of central interest in the social psychology of impression formation and group perception. It provides measures of person-based processing in social perception through the assessment of person memory, measures of category-based processing through the assessment of category memory, and as considered in the main experiment, a measure of reconstructive category use in later retrieval through the guessing parameters. It is unique in making these measures available within one and the same unobtrusive task. This renders alternative accounts of possible dissociations between the different measures in terms of artefacts in measurement procedures such as scale artefacts or order effects unlikely.

PILOT STUDY

In the pilot study, gender categories were employed and discussion groups with eight members were presented. Nine groups of participants differed in the gender composition of the presented discussion groups. Specifically, the ith discussion group was composed of i women and $8 — i$ men ($i = 0, \ldots, 8$), and thus the discussion groups ranged from all-male groups to all-female groups, and all possible intermediate group compositions were realized. Participants observed a succession of written discussion statements on the computer screen, in which each statement was accompanied by its speaker's picture.

Each speaker made six statements. Speakers made their statements in turn. Discussion statements and distractor statements were randomly sampled from a large pool of statements about conditions at the university (cf. Klauer & Wegener, 1998). Since for each participant, statements were randomly assigned to speakers, statement content did not covary systematically with category membership. Some amount of accidental covariation is nevertheless always possible and some residual amount of expectancy-based guessing may therefore contribute to the overall amount of categorization in the pilot study. In the main experiment, the amount of expectancy-based guessing is explicitly assessed separately from category memory. The presentation procedure closely followed that of Klauer and Wegener's (1998) experiments.

From the account by comparative distinctiveness, three predictions can be derived: As relative group size decreases,

- *Hypothesis 1*: item memory increases (Hamilton, Dugan, & Trolier, 1985)
- *Hypothesis 2*: person memory increases
- *Hypothesis 3*: category memory increases.

The analyses were based on 165 participants most of whom were students with different majors (mean age 25.42 years, $SD = 5.67$). They were assigned to the experimental conditions defined by numerical group composition with the restriction that (1) in each experimental condition, there were about as many male as female participants, and (2) proportionately more participants were assigned to experimental conditions with unbalanced discussion groups to compensate for the fact that each participant was shown and later tested on successively fewer minority statements as the minority size decreased.

Because the literature review yielded some evidence for a possible moderating role of ingroup–outgroup status, the data were coded in terms of ingroup and outgroup status of the gender categories *vis-à-vis* the participants' gender. Preliminary tests revealed however that the assignment frequencies were not a function of ingroup–outgroup status. The absence of ingroup–outgroup differences from the memory parameters is a frequent (Beauvais & Spence, 1987; Jackson & Hymes, 1985; Taylor *et al*., 1978) but not ubiquitous finding (Frable & Bem, 1985; Ostrom, Carpenter, Sedikides, & Li, 1993, Experiment 1) in studies using the 'Who said what?' paradigm and gender categories (cf. Lorenzi-Cioldi, Eagly, & Stewart, 1995).

The model was fit to the data from all nine experimental conditions simultaneously with different parameter values for each numerical subgroup size. Minor modifications of the original Klauer and Wegener (1998) model were necessary for the discussion groups with a solo minority.[1] The model achieved a satisfactory goodness of fit as determined by a log-likelihood ratio test, $\chi^2 = 36.02$, $df = 26$, $p = 0.09$.

Note that in discussion groups with solo minorities, the parameter c_1 assessing person memory for the solo member and the parameter d_7 assessing category memory in the majority that faces a solo minority may not unconfound person and category memory, respectively, as clearly as in discussion groups with less extreme compositions. For solo minorities, category memory constitutes identifying information and thus, it directly contributes to the person memory parameter (c_1), which is not the case for other group compositions. Similarly, memory for which statements were made by the solo member may contribute to assigning the other statements to the majority category, that is, person memory for the solo member may directly inflate category memory for the majority (d_7). Note however that identifying what a person said, as in person memory for the solo member (c_1), may be psychologically quite distinct from identifying what a person has not said; a process we just claimed might contribute to category memory for the majority (d_7).

Setting aside these two parameters for the moment, monotonic trends could be imposed on the parameters for item detection (D_1 to D_8), distractor detection ($D_{N,4}$ to $D_{N,8}$),[2] person memory (c_2 to c_8), category memory (d_2 to d_6), and category guessing (a_1 to a_7) without significant loss in goodness of fit. In Figure 1, the parameter estimates are shown. Table 2 presents a summary of the major results of the following model-based hypothesis tests. Log-likelihood ratio tests confirmed that the following monotonic trends were significant: the decrease in category guessing with decreasing subgroup size (a_1 to a_8), $\chi^2 = 240.26$, $df = 6$, $p < 0.01$; the increase in person memory (c_2 to c_8), $\chi^2 = 44.41$, $df = 6$,

[1]When there is a minority of size one, that is, a solo ingroup or outgroup member, the *d* parameter cannot be estimated for that minority because there can be no confusions within the minority. In these experimental conditions, all branches involving and following the *d* parameter were deleted from Figure A1. Furthermore, when there are only ingroup or only outgroup members, all branches involving or following a *d* parameter must be deleted along with all branches involving and following the *a* parameter. Further details about these modifications, technical assumptions, and statistical details on the model tests discussed below can be obtained from the first author.

[2]Distractors are new statements that never occur in the course of the group discussion. They do not belong to a particular subgroup in the sense old items do, because they have no speaker. In other words, it makes no sense to have distractor detection vary as a function of subgroups. Rather, distractor detection can only be a function of the numerical composition of the *entire* group, which we indexed by the size of the majority group. Ranging from balanced groups to maximally unbalanced groups, there were therefore five distractor detection parameters labelled $D_{N,4}$ to $D_{N,8}$.

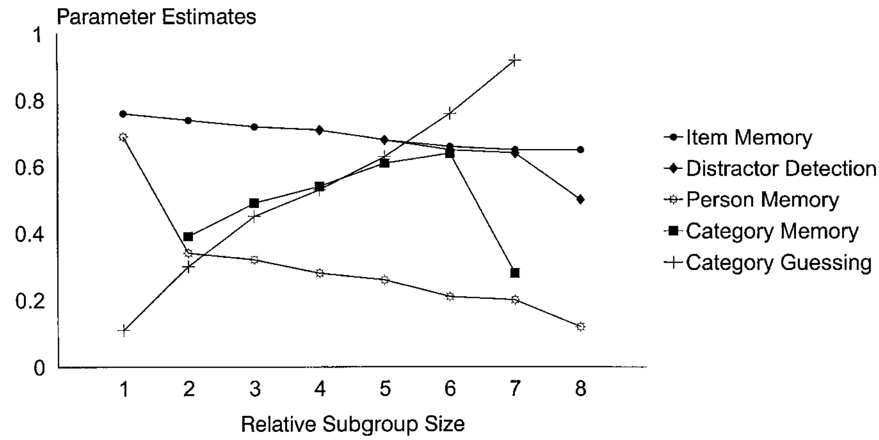

Figure 1. Parameter estimates of the model with monotonic trends in the pilot study. Subgroup size refers to ingroup and outgroup combined for the different memory parameters D, D_N, c, and d, and to the size of the outgroup for the category-guessing parameter a

$p < 0.01$; the increase in item detection (D_1 to D_8), $\chi^2 = 27.76$, $df = 7$, $p < 0.01$; the increase in distractor detection ($D_{N,4}$ to $D_{N,8}$), $\chi^2 = 14.71$, $df = 4$, $p < 0.01$. However, there was no evidence for significant differences in category memory (d_2 to d_6), $\chi^2 = 5.70$, $df = 4$, $p = 0.22$.

Turning back to the parameters c_1 and d_7, it can be seen in Figure 1 that the parameter c_1 indeed achieved an unusually large value as surmised earlier. It was significantly larger than c_2, $\chi^2 = 55.18$, $df = 1$, $p < 0.01$. Contrary to expectations, the parameter d_7 was unusually small and could not be set equal to d_6, $\chi^2 = 4.03$, $df = 1$, $p = 0.04$. Additional tests showed that the significant monotonic trend in distractor detection largely reflects particularly poor distractor detection in groups without gender variation. This suggests — in line with the other memory measures — that the all-male and all-female groups may have been comparatively uninteresting and did not elicit the kind of elaborative encoding and/or effortful retrieval strategies that underlie distractor detection (Strack & Bless, 1994).

To summarize, the substantial effects of group composition on the pattern of assignment frequencies could be parsimoniously represented by the multinomial model. The major findings were as follows: Both person and item memory strengths increased as relative group size decreased and thus, Hypotheses 1 and 2 could be maintained. Hypothesis 3, on the other hand, did not receive significant support: Differences in category memory appear to have played a minor role in the present data set: the only significant finding suggested unusually poor category memory for majorities facing a solo minority. Distractor detection decreased significantly as the group composition became more unbalanced, reflecting particularly poor distractor detection in groups without gender variation. Guessing bias favoring a subgroup increased with subgroup size.

EXPERIMENT

The results of the pilot study provided partial support for the account by comparative distinctiveness. The main experiment was a conceptual replication and extension using age categories rather than gender categories. It extended the pilot study in two ways: (1) by implementing a rating measure of stereotyping, and (2) by assessing participants' stereotype-based guessing strategies in the 'Who said what?' task.

Each speaker made three conservative and three more progressive statements in a discussion about gender roles in our society. When uncertain about a statement's speaker, participants are likely to guess a speaker. When processing is category-based they often employ stereotype consistency as one cue in guessing (cf. Klauer & Wegener, 1998, Experiment 4; van Knippenberg *et al*., 1994): For example, in the case of age categories, participants might be inclined to assign conservative statements to older group members and progressive statements to younger group members when guessing the speaker. This kind of guessing strategy will be termed *stereotype-based* guessing. To summarize, the experiment allowed the assessment of three different category-based processing modes: Category memory, reconstructive use of stereotype-based expectancies in guessing, and stereotyping in ratings.

The discussion groups were composed of young men in their early twenties and older men in the midforties. The experimental design comprised two factors: (1) the age composition of the discussion groups, which was varied in three steps (2 : 6, 4 : 4, and 6 : 2), and (2) the kind of discussion statement (conservative versus progressive).

Over and above Hypotheses 1 to 3, the following additional predictions can be derived from the account by comparative distinctiveness:

- *Hypothesis 4:* Stereotype-based guessing becomes more pronounced in the presence of a minority.
- *Hypothesis 5:* As relative group size decreases, the rating measure of stereotyping indicates more stereotyping.

Method

The group discussion was presented on the computer screen as a succession of written discussion statements. Each statement was accompanied by its speaker's picture. Following the 'Who said what?' assignment task, stereotype ratings were obtained for the members of the discussion groups. Then, participants were asked to estimate the frequencies with which the different kinds of statements had been made by young speakers. Finally, participants were asked to give their age, sex, and university major, as well as their hypotheses about the purpose of the experiment, followed by debriefing. One session required about 30 min.

Participants

There were 104 participants, who were students with different majors of the University of Bonn. Their mean age was 24.8 years, $SD = 3.24$, so that the young speakers constituted the ingroup. They were assigned to the three experimental conditions with differently composed discussion groups using a quasi-random assignment procedure. Two participants were excluded from the analyses because during debriefing they stated that they were familiar with the 'Who said what?' paradigm and its purpose. One participant was excluded because with 37 years, her ingroup–outgroup status seemed unclear. Finally, one participant was excluded because of an unusually high number of errors in statement recognition: the participant classified all statements as new.

There remained 100 participants of which 20 (9 male, 11 female) were assigned to the condition with balanced discussion group and 40 (20 male, 20 female) to each of the two conditions with unbalanced (6 : 2 and 2 : 6) discussion groups. Proportionately more participants were assigned to experimental conditions with unbalanced discussion groups to compensate for the fact that each participant was shown and later tested on fewer statements from the minority group as compared to statements from subgroups in balanced conditions.

Pool of Statements and Assignment of Statements to Speakers

The discussion statements referred to gender roles in society. The statements addressed six different subtopics: child care, sharing of domestic work, job and finances, organization of leisure activities, partnership and faithfulness, and gender-specific accessories. For each subtopic 16 pairs of parallel statements were constructed. The two members of each pair addressed the same issue, while one statement expressed a conservative point of view and the other a more progressive point of view. For example, with respect to child care, the following statements formed a pair: 'The upbringing of children is above all the woman's privilege' and 'The upbringing of children is by no means only the woman's obligation'. Thus, there were 96 conservative and 96 more progressive statements.

For each participant, speakers of a given discussion group were each assigned three conservative and three progressive statements randomly sampled from this pool without replacement and the following additional restrictions: Only one member of a pair of parallel statements was used in a given discussion or as distractor item, and the six statements selected for a given speaker had to refer to different subtopics. This latter restriction was to ensure that a given speaker would not be perceived as holding contradictory views with respect to the same or a very similar issue.

Age Composition of the Discussion Groups

There were three kinds of discussion groups which differed in age composition. Groups with a young minority consisted of two young speakers and six older speakers; balanced groups of four young speakers and four older speakers; groups with a young majority of six young speakers and two older speakers. For each participant, the speakers' pictures were randomly sampled from six pictures of men in their early twenties and six pictures of men in their midforties.

'Who Said What?' Task

The speakers made their statements in turn, so that all eight speakers first made their first statement, then there was a round of second statements, and so forth. The sequence of speakers within any one round was randomized. All randomizations discussed in this section were performed by the computer for each participant anew. The duration of stimulus presentation was shortened to 4 s (usually 7 s in Klauer & Wegener, 1998) in order to provoke a sizeable amount of item recognition errors (and thereby favorable conditions for estimating guessing tendencies; cf. the discussion in Klauer & Wegener, 1998).

In the subsequent test phase, participants were presented statements one by one. The statements comprised the statements presented in the discussion as well as 24 conservative and 24 progressive distractor statements sampled randomly without replacement from the pool of statements. Distractors and old statements were pooled and presented in a random sequence. Participants were instructed to click, using the computer mouse, a button labelled 'old' if they remembered the statement from the discussion; they were instructed to respond 'new', if the item had not been presented before. The 'old' and 'new' buttons were presented below the statement and could switch places from trial to trial as determined by a random procedure.

If the participant's response was 'new', the next statement was presented; if the response was 'old', the eight speakers' photographs were shown, located in two rows of four pictures each, headed by the statement. The speakers' pictures were randomly distributed over the eight locations reserved for the photographs, and a new random assignment of pictures to locations was computed for each trial.

Table 1. Mean differences in ratings of age typicality

Age	Evaluation	Trait	*M*	*SD*
Twenties	Positive	Flexible	2.63	2.17
		Enthusiastic	2.50	1.94
Twenties	Negative	Impudent	3.27	1.94
		Reckless	2.14	1.57
Midforties	Positive	Wise	−2.63	1.88
		Wealthy	−3.41	1.97
Midforties	Negative	Powerful	−2.57	1.36
		Bourgeois	−3.57	1.83

Note: Positive values imply higher typicality for the young group.

Participants were asked to click the speaker's picture if they remembered the speaker; otherwise, they were instructed to guess the correct speaker and click his picture.

Stereotype Ratings

For each age group, two positive and two negative stereotypical trait adjectives were selected in a procedure with three steps involving 64 raters who were not participants in the experiment itself. The items finally selected produced the largest mean differences in ratings of age typicality between the two age groups, comprised positive and negative traits, and were semantically dissimilar. Table 1 shows these traits and the mean differences in ratings of age typicality from the third step of the selection procedure. In that step, 30 students rated a larger set of 20 traits with respect to (a) how typical they were for the young age group and (b) how typical they were for the older age group on ten-point rating scales. The order of the two blocks of ratings, (a) and (b), was counterbalanced across raters. The traits 'conservative' and 'progressive' were intentionally excluded from the entire procedure because we wanted to see whether there would be stereotypic effects on dimensions other than those directly addressed by the observed discussion.

Following the 'Who said what?' task, each participant rated two younger speakers and two older speakers randomly sampled from his or her discussion group with respect to how well each of the eight traits described the speaker on ten-point rating scales. The four speakers were grouped into two pairs each comprising a young and an old speaker. The ratings alternated between a given pair's young and old member for each trait, and they alternated between pairs of speakers from trait to trait until all trait ratings for all four speakers had been obtained. Finally, participants were asked to estimate how many of the 24 progressive discussion statements and how many of the 24 conservative discussion statements were made by young men.

Results

Do the Assumptions of the Baseline Model Hold?

The model was fit to the data from all six experimental conditions spanned by age composition and kind of statement (conservative versus progressive) simultaneously. Different parameter values were admitted for each experimental condition. Like in the pilot study, the same memory parameters were used for age ingroup and outgroup. Thus, the memory parameters could vary as a function of group size and kind of statement (coded as category-consistent versus category-inconsistent with respect to the speaker's category membership for this analysis). In addition, the *b* parameters for guessing the item status (old versus new) in item detection were set equal over experimental conditions with differently composed discussion groups to avoid overparametrization (cf. Klauer & Wegener, 1998, Experiment 6). Thus, there were only two *b* parameters, one for the conservative statements and one for the progressive statements.[3] The baseline model is identified and reserves 20 degrees of freedom for the goodness of fit. The log-likelihood-ratio statistic for testing the model assumptions was 20.88, corresponding to an excellent goodness of fit, $p = 0.60$. Note in particular that there do not appear to be major differences in the memory parameters as a function of ingroup versus outgroup status. In an explicit test for ingroup-outgroup differences, we compared the fit of this model with one that allows for such differences. The model comparison revealed that there were no significant differences as a function of ingroup versus outgroup status ($\chi^2 = 16.30$, $df = 18$, $p = 0.57$).[4]

Are There Effects of the Factor 'Kind of Statement'?

Then we tested whether a second major simplification could be introduced, that is, whether it was necessary to assume differences in memory parameters and in the *b* parameters for guessing item status (old versus new) as a function of kind of statement, or whether the same item memory, person memory, category memory, and item-detection bias parameters could be assumed irrespective of the kind of statement on which their estimates were based. This simplification was possible, $\chi^2 = 12.03$, $df = 11$, $p = 0.36$, indicating that kind of statement did not have a major impact on the different memory parameters.

Alternatively, it turned out as expected that the attempt to set equal the *a* parameters for category guessing across kind of statement led to a significant deterioration in goodness of fit, $\chi^2 = 18.26$, $df = 3$, $p < 0.01$, even when all other parameters were allowed to vary as a function of kind of statement. We elaborate on this finding below.

The simplified model, which allowed for differences in category guessing as a function of kind of statement but otherwise used the same parameters for both kinds of statement, achieved an excellent overall goodness of fit: $\chi^2 = 32.91$, $df = 31$, $p = 0.37$.

Are There Effects of Relative Group Size on the Different Kinds of Parameters?

The parameter estimates and their standard errors are shown in Table 3. A compact summary of the major results of the following model-based hypothesis tests is provided in Table 2. Descriptively, item memory and person memory deteriorate as the relative group size increases whereas category memory increases. Distractor detection was better in the unbalanced conditions than in the balanced condition. The probability to guess the young (rather than old) subgroup increases as the young group's size increases. It is higher for progressive than for conservative statements in each size condition,

[3]For the same reason, it was necessary to anchor the distractor detection parameter D_N for conservative statements as well as the one for progressive statements to one of the parameters for target detection within the balanced group. We set equal the former distractor detection parameter to the parameter for detecting consistent target items, the latter to the one for inconsistent target items. These restrictions clearly have an undesirable subjective element. We repeated the analyses with the other possible choices for anchoring the distractor detection parameters within the balanced group and found that the pattern of results did not depend on the choice.

[4]Given that the discussion statements in the experiment referred to gender roles in society, we checked for effects of participants' gender in a similar manner. It was found that male and female participants differed significantly in item memory such that female participants exhibited generally higher item memory than male participants across conditions. Since this was the only significant difference, participants' gender was not retained as a factor in the subsequent model analyses.

Table 2. Summary of major results of pilot study and experiment

Dependent variable	Par.	Prediction[a]	Results: Pilot study	Results: Experiment
Item memory	D	H_1: increase	Significant increase	Non-significant increase
Distractor detection	D_N	none	Poor in one-gender groups	Better in unbalanced than in balanced groups
Person memory	c	H_2: increase	Significant increase	Significant increase
Category memory	d	H_3: increase	Non-significant decrease	Non-significant decrease
Stereotype-based category guessing	Δa^b	H_4: increase	n.a.	Most pronounced in balanced groups
Stereotyping in ratings	n.a.	H_5: increase	n.a.	Significant increase in minority groups

Notes: Par. = parameter; n.a. = not applicable.
[a]As relative group size decreases. Prediction according to the account by distinctiveness.
[b]Measured by the discrepancy between $a_{\text{progressive}}$ and $a_{\text{conservative}}$.

indicating stereotype-based guessing. That is, when uncertain about a statement's group origin, participants were more inclined to assign the statement to a young speaker if it was a progressive statement than if it was a conservative statement. Equivalently, they were more inclined to assign a statement to an old speaker if it was conservative than if it was progressive.

Table 3 also shows the log-likelihood-ratio statistic, its degrees of freedom, and associated p-value for testing for equality across relative subgroup size for each kind of parameter. The observed monotone trend in person memory as well as the observed monotone trends in category guessing are significant whereas the observed increase in category memory is not. As already mentioned (see section on the effects of the factor 'kind of statement'), the guessing parameters for conservative statements differed significantly from those for progressive statements, implying a significant amount of stereotype-based guessing. Finally, as shown in Table 3, distractor detection was significantly better in the unbalanced groups than in the balanced group.

Table 3. Parameter estimates, their standard errors, and tests for equality over subgroup size in the experiment

		Subgroup size								
		2		4		6				
Dependent variable	Par.	Est.	*SE*	Est.	*SE*	Est.	*SE*	χ^2	*df*	*p*
Item memory	D	0.75	0.02	0.73[a]	0.02	0.71	0.02	4.13	2	0.13
Distractor detection	D_N	0.80[b]	0.02	0.73[a]	0.02	0.80[b]	0.02	5.02	1	0.03
Person memory	c	0.43	0.03	0.34	0.02	0.24	0.01	40.01	2	< 0.01
Category memory	d	0.48	0.05	0.51	0.05	0.55	0.08	0.42	2	0.81
Category guessing[c]	$a_{\text{progressive}}$	0.20	0.03	0.56	0.04	0.78	0.03	70.19	2	< 0.01
Category guessing[d]	$a_{\text{conservative}}$	0.15	0.03	0.31	0.04	0.69	0.04	59.66	2	< 0.01

Notes: Par. = parameter; Est. = estimate. For the a-parameter, estimates give the probabilities of guessing the young group, where subgroup size refers to the young group in this case. Probability of guessing old: $b = 0.29$, $SE(b) = 0.03$.
[a]Parameters set equal.
[b]Parameters set equal (cf. footnotes 2 and 3).
[c]For progressive statements.
[d]For conservative statements.

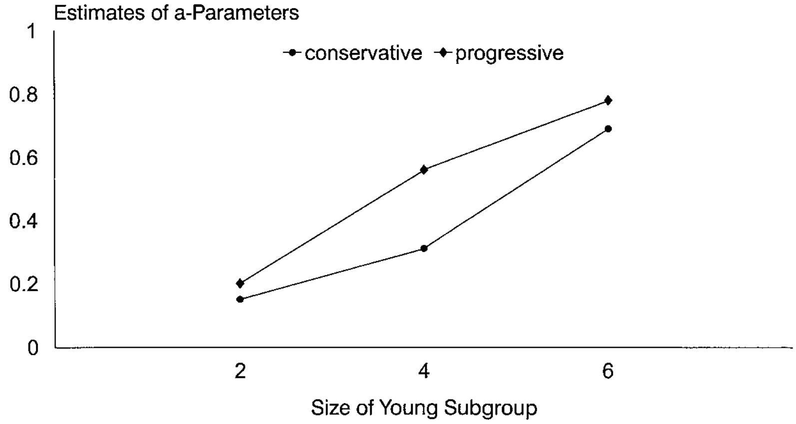

Figure 2. Estimated probabilities of guessing that a statement was made by a young man as a function of the size of the set of young men and the kind of statement (conservative versus progressive)

Is There an Interaction of Relative Group Size and Stereotype Consistency in the Guessing Parameters?

Hypothesis 4 states that stereotype-based guessing should be most pronounced in the presence of a minority. In other words, kind of statement and group size should interact in determining the guessing parameters. In Figure 2, the *a* parameters a_{xy} for guessing the category of younger speakers are presented as a function of kind of statement x ($x =$ conservative, progressive) and size of the young subgroup y ($y = 2, 4$, and 6).

An interesting interaction is suggested by Figure 2 although it differs from the one postulated in Hypothesis 4. The stereotypical tendencies, measured by the discrepancy between the *a* parameters for the two kinds of statement, seem to be most pronounced in the balanced group. As a formal test of this interaction, we tested whether the guessing parameters a_{xy} could be decomposed multiplicatively into independent contributions of kind of statement α_x and category size, β_y, that is, whether a log-linear main-effects model, $a_{xy} = \alpha_x \beta_y$, can be maintained or whether the two factors 'kind of statement' and 'category size' must be assumed to interact on a logarithmic scale. The main-effects model without interaction must be rejected, $\chi^2 = 9.66$, $df = 2$, $p < 0.01$, indicating that the effects of kind of statement differ as a function of category size. As can be seen in Figure 2, stereotypical expectancies were relied on most strongly in the balanced group.

Do the Rating Data Exhibit Evidence for Stereotyping?

Trait ratings were averaged over the two rated speakers of each age category and over traits, separately for the stereotypically young and the stereotypically old traits. Next, ratings on incongruent traits were subtracted from ratings on congruent traits to obtain a difference score reflecting stereotyping strength. As can be seen in Figure 3, the smallest amount of stereotyping was obtained in balanced groups, an intermediate amount of stereotyping for majority members, and a maximum of stereotyping for minorities.

The stereotype index was subjected to an analysis of variance with within-participants factor age (young versus old speakers) and between-participants factor group composition (two young and six

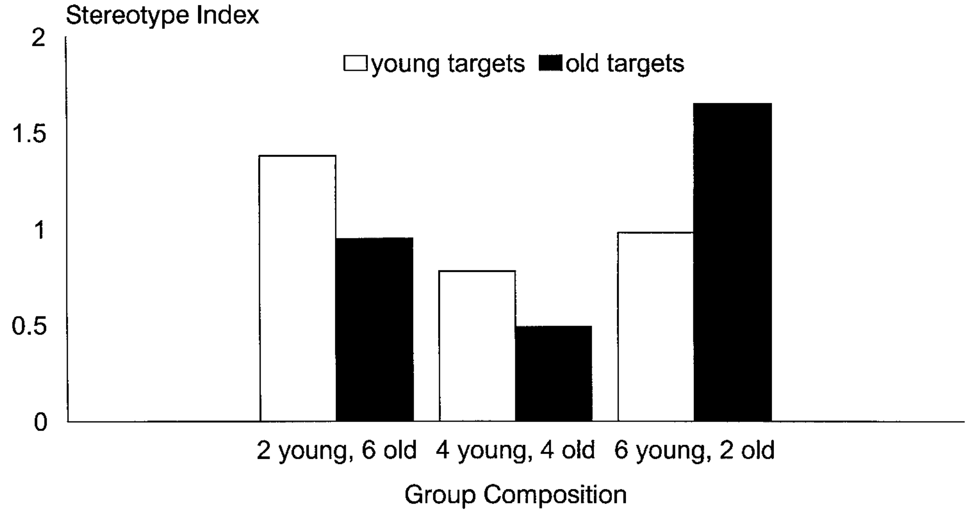

Figure 3. Stereotype index as a function of group composition and age group of targets

old, four young and four old, six young and two old). In this analysis, increased minority stereotyping should result in an age by group composition interaction such that there is more stereotyping of young speakers in the condition with two young and six old speakers, and more stereotyping of old speakers in the condition with six young and two old speakers.[5]

While the main effects for age and group composition were not significant ($F(1, 97) = 0.01$, $p = 0.93$, and $F(2, 97) = 2.77$, $p = 0.07$, respectively), the interaction of age and group composition reached significance: $F(2, 97) = 5.00$, $p < 0.01$. Individual t tests revealed that the difference between congruent and incongruent ratings was significantly larger for minorities than for balanced subgroups ($t(98) = 2.68$, $p < 0.01$, note that this is a between-subjects comparison), and significantly larger for minorities than for majorities ($t(79) = 3.03$, $p < 0.01$; a within-subject comparison), whereas balanced subgroups and majorities did not differ significantly in this respect ($t(98) = -1.04$, $p = 0.30$; a between-subjects comparison). To summarize, stereotyping was significantly more pronounced for minority members than for members of balanced groups as well as of majority groups, while members of the latter two groups were stereotyped to comparable degrees.[6]

Finally, participants were asked to estimate how frequently the different kinds of statements were made by young men. This allowed us to probe for expectancy-based illusory correlations (Hamilton & Rose, 1980) between kind of statement and category membership. Based on the frequency estimates, we computed (Z-transformed)ϕ-coefficients for each group composition, and found no evidence for illusory correlations. That is, none of the ϕ-coefficients was significantly different from zero, nor were there effects of the factor group composition on the coefficients in an analysis of variance.

[5]A preliminary analysis included participants' gender as additional factor and revealed that this factor did not give rise to any significant effects or interactions.

[6]Separate within-participants comparisons for each group composition revealed a tendency for the minority ingroup to be stereotyped more strongly than the majority outgroup ($t(39) = 1.79$, $p = 0.08$), no differences in the extent of stereotyping between equal-sized ingroups and outgroups ($t(19) = 0.76$, $p = 0.46$), and significantly more stereotyping for the minority outgroup than the majority ingroup ($t(39) = 2.45$, $p = 0.02$).

Discussion

The experiment replicated major findings of the pilot study. The multinomial model provided a parsimonious and compact description of the substantial impact of group composition on the pattern of assignment frequencies in the 'Who said what?' task, and there was no evidence for major differences in the memory representations as a function of ingroup versus outgroup status. In addition, the two item pools containing conservative and progressive statements, respectively, did not give rise to substantially different memory parameters, indicating that closely comparable item pools had successfully been constructed.

As in the pilot study, person memory decreased as relative group size increased (Hypothesis 2). Simultaneously, item memory tended to decrease (Hypothesis 1), and category memory tended to increase contrary to Hypothesis 3 although both of these trends did not reach significance. Distractor detection was better for the unbalanced groups than for the balanced group.

Participant's guessing was sensitive to subgroup size and it was affected by stereotypical expectancies. Contrary to Hypothesis 4, stereotype-based guessing was most pronounced in the balanced group; a finding that will be discussed later.

As predicted by Hypothesis 5, the rating data exhibited the largest amount of stereotyping for minorities, irrespective of whether the minority was old or young. Although this result is as predicted by the account by comparative distinctiveness, there is an alternative account of the rating data that was pointed out by an anonymous reviewer. For the rating task, four speakers were grouped into two pairs comprising one young and old speaker each, and the ratings alternated between a given pair's young and old member for each trait. By juxtaposing young and old members in this manner we intended to stress the intergroup perspective in the rating task. However, in the balanced and majority groups, participants thereby rate only a sample of the category members creating the possibility of an internal contrast within the category. For example, when judging whether a majority member is reckless, participants may be sensitive to the fact that there are other majority members who were not sampled whom they perceive to be even more reckless. No identical comparison is available for the minority; more precisely, participants can express it in their ratings if they feel that the other sampled exemplar is more reckless. It is therefore possible that the sampling enhanced intracategory comparisons in the balanced and majority groups undermining intergroup comparisons and moderating stereotypic judgements for these groups. In fairness, it should be noted that the alternatives, that is, to rate all members, to take a sample from each category, or to rate the categories instead of their members, each have their own problems, and that there probably is no optimal procedure for obtaining rating measures of stereotyping when group composition is unbalanced. When all members are rated, distinctiveness may operate during the judgment phase rather than during perception of the group discussion. The same problem exists when equal-probability samples are drawn from each category. If categories instead of members were rated, the level of intergroup comparisons would be stressed rather obtrusively limiting the interpretability of findings.

There were no effects of stereotype congruency versus incongruency on the different memory parameters (paralleling similar findings in this and a related recognition paradigm by Klauer & Wegener, 1998, Experiment 4, and Klauer & Meiser, 2000; cf. also Bayen, Nakamura, Dupuis, & Yang, in press). We attribute the absence of (in)congruency effects to the low strength of the realized expectancies (Stangor & Ruble, 1989): Although young men in their twenties are perceived as *relatively* more likely to hold progressive views than men in the midforties, many members of the latter category are also known to hold progressive views while many members of the former category endorse conservative attitudes. A similar explanation may account for the absence of expectancy-based illusory correlations in the frequency estimates.

GENERAL DISCUSSION

The account by comparative distinctiveness claims that minorities draw attention by virtue of their comparative rarity, leading to increased person-based as well as category-based processing (Taylor, 1981). The account received partial support. Considering first the memory measures, person memory significantly decreased with increasing subgroup size in both the pilot study and the main experiment, as predicted by the distinctiveness hypothesis. This finding replicates the similar finding by van Twuyver and van Knippenberg (1999) reviewed earlier. Simultaneously and as predicted, item detection decreased with increasing subgroup size in both studies (Hamilton *et al.*, 1985), an effect that was significant in the pilot study, indicating that statements made by members of small subgroups are remembered better. Considering the rating data, stereotyping was stronger for members of the minority than for members of balanced groups and majority members as predicted. However, category memory was not significantly affected by relative group size and descriptively, it tended to increase, rather than decrease, with increasing subgroup size. This means that a central prediction of the account by comparative distinctiveness for category salience in memory is not supported by the present data.

A different pattern of results emerged for stereotype-congruent guessing. First, guessing the speaker's category under uncertainty was found to be sensitive to group size. In both experiments, guessing in favor of a given category became more pronounced as category size increased. Second, guessing was guided by stereotypical expectancies triggered by the statement content. Third, this category-based processing mode was significantly more pronounced in the balanced group than in the unbalanced groups.

The guessing parameters measure reconstructive category use conditional on the event that the speaker of a statement and the speaker's category membership were not retrieved. Thus, there is little logical basis for expecting a direct relationship between the guessing measure and the memory measures. In reconstructive guessing, the weight given to different kinds of information such as group size and statement content is likely to be determined by how useful these different cues are perceived to be. In balanced groups, group size is not at all helpful because it does not distinguish one of the subgroups as the more likely source of a given statement. Stereotype congruency, on the other hand, distinguishes half of the group members as potential speakers when the group composition is balanced, thereby reliably reducing uncertainty by 50% for each statement. Thus, it seems likely that stereotype congruency was perceived as particularly helpful in balanced groups, where differential group size is completely useless as an alternative cue for reducing uncertainty. Comparing the memory results with the guessing data, it can be concluded that different laws appear to govern how relative group size affects memory for individual items' speakers and how it affects reconstructive assignment strategies in later retrieval (cf. Garcia-Marques & Hamilton, 1996; Sherman & Bessenoff, 1999). Whereas person memory appears to be governed by rarity-based distinctiveness, reconstructive guessing may be guided by the perceived usefulness of different kinds of information.

How well do alternative accounts of the role of relative group size as proposed in the literature account for the present findings? Like Taylor (1981; Taylor *et al.*, 1978), Mullen (1991) postulates that minorities draw attention by virtue of their relative size. However, as explained in the Introduction, Mullen's model implies increasing category-based processing and *decreasing* person-based processing as subgroup size decreases. This latter assumption distinguishes Mullen's model from the account by comparative distinctiveness, and it is inconsistent with the finding of the present studies that person memory improves as subgroup size decreases.

Finally, according to the account by Oakes and Turner (1986) as detailed in the Introduction, an inverted U-shaped relationship of relative subgroup size and category-based processing is predicted with a maximum of category-based processing in balanced groups. Since a functional antagonism is assumed to exist between person-based and category-based processing in this approach, the opposite prediction follows for person-based processing: Person-based processing should be minimized in balanced groups and should increase as relative subgroup size departs from 50%; a prediction that is inconsistent with the present pattern of findings for person memory. Nor is there a pronounced U-shaped pattern in category memory or stereotype ratings. Stereotype-based guessing was however maximal in balanced groups.

Oakes and Turner (1986) predict a maximum of category-based processing in balanced groups, and they test their prediction in an experiment in which participants perceive group discussions of male and female group members. Like in the present studies, there was no systematic covariation between category membership and content of the discussion statements, that is, in Oakes' (1987) terms there was little fit. In later work, Oakes (1987; cf. Oakes *et al*., 1994) stresses the role of fit for category salience and argues that fit along with accessibility determines category salience. From this perspective, the finding that group composition had no effect on category salience in memory is not unexpected, given that all experimental conditions realized equally low levels of fit.

CONCLUSION

The present results suggest that over and above distinguishing person-based from category-based processing, different modes of category-based processing can themselves be meaningfully distinguished and follow different laws. While stereotyping in ratings was most pronounced in minority subgroups, reconstructive category guessing was maximal in balanced groups. In contrast, category memory was not affected by subgroup size. While it is clearly premature to conclude on the basis of only two studies that different modes of category-based processing dissociate, the present work converges with several different lines of research that demonstrate functional dissociations between different modes of category-based processing; for example, between stereotype activation and application (e.g. Devine, 1989; Gilbert & Hixon, 1991), categorization and stereotyping (e.g. Lepore & Brown, 1997), and different functions of social categorization (e.g. Higgins, 1996) such as organizing memory and reconstructive stereotype use in filling memory gaps. When different category-based processing modes are distinguished, differences in theoretical accounts and in reported findings may in particular be found to be more apparent than real. For example, stereotyping *and* person memory appeared to be governed by the attentional factors implied by comparative distinctiveness, indicating that relative group size modulates the extent of what Brewer and Harasty Feinstein (1999) called category-based individuation. Alternatively, reconstructive category use was much better described by Oakes and Turner's (1986) functional analysis that emphasizes the perceived usefulness of different kinds of information. The dissociations observed between these different processing modes may thereby further our understanding of apparently contradictory theoretical positions and empirical findings.

ACKNOWLEDGEMENTS

The authors wish to thank Christiane Eichenberg for her help in sampling the data of Experiment 1. Thanks are also due to Edgar Erdfelder, Thorsten Meiser, and Theresa Vescio for their comments on a previous draft of this manuscript. The research reported in this paper was supported by grant Kl 614/7-1 from the Deutsche Forschungsgemeinschaft to the first author.

REFERENCES

Batchelder WH, Riefer DM. 1999. Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review* **6**: 57–86.

Bayen UJ, Nakamura GV, Dupuis SE, Yang C-L. in press. The use of schematic knowledge about sources in source monitoring. *Memory and Cognition*.

Beauvais Ch, Spence JT. 1987. Gender, prejudice, and categorization. *Sex Roles* **16**: 89–100.

Biernat M, Vescio TK. 1993. Categorization and stereotyping: Effects of group context on memory and social judgment. *Journal of Experimental Social Psychology* **29**: 166–202.

Brewer MB. 1993. Social identity, distinctiveness, and in-group homogeneity. *Social Cognition* **11**: 150–164.

Brewer MB, Harasty Feinstein AS. 1999. Dual processes in the cognitive representation of persons and social categories. In *Dual Process Theories in Social Psychology*, Chaiken S, Trope Y (eds). Guilford Press: New York; 255–270.

Brewer MB, Weber JG, Carini B. 1995. Person memory in intergroup contexts: Categorization versus individuation. *Journal of Personality and Social Psychology* **69**: 29–40.

Devine PG. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* **56**: 5–18.

Fiske ST, Lin M, Neuberg SL. 1999. The continuum model, ten years later. In *Dual Process Theories in Social Psychology*, Chaiken S, Trope Y (eds). Guilford Press: New York; 231–254.

Frable DE, Bem SE. 1985. If you are gender schematic, all members of the opposite sex look alike. *Journal of Personality and Social Psychology* **49**: 459–468.

Garcia-Marques L, Hamilton DL. 1996. Resolving the apparent discrepancy between the incongruency effect and the expectancy-based illusory correlation effect: The TRAP model. *Journal of Personality and Social Psychology* **71**: 845–860.

Gilbert DT, Hixon JG. 1991. The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology* **60**: 509–517.

Hamilton DL, Dugan PM, Trolier TK. 1985. The formation of stereotypic beliefs: Further evidence for distinctiveness-based illusory correlations. *Journal of Experimental Social Psychology* **48**: 4–17.

Hamilton DL, Rose TL. 1980. Illusory correlation and the maintenance of sterotypic beliefs. *Journal of Personality and Social Psychology* **39**: 832–845.

Higgins ET. 1996. Knowledge activation: accessibility, applicability, and salience. In *Social Psychology. Handbook of basic principles*, Higgins ET, Kruglanski AW (eds). Guilford Press: New York; 133–168.

Jackson AL, Hymes RW. 1985. Gender and social categorization: familiarity and ingroup polarization in recall and evaluation. *Journal of Social Psychology* **125**: 81–88.

Klauer KC, Meiser T. 2000. A source-monitoring analysis of illusory correlations. *Personality and Social Psychology Bulletin* **26**: 1074–1093.

Klauer KC, Wegener I. 1998. Unraveling social categorization in the 'Who said what?'-paradigm. *Journal of Personality and Social Psychology* **75**: 1155–1178.

Lepore L, Brown R. 1997. Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology* **72**: 275–287.

Lorenzo-Cioldi F, Eagly AH, Stewart TL. 1995. Homogeneity of gender groups in memory. *Journal of Experimental Social Psychology* **31**: 193–217.

Mullen B. 1991. Group composition, salience, and cognitive representations: The phenomenology of being in a group. *Journal of Experimental Social Psychology* **27**: 297–323.

Oakes P. 1987. The salience of social categories. In *Rediscovering the social group: A self-categorization theory*, Turner JC, Hogg MA, Oakes PJ, Reicher SD, Wetherell MS (eds). Blackwell: Oxford; 117–141.

Oakes PJ. 1994. The effects of fit versus novelty on the salience of social categories: A response to Biernat and Vescio. *Journal of Experimental Social Psychology* **30**: 390–398.

Oakes PJ, Haslam SA, Turner JC. 1994. *Stereotypes and Social Reality*. Blackwell: Oxford.

Oakes PJ, Turner C. 1986. Distinctiveness and the salience of social category memberships: Is there a perceptual bias towards novelty? *European Journal of Social Psychology* **16**: 325–344.

Ostrom TM, Carpenter SL, Sedikides C, Li F. 1993. Differential processing of in-group and out-group information. *Journal of Personality and Social Psychology* **64**: 21–34.

Sherman JW, Bessenoff GR. 1999. Stereotypes as source-monitoring cues: On the interaction between episodic and semantic memory. *Psychological Science* **10**: 106–110.

Simon B, Aufderheide B, Hastedt C. 2000. The double negative effect: The (almost) paradoxical role of the individual self in minority and majority members' information processing. *British Journal of Social Psychology* **39**: 73–93.

Simon B, Brown RJ. 1987. Perceived intragroup homogeneity in minority-majority contexts. *Journal of Personality and Social Psychology* **53**: 703–711.

Simon B, Hastedt C. 1997. When misery loves categorical company: Accessibility of the individual self as a moderator in category-based representation of attractive and unattractive ingroups. *Personality and Social Psychology Bulletin* **23**: 1254–1264.

Stangor C, Ruble DN. 1989. Strength of expectancies and memory of social information: What we remember depends on how much we know. *Journal of Experimental Social Psychology* **25**: 18–35.

Strack F, Bless H. 1994. Memory for nonoccurences: Metacognition and presuppositional strategies. *Journal of Memory and Language* **33**: 203–217.

Taylor SE. 1981. A categorization approach to stereotyping. In *Cognitive processes in stereotyping and intergroup behavior*, Hamilton DL (ed.). Erlbaum: Hillsdale, NJ; 83–114.

Taylor SE, Fiske ST, Etcoff NJ, Ruderman AJ. 1978. Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology* **36**: 778–793.

Turner JC. 1987. A self-categorization theory. In *Rediscovering the social group: A self-categorization theory*, Turner JC, Hogg MA, Oakes PJ, Reicher SE, Wetherell MS (eds). Blackwell: Oxford; 42–67.

van Knippenberg A, van Twuyver M, Pepels J. 1994. Factors affecting social categorization processes in memory. *British Journal of Social Psychology* **33**: 419–431.

van Twuyver M, van Knippenberg A. 1999. Social categorization as a function of relative group size. *British Journal of Social Psychology* **38**: 125–156.

APPENDIX

Multinomial models represent participants' response patterns by means of a number of different cognitive processes, each of which occurs with a certain probability. A particular combination of cognitive processes results in a certain response, and the relative frequency of the corresponding response option is predicted by the product of the probabilities of the involved cognitive processes. On the basis of these equations for the expected response frequencies, and given observed frequencies, the unknown probabilities of the different cognitive processes can be estimated

The 'Who Said What?' Model

Figure A1 shows the multinomial model proposed by Klauer and Wegener (1998) for the modified 'Who said what?' paradigm. The model assumptions are represented by means of three processing trees, one for each kind of statement, that is, for statements that were made by a speaker from Category A, for those by a speaker from Category B, and for new statements.

Consider first the tree for Category A items. The first branching of that tree models the process of item memory. With probability D_A, the participant detects a presented A item as old, whereas the item is not detected with complementary probability $1-D_A$. If the item is detected, the process of person memory is considered next in the processing-tree representation. With probability c_A the retrieved information about statement and speaker is sufficient to identify the speaker. In this case, the participant responds with the correct assignment. With probability $1-c_A$, however, the information does not suffice to identify the speaker. If the speaker's social category has been encoded along with the statement, the participant may then still remember that category. The probability for category memory is modeled by the parameter d_A. If the category is recalled, but not the individual speaker, then the model assumes that one of the persons of category A is guessed to be the speaker. With a fixed probability $1/n_A$, guessing results in a correct assignment, whereas with complementary probability

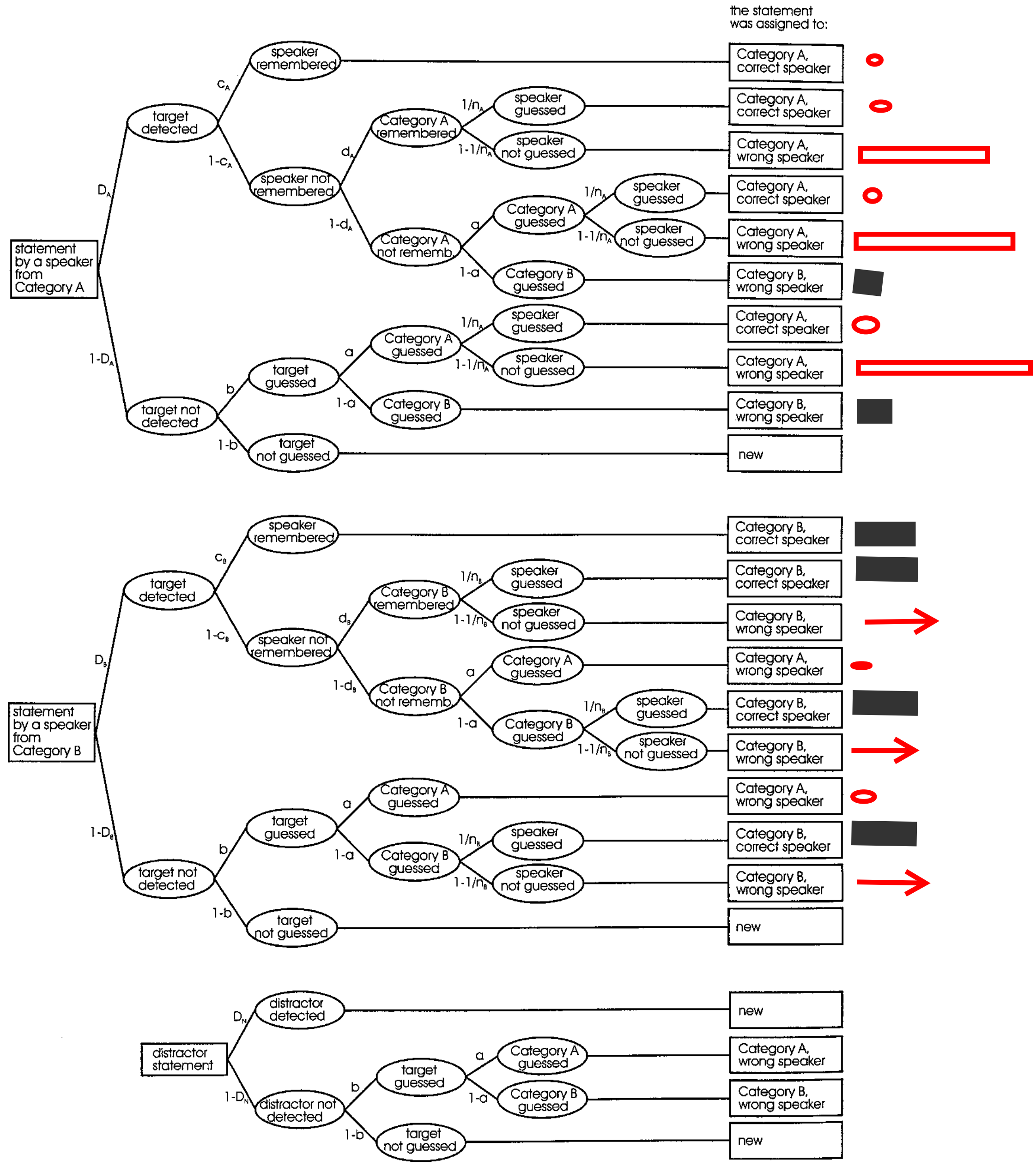

Figure A1. The two-high-threshold multinomial model of social categorization in the modified 'Who said what?' paradigm. D_A = probability of detecting a statement made by a speaker from Category A, D_B = probability of detecting a statement made by a speaker from Category B, D_N = probability of detecting that a distractor is new; c_A = probability of correctly discriminating the speaker of a statement made by a speaker from Category A, c_B = probability of correctly discriminating the speaker of a statement made by a speaker from Category B; d_A = probability of correctly discriminating the category of a statement made by a speaker from Category A, d_B = probability of correctly discriminating the category of a statement made by a speaker from Category B; a = probability of guessing that a statement was made by a speaker from Category A; b = probability of guessing that a statement is old; n_A = size of Category A, n_B = size of Category B

$1-1/n_A$, a within-category error is made, where n_A is the size of Category A. The further branches and partial trees can be read analogously as elaborated by Klauer and Wegener (1998). Although person memory comes before category memory in the tree representation of Figure 1, an assumption about the psychological order of these processes is not implied by the model. In particular, an equivalent tree representation with reversed order of these processes can be given.

Parameter Estimation

For the statistical analysis of a multinomial model the data are aggregated over all participants of a given experimental condition. For every partial tree and response option, the frequency with which the response option was chosen by the participants is obtained.

As has been said, the probability with which a response option is chosen can be expressed as the product of the probabilities of the involved cognitive processes along a given branch or as the sum of joint probabilities of this kind if more than one branch leads to the same response option. Hence the whole multinomial model can be described as a system of equations: for every response option of each partial tree one equation is obtained.

Because the frequencies of the response options can be observed and the same parameters appear in several independent equations, the resulting system of equations can be solved for the unknown parameters under certain conditions, and the probabilities of the involved cognitive processes can thereby be estimated. The job of estimating the probabilities is performed by a computer program by means of the maximum likelihood method (Batchelder & Riefer, 1999).

Testing Hypotheses

In applying a model, a global goodness-of-fit test of the model is conducted first. A goodness-of-fit test evaluates the differences between the observed and the estimated expected response frequencies. To test the goodness of fit, the likelihood ratio statistic is computed, which is asymptotically χ^2 distributed. If this test does not yield a significant result, given a satisfactory statistical test power, it can be concluded that the ensemble of assumptions defining the model can be maintained. In other words, the model fits the data. Note that for the goodness-of-fit test the degrees of freedom have to be computed. As a rule of thumb, they are given by the number of independent response options minus the number of parameters to be estimated.

Special hypotheses can be tested by restricting the model. Usually, a hypothesis assumes that a certain treatment affects a certain cognitive process and hence will cause the probability of this process to change. Fitting the model to the experimental condition should thus yield an estimated probability of this process which differs from the one obtained by fitting the model to the control condition. To test this hypothesis statistically, the model is doubled and treated as one joint model for both experimental and control condition. The parameters of the resulting model are indexed by condition, thereby allowing for different probabilities in the two situations, especially allowing for different probabilities of the cognitive process that is supposed to have been manipulated. Then the change in these latter probabilities can be tested statistically by using a goodness-of-fit test again. If there is indeed a change, it should be impossible to restrict the model to use only one parameter for both probabilities, thereby forcing them to be equal. Thus, if the restriction leads to a significant loss of goodness of fit, the treatment can be concluded to have affected the cognitive process in question. The loss of goodness of fit is the difference of the goodness of fit of the restricted model minus that of the unrestricted model. The degrees of freedom for testing the significance of this χ^2 distributed loss value are given by the difference in the degrees of freedom of the two models.

