Journal of Personality and Social Psychology
1998, Vol. 75, No. 5, 1155–1178

0022-3514/98/$3.00

Unraveling Social Categorization in the "Who Said What?" Paradigm

Karl Christoph Klauer and Ingo Wegener
Rheinische Friedrich-Wilhelms-Universität Bonn

A multinomial model of the "Who said what?" paradigm (S. E. Taylor, S. T. Fiske, N. J. Etcoff, & A. J. Ruderman, 1978) explains the pattern of participants' assignment errors by means of the joint operation of several processes. Specifically, memory for discussion statements, person memory, category memory, and 3 different guessing processes can be accommodated by the model. The model's ability to disentangle these processes is validated in a series of 5 experiments. The model thereby enables a more refined use of the "Who said what?" paradigm in testing theories of social categorization. This is demonstrated in a 6th experiment in which the validated model is applied to the study of the effects of cognitive load on categorization.

Fundamental to the process of stereotyping is the act of social categorization (Taylor, 1981), and the role of categorization in stereotyping has been recognized for many decades (Allport, 1954; Lippmann, 1922). Moreover, it has been argued that the process of categorization is central to impression formation (Brewer, 1988; Fiske & Neuberg, 1990), and it is often assumed to have priority over person-based processes.

Most current theories of stereotyping and impression formation regard social categories as cognitive structures organizing information about the members of the category, although there is some debate about the exact organization and content of such knowledge structures (Hamilton & Sherman, 1994). Categorizing a person as a member of a social category activates this stored information. In addition, the categorization process has been argued to have a number of general cognitive and motivational consequences (e.g., Fiske & Neuberg, 1990; Macrae, Milne, & Bodenhausen, 1994; Tajfel, 1969, 1972; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987).

Much theoretical debate and empirical research is clustered around the following questions (Hamilton & Sherman, 1994): Why do people categorize? What makes them prefer one category over another, and what is the relationship of person-based and category-based processes?

The so-called "Who said what?" paradigm introduced by Taylor, Fiske, Etcoff, and Ruderman (1978, Experiments 1 and 2) has been of prime importance in addressing these and related questions. In Taylor et al.'s Experiment 1, participants were asked to listen to several tape-recorded statements of six people who were engaged in a discussion. Each statement was presented along with the speaker's photograph. Three speakers were Black persons, three were White persons. On presentation of all statements, a list of the statements and the speakers' pictures were shown, and participants were asked to assign each statement to the person who had made it.

Errors in this task can be classified into different kinds. Within-category errors occur if a statement is assigned to a wrong person who is, however, a member of the same category as is the speaker of the statement (e.g., a statement made by a Black person is assigned to another Black person). Between-categories errors occur if a statement is assigned to a person from the wrong category. In the Taylor et al. (1978) experiments, within-category errors were found to be more frequent than between-categories errors (a chance correction of observed frequencies is necessary as explained below), suggesting that category membership was encoded and used in assigning statements.

The finding of more within-category than between-categories errors turns out to be extraordinarily stable; it has been observed for many different categories, and the difference in error rates has frequently been used as a dependent variable to measure the salience of social categories. An important rationale for this practice has been the traditional assumption that as a result of social categorization, within-group differences are minimized and between-groups differences are exaggerated (e.g., Taylor et al., 1978), leading one to expect that the strength of the categorization process is reflected in the differential likelihood of confusions within versus between categories.

Table 1 gives an overview of the 50 studies that have employed the "Who said what?" paradigm (including closely related name-matching and picture-matching tasks). For each experiment, the question and the social categories that were used are listed. Most of the studies in Table 1 used the difference of within-category and between-categories errors as the dependent variable, or equivalently conducted analyses of variance (ANOVAs) with kind of error as a within-subjects factor. Prior to data analysis, the between-categories errors were generally multiplied by a constant, $(n - 1)/n$, in order to take into account that there are more possibilities for confusions between categories (i.e., all n members of the other category) than within categories (all members of the category minus the speaker: $n - 1$).

As can be seen, the "Who said what?" paradigm has been applied to most of the research issues in the field. The paradigm

Karl Christoph Klauer and Ingo Wegener, Psychologisches Institut, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany.

We thank Wolfgang Wasel for the pictures of male and female speakers. Thanks also go to Fetneh Schimmer for collecting the data of the sixth experiment. Mathias Blanz, Markus Brauer, Edgar Erdfelder, and Russell Spears provided helpful feedback on a first draft of the article.

Correspondence concerning this article should be addressed to Karl Christoph Klauer, Psychologisches Institut, Rheinische Friedrich-Wilhelms-Universität Bonn, Römerstr. 164, D-53117 Bonn, Germany. Electronic mail may be sent to christoph.klauer@uni-bonn.de.

Table 1
"Who Said What?"—A List of Studies

Experiment	Question	Categories
Arcuri (1982)	How do multiple categories interact?	Sex, academic status
Beauvais & Spence (1987)	Can Taylor and Falcone (1982) be replicated?	Sex
Biernat & Vescio (1993), Studies 1 and 2	Is categorization a function of distinctiveness? Is there categorization on the basis of attitude position?	Race, attitude position
Blanz (1996), Study 2	What is the interaction of self and social identity, and self-identification?	Educational status, hometown
Blanz (1997)	What is the role of fit and category membership?	University major, university town
Blanz (in press), pretest	Is there categorization for a number of weakly accessible categories?	Educational status, hometown
Blanz (in press), Study 1	What is the role of fit for weak categories?	Educational status, hometown
Blanz (in press), Studies 2 and 3	Is category salience a function of Accessibility × Fit?	Educational status, hometown
Blanz & Aufderheide (in press)	What is the effect of fit on category salience?	University major, university town
Brewer, Weber, & Carini (1995), Study 1	Is categorization a function of perceived meaningfulness of the category?	Color of clothing
Brewer et al. (1995), Study 2; cf. Brewer & Harasty (1996)	Is categorization a function of category membership and intergroup competition?	Color of clothing
Brewer et al. (1995), Study 3; cf. Brewer & Harasty (1996)	What is the effect of minority versus majority status of in- and out-group?	Color of clothing
Frable & Bem (1985), Study 1	Is sex categorization a function of sex role and category membership?	Sex
Frable & Bem (1985), Study 2	Is race categorization a function of sex role and category membership?	Race
Hewstone, Hantzi, & Johnston (1991), Study 1	Is there more categorization if the discussion topic is relevant, and what is the role of category membership?	Race
Hewstone et al. (1991), Study 2	Is there less categorization if interaction with discussion participants is expected?	Race
Jackson & Hymes (1985)	Is category salience enhanced by issue relevance?	Sex
Johnston, Hewstone, Pendry, & Frankish (1994), Study 3	Is there categorization based on stereotype incongruency (subtyping)?	Stereotype congruency
Judd & Park (1988)	What is the basis of the out-group homogeneity effect?	Minimal groups
Lorenzo-Cioldi (1993)	Can out-group homogeneity be shown in memory-based categorizations?	Sex, abstract stimuli
Lorenzo-Cioldi, Eagly, & Stewart (1995)	Is the effect of out-group homogeneity mediated by group status?	Sex, color label
Miller (1987)	What is the relationship of categorization and stereotyping?	Sex
Miller (1988), Study 1	Is attractiveness encoded and retrieved?	Attractiveness
Miller (1988), Studies 2–5	What is the relationship to stereotyping?	Attractiveness
Ostrom, Carpenter, Sedikides, & Li (1993), Studies 1 and 2	Is in-group information processed differently from out-group information?	Sex
Ostrom et al. (1993), Study 3	Is in-group information processed differently from out-group information?	University major
Simon & Hastedt (1997a)	Do members of majorities categorize more strongly than members of minorities?	Preference for urban versus rural life
Simon & Hastedt (1997b), Studies 1 and 2	What is the relationship of social and personal identity?	Preference for urban versus rural life
Spears & Haslam (1997), Studies 1 and 2	What is the role of processing load?	Sex
Spears, Haslam, & Jansen (1996)	What is the role of processing load?	Sex
Stangor, Lynch, Duan, & Glass (1992), Study 1	How do multiple categories interact? What is the effect of category priming?	Sex, race
Stangor et al. (1992), Study 2	What is the role of attention and of physical similarity?	Sex, race
Stangor et al. (1992), Study 3	Is there more categorization for prejudiced perceivers?	Sex, race
Stangor et al. (1992), Study 4	Is categorization goal-dependent? Is there categorization based on physical features?	Sex, clothing style
Stangor et al. (1992), Study 5	Is there categorization based on salient, but useless physical features?	Sex, color of clothing
Taylor & Falcone (1982)	Is sex categorization a function of sex role?	Sex
Taylor, Fiske, Etcoff, & Ruderman (1978), Study 1	Is race encoded and used at retrieval?	Race
Taylor et al. (1978), Study 2	Is sex encoded and used at retrieval?	Sex
van Knippenberg, van Twuyver, & Pepels (1994)	What is the role of topic relevance and structural and normative fit?	Sex, academic status
van Twuyver (1996, chap. 4)	What is the role of numerical group composition?	Sex, academic status
van Twuyver & van Knippenberg (1995)	What is the role of category priming for weakly accessible categories?	University major, university town
Walker & Antaki (1986)	Is there categorization of sexual orientation, and is it a function of homophobia?	Sexual orientation

continues to attract attention because it provides an elegant and unobtrusive measure of social categorization.

In the present article, the different memory processes and guessing processes that contribute to the participants' assignment errors in the "Who said what?" task are analyzed. It is shown that the conventional analyses of "Who said what?" data are frequently misleading and problematic to interpret. To disentangle the different confounded processes, a mathematical model of the "Who said what?" task is then introduced and empirically validated in a series of five experiments. The model provides a more refined as well as a more valid use of the paradigm in social psychological research. As will be seen, the model not only overcomes the interpretational problems of the conventional analysis but by providing independent measures of a number of different processes also allows one to address new theoretical questions in the framework of the "Who said what?" paradigm that were not easily addressed previously. The objective of the present article is therefore to develop and validate a new tool for the use of social psychologists in the study of social categorization. The model is readily implemented on the basis of available software, and Appendix A contains an easy-to-read guide detailing the steps involved in conducting the new analyses practically.

The usefulness of the new method is demonstrated in a sixth experiment that investigates the effects of cognitive load on social categorization. We conclude the article by discussing further areas of application where the model can help to answer open questions posed by competing theories of impression formation, social categorization, and stereotyping.

Problems of the Conventional Approach

The analysis on the basis of the error-difference measure is often potentially misleading or at least difficult to interpret, because many different memory and guessing processes jointly contribute to the pattern of the participants' assignment errors. For this reason, it is often impossible to trace back effects on the error-difference measure (or on other ad hoc indices of social categorization; see below) to the social-categorization process rather than to one of the other involved processes. This point can be underlined by means of three examples, corresponding to the three confounded processes of (a) item discrimination, (b) person discrimination, and (c) expectancy-based guessing.

Confounding With Item Discrimination

One problem of the previous applications of the paradigm arises because participants do not have the option to say that they do not remember the item at all. From the results of standard recognition tasks it is known, however, that participants frequently fail to discriminate old from new items so that item memory is likely to be less than perfect. If item discrimination fails—that is, if the statement is not remembered—the participant has little choice other than to guess the speaker. When the content of the statements is not correlated with category membership (i.e., in situations without structural fit; Oakes, 1987), such guessing is likely to be blind with respect to the speaker's category, so that both kinds of errors are equally likely after correcting between-categories errors $(n - 1)/n$. Consequently, if item discrimination is poor, differences between within-category and between-categories errors are leveled because the proportion of assignments based on blind guessing is high. This confounding of item discrimination and social categorization in the error-difference measure is likely to cause misleading interpretations when factors are manipulated that also affect the level of item discrimination.

For example, studies manipulating processing load (e.g., manipulating processing pace or the amount of information presented; cf. Table 1) or motivation to process (e.g., Brewer, Weber, & Carini, 1995; Hewstone, Hantzi, & Johnston, 1991; Judd & Park, 1988) are likely to affect not only social categorization but also other factors such as memory for the statements. In such studies, decreasing item discrimination rather than decreasing social categorization may thus explain decreases in the error-difference measure (e.g., Spears & Haslam, 1997; Spears, Haslam, & Jansen, 1996). In the extreme case, where there is no item memory and all assignments are based on category-blind guessing, the error-difference measure cannot be reliably different from zero even if the speakers' social-category membership is highly salient.

Confounding With Person Discrimination

Social perception of individuals is often analyzed as consisting of category-based and individuating, person-based components. Fiske and Neuberg (1990) proposed a model of impression formation in which social perceivers may form an impression of a person anywhere on a continuum from category based to person based. Where the impression lies on this continuum depends, among other things, on the interest the perceiver has in the target and the consequent attention to the target's individuating features. Similarly, Brewer (1988) and Pratto and Bargh (1991) have proposed that perceived persons are represented in memory separately in terms of their individuating features as well as their social-category memberships.

Attending to the speakers' individuating features increases the likelihood that such features will be encoded. In the context of the "Who said what?" paradigm, the extent of person discrimination (i.e., the accuracy of memory for the speaker) may thereby be enhanced. As a consequence, both within-category and between-categories errors decrease, and along with them, the possible range of the error-difference measure decreases. In the extreme case, where there are no erroneous assignments, the error difference is necessarily zero even though the speakers' social-category memberships may be highly salient. For a given level of item discrimination, the error-difference measure is therefore also a function of the confounded process of person discrimination. For example, studies manipulating outcome dependency (e.g., Brewer et al., 1995, Experiment 2; Judd & Park, 1988) are likely to affect not only social categorization but also perceivers' attention to the speakers' individuating features (Fiske & Neuberg, 1990). Different levels of person discrimination rather than differences in the extent of social categorization may thus explain observed effects on the pattern of assignment errors in such studies.

Confounding With Expectancy-Based Guessing

Many of the studies in Table 1 have implemented a correlation between the content of the statements and category membership,

thereby establishing structural fit. If the covariation corresponds to preexisting expectancies (i.e., to stereotypes), so-called normative fit is obtained (Oakes, 1987). If item discrimination or person discrimination or both fail, the participant has little choice other than to guess a speaker. Fit can be exploited by guessing processes to constrain the set of probable speakers for given statements resulting in an increased error-difference measure due to expectancy-based guessing (Spears & Haslam, 1997, p. 183). For example, if Category A members have predominantly made proabortion statements, and Category B members have made anti-abortion statements, then a reasonable guessing strategy is to assign proabortion statements to persons from Category A and anti-abortion statements to persons from Category B.

In the case of normative fit, a reliable positive error difference can even be achieved through stereotype-congruent guessing in the absence of any memory for the individual statements or their speakers. Stereotype-congruent guessing can be used to this effect even by respondents who have not observed the discussion itself and for whom the speakers and statements are new. If, on the other hand, only structural fit is given, the expectancy must have been formed on-line, while the discussion was being observed, to become effective as a response bias in assignments. Response biases based on expectancy are well documented in standard recognition tasks (Stangor & McMillan, 1992), strongly suggesting that similar biases may also contribute to the pattern of assignments in the "Who said what?" paradigm.

Expectancy-based guessing increases the error-difference measure independently from actual memory for the individual statements or their speakers' social-category memberships. Studies that have manipulated the extent of structural and normative fit (cf. Table 1) in particular may have manipulated and then measured the effectiveness of this guessing strategy rather than any real differences in social categorization.

Summary

There is some awareness in the literature of the different confounded processes illustrated in the above examples as reflected in discussions about the appropriateness of the error-difference measure. For example, additional corrections have been proposed to account for the fact that the categories may attract different total numbers of assignments (Taylor et al., 1978). Similarly, it has been suggested to use percentages of errors relative to the total number of errors made by participants (Miller, 1987), to use the ratio rather than the difference of within-category versus between-categories errors (Spears & Haslam, 1997), or to focus only on the confusions within categories (Brewer & Harasty, 1996; cf. Brewer, Weber, & Carini, 1995; Simon & Hastedt, 1997b). The arguments exchanged in this debate recognize the fact that different processes contribute to the pattern of errors in the "Who said what?" paradigm, suggesting different corrections and dependent measures. From the point of view of the participant, at least the following is involved:

> *Item discrimination:* Does the participant remember the statement that is to be assigned?
>
> *Person discrimination:* Provided item discrimination is given, does the participant recall the speaker?
>
> *Category discrimination:* If item discrimination is given, but person discrimination fails, does the participant remember and use the speaker's category membership?
>
> *Guessing:* If the available information is not sufficient to identify the speaker, the participant must guess the speaker from a set of possible candidate speakers. This set may be a subset of the total set of speakers, and, as specified in the following section, different kinds of guessing processes are likely to be engaged in the paradigm.

It is reasonable to assume that the observed error frequencies are determined by the joint operation of these processes. Consequently, it is problematic to consider the error-difference measure or, for that matter, any other ad hoc index as a measure of only one process such as category discrimination even though intuition may suggest that the error-difference measure is particularly closely tied to social categorization. The problems that arise were in fact exemplified in this section by means of three examples presented in terms of the error-difference measure. It is not difficult to see that similar arguments can be leveled against the other ad hoc indices mentioned above.

The objective of the present article is to propose and validate a substantive model of the processes involved in the "Who said what?" paradigm. Based on a small modification of the paradigm, the model aims at disentangling the separate contributions of these different processes, thereby overcoming the problems of the conventional approach.

A Model of the "Who Said What?" Paradigm

As has been said, one of the problems with the conventional approach arises because participants do not have the option to say that they do not remember a statement at all. For this reason, among others, we propose to modify the assignment phase of the paradigm. Apart from the statements that occur in the discussion, henceforth called *targets* or *old statements*, *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.

This small modification of the assignment phase of the "Who said what?" paradigm provides a somewhat richer database than the previous applications of the task. In what follows, it will be shown that the separate impacts of the different processes that jointly cause the pattern of assignment errors can be disentangled on the basis of such data. It is evident, for example, that the use of distractor statements allows one to assess the extent of item discrimination. For this purpose, the old–new judgment can be analyzed by means of a signal-detection model. The old–new discrimination is, however, only part of what is required in the modified assignment task. Those statements recognized as old are subsequently assigned to speakers, yielding additional assignment data to which signal-detection-type models can again be applied to disentangle the separate contributions of person discrimination, category discrimination, and guessing. The model proposed below can be understood as organizing these different signal-detection analyses into one psychologically motivated process model of the "Who said what?" paradigm.

The Data Matrix

Table 2 presents the basic data matrix for the modified paradigm. The three rows of the table specify the origin of the statements. There are (a) statements made by persons from Category A, (b) statements made by persons from Category B, and (c) new statements. The four columns code the participants' assignments according to whether the statement was judged new (column 4 of Table 2) or whether the statement was judged old and then assigned to the correct speaker (column 1), a wrong speaker of Category A (column 2), or a wrong speaker of Category B (column 3).

Because new statements have no speaker, they cannot be correctly assigned to a speaker, and thus the cell in the lower left corner of Table 2 necessarily remains empty. The data matrix therefore constitutes a contingency table with 11 cells, numbered 1 to 11 in Table 2. In the analyses reported below, the cell entries are the frequencies with which the different kinds of assignments are made by the participants.

The original paradigm defines a reduced data matrix that is given by cells 1 to 3 and 5 to 7 of Table 2; that is, the rows and columns for the new statements are deleted. Cells 2 and 7 correspond to within-category errors, cells 3 and 6 to between-categories errors. In the analysis based on the error-difference measure, the data from cells 2 and 7 for within-category errors and from cells 3 and 6 for between-categories errors are pooled, whereas the correct assignments in cells 1 and 5 do not contribute to the analysis at all. In the modified paradigm, the database is supplemented by cells 4 and 8 to 11, and as will be seen, the information from all cells is used.

A Multinomial Model

Figure 1 shows a processing-tree representation (Hu & Batchelder, 1994) of a so-called multinomial model (Riefer & Batchelder, 1988) that describes participants' responses by means of the processes of item discrimination, person discrimination, and category discrimination as well as three guessing processes. It is a model for contingency tables of the form of Table 2, in which the cells contain the frequencies of the different kinds of responses. The model is related to similar multinomial models that have been developed and validated in the context of the source-monitoring paradigm (Batchelder, Hu, & Riefer, 1994; Batchelder & Riefer, 1990; Bayen, Murnane, & Erdfelder, 1996; Erdfelder, Murnane, & Bayen, 1995; Riefer & Batchelder, 1988; Riefer, Hu, & Batchelder, 1994). A comprehensive review of the theory and applications of multinomial modeling was given by Batchelder and Riefer (in press).

Table 2
Data Matrix of the Modified Paradigm

Source of statement	Assignment: To the correct speaker	To a wrong speaker from Category A	To a wrong speaker from Category B	To the set of new statements
A speaker from Category A	1	2	3	4
A speaker from Category B	5	6	7	8
New statement		9	10	11

Note. A new statement that is assigned to a speaker rather than to the set of new statements can be assigned only to a wrong speaker, because it was never made by any speaker in the discussion phase. Therefore, the cell in the lower left corner must remain empty. The other cells are numbered from 1 to 11.

Multinomial models can be characterized as discrete-state models. They are discrete models in the sense that they postulate only a finite number of processing states, represented as the nodes of the processing trees depicted in Figure 1. The above-mentioned link to signal-detection models in particular is therefore established by discrete rather than continuous models of signal detection. The familiar Gaussian model of signal detection that separates hits and false alarms into d' and a response criterion is a continuous model. In contrast, the present discrete-state model assumes discrete processing states for item recognition: There is a state in which the participant correctly detects targets as old, there is another state in which distractors are correctly detected as new, and there are states in which the participant is uncertain about the status of the item and in which guessing processes prevail. Technically, the present model therefore builds on a particular discrete model of signal detection called the *two-high-threshold* model as explained in Appendix B. The relative merits of discrete versus continuous models of signal detection are discussed in Appendix B along with other technical issues.

In the remainder of this section, we present the assumptions of the multinomial model and discuss the techniques for estimating the model parameters and for testing hypotheses. Appendix A elaborates on this information and spells out the steps involved in applying the model in more detail. It is aimed at readers who are not familiar with multinomial modeling. Appendix A also contains a brief discussion of the available software for conducting the analyses.

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. In Figure 1, the item categories are shown on the left side and the response categories in rectangles on the right side. In between, the mediating latent processes are depicted within ellipses. The item categories correspond to the rows of the data matrix, and thus each tree models the responses in one row of that matrix. The response categories, on the other hand, map the columns of the basic data matrix (see Table 2). As can be seen, some of the response categories appear repeatedly because different processes can result in the same observable response.

Consider first the tree for A items. The first branching of that tree models the process of item discrimination. With the 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 discrimination is considered next in the processing-tree representation. With probability c_A, the retrieved information about statement and speaker is sufficient to reconstruct the speaker's identity. 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 partici-

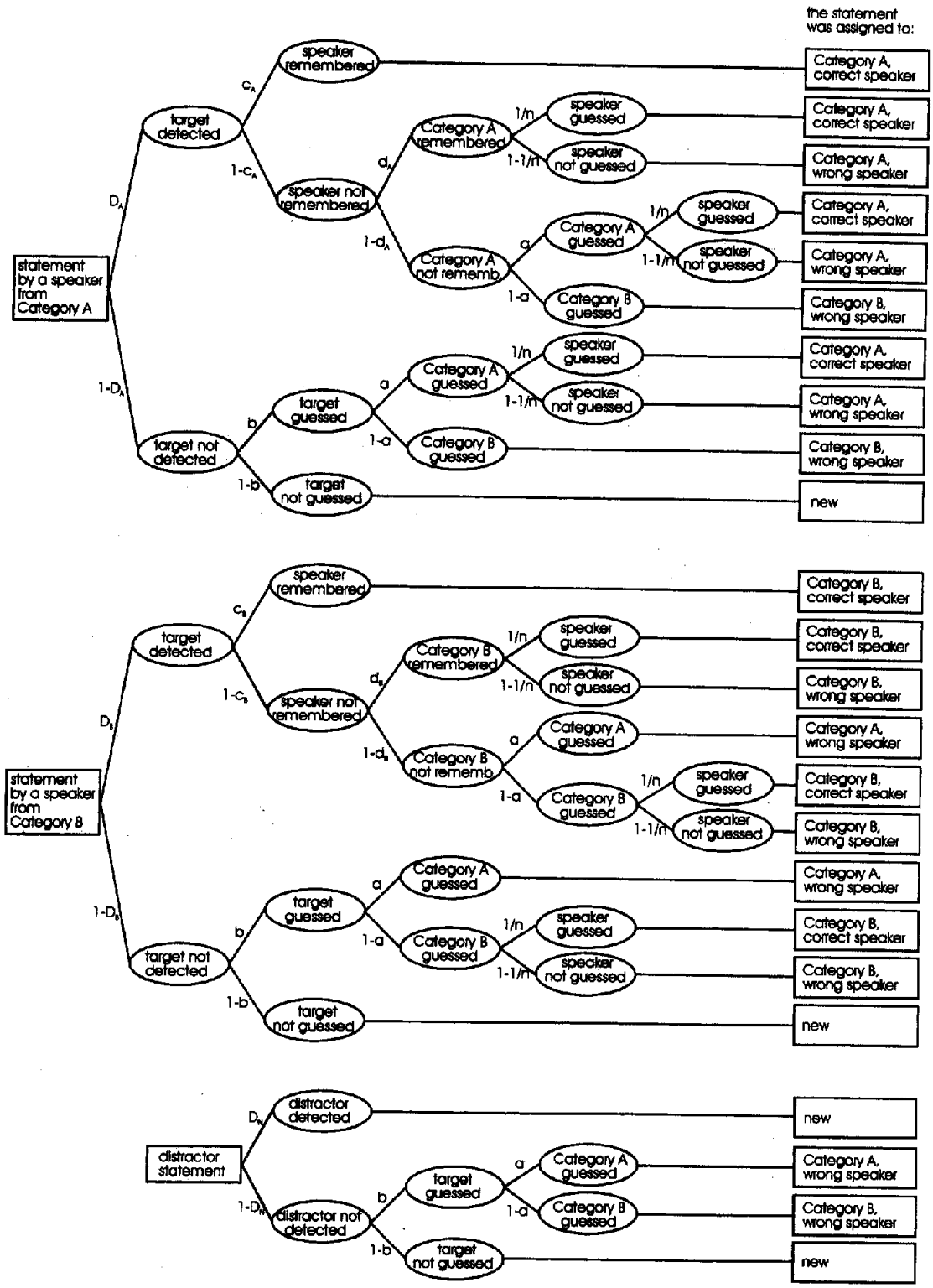

Figure 1. The two-high-threshold multinomial model of social categorization in the modified "Who said what?" paradigm. D_A = probability of detecting a statement made by a speaker from Category A; D_B = probability of detecting a statement made by a speaker from Category B; D_N = probability of detecting that a distractor is new; 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; c_A, c_B = person discrimination parameters; a = probability of guessing that a statement is made by a speaker from Category A; b = probability of guessing that a statement is old.

pant may then still remember that category. The probability for successful category discrimination is modeled by the parameter d_A.[1] If the category is recalled, but not the individual speaker, then the model assumes that one of the n persons of Category A is guessed to be the speaker. With a fixed probability $1/n$, guessing results in a correct assignment, whereas with complementary probability $1 - (1/n)$, a within-category error is made.

With probability $1 - d_A$, category discrimination does not succeed. In this case, the participant can still guess the speaker's category. The probability of guessing Category A rather than B is given by parameter a. Note that this bias parameter ensures that response biases favoring one of the categories over the other can be accomodated by the model. If Category A is guessed, a second guess picks the speaker with probability $1/n$ and a wrong person within that category with probability $1 - (1/n)$ as above. If instead Category B is guessed, the participant necessarily responds with a between-categories error.

Consider now the lower half of the tree for A items. It describes the case that item discrimination fails (probability $1 - D_A$). The model assumes that the participant then guesses whether the item is old or new. The probability of guessing old is given by the bias parameter b. Again, b need not equal .5, so that the model takes into account that participants may be biased toward one of the response alternatives, "old" or "new."

If the participant guesses that the statement is old, the same processes as described above for the case of failing person discrimination and category discrimination are assumed to occur. That is, both a category and a person therein are guessed. If the participant guesses, on the other hand, that the statement is new, he or she falsely responds "new" with probability $1 - b$.

The processing tree for statements made by a speaker from Category B is analogous. Different discrimination parameters, indexed by the category, are assumed, however, so that the strengths of the processes may differ as a function of category.

The tree for distractors is simpler. With probability D_N, the item is correctly discriminated as new. Such discrimination can be based, for example, on so-called autonoetic processes (Strack & Bless, 1994), which are inferences based on peculiarities of distractors that allow the participant to reason that he or she would surely have recognized this particular statement if it had been presented at all. In this case, the participant responds "new." Otherwise, with probability $1 - D_N$, the same subtree of processes follows as in the other trees after failed item discrimination with the only difference being that there can be no correct assignment for distractors.

In summary, the different cells of the data matrix (Table 2) are reached by means of the joint operation of different processes along different processing paths. On the basis of the parameters for the individual processes, expected frequencies for the cells of Table 2 can be computed. It is possible in particular to express the conventional error-difference measure as a function of these parameters to demonstrate its complex dependence on these different processes.[2] Conversely, on the basis of the observed frequencies, the parameters can be estimated by means of the maximum-likelihood method; the model fit can be assessed by comparing observed and expected frequencies, and hypotheses about parameter values can be tested (Hu & Batchelder, 1994; Riefer & Batchelder, 1988) as explained in Appendix A.

The following, in part conditional processes and corresponding parameters, are considered:

Item discrimination: Parameters D_A, D_B, and D_N according to category of statement.

Guessing the status of the statement (old rather than new): Parameter b.

Person discrimination: Parameters c_A, c_B according to the speaker's category.

Category discrimination: Parameters d_A, d_B according to the speaker's category.

Guessing the category (A rather than B): Parameter a.

Guessing the person within the correct category: Fixed probability of success $1/n$, where n is the category size.

In all, there are nine parameters. The data matrix has 11 cells, and because the three row marginals are fixed, it provides only eight degrees of freedom. Therefore, an additional assumption is needed to obtain an identified model as elaborated below.

The separate estimation of the different process parameters allows for a differentiated evaluation of the role of the different processes. The D parameters assess the extent of item memory, the d parameters quantify the strength of social categorization in memory, the c parameters reflect the strength of individuating information, and the a parameter can be used to evaluate the amount of expectancy-congruent guessing as will be seen. The b parameter, finally, captures response bias in item detection. Remember that the problem with the conventional analysis is that it fails to disentangle the confounded processes of item, category, and person discrimination as well as expectancy-congruent guessing (cf. the Problems of the Conventional Approach section). It is an important objective in this article to show that the model parameters D, d, c, and a provide independent and unconfounded measures of these processes. For this purpose, it is necessary to validate the model and its assumptions empirically.

Validating the Model

Overview

Validating the model has two aspects, one statistical, the other empirical. The goodness of fit of a multinomial model can be

[1] Introducing person discrimination before rather than after category discrimination partials out those cases of correct category discrimination from the measure d of category memory that occur as a consequence of correct memory for the individual speaker. This concurs with the conventional procedure, in which only assignment errors are used in the measure of category salience. The probabilities d_A and d_B are thereby defined as the conditional probabilities of correct category discrimination, given that person discrimination fails.

[2] The error-difference measure (proportion of within-category errors minus proportion of between-categories errors) is the following function of the model parameters:

$$D_A(1 - c_A)d_A[1 - (1/n)] + D_A(1 - c_A)(1 - d_A)a[1 - (1/n)] + (1 - D_A)b\,a[1 - (1/n)] + D_B(1 - c_B)d_B[1 - (1/n)] + D_B(1 - c_B)(1 - d_B)(1 - a)[1 - (1/n)] + (1 - D_B)b(1 - a)[1 - (1/n)] - [D_A(1 - c_A)(1 - d_A)(1 - a) + (1 - D_A)b(1 - a) + D_B(1 - c_B)(1 - d_B)a + (1 - D_B)b\,a]. \quad (1)$$

evaluated by means of a chi-square test for a given data set (Hu & Batchelder, 1994). Although statistical fit is a necessary condition for a justified use of the model in data analyses, it does not guarantee that the different parameters measure those processes that they are intended to operationalize. An additional experimental validation for each process and associated parameter, or group of parameters, is therefore desirable (Bayen et al., 1996; Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995; Erdfelder et al., 1995).

The experimental validation is based on the following rationale: For each process, a validating experiment is conducted in which conditions are realized and compared that can be expected on the basis of prior research and established theories to differ primarily with respect to that process. The model is then fit to the different conditions with separate parameters for each condition. The validation is successful if the differences between conditions are mapped onto the parameters assumed to measure the manipulated process (convergent validity) and if simultaneously there are no substantial differences in the parameters associated with other processes (discriminant validity); in this sense, the experimental validation tests for both convergent and discriminant validity (Campbell & Fiske, 1959) of the different process parameters. The validation rests on the availability of experimental manipulations that are relatively process-pure in the sense that each manipulation primarily affects only one of the involved processes. Of course, for each validating experiment, overall statistical model fit is also a prerequisite for a successful validation of the model.

In the present case, there are five different processes associated with five kinds of parameters: *D*, *c*, and *d* parameters and bias parameters *a* and *b*. In the sequel, five validating experiments are discussed, one for each kind of parameter. We present the experiments in the order in which they were conducted. In all experiments, the chosen significance level was .05, although we also report the individual probability values. To minimize inflation of the significance level due to multiple significance tests, we used omnibus tests whenever possible and in particular in testing for discriminant validity.

The validating experiments aim at using established theories and prior research results to implement specific and focused manipulations of the different processes involved in the "Who said what?" paradigm, and they are not intended to contribute to the development of current theories of social categorization and impression formation. Once the model has been empirically validated, however, it becomes a powerful measurement tool that can be used to test social psychological theories of social categorization and impression formation in a deeper manner than was possible before and that opens up new avenues of research that were not accessible before. We illustrate this use of the validated model below in a first application that addresses the effects of cognitive load on social categorization by means of the new method.

Experiment 1: Guessing "Old" Versus "New"

The bias parameter *b* quantifies response bias in the process of guessing whether the statement is old or new when the participant does not remember the statement or cannot identify the statement as a distractor. Manipulating this guessing process is relatively straightforward. It is well-known that bias of this kind varies as a function of the proportion of distractors that is presented. With increasing proportion of distractors, participants tend to become more cautious in guessing that a statement is old, and the *b* parameter should thus decrease (e.g., Buchner et al., 1995).

In Experiment 1, two groups of participants underwent the modified "Who said what?" paradigm based on gender categories. Participants in the first group judged equal numbers of old and new statements; the second group received three times as many new statements as old statements. The instructions for the assignment phase informed participants about the proportion of new statements they were about to judge. These were the only differences between groups, and the hypothesis was that we would obtain a substantial and significant difference between groups in the *b* parameter and no substantial differences in the other model parameters.

Method

Participants. Participants were 21 male and female students from different faculties of the University of Bonn. All participants were native speakers of German, and they received either partial course credit or a small monetary reward for their participation. Eleven participants were assigned to the condition with (comparatively) few distractors and 10 to the condition with many distractors in a quasi-random assignment procedure.

Pool of statements. A pool of statements was compiled on the basis of statements made by other students in an evaluation of the university and the Psychological Institute. Statements referring to special conditions at the Psychological Institute were, however, excluded. The remaining 202 statements addressed a wide range of topics, notably conditions at the library (e.g., "The library should have longer opening hours"), examinations (e.g., "Examinations are unfair because of their subjective character"), rooms and facilities (e.g., "The ugly furniture in seminar rooms reduces the motivation to study"), the relationship of lecturers and students (e.g., "Lecturers should offer more than one consultation hour per week"), the body of student representatives ("The student representatives' budget is too small"), and so forth.

Procedure. Participants underwent the experiment in individual sessions of about 25 min. Each participant was seated in a separate cubicle in front of a personal computer. The instructions were displayed on a computer monitor and told participants that they were to observe a group of eight students engaging in a discussion about the conditions at the university and that their task was to form an impression of the group as a whole.

They then watched a succession of statements. Each statement appeared on the computer screen along with the speaker's photograph. The photograph was displayed in the center of the screen. The height and width of the speakers' pictures were 7 and 5.5 cm, respectively. They were black-and-white head and shoulder portraits. A statement was written below the speaker's photograph in a large (20 pt) type font. Presentation duration was 7 s, and interstimulus interval was 0.5 s.

The eight speakers—four male, four female—were of the same age range as the participating students. Each speaker made 6 statements. For each participant, the 48 statements were randomly sampled without replacement from the pool of statements. The speakers made their statements in turns, so that all eight speakers first made their 1st statement, then there was a round of 2nd statements, and so forth. The sequence of speakers within any one round was randomized. All randomizations were performed by the computer for each participant anew.

In the subsequent test phase, statements were presented to participants one by one. The statements comprised the statements presented in the discussion as well as new statements sampled randomly without replacement for each participant from the pool of statements. The group with

(relatively) few distractors saw 48 new statements; the group with many distractors saw three times as many, namely, 144 distractors. Distractors and old statements were pooled and presented in a random sequence that was newly determined for each participant.

The statements were shown in the same location and type as during the discussion. Instead of the photograph, two buttons labeled "old" and "new" were, however, seen, and participants were instructed to click, using the computer mouse, the button labeled "old" if they remembered the statement from the discussion; they were instructed to respond "new" if the item had not been presented before.

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

After the test phase, the computer presented a short questionnaire in which participants were asked to give their age, sex, and university major, as well as their hypotheses about the purpose of the experiment. On completion of the questionnaire, participants were thanked and debriefed.

Results and Discussion

In Appendix C, the data matrix for Experiment 1 is shown. It actually consists of two data matrices of the form of Table 2, one for each group. The Categories A and B correspond to the gender categories of female and male speakers, respectively.

The model was fit, using the computer program by Hu (1991), to the data from both groups simultaneously, where different parameter values were admitted for each group. In each group, it is assumed that $D_A = D_B = D_N$. This assumption was made for two reasons (cf. Bayen et al., 1996). First, versions of the model in which the item discrimination parameter D_N is allowed to vary freely are not identified and thus cannot be estimated (cf. *A Multinomial Model* section). Second, the restriction embodies the standard assumption made for two-high-threshold models that the probabilities of correctly identifying targets and distractors are equal. Snodgrass and Corwin (1988) showed that two-high-threshold models that include this assumption compare favorably with signal-detection models of simple item detection. The restriction is tested empirically through the chi-square goodness-of-fit test of the model that is reported below. If the assumption has to be rejected, it is possible to realize situations in which the D parameters can be estimated separately as explained in the General Discussion.

The data matrix of each group provides 8 degrees of freedom as explained above, so that there are 16 degrees of freedom in all. The restricted model uses seven parameters for each group as shown in Table 3. The degrees of freedom for the approximately chi-square-distributed log-likelihood ratio statistic G^2 are therefore $16 - 2 \times 7 = 2$. The observed value of G^2 was 1.06, indicating a very satisfactory goodness of fit ($p = .59$). Table 3 shows the estimates of the different parameters and their 90% confidence intervals.

As can be seen from the confidence intervals, the d parameters for category discrimination are reliably larger than zero according to a one-sided significance test. As the validation of these parameters succeeded in Experiment 3, sex appears to be encoded and used in the assignments.

Table 3
Parameter Estimates and 90% Confidence Intervals (CIs) for Experiment 1

	Few distractors		Many distractors	
Parameter	Estimate	CI	Estimate	CI
D	.82	.79–.85	.77	.74–.80
d_A	.57	.35–.80	.53	.36–.70
d_B	.62	.45–.78	.32	.03–.61
c_A	.17	.10–.24	.23	.16–.31
c_B	.36	.29–.44	.33	.26–.41
a	.59	.45–.74	.43	.29–.57
b	.33	.25–.41	.10	.07–.13

Note. A = female speakers; B = male speakers; D = item discrimination parameter; d_A, d_B = category discrimination parameters; c_A, c_B = person discrimination parameters; a = probability of guessing Category A; b = probability of guessing a statement is old.

The attempt to set the bias parameters b equal over the two groups results in a substantial loss of goodness of fit ($\Delta G^2 = 25.25$, $df = 1$, $p < .01$). The bias parameters thus differ significantly: As can be seen in Table 3, participants became more cautious in guessing "old" as the number of distractors increased. All other parameters can be set equal over the two groups without substantial loss of goodness of fit ($\Delta G^2 = 12.16$, $df = 6$, $p > .05$). Thus, there is both convergent and discriminant validity. The simplified model that restricts all parameters other than the b parameters to be equal across groups achieves a satisfactory overall goodness of fit ($G^2 = 13.22$, $df = 8$, $p = .10$).

Experiment 2: Item Discrimination

In the second experiment, a manipulation of distractor similarity aimed at affecting the process of item discrimination. Specifically, each statement of the pool of 202 statements used in Experiment 1 was paraphrased in a 2nd statement conveying the same or a very similar meaning. For example, the statement "Modern media (e.g., computers) should be employed more often" was paraphrased as "The possibilities of modern media (e.g., computers) should be used better."

Two groups of participants differed in the distractors they received. For members of the first group, equal numbers of old and new statements were randomly sampled without replacement from the original pool of 202 statements as in Experiment 1. For the second group, pairs of meaning-equivalent statements were sampled. One member of each pair was used as target, the other as distractor, resulting in sets of targets and distractors that should be very difficult to discriminate because of their content similarities. An analogous manipulation was applied successfully in the context of validating a related source-confusion model (Bayen et al., 1996).

Method

Participants. Participants were a new sample of 20 male and female students from different faculties of the University of Bonn. All participants were native speakers of German, and they received either partial course credit or a small monetary reward for their participation. Ten

participants were assigned to each group by means of a quasi-random assignment procedure.

Procedure. The same procedure as in Experiment 1 was employed.

Results and Discussion

In Appendix C, the data matrix for Experiment 2 is shown. In fitting the model, $D_A = D_B = D_N$ was again assumed for the first group of participants with low target–distractor similarity; an assumption that could be empirically upheld in the first experiment. For the second group, we reasoned that high target–distractor similarity might differentially affect the ability to detect targets $D_A = D_B = D$ and to detect distractors D_N. Therefore, we allowed D_N to vary freely in this group.

The resulting model for the two groups is thereby not identified, however, and an additional assumption is required to make it so. Only if the b parameters are set equal over both groups does an identifiable model result, referred to as Model 0. The model has a very satisfactory goodness of fit ($G^2 = 1.32$, $df = 2$, $p = .51$). Table 4 shows the parameter estimates and confidence intervals. As can be seen, there are pronounced differences in the item discrimination parameters between the two groups, and the ability to detect distractors D_N is disrupted particularly strongly by high target–distractor similarity. The d parameters are again significantly larger than zero, indicating the use of sex as a social category in memory, presupposing the validation of the d parameters in Experiment 3.

Models that admit different b parameters for both groups but constrain the D_N parameter (Model 1) or the $D_A = D_B = D$ parameter (Model 2) to be equal in both groups are also identified. Both models cannot, however, be fit to the data: For Model 1, $G^2 = 42.12$, $df = 2$, $p < .01$; for Model 2, $G^2 = 20.87$, $df = 2$, $p < .01$. Thus, the data cannot be explained by models that assume equal item discrimination parameters between experimental groups.

Finally, a model in which all parameters except the item discrimination parameters are set equal across groups does not entail a sizable loss in goodness of fit with respect to Model 0 ($\Delta G^2 = 3.10$, $df = 5$, $p = .68$). The goodness of fit of this simplified model is very satisfactory ($G^2 = 4.42$, $df = 7$, $p = .73$). In summary, there is again convergent and discriminant validity.

Table 4
Parameter Estimates and 90% Confidence Intervals (CIs) for Experiment 2

	Low similarity		High similarity	
Parameter	Estimate	CI	Estimate	CI
$D_A = D_B$	.76[a]	.72–.79	.60	.55–.66
D_N	.76[a]	.72–.79	.00	−.26–.26
d_A	.48	.21–.76	.61	.42–.80
d_B	.61	.42–.80	.73	.56–.91
c_A	.30	.22–.38	.28	.19–.37
c_B	.31	.23–.38	.28	.18–.37
a	.56	.40–.71	.48	.41–.56
b	.26[b]	.20–.32	.26[b]	.20–.32

Note. A = female speakers; B = male speakers; N = distractors; D_A, D_B, D_N = item discrimination parameters; d_A, d_B = category discrimination parameters; c_A, c_B = person discrimination parameters; a = probability of guessing Category A; b = probability of guessing a statement is old.
[a] Parameters set equal. [b] Parameters set equal.

Experiment 3: Category Discrimination

Experiment 3 is a central experiment in the present series because it addresses the process of category discrimination that is of prime interest to social psychologists. In Experiment 3, old and new statements were again sampled from the pool of statements used in Experiment 1. Only male speakers appeared, however, and the categorization was based on either hometown (student from Aachen vs. student from Münster) or academic status (student vs. lecturer). The kind of categorization (hometown vs. academic status) was manipulated as a between-subjects factor.

For several reasons, we expected the kind of categorization to affect the likelihood of category discrimination. Many of the studies listed in Table 1 have obtained evidence for the role of category accessibility. We assumed that the categorization based on academic status is more accessible in our student population than the categorization based on the two rather similar hometowns (cf. notably Blanz, in press; van Knippenberg, van Twuyver, & Pepels, 1994; van Twuyver & van Knippenberg, 1995). Furthermore, the statements discussed conditions at the university and thus topic relevance was given only for the categorization based on academic status. Topic relevance has also been argued to contribute to the likelihood of categorization (Oakes, 1987; Oakes, Haslam, & Turner, 1994; van Knippenberg et al., 1994). In addition, the pictures of lecturers (corresponding to the students from Münster in the other experimental condition) represented slightly older men, and thus there was some amount of normative fit between academic status and age, in terms of Oakes's (1987) analysis, in the first experimental condition but no normative fit (only structural fit) between hometown and age in the second condition. Finally, the distinction between lecturer and student was introduced somewhat more visibly than the distinction based on hometown through the kind of captions that accompanied the speakers' pictures. Whereas the captions had the format "Name, hometown" in the hometown condition, the status condition used two different formats, namely, "Name, student" and "Dr. Name, lecturer."

Method

Participants. Participants were a new sample of 20 male and female students from different faculties of the University of Bonn. All participants were native speakers of German, and they received either partial course credit or a small monetary reward for their participation. They were assigned to two groups of 10 persons each by means of a quasi-random assignment procedure.

Categories. All eight speakers were male, and their pictures were accompanied by captions. In the first experimental group, the caption gave the speaker's first name and the addendum "Aachen" or "Münster," thereby specifying the speaker's hometown—for example, "Frank, Münster." In the second group, the caption gave either the speaker's first name and the addendum "student" or the speaker's last name preceded by the title "Dr." and the addendum "lecturer"—for example, "Dr. Salk, lecturer."

In both groups, the speakers thereby formed two categories of four persons each. In the first group, the categories were defined by hometown; in the second by academic status. The same pictures were used

in both groups, and they were partitioned into categories in the same way in both groups. Thus, the only differences between the two groups were the different kinds of captions.

The statements from which targets and distractors were sampled were originally generated by students as explained above. Some minor editing of the statement pool removed explicit references to the student status of the authors of the statements.

Procedure. The procedure followed that of the previous experiments.

Results and Discussion

In Appendix C, the data matrix for Experiment 3 is shown. In modeling the data, we again assumed $D_A = D_B = D_N = D$ in each group. The resulting model achieves an acceptable goodness of fit ($G^2 = 4.60$, $df = 2$, $p = .10$), and Table 5 shows the parameter estimates and confidence intervals. As can be seen, the category discrimination parameters are much smaller in the group with categorization based on hometown than in the group using academic status. Confidence intervals indicate that in fact the d parameters for the hometown condition did not differ significantly from zero.

The attempt to restrict the d parameters to be equal across groups implies a significant loss in goodness of fit ($\Delta G^2 = 40.82$, $df = 2$, $p < .01$). All other parameters can be set equal without such substantial loss ($\Delta G^2 = 7.23$, $df = 5$, $p = .20$). The resulting simplified model itself attains an acceptable goodness of fit ($G^2 = 11.83$, $df = 7$, $p = .10$).

Again, convergent and discriminant validity can be demonstrated, this time in the particularly important attempt to validate the process of category discrimination. The experiment also shows that categorization is not dependent on superficial physical similarities between the speakers, because both experimental groups were completely comparable with respect to such features.

Experiment 4: Response Bias in Guessing Category Membership

In the fourth experiment, we attempted to manipulate response preferences for one category over the other. On the basis of prior research, it seemed promising to implement a covariation of the content of the statements and category membership for this purpose, thereby realizing structural fit. Two pools of statements were constructed, one containing negative statements that criticized conditions at the university, the other containing comparatively positive statements that lauded conditions or presented them as satisfactory.

Two groups of participants worked under different conditions. In the group with critical men, male speakers made 75% negative and 25% positive statements, and female speakers made 25% negative and 75% positive statements. In the group with critical women, these proportions were reversed. Thus, in both groups there was the same amount of structural fit, albeit reversed in terms of the content of the statements.

It was assumed that participants would form an impression of the association of the members of each gender category with predominantly positive versus negative statements as they observed the discussion (Rothbart, Fulero, Jensen, Howard, & Birrell, 1978) or, more likely, as they attempted to retrieve information about the group members for a number of subsequent trait ratings (McConnell, Sherman, & Hamilton, 1994). It is likely that this impression would then be used in the later assignment phase. For example, in the group with critical men, participants may be biased to assign negative statements to male speakers and positive statements to female speakers if they are uncertain about the speaker and the speaker's category membership. Because of the reversed direction of the implemented structural fit, the resulting response bias should differ between groups.

The data matrix for this experiment comprises four basic data matrices like Table 2, because the assignment frequencies were counted separately for positive and negative statements within each group. The model was fit with separate parameters for each group and type of statement. In terms of the response bias parameter a, we thus expected a greater tendency to assign negative statements to male speakers in the group with critical men than in the group with critical women, whereas we expected the opposite relationship for positive statements.

What effects of the manipulation of structural fit can be expected on the other processes involved in the "Who said what?" paradigm? If participants should already form an on-line impression of the association of gender and kind of statement, the 75% items can be considered expectancy-congruent, the 25% items expectancy-incongruent for each gender category. There is a complex body of literature on memory for expectancy-congruent and expectancy-incongruent information that bears on the present situation under this premise. Specifically, guessing strategies have been found to cause response bias in recognition measures, and there is often differential memory for congruent and incongruent information. For recognition measures, the pattern of results has been argued to suggest that "guessing strategies are frequently used and that these strategies may override the preferential encoding of incongruent information" (Stangor & McMillan, 1992, p. 55), supporting the present prediction of differential response bias. In addition, the present study combined a number of conditions in which, according to the meta-analysis of Stangor and McMillan (shown in their Table 4), differential memory in terms of recognition sensitivity, corrected for response bias, tends to be low: (a) Expectations, if any, are formed in the course of the experimental session; (b) the processing goal is to form an impression rather than to

Table 5
Parameter Estimates and 90% Confidence Intervals (CIs) for Experiment 3

	Hometown		Academic status	
Parameter	Estimate	CI	Estimate	CI
D	.69	.66–.73	.73	.69–.77
d_A	.06	−.26–.37	.60	.41–.79
d_B	.00	−.33–.33	.61	.41–.80
c_A	.09	.03–.16	.19	.11–.27
c_B	.21	.14–.28	.22	.14–.29
a	.50	.37–.64	.49	.36–.62
b	.22	.17–.28	.32	.25–.38

Note. A = Münster and teachers, respectively; B = Aachen and students, respectively; D = item discrimination parameter; d_A, d_B = category discrimination parameters; c_A, c_B = person discrimination parameters; a = probability of guessing Category A; b = probability of guessing a statement is old.

memorize; and (c) the processing target is a group rather than an individual.

We reasoned, however, that congruency would not be a major factor in the present case because participants would be unlikely to form such on-line expectancies. The statements used here, being evaluations of a wide range of different states of affairs and issues at the university, usually exhibit high distinctiveness and high consistency as well as some amount of consensus in terms of Kelley's (1967) analysis and tend to be attributed as much to the conditions themselves as to the speakers' characteristics. Thus, in McConnell, Sherman, and Hamilton's (1997) terms, perceivers expected little entitativity in this situation and consequently were unlikely to engage in the kind of active integrative processing that is instrumental in inducing on-line expectancies in the first place.

For these reasons, we expected comparatively little impact of the manipulation on the item and person discrimination parameters. Moreover, because the same amounts of structural fit were implemented in both groups, category discrimination that may be a function of structural fit (Oakes, 1987) should not be differentially affected. Finally, there was no reason to expect an effect of the group manipulation on bias b in item detection. As elaborated above, however, guessing strategies based on the content of the statements are assumed to cause differences in response bias a in guessing category membership.

Method

In a departure from the procedure of the previous experiments, participants were asked to rate speakers on a number of traits after observing the discussion and before the assignment phase. The rating task served two purposes. Introducing a so-called filled delay, it was likely to reduce memory, and for the purpose of detecting response bias, conditions with less than perfect memory are propitious. Furthermore, the rating task makes it necessary for participants to form memory-based overall impressions (McConnell et al., 1997) and thereby may induce overall representations of the association between gender and kind of statement.

More students were asked to participate than in the previous experiments because each participant contributed fewer data per cell of the data matrix that comprises four basic contingency tables like Table 2 as discussed above. For each participant, four male and four female speakers were randomly sampled from larger sets of pictures of eight male and female persons, respectively.

Participants. Participants were a new sample of 54 male and female students from different faculties of the University of Bonn. All participants were native speakers of German, and they received either partial course credit or a small monetary reward for their participation. Participants were assigned to two groups of 27 persons each by means of a quasi-random assignment procedure.

Pools of statements. Statements that were not clearly critical of the conditions at the university were deleted from the pool of statements used in the previous experiments, resulting in a reduced pool of 180 negative statements. For each such statement (e.g., "The library should have longer opening hours"), a parallel positive statement was constructed that claimed that the conditions were already satisfactory in this respect (e.g., "The opening hours of the library are quite long"), which yielded a parallel pool of 180 more positive statements. In sampling statements for an experimental session, parallel positive and negative statements were never included in a given sample of targets and distractors.

Rating measure. Having observed the discussion, participants were asked to rate speakers with respect to a number of traits. The 20 traits were taken from Rosenberg, Nelson, and Vivekanathan (1968), and five traits each represented the socially good and socially bad, the intellectually good and intellectually bad poles identified in that study. For each trait, participants were asked to rate a randomly sampled male speaker as well as a female speaker on separate 10-point scales that were simultaneously presented.

Procedure. Participants first observed the discussion. Then they made the ratings, and finally they underwent the assignment phase.

Members of the first group of participants saw male speakers making 75% negative and 25% positive statements, whereas these proportions were reversed for female speakers. Specifically, two male (female) speakers made five negative (positive) and one positive (negative) statement, and the other two male (female) speakers made four negative (positive) and two positive (negative) statements. In all, there was again a total of 48 targets. All target selections were determined randomly and separately for each participant.

For members of the second group, the roles of male and female speakers were exchanged in these assignments, so that the majority of negative statements were made by female speakers. Distractors always comprised 24 positive and 24 negative statements.

Results and Discussion

Manipulation check. The negative items were originally produced by University of Bonn students as part of an evaluation of the university and the Psychological Institute. Therefore it seemed likely that the participating students would tend to endorse these statements and would tend to reject the more positive statements. If the content manipulation of the structural fit is perceived by the participants, greater liking should result for speakers with many negative statements because of greater perceived attitude similarity (e.g., Newcomb, 1956): We expected an interaction of group and gender so that male speakers obtain better evaluations when they make mostly negative rather than positive statements, that is, in the group with critical men, whereas in the group with critical women, female speakers obtain better evaluations.

For each participant, ratings were coded so that greater numbers indicated a more positive evaluation, and the ratings were averaged over the 20 traits separately for male and female speakers. In the group with critical men, the average evaluations of male and female speakers were 6.9 and 6.3, respectively. In the group with critical women, these evaluations amounted to 6.2 and 6.3, respectively. An ANOVA with between-subjects factor group and within-subjects factor gender revealed an overall effect of gender, $F(1, 52) = 4.41$, $p = .04$, that was moderated by the expected interaction, $F(1, 52) = 11.03$, $p < .01$. Thus, the manipulation appears to have had the expected effect, particularly in the group with critical men, although considering the numerical size of the differences, the effect is far from dramatic.

Model analyses. The data matrix for Experiment 4 is given in Appendix C. The model was fit with different parameters for each kind of statement and group. To assess the possible effects of statement congruency versus incongruency on item memory, we allowed D_A, D_B, and D_N to vary freely for each kind of statement and group. Because the resulting model is not identified, a number of restrictions must be introduced. As in Experiment 2, it was necessary (a) to set equal the b parameters across groups and kind of statement and (b) to anchor one of the D_N parameters for distractor detection by setting it equal to a D_A or D_B parameter from that experimental condition. We set $D_A = D_N$ in the group with critical men for positive statements (cf. Table 6). Obviously, restriction (b) is somewhat arbitrary as

Table 6
Parameter Estimates and 90% Confidence Intervals (CIs) for Experiment 4

	Group with critical men				Group with critical women			
	Postitive statements		Negative statements		Positive statements		Negative statements	
Parameter	Estimate	CI	Estimate	CI	Estimate	CI	Estimate	CI
D_A	.74[a]	.71–.77	.76	.69–.82	.69	.62–.77	.68	.64–.73
D_B	.74	.67–.81	.77	.73–.81	.72	.68–.76	.69	.62–.77
D_N	.74[a]	.71–.77	.64	.57–.77	.55	.41–.70	.66	.54–.78
d_A	.30	−.10–.69	.56	.39–.74	.56	.38–.74	.39	.15–.62
d_B	.63	.48–.78	.53	.31–.75	.35	.13–.58	.70	.53–.88
c_A	.20	.14–.25	.32	.23–.41	.15	.06–.24	.28	.22–.33
c_B	.27	.18–.36	.33	.27–.38	.29	.24–.35	.30	.19–.40
a	.72	.59–.85	.36	.24–.48	.44	.33–.55	.55	.42–.67
b	.19[b]	.14–.24	.19[b]	.14–.24	.19[b]	.14–.24	.19[b]	.14–.24

Note. A = female speakers; B = male speakers; N = distractors; D_A, D_B, D_N = item discrimination parameters; d_A, d_B = category discrimination parameters; c_A, c_B = person discrimination parameters; a = probability of guessing Category A; b = probability of guessing a statement is old.
[a] Parameters set equal. [b] Parameters set equal.

there are many ways in which it can be satisfied. For this reason, we repeated the analyses with all possible other restrictions (b) and found that the overall sizes of some of the parameter estimates, but not their pattern, depended slightly on the particular restriction realized. As one would expect, by far the largest variability was found in the D_N parameters (maximal variation .13). However, neither the values of the test statistics reported below nor the estimates of the response bias a, which is under scrutiny in this experiment, depended on the restriction.

The model with restrictions (a) and (b) is saturated, that is, it uses as many parameters as there are degrees of freedom. The saturated model fitted the data perfectly ($G^2 = 0$) as it should whenever the parameter estimates do not converge against the boundary values 0 and 1 (Riefer & Batchelder, 1988). Table 6 shows the parameter estimates and confidence intervals. As predicted, the response bias parameter a for the tendency to assign positive statements to women is larger in the group with critical men than in the group with critical women. Conversely, and as predicted, the response bias parameter a for assigning negative statements to women is larger in the group with critical women than in the group with critical men. It is interesting to note that the pattern of a values mirrors the rating data, in that the differences between male and female speakers are more pronounced in the group with critical men than in the group with critical women. Again, the d parameters are, with one exception, significantly larger than zero.

The attempt to restrict the a parameters to be equal across groups, that is, to vary only as a function of kind of statement, results in a significant loss of goodness of fit ($\Delta G^2 = 9.62$, $df = 2$, $p < .01$). Thus, the differences in response tendencies between groups are significant. On the other hand, restricting analogously D, c, and d parameters for male and female speakers as well as the D parameters for distractors to be equal across groups—that is, to vary only as a function of kind of statement—does not entail a significant loss in goodness of fit ($\Delta G^2 = 19.39$, $df = 14$, $p = .15$), indicating as expected that the effects of possibly perceived item congruency versus incongruency on the different memory processes were slight. The value $G^2 = 19.39$ is also the goodness-of-fit value of the resulting simplified model, which accordingly exhibits a satisfactory goodness of fit ($df = 14$, $p = .15$).

Discussion. In summary, convergent and discriminant validity is again obtained. The fact that the manipulation was more effective for the group with critical men in terms of both the rating data and the response tendency parameters may reflect a preexisting stereotypical bias to associate critical statements more strongly with men than with women. Under this interpretation, the manipulation succeeded in eliminating the stereotypical tendency in the group with critical women and enhanced that tendency in the group with critical men.

Experiment 5: Person Discrimination

In Experiment 5, the process of person discrimination was to be manipulated. As in Experiment 2, we used a similarity manipulation for this purpose. Two experimental groups were realized. For members of the group with similar men, the male speakers used in previous experiments were replaced by more similar-looking men. For members of the group with similar women, the original female speakers were replaced by more similar-looking women. In the group with similar-looking men, a reduced person discrimination parameter for male speakers was expected, whereas in the group with similar-looking women a reduced person discrimination parameter for female speakers was expected.

Method

Participants. Participants were a new sample of 20 male and female students from different faculties of the University of Bonn. All participants were native speakers of German, and they received either partial course credit or a small monetary reward for their participation. They were assigned to two groups of 10 persons each by means of a quasi-random assignment procedure.

Pictures. Pictures were generated on the basis of the eight pictures of four men and four women used in Experiment 1. A set of more similar-looking male (female) portraits was constructed by using one man's (woman's) nose, mouth, and eyes to replace those of the other three men (women). This was done by means of a commercially avail-

able computer program called Kai's PowerGoo (1996), which specializes in processing digitalized photographs of faces.

Procedure. The group with similar men saw the similar men as speakers as well as the original set of female speakers. The group with similar women saw the similar women as speakers and the original male speakers. In all other respects, the procedure followed that of Experiment 1 (with equal numbers of targets and distractors).

Results and Discussion

The raw frequencies of the different assignments are shown in Appendix C. In fitting the model, it was again assumed that $D_A = D_B = D_N = D$ in each group. The resulting model exhibited a very satisfactory goodness of fit ($G^2 = 0.66$, $df = 2$, $p = .72$). Table 7 shows the parameter estimates and confidence intervals. As can be seen, person discrimination for female speakers was much reduced in the group with similar women, whereas there was no such reduction for male speakers in the group with similar men. Note that again the categorization parameters d are significantly larger than zero.

The attempt to set the c parameters equal across groups resulted in a significant loss of goodness of fit ($\Delta G^2 = 13.86$, $df = 2$, $p < .01$), confirming that the c parameters differ between groups. According to the confidence intervals shown in Table 7, the effect is concentrated on the discrimination of female speakers. All other model parameters could be set equal without significant loss in goodness of fit ($\Delta G^2 = 5.19$, $df = 5$, $p > .39$), and the resulting model also achieved a very good overall goodness of fit ($G^2 = 5.85$, $df = 7$, $p = .56$).

In summary, the manipulation affected only the c parameters, whereas the parameters that map the other processes involved in the "Who said what?" paradigm did not differ significantly between experimental groups. Unexpectedly, however, participants were quite capable of distinguishing the similar men—that is, the manipulation was effective only for the female speakers. The finding of reduced person discrimination for female speakers in the group with similar women, but not for men in the group with similar men, fits Lorenzi-Cioldi's (1993) findings that women in general are less discriminable than men.

Experimental Validation: Summary and Discussion

Multinomial models are powerful theories of the decision processes involved in the "Who said what?" paradigm. One of the valuable features of these theories is that they provide several independent parameters for explaining the pattern of assignments by means of different processes. Thereby, they offer the possibility of unconfounding memory for the item from memory for the category and speaker and of measuring category salience independently of bias.

The validating experiments addressed, in turn, the processes of guessing bias in item detection, sensitivity in item memory, category memory, bias for category, and person memory, and they demonstrated that the respective model parameters adequately and independently measure these processes. In each case, the experimental manipulation that aimed at affecting the process under scrutiny mapped onto the appropriate process parameters and did not substantially influence parameters associated with other processes.

A number of general criticisms can be raised against the statistical and substantive assumptions underlying multinomial models. We postpone the discussion of such points until the General Discussion and Appendix A. A specific criticism of the present experiments is, however, that the aspect of discriminant validity (Campbell & Fiske, 1959)—that is, the lack of influence of process manipulations on parameters associated with other processes than the manipulated one—rests on accepting a null hypothesis, which raises the issue of statistical test power. In other words, it is possible that if statistical power had been greater, reliable differences might also have been found in other parameters.

For two reasons, we believe that this justified criticism does not present a fundamental problem. First, we consider it unlikely that any experimental manipulation is totally process-pure. Rather, it seems likely that most experimental manipulations would affect several of the processes involved in the "Who said what?" task in one way or another, so that we would actually expect real, if comparatively small and uninteresting, effects on a range of parameters that might be detected given sufficient test power. Second, the present model, like most substantive models, is unlikely to provide a totally valid and complete account of the processes involved (cf. also Appendix A). It is more reasonable to consider the model a useful first approximation of the true state of affairs that is able to provide independent assessments of the major contributions of different processes. In the next two sections, we illustrate this aspect of the validated model in an empirical application.

Table 7
Parameter Estimates and 90% Confidence Intervals (CIs) for Experiment 5

	Similar men		Similar women	
Parameter	Estimate	CI	Estimate	CI
D	*.68*	*.64–.71*	*.69*	*.65–.73*
d_A	.45	.21–.69	.45	.23–.68
d_B	.55	.30–.80	.58	.37–.79
c_A	.34	.25–.42	.11	.03–.18
c_B	.37	.28–.45	.31	.23–.40
a	.46	.34–.59	.52	.41–.63
b	.26	.21–.32	.36	.30–.42

Note. A = female speakers; B = male speakers; D = item discrimination parameter; d_A, d_B = category discrimination parameters; c_A, c_B = person discrimination parameters; a = probability of guessing Category A; b = probability of guessing a statement is old.

Applying the Validated Model

The Role of Processing Load in Categorization

There is a large body of data suggesting that cognitive load facilitates stereotyping (e.g., Bodenhausen, 1988, 1990; Bodenhausen & Lichtenstein, 1987; Gilbert & Hixon, 1991; Macrae, Hewstone, & Griffiths, 1993; Macrae et al., 1994; Pendry & Macrae, 1994; Pratto & Bargh, 1991; Stangor & Duan, 1991). Manipulations of cognitive load have been based on processing pace (e.g., Pratto & Bargh, 1991), on the amount of information presented (e.g., Stangor & Duan, 1991), on concurrent task demands (e.g., Macrae et al., 1994), on task complexity (Bodenhausen & Lichtenstein, 1987), and on mood state (Stroessner, Hamilton, & Mackie, 1992). Under cognitive load, people tend to recall relatively more stereotype consistent and less stereotype

inconsistent information (e.g., Macrae et al., 1993), and their judgments are more strongly based on stereotypes (e.g., Bodenhausen & Lichtenstein, 1987). Theoretically, the assumption that category-based processing is preferred when cognitive load is high is an important component of the continuum model of Fiske and Neuberg (1990), and stereotypes and categories are often conceived as simplifying heuristics and energy-saving devices (e.g., Macrae et al., 1994).

Recently, Spears and Haslam (1997) provided a critical review of this literature in which they argued that categorization and stereotyping can themselves be effortful processes that require mental resources (cf. Gilbert & Hixon, 1991) and therefore can be disrupted under conditions of high cognitive load. Spears and Haslam questioned the assumption that categorizing people as individuals invokes processes that differ qualitatively from those involved in perceiving them in terms of their social-category memberships. They assumed that categorization effects are not primarily a product of load but are determined by the perceived appropriateness and meaning of the categorization prescribed by the context. Departing from this basic assumption, Spears and Haslam derived detailed predictions for the pattern of errors in the "Who said what?" paradigm as a function of load. In particular, in the case of low or no fit,

> people may be less likely to use accessible social categories as a strategy for making sense of the stimulus array when (a) performance on the recall task is easy allowing retention of all identifying information (i.e., under low memory demands or low load), or (b) attention to all stimulus relevant information is undermined because the task is too debilitating (i.e., under high memory demands, or high load). Thus where there is no clear fit between stimuli and category both high and low load may, for different reasons, reduce the ratio of intra-category to inter-category recall errors. Under moderate load conditions, however, it may be both (1) *possible* to organize information according to social category cues (although not so easily as when the fit is high), and (2) *meaningful* and useful to do so. (Spears & Haslam, 1997, p. 183)

They went on to predict a curvilinear relationship such that social categorization in terms of the ratio of within-category to between-categories errors increases as the cognitive load increases from low levels, reaches a maximum at moderate levels of load, and decreases again as cognitive load is further increased. In support of their view, they reported several studies using the "Who said what?" paradigm with low fit (i.e., without covariation of the content of the statements and their speakers' category memberships) and different manipulations of load to show that the error-difference as a measure of degree of categorization decreases not only for low levels but also for high levels of load (Spears & Haslam, 1997; Spears et al., 1996).

For example, Spears et al. (1996) reported an experiment in which processing pace was varied in four steps in a "Who said what?" study based on gender categories. Statements were presented at a rate of one statement per 10, 7.5, 5, or 2.5 s. The error-difference measure decreased strongly under the high level of load (2.5 s), suggesting that categorization decreases when cognitive demands are high.

Experiment 6: Cognitive Load

The sixth experiment manipulated processing pace to illustrate the use of the model in assessing the effects of cognitive load on categorization.

Method

Participants. Participants were a new sample of 40 male and female students from different faculties of the University of Bonn. All participants were native speakers of German. Participation was voluntary. Participants were assigned to four groups of 10 persons each by means of a quasi-random assignment procedure.

Procedure. The procedure followed that of Experiment 1 (equal numbers of targets and distractors) with the exception that presentation time per stimulus was either 10, 7.5, 5, or 2.5 s. Processing pace was realized as a between-subjects factor.

Results

Conventional analysis. Figure 2 presents the results of the conventional analysis in terms of within-category and between-categories errors. Between-categories errors were multiplied by .75 to take into account that there are more opportunities for errors of this kind (four candidate speakers) than for within-category errors (three candidate speakers). As can be seen, the error difference follows the curvilinear pattern expected by Spears and Haslam (1997) and in particular declines at high levels of cognitive load, thereby replicating the findings of Spears et al. (1996).[3]

An ANOVA of these error frequencies with between-subjects factor processing pace and within-subjects factor kind of error revealed a main effect of kind of error, $F(1, 36) = 165.83$, $p < .01$, indicating a positive error difference. This main effect was moderated by processing pace, $F(3, 36) = 2.94$, $p < .05$, so that the effect of processing load on the error-difference measure was significant. A polynomial trends analysis partitioned this effect into a linear, a quadratic, and a cubic component. There was a tendency for a linear trend in the error-difference measure, indicating that it tended to decrease with increasing levels of load, $F(1, 36) = 3.66$, $p = .06$, whereas there was no evidence for a cubic trend, $F(1, 36) = 1.32$, $p = .26$. A significant quadratic trend in the direction predicted by Spears and Haslam (1997) emerged, however, so that the error-difference has a maximum at moderate levels of cognitive load, $F(1, 36) = 3.84$, $p = .03$, one-tailed.

Model analyses. The data matrix for Experiment 6 is shown in Appendix C. For the model analyses, we assumed $D_A = D_B = D_N = D$ for each group as before. The resulting model achieved an excellent goodness of fit ($G^2 = 2.28$, $df = 4$, $p = .68$). Table 8 shows the parameter estimates and confidence intervals. As can be seen, the parameters for item and person memory show a pronounced decline as processing load increases, whereas there is no substantial decrease in the category discrimination parameters, even for the extraordinarily short presentation duration of 2.5 s per statement.

Setting equal the item discrimination parameter D across groups led to a large loss in goodness of fit ($\Delta G^2 = 145.64$, $df = 3$, $p < .01$). Similarly, the person discrimination parameters c_A and c_B could not be constrained to be equal over the presentation durations without significant loss in goodness of fit ($\Delta G^2 = 14.05$, $df = 6$, $p < .05$). Furthermore, the differences

[3] The numbers of errors shown in Figure 2 declined with increasing processing pace. The total number of errors, including the erroneous classification of old statements as distractors, of course increased strongly as shown in Appendix C.

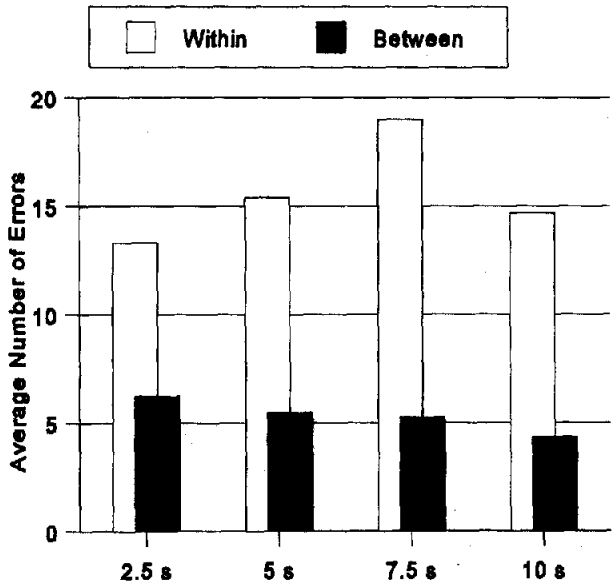

Figure 2. Within-category and corrected between-categories errors as a function of presentation duration per statement.

in the response bias parameter b for item detection are significant ($\Delta G^2 = 15.45$, $df = 3$, $p < .01$), indicating that participants became more cautious in responding "old" as the processing load decreased.

There was no evidence, however, for significant differences in the category discrimination parameters d_A and d_B as a function of presentation duration ($\Delta G^2 = 1.51$, $df = 6$, $p = .96$) or in the response bias parameter a for guessing the category ($\Delta G^2 = 0.35$, $df = 3$, $p = .95$).

Discussion

The model analyses decomposed the effects of processing pace on the pattern of errors into several basic relationships. In particular, as processing pace increased, item discrimination and person discrimination were clearly disrupted, indicating that processing load was successfully manipulated, whereas there is no evidence for a comparable decline of category discrimination. The results of the significance tests showed that the complex curvilinear function relating the error-difference measure to processing pace was largely caused by the interaction of processes other than category discrimination (cf. the Problems of the Conventional Approach section). For the lowest level of load, the increase in person discrimination contributed to the decrease in the error-difference measure. On the other hand, the smaller error difference obtained under high levels of load largely reflects the pronounced decrease in item discrimination.

In fact, the information about category discrimination per se was in some instances comparatively low in these data sets as evidenced by the relatively large confidence intervals for some of the d parameters. The factors that determine measurement accuracy (i.e., the size of standard errors of the estimates and their confidence intervals) are discussed in Appendix A. In summary, the present results do not support the hypothesis of Spears et al. (1996) and Spears and Haslam (1997) that categorization is disrupted by high cognitive load. Note also that the conventional analysis misleadingly suggests the opposite conclusion.

The so-called efficiency of a process is one of the attributes required for automaticity (Bargh, 1994). A process is efficient to the extent that it does not depend on limited mental resources, and efficiency is often operationalized as the robustness of the process against increases in processing load (e.g., Andersen, Spielman, & Bargh, 1992; Bargh & Tota, 1988; Lupfer, Clark, & Hutchinson, 1990; Winter, Uleman, & Cunniff, 1985). In addition, a second "horseman" of Bargh's (1994) four horsemen of automaticity, namely, lack of intentionality, is one of the prime characteristics of the "Who said what?" paradigm itself. Consequently, the present results, securing efficiency and lack of intentionality, lend new and independent support to the point of view that a number of basic categories such as race and gender are activated and encoded automatically (Blair & Banaji, 1996; Devine, 1989; Fiske & Neuberg, 1990; Taylor et al., 1978). The automaticity of gender-categorization in fact seems to be so firmly established that even in situations in which item and person discrimination are very heavily disrupted, categorization is not affected. In line with the continuum model of Fiske

Table 8
Parameter Estimates and 90% Confidence Intervals (CIs) for Experiment 6

	Presentation duration							
	2.5 s		5 s		7.5 s		10 s	
Parameter	Estimate	CI	Estimate	CI	Estimate	CI	Estimate	CI
D	.43	.38–.47	.67	.63–.71	.82	.79–.85	.73	.69–.76
d_A	.37	.04–.71	.40	.09–.72	.40	.06–.75	.53	.28–.77
d_B	.68	.45–.91	.63	.46–.80	.73	.59–.87	.68	.49–.88
c_A	.15	.04–.26	.25	.17–.33	.25	.18–.33	.30	.22–.38
c_B	.17	.06–.28	.22	.14–.30	.19	.12–.26	.38	.30–.47
a	.57	.48–.66	.59	.45–.74	.59	.39–.79	.52	.36–.68
b	.32	.28–.36	.20	.15–.25	.20	.13–.27	.19	.14–.24

Note. A = female speakers; B = male speakers; D = item discrimination parameter; d_A, d_B = category discrimination parameters; c_A, c_B = person discrimination parameters; a = probability of guessing Category A; b = probability of guessing a statement is old.

and Neuberg (1990), it can be concluded that categorization is an extremely energy-saving process.

General Discussion

In this article, we have presented and validated a theoretical model of the processes involved in the "Who said what?" paradigm. The model is a member of the class of multinomial models that have seen a growing range of effective applications to a number of phenomena from cognitive (e.g., Riefer & Batchelder, 1988) and social psychology (e.g., Klauer & Batchelder, 1996) as reviewed by Batchelder and Riefer (in press).

Using the techniques explained in Appendix A, one can easily obtain estimates and confidence intervals for the model's process parameters. Tests of hypotheses relating experimental manipulations to changes in the underlying processes can also be easily conducted within the multinomial framework.

Apart from the substantive assumptions detailed above, multinomial models rest on a number of general decision-theoretical and statistical assumptions. The ensemble of these assumptions defines the multinomial model and as such has received considerable empirical support in the present series of validating experiments. Nevertheless, some of these assumptions are potentially problematic, and they are discussed in Appendix A.

The new approach has been shown to provide a more valid and at the same time more refined use of the "Who said what?" paradigm. It permits one to disentangle and evaluate separately the confounded processes that operate in the "Who said what?" task, thereby assessing category discrimination independently from person and item discrimination and separating these processes from response biases. It is useful to consider possible applications of the model in order to clarify the substantive interpretation of the different process parameters and to illustrate the potential of the new method for advancing social psychological research. To our minds, promising applications of the model can be seen in the investigation of cognitive load, of normative and structural fit, and of distinctiveness, as well as in the study of in-group versus out-group differences, of crossed categorizations, of the effects of intergroup interaction and outcome dependency, and so forth. We elaborate briefly on the first three topics. Generally speaking, the model can make three different kinds of contributions in these contexts: (a) It can help to avoid reaching wrong conclusions by controlling for confounded factors, (b) it opens the door to addressing new theoretical questions that could not be easily addressed previously, and (c) existing theoretical controversies can in some cases be resolved empirically by the new method.

Cognitive Load

The new approach is well suited to study the impact of cognitive load on categorization, because in that field it seems imperative to disentangle the effects of load on item memory, on person memory, and on category memory (cf. Spears & Haslam, 1997). As illustrated above, failing to control for the effects of cognitive load on the confounded processes can lead to seriously flawed conclusions.

The modified "Who said what?" paradigm also opens the door to addressing new theoretical questions in this area. For example, the model allows one to separately assess memory for individuating (c parameters) versus category-based (d parameters) information. This raises the possibility of testing over different levels of load whether there is in fact an antagonistic relationship between person-based and category-based representations as has sometimes been assumed (e.g., Turner, 1987) or whether both kinds of processes operate in a more independent fashion (Pratto & Bargh, 1991) as seemed to be the case in Experiment 6.

Finally, the differentiated evaluation provided by the model in some instances allows one to resolve open theoretical debates. Results obtained by Gilbert and Hixon (1991) suggest that cognitive load may hinder the activation of a category and associated stereotype but increase application of category and stereotype once activated. A related, but different distinction was proposed by Spears and Haslam (1997, p. 183). Spears and Haslam argued that if fit is detected, perceivers may employ rational guessing strategies to go beyond the information given. In the case of the "Who said what?" paradigm, perceivers may use fit to attribute information to categories if the origin of the information is not remembered, thereby increasing the error-difference measure (cf. the Problems of the Conventional Approach section). In the case of high fit, the effect of load on categorization in the error-difference measure is thus likely to depend heavily on fit detection, according to Spears and Haslam, and load may block the detection of fit rather than the activation and accessibility of a social category and the associated stereotype.

Both theoretical positions, that of Gilbert and Hixon (1991) as well as that of Spears and Haslam (1997), lead to the same prediction that given substantial fit, the error-difference measure should decrease as cognitive load increases (Spears & Haslam, 1997, p. 183), if for different reasons. The new method allows one to distinguish between the two points of view empirically. Remember that guessing strategies based on fit can be evaluated by means of the response bias parameters a for guessing the category as demonstrated in Experiment 4, whereas the activation and accessibility of the social category in memory is measured by the category discrimination parameters d. Thus, under conditions of high fit, Spears and Haslam's argument leads one to expect effects of load primarily on the a parameters, whereas the d parameters should not be affected. In contrast, the activation hypothesis of Gilbert and Hixon also predicts effects on the d parameters.

The Role of Distinctiveness

A similar contribution is made by the new model in studying the effects of distinctiveness. Taylor (1981; cf. Taylor & Fiske, 1978; Taylor et al., 1978) has proposed that distinctiveness of a category directs attention to that category, raising its salience and its likelihood of being encoded and used in interpreting the social environment. In particular, Taylor et al. assumed that a category becomes more salient when the members of the category form a minority and that salience increases as the size of the minority decreases.

Oakes and Turner (1986) questioned this hypothesis and the experimental evidence for it (Taylor et al., 1978, Experiment 3). According to Oakes and Turner and Oakes et al. (1994, Chap. 3), a balanced composition of the group leads to highest category salience.

The "Who said what?" paradigm offers an unobtrusive method of investigating these two opposing hypotheses that was in addition accepted as a particularly suitable tool for doing so by both Taylor et al. (1978) and Oakes (1994). Previous experiments applying the paradigm to this question have compared so-called solo minorities with balanced groups (Biernat & Vescio, 1993; Taylor, Fiske, Close, Anderson, & Ruderman, 1975; but see Simon & Hastedt, 1997a; van Twuyver, 1996).

Again, the new approach has important advantages that recommend it for addressing this open theoretical debate. It naturally accomodates and corrects for differential sizes of categories in the paradigm. In addition, because it has been argued that distinctiveness also increases item memory (Hamilton, Dugan, & Trolier, 1985; Hamilton & Gifford, 1976), an analysis that partials out possible differences in item memory is required. Furthermore, response biases may favor the majority over the minority in the case of uncertainty, which is a third confounded factor that is controlled for in a validated fashion by the new approach. Finally, distinctiveness might increase the salience of individuating information for the distinct persons in addition to or instead of raising the salience of their category membership. As illustrated above, failing to control for these confounded factors entails the risk of reaching seriously flawed conclusions.

The Role of Fit

A second influential theory of the salience of social categories has used the concept of fit rather than distinctiveness (Oakes, 1987). Many of the studies in Table 1 have implemented a covariation of category membership and content of the statements, thereby generating fit. In these studies, the error-difference measure has often increased with increasing levels of fit as expected on the basis of Oakes's theory. However, fit may lead to enhanced categorization (a) on the basis of increased category memory or (b) through the use of the reasonable guessing strategy to attribute given statements to the category that fits their content. Because the effectiveness of this guessing strategy depends on the actual level of fit, studies that manipulate fit run the risk of trivially measuring this manipulated effectiveness rather than any real differences in the salience of the social categories. Again, by explicitly disentangling guessing from memory, the new method can help to avoid erroneous interpretations of this kind.

The separate process evaluation also opens up new lines of theoretical and empirical enquiry. From a social psychological point of view, Oakes's (1987) intriguing Accessibility × Fit analysis would of course by no means lose theoretical interest or practical impact if the effects of fit were shown to be mediated in a top-down manner (i.e., by guessing biases based on fit) rather than in a bottom-up manner (i.e., by enhanced statement-category associations in memory). Nevertheless, it is important to evaluate the separate contributions of both factors for theoretical reasons. First, functional dissociations between response biases and memory-sensitivity parameters are known to occur (e.g., Stangor & McMillan, 1992), indicating that both are differentially affected by the same antecedent variables. Second, social judgments are, in turn, differentially affected (a) by overall impressions, on which response biases are based, and (b) by memory for the presented items and their features such as the speaker's category membership (e.g., Garcia-Marques & Hamilton, 1996; Hastie & Park, 1986).

On the basis of this differentiated evaluation and taking into account the level of fit, the new approach might, for example, help to clarify the open question of why social categorization in the "Who said what?" paradigm has sometimes been found to show little relationship with relevant social judgments (e.g., Taylor & Falcone, 1982). In fact, theories of the memory–judgment link (Garcia-Marques & Hamilton, 1996; Hastie & Park, 1986) suggest that when perceivers are instructed to form an on-line impression of the discussion group, as they usually are in this task, the perceivers' ratings can be expected to covary with guessing bias based on overall impressions of the category subgroups (a parameter) more strongly than with the statement-category associations in memory (d parameters). Both aspects are, however, confounded in the conventional analysis, and guessing bias in particular can only contribute to the error-difference measure, provided there is some actual fit. Thus, the present analysis predicts that the correlation between the error-difference measure and social judgments is (a) moderated by level of fit and (b) mediated in a top-down manner by category-based expectancies. Hence, any manipulation affecting these expectancies and the associated a parameter of the present model is hypothesized to cause parallel effects in the rating measures as exemplified in Experiment 4 for a manipulation of fit.

Conclusion

An important issue in the article has been to demonstrate the ability and validity of the multinomial model to disentangle and isolate the sometimes opposing processes that jointly determine the pattern of errors in the "Who said what?" paradigm. Applying the model to measure these different processes requires a small procedural extension of the conventional "Who said what?" task by incorporating distractors in the test phase of the paradigm. We believe that this is a small price to pay for the detailed process evaluations obtained in return.

In addition to the validation of the model and its application to studying the effects of processing pace presented in the article, three further areas of application were sketched in which the model may provide a new perspective on resolving a number of open questions and debates of long standing. The model is able to do so largely because its complexity begins to match the complexity of the social psychological theories that have been brought to bear on the field of social categorization. We believe that such a match is a favorable, and quite possibly necessary, condition for scientific progress.

References

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

Andersen, S. M., Spielman, L. A., & Bargh, J. A. (1992). Future-event schemas and certainty about the future: Automaticity in depressives' future-event predictions. *Journal of Personality and Social Psychology, 63,* 711–723.

Arcuri, L. (1982). Tree patterns of social categorization in attribution memory. *European Journal of Social Psychology, 12,* 271–282.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer &

T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 1–40). Hillsdale, NJ: Erlbaum.

Bargh, J. A., & Tota, M. E. (1988). Context-dependent automatic processing in depression: Accessibility of negative constructs with regard to self but not to others. *Journal of Personality and Social Psychology, 54,* 925–939.

Batchelder, W. H., Hu, X., & Riefer, D. M. (1994). Analysis of a model for source monitoring. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 51–65). New York: Springer.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97,* 548–564.

Batchelder, W. H., & Riefer, D. M. (in press). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review.*

Batchelder, W. H., Riefer, D. M., & Hu, X. (1994). Measuring memory factors in source monitoring: Reply to Kinchla. *Psychological Review, 101,* 172–176.

Bayen, U., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 197–215.

Beauvais, C., & Spence, J. T. (1987). Gender, prejudice, and categorization. *Sex Roles, 16,* 89–100.

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

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70,* 1142–1163.

Blanz, M. (1996, September). *Ein erweitertes accessibility × fit—Modell zur Salienz sozialer Kategorien* [An extended Accessibility × Fit—model for the salience of social categories]. Paper presented at the meeting of the German Psychological Society, Munich, Germany.

Blanz, M. (1997). Soziale Kategorisierung und soziale Diskriminierung bei Exklusion versus Inklusion der eigenen Person [Social categorization and social discrimination with exclusion versus inclusion of self]. *Zeitschrift für Sozialpsychologie, 28,* 265–279.

Blanz, M. (in press). Accessibility and fit as determinants of the salience of social categorizations. *European Journal of Social Psychology.*

Blanz, M., & Aufderheide, B. (in press). Social categorization and category attribution: The effects of comparative and normative fit on memory and social judgment. *British Journal of Social Psychology.*

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology, 55,* 726–737.

Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science, 1,* 319–322.

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology, 52,* 871–880.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.

Brewer, M. B., & Harasty, S. (1996). Seeing groups as entities: The role of perceiver motivation. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (pp. 347–370). New York: Guilford Press.

Brewer, M. B., Weber, J. G., & Carini, B. (1995). Person memory in intergroup contexts: Categorization versus individuation. *Journal of Personality and Social Psychology, 69,* 29–40.

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124,* 137–160.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

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

Erdfelder, E., & Buchner, A. (1998). Process dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127,* 83–96.

Erdfelder, E., Murnane, K., & Bayen, U. J. (1995). Die Messung kognitiver Prozesse im Paradigma der Quellendiskrimination [The measurement of cognitive processes in the source-discrimination paradigm]. In K. Pawlik (Ed.), *Bericht über den 39. Kongreß der Deutschen Gesellschaft für Psychologie in Hamburg, 1994* (pp. 541–547). Göttingen, Germany: Hogrefe.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23,* 1–74.

Frable, D. E., & Bem, S. E. (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, D. L. (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, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60,* 509–517.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Hamilton, D. L., Dugan, P. M., & Trolier, T. K. (1985). The formation of stereotypic beliefs: Further evidence for distinctiveness-based illusory correlations. *Journal of Experimental Social Psychology, 48,* 4–17.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392–407.

Hamilton, D. L., & Sherman, J. W. (1994). Stereotypes. In R. W. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 2, pp. 1–68). Hillsdale, NJ: Erlbaum.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93,* 258–268.

Hewstone, M., Hantzi, A., & Johnston, L. (1991). Social categorization and person memory: The pervasiveness of race as an organizing principle. *European Journal of Social Psychology, 21,* 517–528.

Hu, X. (1991). Statistical inference program for multinomial binary tree models [Computer program]. Irvine: University of California at Irvine.

Hu, X. (1997). *GPT.EXE: Software for general processing tree models* [On-line]. Available WWW: http://141.225.14.108/gptgateway.htm

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59,* 21–47.

Jackson, A. L., & Hymes, R. W. (1985). Gender and social categorization: Familiarity and ingroup polarization in recall and evaluation. *Journal of Social Psychology, 125,* 81–88.

Johnston, L., Hewstone, M., Pendry, L., & Frankish, C. (1994). Cognitive models of stereotype change: IV. Motivational and cognitive influences. *European Journal of Social Psychology, 24,* 137–265.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology, 54,* 778–788.

Kai's PowerGoo [Computer software]. (1996). Hamburg, Germany: MetaTools.

Kelley, H. H. (1967). Attribution theory in social psychology. In D.

Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 192–238). Lincoln: University of Nebraska Press.

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101,* 166–171.

Klauer, K. C., & Batchelder, W. H. (1996). Structural analysis of subjective categorical data. *Psychometrika, 61,* 199–240.

Lippmann, W. (1922). *Public opinion.* New York: Harcourt Brace.

Lorenzo-Cioldi, F. (1993). They all look alike, but so do we . . . sometimes: Perceptions of in-group and out-group homogeneity as a function of sex and context. *British Journal of Social Psychology, 32,* 111–124.

Lorenzo-Cioldi, F., Eagly, A. H., & Stewart, T. L. (1995). Homogeneity of gender groups in memory. *Journal of Experimental Social Psychology, 31,* 193–217.

Lupfer, M. B., Clark, L. F., & Hutchinson, H. W. (1990). Impact of context on spontaneous trait and situational attributions. *Journal of Personality and Social Psychology, 58,* 239–249.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin, 107,* 401–413.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98,* 185–199.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23,* 77–87.

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.

McConnell, A. R., Sherman, S. J., & Hamilton, D. L. (1994). On-line and memory-based aspects of individual and group target judgments. *Journal of Personality and Social Psychology, 67,* 173–185.

McConnell, A. R., Sherman, S. J., & Hamilton, D. L. (1997). Target entitativity: Implications for information processing about individual and group targets. *Journal of Personality and Social Psychology, 72,* 750–762.

Miller, C. T. (1987). Categorization and stereotypes about men and women. *Personality and Social Psychology Bulletin, 12,* 502–512.

Miller, C. T. (1988). Categorization and the physical attractiveness stereotype. *Social Cognition, 6,* 231–251.

Newcomb, T. M. (1956). The prediction of interpersonal attraction. *Psychological Review, 60,* 393–404.

Oakes, P. (1987). The salience of social categories. In J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, & M. S. Wetherell (Eds.), *Rediscovering the social group: A self-categorization theory* (pp. 117–141). Oxford, England: Blackwell.

Oakes, P. J. (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, P. J., Haslam, S. A., & Turner, J. C. (1994). *Stereotypes and social reality.* Oxford, England: Blackwell.

Oakes, P. J., & 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, T. M., Carpenter, S. L., Sedikides, C., & Li, F. (1993). Differential processing of in-group and out-group information. *Journal of Personality and Social Psychology, 64,* 21–34.

Pendry, L. F., & Macrae, C. N. (1994). Stereotypes and mental life: The case of the motivated but thwarted tactician. *Journal of Experimental Social Psychology, 30,* 303–325.

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology, 27,* 26–47.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318–339.

Riefer, D. M., & Batchelder, W. H. (1991). Statistical inference for multinomial processing tree models. In J. P. Doignon & J. C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 313–336). New York: Springer.

Riefer, D. M., Hu, X., & Batchelder, W. H. (1994). Response strategies in source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 680–693.

Rosenberg, S., Nelson, C., & Vivekanathan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9,* 283–294.

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology, 14,* 237–255.

Simon, B., & Hastedt, C. (1997a). *The interplay of the individual self and the collective self in numerically defined minority and majority groups.* Unpublished manuscript, Westfälische Wilhelms-Universität Münster, Germany.

Simon, B., & Hastedt, C. (1997b). 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.

Snodgrass, J. W., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.

Spears, R., & Haslam, S. A. (1997). Stereotyping and the burden of cognitive load. In R. Spears, P. J. Oakes, N. Ellemers, & S. A. Haslam (Eds.), *The social psychology of stereotyping and group life* (pp. 171–207). Oxford, England: Blackwell.

Spears, R., Haslam, S. A., & Jansen, R. (1996, August). *Do social categories make light at the end of the load?* Paper presented at the International Congress of Psychology, Montreal, Quebec, Canada.

Stangor, C., & Duan, C. (1991). Effects of multiple task demands upon memory for information about social groups. *Journal of Experimental Social Psychology, 27,* 357–378.

Stangor, C., Lynch, L., Duan, C., & Glass, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology, 62,* 207–218.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin, 111,* 42–61.

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

Stroessner, S. J., Hamilton, D. L., & Mackie, D. M. (1992). Affect and stereotyping: The effect of induced mood on distinctiveness-based illusory correlation. *Journal of Personality and Social Psychology, 62,* 564–576.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Social Issues, 25,* 79–97.

Tajfel, H. (1972). La catégorisation sociale. [Social categorization]. In S. Moscovici (Ed.), *Introduction à la psychologie sociale* (Vol. 1, pp. 272–302). Paris: Larouse.

Taylor, S. E. (1981). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 88–114). Hillsdale, NJ: Erlbaum.

Taylor, S. E., & Falcone, H. (1982). Cognitive bases of stereotyping.

The relationship between categorization and prejudice. *Personality and Social Psychology Bulletin, 8,* 426–432.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top-of-the-head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 250–289). New York: Academic Press.

Taylor, S. E., Fiske, S. T., Close, M., Anderson, C., & Ruderman, A. (1975). *Solo status as psychological variable: The power of being distinctive.* Unpublished manuscript, Harvard University, Cambridge, MA.

Taylor, S. E., Fiske, S. T., Etcoff, N. J., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology, 36,* 778–793.

Turner, J. C. (1987). A self-categorization theory. In J. C. Turner, M. A. Hogg, P. J. Oakes, S. E. Reicher, & M. S. Wetherell (Eds.), *Rediscovering the social group: A self-categorization theory* (pp. 42–67). Oxford, England: Blackwell.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* Oxford, England: Blackwell.

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. (1996). *Factors affecting social categorization processes in memory.* Unpublished doctoral dissertation, Katholieke Universiteit Nijmegen, Nijmegen, the Netherlands.

van Twuyver, M., & van Knippenberg, A. (1995). Social categorization as a function of priming. *European Journal of Social Psychology, 25,* 695–701.

Walker, P., & Antaki, C. (1986). Sexual orientation as a basis for categorization in recall. *British Journal of Social Psychology, 25,* 337–339.

Winter, L., Uleman, J. S., & Cunniff, C. (1985). How automatic are social judgments? *Journal of Personality and Social Psychology, 49,* 904–917.

Yonelinas, A. P., & Jacoby, L. L. (1996). Response bias and the process dissociation procedure. *Journal of Experimental Psychology: General, 125,* 422–434.

Appendix A

A Practical Guide to Multinomial Modeling

The basic assumption of a multinomial model like the one proposed in the present article is that the observed response patterns can be seen as the final product of a number of different cognitive processes, each of which occurs with a certain probability. In tasks such as the "Who said what?" task, participants have a number of separate response options. The combination of cognitive processes determines how often each response option is chosen: In mathematical terms, the relative frequency of a response option is modeled by the product of the probabilities of the involved cognitive processes. The unknown probabilities of the different cognitive processes can therefore be estimated on the basis of these response frequencies as explained below. The present introduction to multinomial modeling is based on the pioneering work of Riefer and Batchelder (1988; Batchelder & Riefer, 1990). A comprehensive review of technical issues and applications was given by Batchelder and Riefer (in press).

The Structure of Multinomial Models

The different combinations of processes resulting in the responses are visualized in a multinomial processing-tree representation like the one in Figure 1. On the left side, there are three starting points in this model, representing the three situations a participant may be confronted with in each trial of the assignment task: The presented statement may be a statement made by a member of Category A, a statement made by a member of Category B, or a new statement. The combination of cognitive processes involved may be different for each of the three situations. Hence, the model comprises three different trees.

On the right side, all possible response options are depicted in Figure 1. Note that some of the response options occur several times in each tree, and thus there can be different combinations of cognitive processes resulting in the same response. In contrast to the observable starting points and the observable responses, the mediating cognitive processes are not observable. This fact is visualized by depicting the observable events in rectangles and the latent processes in ellipses.

Every cognitive process takes place with a certain probability, and these probabilities are represented by the parameters attached to the connecting lines between the ellipses. In the present model, there are always two alternatives at each branching: Either a cognitive process takes place with a certain probability or it does not take place with the complementary probability. Thus, the present model is called a *binary model,* but other models may sometimes have more than two options branching out from a given node.

Because the branches of each partial tree represent a combination of cognitive processes and because each cognitive process takes place with a certain probability, the probability for each branch is the product of the probabilities of all processes that constitute the branch. Consequently, the response option a branch ends in will be chosen with the joint probability of all involved cognitive processes. Because some branches end in the same response option, that option will ultimately be chosen with probability equaling the sum of several joint probabilities.

Note that not only response options but also cognitive processes can appear repeatedly in the same tree. Consider, for example, the process of guessing the category that is associated with parameter a. As can be seen, the process occurs twice in the first tree in Figure 1.

Sometimes it is assumed that the same cognitive process even occurs in different trees of a model as is the case for the guessing process associated with parameter a, among others, in the present model. Thus, we assume that the process of guessing that a statement was made by a speaker from Category A is the same regardless of whether the statement was actually made by a speaker from Category A, was made by a speaker from Category B, or was not presented in the discussion phase at all. Parameter a therefore appears in all partial trees of the present model.

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. Appendix C gives these frequencies for the experiments reported in the present article, broken down by experimental conditions.

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 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. In the present multinomial model, for example, the term

$$[D_A(1 - c_A)(1 - d_A)(1 - a)] + [(1 - D_A)b(1 - a)] \quad (2)$$

equals the response probability of assigning a statement made by a speaker from Category A to a (necessarily) wrong person of Category B (response option "Category B, wrong speaker" in the first tree in Figure 1). In words, this response can arise as the result of two different combinations of cognitive processes. First, it is possible that a statement is recognized as old (probability D_A) but neither the person who made it ($1 - c_A$) nor the speaker's social category can be remembered ($1 - d_A$), so that the respondent has to guess and, in this case, guesses a speaker belonging to the wrong category ($1 - a$). Second, it is possible that the statement is not recognized as old ($1 - D_A$) but is guessed to be old (b) and then guessed to originate from a person of Category B ($1 - a$). The sum of these two joint probabilities yields the overall probability of responding with a between-categories error given a statement made by a member of Category A.

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. In this iteration process, several combinations of probabilities are processed until the best solution is found, that is, until the difference between the observed frequencies and those that would be expected on the basis of the estimated parameters is minimized, where the difference is measured in terms of the ratio of the likelihoods of the observed and the expected frequencies.

Testing Hypotheses

In applying a model, a global goodness-of-fit test of the model should be conducted. The question of whether the observed response pattern can be explained by the model is answered using a goodness-of-fit test that evaluates the differences between the observed and the estimated response frequencies. To test the goodness of fit, one computes the likelihood ratio statistic G^2, which is asymptotically chi-square distributed. If this test does not yield a significant result, given a satisfactory statistical test power, it can be concluded that there are no substantial deviations from the model. In this case, the model fits the data. Note that the degrees of freedom have to be computed for the goodness-of-fit test. As a rule of thumb, they are given by the number of independent response options minus the number of parameters to be estimated. See the body of the article for examples.

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 that differs from the one obtained by fitting the model to the control condition. This hypothesis is statistically tested by doubling the model and treating it 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 chi-square-distributed loss value are given by the difference in the degrees of freedom of the two models.

Before a new model can be used to examine the cognitive processes involved in a given task, its substantive interpretation should be validated properly. The global goodness-of-fit test examines whether a certain probability structure can explain the frequencies with which the different response patterns were observed. Because many different models with different substantive implications can usually be fitted to the same set of data, the goodness-of-fit test does not imply that the parameters really measure the intended cognitive processes. Hence, a series of experiments should aim at validating the different process parameters as discussed at length in the body of the article.

Software

Computer software for estimating parameters and testing the models is available on the Internet (http://xhuoffice.psyc.memphis.edu/gpt/index.htm) from Xiangen Hu. There are two programs available: an older DOS version called the statistical inference program for multinomial binary tree models (MBT; Hu, 1991) and a Windows95/NT version called the statistical inference program for general processing tree models (GPT; Hu, 1997).

For working with MBT, an ASCII document containing the system of equations of the model and an ASCII document containing the aggregated data are needed. After reading the two files, the program asks for restrictions of the model. For example, parameters can be constrained to be equal, or they can be fixed at a user-determined value. Then the parameters are estimated, and an ASCII document is created containing parameter estimates, their 90% confidence intervals, the chi-square value of goodness of fit, and the estimated and the empirical frequency distribution as well as the so-called Fisher information matrix from which the confidence intervals are computed (Rao, 1973). In addition, the program checks the identifiability of the model, that is, whether the system of equations can be solved uniquely. For this purpose, several parameter estimations are conducted from different starting values of the parameters, and it is necessary to check whether the final estimates for each parameter are always the same..

GPT can work with ASCII documents containing equation systems and aggregated data or, alternatively, the processing-tree representation of the multinomial model can be drawn in which case the data have to be entered via the keyboard. Additional options like analysis of power and model simulations are integrated. Whereas MBT can handle models with up to only 61 branches, GPT is able to deal with very large models.

Appendix B

Model Assumptions and Technical Issues

Threshold Theories Versus Statistical Decision-Theory Models

Most current theories of recognition make use of a model of decision that is derived from Green and Swets's (1966) continuous statistical decision-theory model of signal detection, whereas the item detection part of the present model is a discrete, two-high-threshold model of signal detection. As explained in the body of the article, the present model postulates a finite number of discrete processing states for item recognition: There is a state in which the participant correctly detects targets as old, there is another state in which distractors are correctly detected as new, and there are states in which the participant is uncertain about the status of the item and in which guessing processes prevail. In a two-high-threshold model, there are two thresholds that divide the decision space into three discrete areas corresponding to these states. If either threshold is crossed on presentation of a test item, the item is detected as either old or new depending on which threshold was crossed. If neither threshold is crossed, the item is in an undetected state and the participant responds either "old" or "new" depending on a guessing probability b (Bayen et al., 1996, p. 201). In two-high-threshold models, it is assumed that only old items can cross the detect-as-old threshold and only new items can cross the detect-as-new threshold. In one-high-threshold models, there is only one threshold, namely, the detect-as-old threshold, which only old items can cross.

The relative value of statistical decision-theory models and one-high-threshold models of the decision component has been discussed by Kinchla (1994) and Batchelder, Riefer, and Hu (1994) and more recently by Yonelinas and Jacoby (1996) and Erdfelder and Buchner (1998) in the context of a related multinomial model. The one-high-threshold version of the present model is obtained under the assumption that the ability to detect distractors is zero ($D_N = 0$).

Although the one-high-threshold model for simple item detection is generally considered inferior to the statistical decision-theory model, both so-called low-threshold and two-high-threshold models of simple item detection fare very well when compared to models derived from statistical decision theory (Macmillan & Creelman, 1990; Macmillan & Kaplan, 1985; Snodgrass & Corwin, 1988; Swets, 1986). As has been said, the present model builds on a two-high-threshold model of item detection.

Data Collection Procedure

A standard assumption in most areas of modeling as well as in multinomial modeling is that the observations are realizations of independently and identically distributed random variables. In practice, data are collected from several participants, and several observations are usually obtained from each participant. This approach runs the risk of violating the assumption of identical and independent distributions. In particular, if there are large individual differences and within-subject dependence, the estimated confidence intervals may be too narrow.

Some amount of Monte Carlo study has been done to evaluate the severity of this problem and has tended to indicate that the models are fairly robust under small violations in the sampling assumptions (Riefer & Batchelder, 1988, 1991; Riefer et al., 1994). In particular, the size of biases in parameter estimates and confidence intervals was found to decrease very quickly as the number of data observations with different underlying parameters increased, indicating that the effects of individual differences cancel out. Moreover, in the present experiments, we took care to ensure that participants were sampled from a relatively homogeneous population, so that we expected minimal individual differences in memory processes: All participants were students, and all of them were native speakers of German. With respect to sampling from less homogeneous populations, the Monte Carlo results suggest using somewhat larger samples of participants while collecting fewer data observations per participant, thereby quickly reducing the impact of individual differences.

Sample Size and Accuracy of Estimation

A related question concerns the appropriate number of data observations required for reliable estimation of parameters. The statistical theory on which the estimation is based is a so-called large-sample theory, that is, the confidence intervals and chi-square tests are only approximations that become successively more accurate as the sample size increases. This raises the question of how large an N is required. No general answer has been given, but the rule of thumb that the expected cell frequencies of the data matrix should generally exceed the value five was upheld in the present experiments.

Sample size is also the major factor influencing the size of the standard errors and confidence intervals of parameter estimates. For a fixed sample size, on the other hand, measurement accuracy for a given parameter is also influenced by substantive boundary conditions. For example, in situations with poor item memory or good person discrimination, the amount of information about the category discrimination parameters in the data is low, if for different reasons. When item memory is poor, the proportion of assignments based on guessing for forgotten statements increases, lowering the impact of category discrimination on the pattern of errors as discussed above (cf. the Problems of the Conventional Approach section). When person discrimination is excellent, on the other hand, few assignment errors will occur for statements labeled "old" by the participant, again lowering the impact of category discrimination on the pattern of errors. Thus, in terms of the amount of measurement error in estimating category discrimination parameters, optimal conditions are realized when item memory is relatively good and person discrimination is far from perfect. If the focus is on other processes, however, other sets of conditions can be more favorable. For example, to assess response bias in guessing categories, less than perfect item discrimination is favorable (cf. Experiment 4).

Differential Item Memory

In many of the above model analyses, the assumption $D_A = D_B = D_N = D$ was made. In particular, statement memory was assumed not to be a function of category, an assumption that was tested by the goodness-of-fit test of the model (cf. Batchelder & Riefer, 1990) and that could in fact be maintained in all cases.

In some situations, however, there are theoretical reasons to believe that item memory may vary as a function of category. For example, a number of studies have directly or indirectly manipulated motivation to attend to one category rather than the other (e.g., Brewer et al., 1995; Simon & Hastedt, 1997b; cf. Table 1; cf. also Experiment 4) and have in part obtained evidence for differential processing as a function of category.

In these cases, it is desirable to be able to obtain separate estimates of D_A and D_B. Both the model with $D_A = D_N = D$, and separate D_B, as well as the model with $D_B = D_N = D$, and separate D_A, can be shown to be identified, and they provide separate estimates of D_A and D_B (cf. Bayen et al., 1996; cf. Experiment 4). Both models are saturated—that is, they have as many parameters as there are degrees

of freedom in the data. In situations with several groups, alternative sets of restrictions are possible to make the model identified as exemplified in Experiment 4. Finally, it is possible to enlarge the database structurally to obtain more degrees of freedom. For example, as in Experiment 1, a manipulation of response bias can be implemented to realize conditions that differ only with respect to detection bias b. A model that allows for different b parameters and sets equal all other parameters over the conditions that differ only with respect to the response-bias manipulation is identified even if D_A, D_B, and D_N are allowed to vary freely otherwise.

Appendix C

Data Matrices of Experiments

		Cells of basic data matrix										
Experiment	Group	1	2	3	4	5	6	7	8	9	10	11
1	Few distractors	77	120	38	29	105	40	83	36	19	13	496
	Many distractors	70	80	42	48	83	38	67	52	15	20	1405
2	Low similarity	79	78	35	48	83	36	80	41	15	12	453
	High similarity	65	72	34	69	65	25	77	73	61	65	354
3	Academic status	61	88	38	53	71	38	95	36	20	21	439
	Hometown	37	66	78	59	54	77	54	55	15	18	447
4	Critical men, +	133	187	64	102	49	29	50	34	23	9	616
	Critical men, −	55	47	28	32	178	51	167	90	16	28	604
	Critical women, +	36	57	29	40	150	81	144	111	24	31	593
	Critical women, −	139	142	80	125	51	18	53	40	23	19	606
5	Similar men	76	64	43	57	82	31	69	58	19	22	439
	Similar women	49	94	53	44	76	38	75	51	28	26	426
6	2.5 s	36	64	41	99	41	42	69	88	50	38	392
	5 s	64	74	35	67	63	38	80	59	19	13	448
	7.5 s	80	88	40	32	71	30	102	37	10	7	463
	10 s	78	76	34	52	90	24	71	55	13	12	455

Note. + = positive statements; − = negative statements.

Received September 30, 1997
Revision received April 16, 1998
Accepted May 20, 1998 ■