Personality and Social Psychology Review
17(1) 72–86

DOI: 10.1177/1088868312461308
http://pspr.sagepub.com

Accuracy in Categorizing Perceptually Ambiguous Groups: A Review and Meta-Analysis

Konstantin O. Tskhay[1] and Nicholas O. Rule[1]

Abstract

Since the 1940s, social psychologists have conducted research testing whether it is possible to accurately identify members of perceptually ambiguous groups. This study quantitatively reviews the research on the perception of ambiguous groups to better understand the human capacity to accurately identify others based on very subtle nonverbal cues. Standard random-effects meta-analytic techniques were used to examine the distinctions between different target groups in terms of their identifiability, as well as to compare rates of accuracy across perceptual modalities (e.g., photographs, audio, video) and other study design differences. Overall, the accuracy of identifying targets was significantly better than chance guessing (i.e., 64.5%). Furthermore, stimulus modality was found to be a moderator of accuracy. Other moderators (e.g., time of exposure, analytic approach) were identified and examined. These data help to document and characterize broad trends in the proliferating and expanding study of the perception and categorization of ambiguous social groups.

Keywords

person perception, social categorization, nonverbal behavior, accuracy, meta-analysis

One of the fundamental questions in person perception research concerns the processes that are involved in the categorization of people into groups (Macrae & Quadflieg, 2010). In their review of social categorization, Macrae and Bodenhausen (2000) made a compelling case for person perception as a rapid and automatic process. Indeed, when two individuals meet for the first time, they immediately begin to make inferences about each other based on physical characteristics: clothing, hairstyle, body type, and even individual facial features. For example, Berry and McArthur (1985) found a significant correlation between facial appearance and the personality traits ascribed to a person. Individuals who appeared more babyfaced were perceived to possess warm traits (e.g., honesty, naivete, kindness; Berry & McArthur, 1985), whereas individuals with more mature faces tended to be seen as more powerful (e.g., competent, shrewd; Zebrowitz, 1997).

People infer much more than just personality traits at first acquaintance, however. Indeed, some of the first things that people immediately and automatically assess when seeing another person are age (Wright & Stroud, 2002), race (Richeson & Trawalter, 2005), and sex (Macrae & Martin, 2007; see Brewer, 1988, for additional review). These groups are obvious—we can easily identify these groups in our environment. Furthermore, we are rarely confused about these identities.

Although much research in person perception has concerned the perception of groups with physical markers that are perceptually obvious, relatively little work has been done to explore the categorization of groups that are perceptually ambiguous. Perceptually ambiguous group members are those who, though belonging to a certain social category, may not naturally express any visible markers that would identify them as such, yet may still have features that could be utilized for correct identification. The majority of groups to which we may belong (e.g., professions, religious groups, political parties) are ambiguous, yet research has nonetheless shown that many of these distinctions are perceptible. The present work therefore provides a comprehensive review and meta-analysis of this research with the goal of understanding the antecedents, processes, and outcomes of perceiving group membership from nonobvious cues. In doing so, we seek to consolidate the findings of the past decades and to provide an indication of the magnitude (effect size) with which people are able to make these judgments.

[1]University of Toronto, Ontario, Canada

Corresponding Author:
Konstantin O. Tskhay, Department of Psychology, University of Toronto, 100 St. George Street, Toronto, Ontario, M5S 3G3, Canada
Email: konstantin.tskhay@mail.utoronto.ca

The Roots of Ambiguous Group Research: The Categorization of Jewish Group Membership

Since the Holocaust, social psychologists have keenly studied the nature and effects of prejudice (e.g., Allport, 1954). The prejudice framework was largely based on the idea that the mind creates categories and constructs stereotypes as heuristics that ease and enhance interaction with the environment (Bodenhausen, 1990; Fiske & Taylor, 1984). Specifically, Allport (1954) theorized that prejudice against a particular group is a product of affective states and categorical features associated with that group. In this view, prejudice requires the capacity to categorize individuals, which rests on the necessity of perceiving cues that are rather salient—such as those demarcating race and sex (Allport, 1954). To date, multiple studies have examined the ability to discriminate Jewish people from those belonging to other groups (see below).[1]

Historically, Jewish individuals have been the targets of an enormous amount of prejudice. Even though Jews might possess some features that distinguish them from non-Jews, these markers are not always diagnostic of identity; therefore, Jews were required to identify themselves by wearing armbands and patches on their clothing during their public persecution in Nazi Germany (Allport, 1954). Even so, many people at the time believed that they had an inborn ability to correctly identify who was Jewish, even in the absence of these explicit identifiers. This folk belief was empirically investigated.

Allport and Kramer (1946) randomly presented 20 yearbook photographs of Jews and Non-Jews to 223 undergraduate students for 15 s each and asked them to categorize the person in each photograph as Jewish or non-Jewish, or to pass on the trial by indicating a lack of knowledge. The reported median identification for the sample was slightly above chance (55.5%; Allport & Kramer, 1946). Moreover, they found that highly prejudiced people were more accurate at distinguishing Jews from non-Jews. Lund and Berg (1946) used a different procedure to investigate this belief. The researchers asked 18 judges to guess the religious affiliation of 2,875 subjects (ranging from preschoolers to university students) passing through two rooms. In the first room, the subjects were judged solely on physical appearance. In the second room, however, the subjects were interviewed, allowing the judgment to be based on physical appearance and speech. The results showed that the live presentation of subjects to judges produced astonishing results: the judges were accurate in the categorization of approximately 87% of targets (see also Savitz & Tomasson, 1959, for similar results). This suggests that dynamic information from live presentation and vocal cues could aid identification of Jewish group membership.

Carter (1948) investigated Allport and Kramer's (1946) claim that Jews could be discriminated from non-Jews by proposing a more standardized method of obtaining target stimuli and testing whether Jews might be confused with individuals of Mediterranean background. Even though Carter was able to show that Jews were indeed miscategorized with targets of Mediterranean descent, the results still suggested that Jews were categorized better than chance overall. Lindzey and Rogolsky (1950) used additional standardization of target stimuli to ensure that the original study results replicated. The researchers created a picture database by selecting 200 photographs from yearbooks of Jewish and non-Jewish student organizations (i.e., 100 each). They further standardized the sample by asking an independent group of judges to rate each picture on how "Jewish" the person appeared to them. The photographs from both groups were then rank ordered, and every tenth picture was chosen for use as a stimulus, thus yielding a set of 10 pictures for each group that represented varying degrees of differentiation. In addition, the researchers asked raters to select six pictures from each group that had the most "Jewish" and "non-Jewish" appearance. The three pictures that were most frequently chosen from each group were assumed to represent extreme cases. A separate sample of 685 students judged the 26 faces as "Jewish" or "non-Jewish," or abstained from judgment by indicating that they did not know. Accuracy was again significantly above chance, replicating the previous results.

Although the effects for distinguishing Jewish and non-Jewish targets seemed to replicate well, social psychologists at the time did not account for guessing or "response bias" (i.e., the tendency for participants to respond in a certain way, such as overestimating the number of non-Jewish targets). Elliott and Wittenberg (1955) therefore attempted to control participants' response bias by using different ratios of Jewish to non-Jewish targets (i.e., 15:5, 10:10, 5:15), tested across three independent samples of participants. The findings contradicted previous conclusions, as accuracy for the three samples was at chance. Following these studies, Scodel and Austrin (1957) conducted a study that also attempted to account for response bias by using a sample of targets with a 30 Jewish to 70 non-Jewish ratio. They too found that Jews could not be accurately discriminated from non-Jews.

Later studies that also accounted for guessing biases provided further indication for above-chance accuracy, however. Pulos and Spilka (1961) selected photographs of 20 Jewish and 20 non-Jewish individuals from a set of 82 yearbook photographs of men in cap and gown, thereby limiting cues from attire that might communicate information about group membership. The 40 selected targets had been pre-rated to represent the middle of the distribution of how "Jewish" they appeared. Furthermore, the researchers attempted to statistically control for response bias by considering the participants' response tendencies (i.e., favoring one category over another) as a covariate in their analyses. They reported that high- and low-prejudiced people could discriminate between Jewish and non-Jewish targets better than

chance. To increase the generalizability of these results, Himmelfarb (1966) tested male and female targets. In his first study, he found that the overall accuracy of identifying Jewish targets was better than chance, with subjects scoring higher for male photographs than for female photographs. The study also demonstrated a moderately-sized positive correlation between the participants' scores on a measure of anti-Semitism and the accuracy of their categorizations, similar to Allport and Kramer (1946). In his second study, he used a forced-choice technique in which participants had to choose which of two photographs projected on a screen was of a Jewish person. This ostensibly allowed him to obtain an accuracy score that was free of response bias. Again, participants categorized targets better than chance, replicating earlier work. Finally, Dorfman, Keeve, and Saslow (1971) controlled for response bias by using analyses based on signal detection theory (see Sporer, 2001, for a review). Whereas the traditional percent-correct analysis takes only hits and correct rejections into account, signal detection estimates consider hits alongside false alarms, allowing for a more precise examination of accuracy while accounting for response bias—systematic and random guessing. As such, signal detection analysis gives researchers an opportunity to examine the true accuracy of the participants, as well as a statistically derived estimate of response bias. Thus, even after Dorfman et al. (1971) controlled for response bias, the accuracy of identifying group membership was again above chance (see also Quanty, Keats, & Harkins, 1975).

The heterogeneity of these results across several decades of research ultimately led to two meta-analytic reviews focusing on the accuracy of judgments of Jewish group membership (Rice & Mullen, 2003) and the relationship between perceptual accuracy and prejudice (i.e., anti-Semitism; Andrzejewski, Hall, & Salib, 2009). Rice and Mullen (2003) found a small, yet significant, effect for the accuracy of categorizations: overall, judges were able to discriminate between Jewish targets and non-Jewish targets better than chance. More recently, Andrzejewski et al. (2009) not only provided a meta-analytic review of the relationship between anti-Semitism and accuracy in judging Jewish group membership, but they also conducted five new studies testing the accuracy of identifying Jewish and non-Jewish individuals using signal detection analyses. Their findings supported the general conclusions of the previous research, showing that accuracy was above chance for judgments made from still photographs. Overall, both meta-analyses showed that the effect of accurately categorizing individuals as Jewish and non-Jewish was small, yet greater than chance (Andrzejewski et al., 2009; Rice & Mullen, 2003).

"Gaydar"

Similar to the once popular belief that Jews and non-Jews could be distinguished, the folk belief in "gaydar" is a pervasive one in society today. "Gaydar" (derived from a combination of the words *gay* and *radar*) describes the ability to accurately distinguish others' sexual orientation based on numerous indirect, often nonverbal, cues (Shelp, 2002). Given the highly contentious political environment surrounding gay rights in recent decades, the question of whether sexual orientation can be accurately perceived has held a high degree of relevance (e.g., the US military's recently repealed "don't ask, don't tell" policy; Policy Concerning Homosexuality in the Armed Forces, 1993). Thus, identifying someone who is gay holds numerous social implications, such as rejection, discrimination, and prejudice (Herek, 2004), with a high potential impact for the target's safety and well-being (e.g., Herek, Cogan, & Gillis, 2002).

Berger, Hank, Rauzi, and Simkins (1987) set out to investigate if sexual orientation could be identified at better than chance levels. Video stimuli were created from brief interviews with 24 gay and straight men and women (i.e., 6 targets in each gender × sexual preference combination) presented to 143 judges. The authors reported that targets' sexual orientations could not be identified any better than chance. However, a reanalysis by Hallahan (1998) showed that if the unit of analysis was judges instead of targets, the results would achieve statistical significance due to increased sample size and, consequently, greater statistical power (see Ambady, Hallahan, & Conner, 1999, for an additional review). Ambady et al. (1999) further examined whether sexual orientation could be identified from "thin slices" of behavior. Similar to Berger et al. (1987), the researchers relied primarily on video clips, but they showed only brief segments (10 s and 1 s) of the videos to the participants when making their judgments. In addition, some participants viewed only still frames taken from the videos. Judgments of the targets' sexual orientations were significantly correlated with the targets' self-reports, with viewing modality (i.e., video versus still photograph) providing a moderating effect. Accuracy for identifying sexual orientation, even though greater than chance for all of the conditions, was greater in the 10-s video clip than in the 1-s video clip, both of which were greater than still photographs.

Equipped with such strong evidence, research on the perception of sexual orientation expanded to identifying various aspects of the categorization process, such as the minimum amount of time necessary to make a correct identification, and examination of the features that people attend to when correctly perceiving targets' sexual orientations. Rieger, Linsenmeier, Gygax, Garcia, and Bailey (2010), for example, explored the possibility that perceivers use sex atypicality as a cue to determine targets' sexual orientation. Replicating Ambady et al.'s (1999) study, they found that sexual orientation was rated significantly better than chance from judgments of movement, speech, appearance, and targets' stated interests. Moreover, they found that individuals perceived as gay were also perceived to be sex atypical, suggesting that people use stereotypes of sex inversion to

identify who is gay and who is straight (see also Johnson, Gill, Reichman, & Tassinary, 2007). Extending this work, Sylva, Rieger, Linsenmeier, and Bailey (2010) conducted a study to see whether gays and lesbians could effectively conceal their sex atypicality to hide their sexual orientation. The results showed that gay and lesbian targets were distinguishable from straight targets whether or not they had deliberately attempted to conceal their sexual orientation. Furthermore, Freeman, Johnson, Ambady, and Rule (2010) found that accurate judgments of sexual orientation from faces were also largely moderated by gendered facial cues such that men with stereotypically female face shapes and skin textures were judged as gay more often than straight. Even though sex atypicality appears to be an effective indicator of individuals' sexual orientations, it is not the only cue available to perceivers (Freeman et al., 2010).

Other research studies found that male sexual orientation could be judged better than chance from separate facial features (e.g., mouth, eyes, hair; Rule, Ambady, Adams, & Macrae, 2008) and that face-based judgments appear to occur automatically (Rule, Macrae, & Ambady, 2009). Rule, Ambady, and Hallett (2009) further expanded the findings to female targets, showing that female sexual orientation was judged at better than chance levels from photographs of eyes and that deliberation about categorization disrupted accuracy, again suggesting that judgments of sexual orientation occur automatically. Moreover, gay and straight perceivers show better memory for gay and straight targets, respectively (Rule, Ambady, Adams, & Macrae, 2007), which suggests that after the features relevant to sexual orientation are identified, ingroup members are allocated more cognitive resources than are outgroup members. Interestingly, Rule and Ambady (2008) found that sexual orientation was identified from photographs with greater than chance accuracy from very "thin slices" (e.g., 50 ms), which suggests that the process of collecting and integrating the cues needed for categorization occurs rapidly.

Visual cues are not the only ones that allow for accurate categorization. The voice seems to provide another salient set of cues that aids people in correctly distinguishing between gay and straight individuals. For instance, Linville (1998) presented recordings of men speaking (5 gay, 4 straight) to 25 listeners and found that they were able to successfully identify the sexual orientation of the speaker. Furthermore, she also found that the specific vocal features supporting categorization were duration and peak frequency of the "s" sound. This finding suggests that there may be a systematic difference in the way that gay and straight people speak. Similarly, Gaudio (1994) reported that pitch variability could play a role in the perception of sexual orientation, among other phonetic factors. Smyth, Jacobs, and Rogers (2003) attempted to create a gay- and straight-sounding voice bank that could be used in further research on the phonetic correlates of speakers' perceived sexual orientation. In the process, they found that men reading a scientific passage were rated to be more "gay sounding" as compared with other passages they read. Furthermore, the researchers warned that a speech sample alone is probably not the ideal way of judging the sexual orientation of random targets, as many straight speakers in their study were perceived to be gay and the target with the "straightest" voice was actually a gay man. More recently, Zimman (2010) reported that there seems to be no one particular cue that aids listeners in categorizing someone as gay from the voice but, rather, that there are multiple styles that perceivers use, all of which seem to be deviations from the typical straight-sounding voice. Indeed, Munson and Babel (2007) also suggested that there may be multiple vocal cues beyond mere sex atypicality that aid listeners in categorizing voices as gay and straight.

Religious and Political Affiliation

Not unlike "gaydar," members of many religious groups also believe that they can distinguish followers of their faith from nonfollowers based on physical appearance. One such group is the Church of Jesus Christ of Latter-Day Saints, more commonly referred to as "Mormons." Their belief about religious-based intergroup discriminability is largely grounded in the idea that Mormons are spiritually distinct from other groups (Church of Jesus Christ of Latter-Day Saints, 2004). Rule, Garrett, and Ambady (2010a) investigated the idea that Mormons could be distinguished from non-Mormons and found that accuracy was greater than chance among Mormon and non-Mormon judges, regardless of whether the judges were from regions with very high or relatively low numbers of Mormons. Moreover, Mormons and non-Mormons showed significantly better memory for ingroup members' faces, respectively, without being explicitly told any of the targets' religious affiliations or any mention of religion in the task instructions. Further investigation of the facial cues that allowed for accurate categorization found that these judgments were based on perceptions of health (Rule, Garrett, & Ambady, 2010b). Thus, individuals with healthier skin were more likely to be categorized as Mormon, and since Mormons are actually healthier than non-Mormon controls (Enstrom, 1989; Enstrom & Breslow, 2008), these judgments seemed to be the "kernel of truth" that forms the basis for participants' accuracy. Thus, groups defined by differences in beliefs may in some cases be distinguishable.

Religious groups often share more than just their beliefs, however, and it may therefore not be surprising that they might be distinct. Yet there are other ideologically distinct groups that are less insular and can still be distinguished from their appearance. Some evidence has shown that members of the two major US political parties, Democrats and Republicans, can be categorized with greater than chance accuracy based on just their faces (Rule & Ambady, 2010). In contrast, Benjamin and Shapiro (2009) presented participants with 10-s silent video clips of political debates

between Republicans and Democrats and found that political orientation could not be distinguished any better than chance. Further research by Olivola and Todorov (2010), however, speculated that the sample size of Benjamin and Shapiro's study did not allow for sufficient statistical power to yield significant results. They analyzed data from 1,005 participants and found accuracy that was significantly greater than chance, regardless of whether participants were informed of the base rate (e.g., proportion of Democrats) in the target sample.

Similar to this, Samochowiec, Wänke, and Fiedler (2010) also reported that left versus right differences in political orientation, more generally, could be identified with above-chance accuracy from 15-s silent video clips and photographs of European political candidates. Therefore, this effect is not specific to American politicians, but may represent a more universal difference in liberal versus conservative ideologies. Indeed, several studies have shown that perceptions of personality dimensions can predict political candidates' electoral success across multiple cultures (Antonakis & Dalgas, 2009; Poutvaara, Jordahl, & Berggren, 2009; Rule, Ambady, et al., 2010), and Rule and Ambady (2010) found that stereotypes of personality traits (i.e., warmth and power) related to particular political groups moderated the effects of accurate categorizations. Thus, it is possible that these categorizations are based on stereotypes of inferred personality traits that otherwise contain a kernel of truth. This again supports the notion that social categorization is largely based on heuristics and stereotypes about various groups in the environment (Bodenhausen, 1990; Macrae & Bodenhausen, 2000).

Present Work

Categorization is a rapid and automatic process (Macrae & Bodenhausen, 2000). Many studies have demonstrated near-perfect accuracy in categorizing individuals belonging to perceptually obvious groups (e.g., 99.2% for race; Remedios, Chasteen, Rule, & Plaks, 2011). Ambiguous groups, however, are less studied and may provide valuable insight into a broader and more comprehensive understanding of the processes undergirding person perception. That is, although categorization by race, for example, is an automatic process and largely relies on the perception of skin tone (e.g., Maddox, 2004), it does not allow for an in-depth study of other cues that may be involved in person perception. As such, the study of ambiguous groups is a gateway into unraveling the encoding and processing of other cues that may illuminate how we are able to make rapid, accurate, and automatic categorizations of all social groups.

We therefore conducted a meta-analysis with the hope of summarizing the work on perceptually ambiguous groups that has been conducted to date. Meta-analysis is a quantitative review of the research literature and is used to summarize research findings. As such, meta-analysis provides effects that are representative of multiple studies and is derived through careful examination of research results. Furthermore, this statistical technique not only allows for consolidation of the findings reported in the scientific literature, but it also reveals future directions for empirical work in the field. The overall effect size derived from the application of such a powerful and precise method is robust and could be used for power analysis and sample size estimation a priori, and the variables found to moderate these effects should be considered in the design and hypotheses of future studies, as well as to suggest new avenues of research.

Meta-analytic techniques have many advantages (Rosenthal, 1991). One of the benefits of meta-analysis is that it addresses the dichotomous nature of statistical significance by focusing on the size of the effect instead of on the relatively arbitrary nature of *p*-value thresholds. Many studies may yield nonsignificant results and may not be accepted for publication or even drafted into manuscripts, as is common with studies that were unable to collect a sample sufficient to reach statistical significance. Although some studies may not reach a *p* value of .05, they may nevertheless replicate the effect sizes of other, published studies. It is of great importance to include these studies in meta-analytic reviews, as they also demonstrate an effect. Another advantage of meta-analysis is that it can reveal moderating effects present across studies that may not otherwise be captured in any single study, highlighting potentially important trends in methodological differences (Rosenthal & DiMatteo, 2001; see also Lipsey & Wilson, 2001). Last, it is important to mention that meta-analysis is a much more precise approach to the reviewing of scientific literature than are descriptive reviews because it is more objective and robust against the biases that the authors of such reviews may hold.

There are two main types of meta-analysis that can be conducted: fixed effects and random effects (Rosenthal 1991). In this review, we focus on random-effects meta-analysis in line with suggestions from Hunter and Schmidt (2004). The random-effects approach was chosen because it is a more conservative analysis that requires fewer assumptions. This approach is also broader because it attempts to encompass similar, yet hypothetical, studies by statistically assuming that the sample of studies included in the meta-analysis was drawn from a population of similar studies. Furthermore, considering that the assumptions of the fixed-effects meta-analysis are rarely met in practice, random-effects analyses tend to be more appropriate and desirable, as well as more statistically conservative (Hunter & Schmidt, 2004).

The present meta-analysis aims to estimate the size of the effect for accurately categorizing perceptually ambiguous groups. Even though it seems that ambiguous groups can be identified with greater than chance accuracy, the estimated effect sizes vary between studies, and it is therefore important to quantitatively estimate the overall effect that would reflect the findings to date. We also want to emphasize the importance of studying ambiguous groups in addition to the now vast literature on the perception of obvious groups.

Even though most published studies examining ambiguous groups have provided effects for accuracy that are significantly greater than chance guessing it is important to consider methodological differences between the studies to understand what features of the methodology increase or decrease the size of the effect. Some studies examining sexual orientation, political orientation, and Jewish group membership have used different modalities for the presentation of targets (e.g., video, audio, still photographs). The researchers also manipulated the number of targets as well as the source from which the targets were obtained. Thus, in this work, we aimed to determine the overall mean effect for accuracy in categorizing ambiguous groups. Moreover, the modality of presentation and other variables were examined as potential cross-study moderators of the effects for these four groups.

Method

Procedure

The search for studies testing the accuracy of categorization of ambiguous groups was conducted using standard literature search techniques. The keywords used in the search of computerized databases (i.e., PsycInfo, The Web of Knowledge, Google Scholar) were *person perception, accuracy, categorization, gay, gaydar, ambiguous groups, Mormon, Democrats, Republicans, politicians, religion, Jewish*, and *Jew*. The studies in the review survey these specific groups and were all written in English. Furthermore, because studies on ambiguous groups are relatively uncommon, the cross-reference technique (Rosenthal, 1991) was used to obtain additional studies. Thus, we examined ascending and descending citations to retrieve additional studies that could be included in the analysis. Any studies that tested for accuracy in the perception of ambiguous groups were eligible for further processing. In addition, a request for unpublished data was made via the electronic listserv of the Society for Personality and Social Psychology. Joshua A. Tabak and Vivian Zayas (2011) shared two manuscripts, which yielded five independent effect sizes for inclusion in this study. In sum, 61 articles were identified initially for further processing.

Inclusion criteria. To warrant inclusion in the present meta-analysis, the participants in the studies had to have engaged in an attempt to identify the category membership of an ambiguous group member. Ambiguous group membership was defined as lacking clear expression of salient characteristics that could theoretically aid group identification. As such, studies examining obvious group membership (i.e., age, race, and sex) were not included in the analysis, whereas sexual orientation, political affiliation, Jewish, and Mormon target groups were included in the analysis. Furthermore, the studies had to provide estimates of effect size, statistical measures of accuracy that could be converted to an effect size index, or enough information to compute an effect size. For example, if a study provided some information in accordance with signal detection theory (e.g., hits and correct rejections), the percent correct could be computed by adding the two values and dividing by the total number of targets presented to the participants. Furthermore, the percent correct could be converted into the proportion index π, which shows the proportion of correct categorizations on a scale with a null value of .50, regardless of the number of choices used in the actual study (Rosenthal & Rubin, 1989).

In the next step, the proportion index π was transformed into the standard *r* (Pearson's correlation coefficient) effect-size statistic by using a formula derived from the transformation of the correlation coefficient *r* into the binomial effect-size display (Rosenthal & Rosnow, 2007). Many studies included in the analysis provided the more conservative percent-correct equivalent *A'*, which was converted to a Pearson's correlation coefficient *r* analogously. The positive effect sizes were the estimates of the effect above the null value (i.e., .50), whereas the negative effect sizes provided the estimate of correct identification below the null value. As such, the negative effects could reflect the consistent presence of bias in the participant ratings, methodology, or the analyses. Studies were excluded if information on the accuracy of categorization was omitted (i.e., Quanty et al., 1975). Furthermore, studies were excluded if the accuracy of categorization did not involve actual group membership but, instead, tested for consistency in perception or consensus between the two groups (i.e., Babel, 2007). Studies were also excluded if the targets were extensively manipulated (e.g., presented only a part of the face; Rule et al., 2008, Studies 3 and 4), which would diminish the ecological validity of the findings. If one study provided several effect sizes from non-independent samples, the overall effect size for the study was derived using the standard meta-analytic technique of averaging the Fisher's *Z* transform of the correlation coefficient *r* (e.g., Rosenthal, 1991). Furthermore, if the information regarding the average accuracy of the ingroup and outgroup subsample was present, the effect size was calculated for each independent subsample. Both authors read and coded the published and unpublished articles that met the inclusion criteria. The authors resolved any ambiguity in coding and showed total agreement.

Inclusion criteria did not specify restrictions based on time period or geographic location, which allowed for a broader and more inclusive search. No population restrictions were placed because no study had reported a sample that was drawn from a population that is theoretically different from normal. Forty-seven articles met the criteria for inclusion.

Coding procedure. The following variables were coded from the articles included in the meta-analysis: (a) the authors, (b) the year of publication, (c) the number of citations recorded from the Web of Knowledge or PsycInfo computerized databases for each article, and (d) publication

status (i.e., published, unpublished). For each individual study, we coded the following: (a) the effect size *r*, estimated as described above; (b) the target group at study (e.g., Mormons); (c) the number of judges in each independent subsample if ingroup and outgroup comparisons were made; (d) the number of targets; (e) the modality of presentation (e.g., still photograph, audio clip); (f) whether the time of exposure was constrained or unconstrained (i.e., the stimulus was presented for a fixed amount of time versus the time of exposure was not limited); (g) the source of the target stimuli (i.e., Internet, print, or recruited to a laboratory); and (h) the type of analysis (e.g., signal detection).

Results

The resulting sample of studies consisted of 47 articles providing 131 independent effect sizes from 6,448 judges. The mean effect sizes, standard deviations, number of samples, and 95% confidence intervals for each categorical moderator examined, as well as the overall effect, can be found in Table 1. In addition, refer to Table 2 for the correlations of the effect sizes and the variables of interest.

Table 1. The Combined Descriptive Statistics for the Overall Effect and Categorical Moderators

			95% CI	
Variable	M_{Zr} (SD_{Zr})	*k*	LL	UL
Overall				
Random effect	.29 (.28)	131	0.24	0.33
Target group				
Mormons	.16 (.07)	6	0.08	0.23
Political orientation	.18 (.14)	8	0.07	0.29
Sexual orientation	.29 (.26)	96	0.23	0.34
Jewish group membership	.36 (.42)	21	0.17	0.55
Lab				
Rule	.23 (.16)	56	0.19	0.27
Other researchers	.33 (.34)	75	0.25	0.40
Modality				
Still photograph	.26 (.26)	92	0.21	0.32
Video	.29 (.18)	24	0.22	0.37
Audio	.36 (.52)	12	0.03	0.69
Live presentation	.65 (.25)	3	0.03	1.28
Analysis type				
SDT analysis	.19 (.10)	47	0.16	0.22
Other type of analysis	.34 (.33)	84	0.27	0.41
Target stimuli source				
Internet	.23 (.19)	67	0.18	0.28
Print	.31 (.40)	20	0.12	0.49
Recruited	.36 (.32)	44	0.27	0.46
Judge's group membership				
Ingroup	.28 (23)	20	0.17	0.38
Outgroup	.30 (.37)	42	0.19	0.42
Exposure time				
Constrained	.32 (.35)	60	0.23	0.41
Unconstrained	.26 (.20)	71	0.21	0.31
Publication status				
Published	.28 (.29)	128	0.23	0.33
Unpublished	.39 (.10)	3	0.29	0.48

Note: CI = confidence interval; LL = lower limit; UL = upper limit. SDT = signal detection theory. Variation in the number of studies emerged because not all studies provided relevant information for a given moderation analysis.

General Effects

Most (92%) effect sizes in the sample were positive, meaning that judges were typically more accurate than chance in categorizing targets. A random-effects model was used to compute the combined effect size and to conduct the moderation analyses. Random-effects meta-analysis is more conservative and allows one to conclude that significant findings are generalizable to other similar studies that were or will be conducted outside of those included in this investigation. The analysis revealed that the aggregate effect was positive, moderate-to-small, and statistically significant: $r = .29$, 95% CI = [.24, .33]. The combined (aggregated) effect indicates that judges were more accurate than would be expected by chance in their categorization of members of ambiguous groups. Because the 95% confidence interval around the mean value is narrow and does not include zero, it can be concluded that the estimate of the analysis is rather reliable.

Publication Bias

A file-drawer analysis was conducted to further assess the reliability of the findings. By using the formula supplied by Rosenthal (1979), we found that more than 20,000 studies averaging a null effect size would need to be added to the sample to bring the overall results to barely significant ($p = .05$). The tolerance level (670), or the number of studies that could already exist and average a null effect, was assessed using the ($5k + 10$) formula, where *k* is the number of effects contributing to the current sample. Because the fail-safe *N* (i.e., the number of studies required to be added to the sample to yield nonsignificant results in the overall analysis) is greater than the tolerance level, it is safe to speculate that there are not enough unpublished studies on the accurate categorization of ambiguous groups that average a null effect to render the obtained results nonsignificant. Moreover, we further investigated potential publication bias by constructing a funnel plot (Figure 1), where the *x*-axis represents the effect size estimate and the *y*-axis is an indicator of the sample size (standard error of Fisher's Z_r; Sterne & Egger, 2001). Linear regression was used to assess the graph's asymmetry (i.e., presence of publication bias). The results revealed that the graph was symmetrical, $bias = .23$, $SE_{bias} = .26$, $t(129) = 0.90$, $p = .37$; thus, it is concluded that there was little to no bias in publication of the studies included in this

Table 2. Zero-Order Correlations Between the Moderators

	1	2	3	4	5	6	7	8	9	10
1. Effect size	—	.20*	−.20*	−.08	−.02	−.27**	.08	.19*	−.06	−.16
2. Number of citations	.16	—	−.74***	−.36***	−.13	−.39***	.14	.32***	−.26**	−.26**
3. Year of publication	−.17*	−.37***	—	.55***	.03	.66***	−.36***	−.62***	.46***	.62***
4. Number of targets	.19*	−.10	−.20*	—	.22**	.54***	−.54***	−.73***	.41***	.57***
5. Number of judges	−.05	.13	−.43***	−.05	—	−.10	−.48***	−.30***	.08	−.06
6. Signal detection analysis	−.25**	−.32***	.40***	.04	−.23**	—	−.48***	−.68***	.46***	.87***
7. Modality	.19*	.07	−.15	.18*	−.21*	−.43***	—	.78***	−.51***	−.55***
8. Source	.22**	.31***	−.31***	−.05	−.08	−.66***	.73***	—	−.51***	−.79***
9. Exposure time	−.10	−.25**	.16	.16	−.09	.46***	−.40***	−.50***	—	.58***
10. Lab	−.17*	−.26**	.45***	.04	−.27**	.87***	−.50***	−.77***	.58***	—

Note: Spearman's correlation coefficients are presented above the diagonal. Pearson's correlation coefficients are presented below the diagonal. Signal detection analysis: 1 = *used SDT*, 0 = *did not use SDT*. Exposure time: 1 = *unconstrained*, 0 = *constrained*. Lab: 1 = *Rule Lab*, 0 = *Other Labs*. Source: −1 = *online*, 0 = *print*, 1 = *recruited*.
*$p < .05$. **$p < .01$. ***$p < .001$.

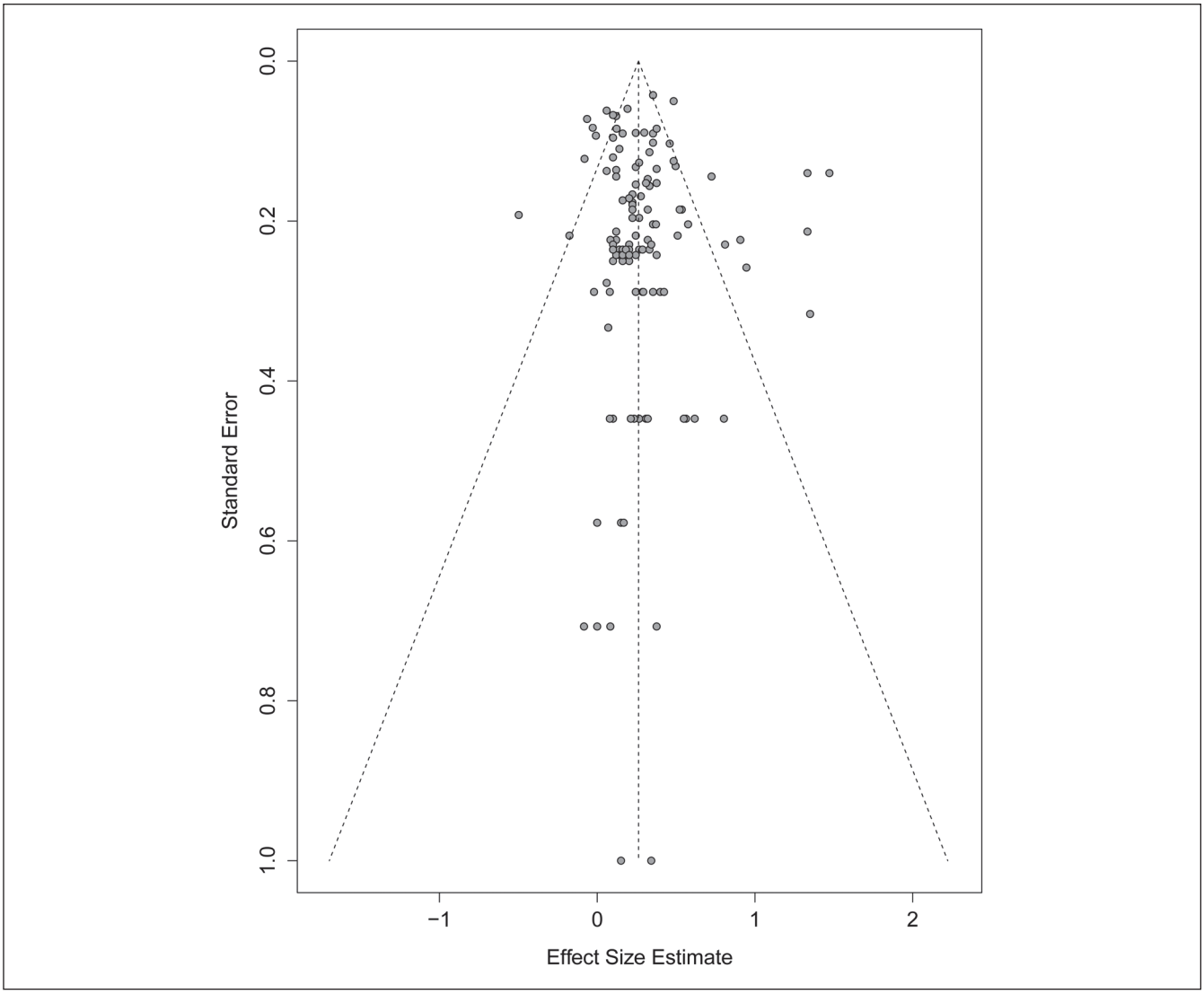

Figure 1. The funnel plot for the studies included in the current meta-analysis
Note: The effect size estimates (x-axis) are expressed as Fisher transformed correlation coefficients *Zr*. Standard errors of the estimates (*y*-axis) represent the sample sizes: they are the inverse of the degrees of freedom of *Zr*, which are estimated from the sample size used in any individual study.

analysis. That is, theoretically, the symmetry of the plot suggests that smaller studies are as likely to be published as those that have obtained larger sample sizes.

In addition, we examined whether published studies were significantly different from unpublished studies. An independent-samples *t*-test comparing the two study types indicated that there were no significant differences in the size of effects between published and unpublished studies in the sample, $t(129) = 0.96$, $p = .34$, $r = .08$. All the tests performed to assess publication bias therefore supported the reliability of the overall effect.

Moderator Analyses

The heterogeneity of the effect sizes was assessed using an omnibus chi-square test. The sample was significantly heterogeneous, $\chi^2(131) = 372.98$, $p < .0001$, $SD_{Zr} = .28$. Such heterogeneity in the sample of effect sizes suggests that there are significant differences between the studies in terms of their effects. Thus, further analyses of moderation were performed to identify potential sources of variation. Table 1 presents descriptive statistics for the categorical moderators that follow.

Target group. We first analyzed differences between the various target groups (e.g., sexual orientation, political orientation, and religious affiliation). A one-way ANOVA showed no significant difference between them: $F(3, 127) = 1.37$, $p = .26$, $\eta^2_{partial} = .03$. Thus, on average, judges had the same level of accuracy for identifying members belonging to any single group in this study or other similar studies.

Laboratory. Given the high representation of effects coming from work done by Rule and colleagues, we next tested for a lab effect. The effect sizes from Rule and colleagues' studies were marginally smaller than those from other laboratories, $t(129) = 1.96$, $p = .052$, $r = .17$. Given the large number of studies from Rule and colleagues in the current sample, this suggests that the overall effect size for the accuracy of categorization in this work may be a conservative estimate.

Modality. The modality of presentation (audio, video, photograph, or live perception) was examined as another potential moderator. The omnibus one-way ANOVA test revealed a marginally significant difference according to presentation modality, $F(3, 127) = 2.25$, $p = .086$, $\eta^2_{partial} = .05$. Even though the omnibus test did not reach significance, we tested for the presence of a linear trend in the distribution of effect sizes. Contrast weights –3, –1, + 1, + 3 were assigned to still photographs, video clips,[2] audio clips, and live presentations, respectively. It was expected that still photographs would provide the least amount of information, followed by videos and live presentations. The audio clips were positioned between the videos and live presentations because speech has been suggested to be a particularly revealing or "leaky" channel (i.e., providing greater accuracy in lie detection; Zuckerman, DePaulo, & Rosenthal, 1981). We expected that accuracy would increase from video clips to audio clips, as the former were mostly silent in the current sample (75%). Furthermore, we expected that the studies using live presentations of targets would yield the greatest effect sizes because the judges saw the targets directly and in a highly naturalistic, interactive environment.

The linear model was significant, $t(128) = 2.49$, $p = .01$, $r = .21$. Post hoc analyses indicated that the trend was mainly driven by a marginally significant difference between still photographs and live presentations, $p = .08$. However, this trend does not seem to be reliable, overall, because there were only three studies that used live targets in the current sample. Furthermore, when these 3 studies were removed from the sample, the linear trend became nonsignificant, $F(1, 126) = 1.22$, $p = .27$, $r = .10$.

Type of analysis used. We also examined whether researchers who controlled for guessing provided effects that were more conservative than those who used other analytic techniques. Thus, we compared the effect sizes that were extracted from analyses using signal detection theory with the effect sizes from the studies that did not. The difference was significant, $t(129) = 2.96$, $p = .004$, $r = .25$, such that researchers who used signal detection theory provided smaller effects. Because signal detection theory analysis accounts for guessing, the effects that came from the studies that did not use this kind of analysis should be interpreted accordingly.

Ingroup membership. In addition, we tested for differences in sensitivity between judges who were ingroup members ($k = 20$) versus those who were outgroup members ($k = 42$). The test revealed no significant difference, $t(60) = .31$, $p = .76$, $r = .04$. Overall, then, ingroup and outgroup members appeared to be equally accurate in judging targets' group membership.

Target stimuli source. The selection and composition of stimuli is an important consideration for yielding unbiased effects in social perception experiments. We therefore wondered if there was a significant difference between the effect sizes obtained from the studies that used target stimuli collected from online sources (e.g., personal advertisements, facebook.com), printed media (e.g., yearbooks, newspapers), and the local community (i.e., recruited to a laboratory). We expected the effects to follow a linear trend, such that the accuracy would change as a function of the target source. Therefore, we proposed that targets who had the most control over self-presentation (i.e., online) would be identified the least accurately, whereas those who had little control over self-presentation (i.e., recruited in the lab setting) would be identified with the greatest accuracy. We expected recruited targets to be the most identifiable because they should have the least amount of control over their presentation, which would have been governed by the experimenters. The linear model achieved a significant fit, $F(1, 129) = 6.38$, $p = .013$, $r = .22$. Although this may support our hypothesis about control of self-presentation, further information about targets' relative impression management efforts across these conditions would be needed to fully test this possibility.

Number of targets. The average number of targets used in the studies was 100.04 (SD = 251.12). Because Fisher transformed Z_r coefficients and the number of targets did not achieve normality after linear transformations, the relationship between the number of targets and accuracy was examined by using the nonparametric Spearman correlation coefficient ρ. The estimated correlation was negative and not significant, Spearman's $\rho(129) = -.08$, $p = .37$. Thus, the number of targets did not covary much with the magnitude of the effect sizes. We repeated this analysis on the studies that presented fewer than 100 targets to judges to ensure that the effect would not be present even when the smaller samples were examined; the correlation was not significant, Spearman's $\rho(129) = .01$, $p = .93$.

Temporal constraint. It was of interest to explore whether accuracy changed as a function of stimulus viewing time. It could be expected that accuracy may decrease if the time of exposure is constrained because (a) the judges did not have enough time to process the features (33 ms; Rule & Ambady, 2008) or (b) the judges felt the time pressure. Many studies constrained the exposure to the stimuli by presenting targets for short durations, whereas other designs allowed for unrestricted exposure. There were no differences between the effect sizes from the participants who were exposed to a stimulus for a fixed amount of time versus those who were not, $t(129) = 1.18$, $p = .24$, $r = .10$.

Number of citations and year of publication. We hypothesized that the size of an effect would be positively related to its impact in the literature, as measured by the number of times the work was cited. To test the hypothesis, we fit a linear regression model in which the number of citations predicted the effect sizes while controlling for the publication year. There was no relationship between the two variables after controlling for the year of publication, $\beta = .11$, $t(128) = 1.18$, $p = .24$. Even though there was no relationship between the number of citations and effect sizes, the era when the study was conducted was significantly and negatively correlated with the magnitude of the effect sizes such that older studies tended to provide greater effects, $r(129) = -.17$, $p = .048$. This could be because of the evolution of analytic methods throughout these studies over the years, principally the increased application of signal detection theory to analyzing categorization accuracy. Indeed, after controlling for the application of signal detection analysis, the year of publication was no longer a significant predictor of the magnitude of the effect size, $\beta = -.09$, $t(128) = -0.93$, $p = .34$. Considering these results, the data again remind us that researchers should account for guessing by using signal detection theory analysis to ensure that the results they obtain are not inflated.

Discussion

Aggregating across tens of studies and hundreds of effects, people were found to be significantly more accurate than chance in categorizing members of perceptually ambiguous groups. The total combined effect suggests that approximately 64.5% of targets would be correctly categorized. Thus, although accuracy for these groups is reliably and significantly better than chance, it is still quite ambiguous as compared with other social groups whose markers of group membership are more obvious (e.g., 99.2% for race; Remedios et al., 2011).

These data are also consistent with those of previous meta-analyses examining the categorization of individuals as Jewish and non-Jewish. The effects obtained in the present work for categorizing Jewish and non-Jewish targets were comparable with the two previous meta-analyses examining those groups (Andrzejewski et al., 2009; Rice & Mullen, 2003). This suggests convergence between the analytic methods used here and those used in the previous studies. Unlike those analyses, though, here we were able to generalize beyond one particular social distinction to three other groups. The present effects are therefore helpful for characterizing the overall capacity for perceivers' judgments of social group membership to give a general sense of our ability to infer information about others' social identities from various cues.

A central goal of this study was to identify possible moderators of accuracy in perceiving social group membership. We found that the type of distinction or target group studied had little effect on the accuracy of categorization. A large share of the work examining the accuracy of judgments of perceptually ambiguous groups has come from research done by Rule and colleagues. Although a random-effects analysis showed no significant difference at $\alpha = .05$ between the size of effects obtained by Rule and colleagues versus those obtained by researchers in other labs, the effect was marginally significant ($p = .052$). The influence of these effects on the overall sample in this work and in other potential studies may therefore serve to provide a more conservative estimate of the magnitude of the overall effect. Furthermore, it is important to note that even though a large portion of the effects came from the same lab, other labs contributed the majority (57%) of effect sizes to the overall sample.

More interesting, it was found that the modality of presentation significantly influenced categorization accuracy. Live presentations seem to provide more information than do still photographs, video, and audio tracks when categorizing targets. This is likely due to the presence of dynamic (e.g., voice, gesticulation) information that might enhance accuracy through the provision of multiple, often simultaneous, perceptual cues. Given that still photographs, audio clips, and silent video clips did not differ when directly compared, it is possible that targets reveal their group membership more readily through live presentation in a naturalistic setting. This finding should be interpreted cautiously, however, as the studies that used live targets comprised only 1.5% of the overall sample of studies.

Furthermore, we observed that the analytic approach used by various researchers also had a significant impact on the sizes of the effects obtained. Studies applying signal detection theory to data analysis provided a more conservative, and possibly more accurate, estimate because of the consideration of random guessing and systematic response biases. As such, we would encourage researchers to use signal detection analyses in their studies when appropriate because signal detection analyses seem to provide not only a more conservative but also a more precise estimate of the true effect size.

In addition to examining analytic techniques that allow for the correction of bias, this work also considered factors of the experiments that might have influenced the studies' results. Although ingroup members are typically considered more expert in perceiving other members of their group (Sporer, 2001), the results of our analysis showed that outgroup judges were just as sensitive to detecting targets' identity as were ingroup members. Furthermore, sensitivity did not significantly vary according to whether the researchers presented targets for a fixed temporal interval or allowed judges unlimited time to make their decisions.

Although judge characteristics did not significantly affect accuracy, some target characteristics did seem to have an effect. The source from which targets were obtained, for example, followed a linear trend. Here, we observed that target stimuli obtained from modern media, which allowed for more control over self-presentation (i.e., the Internet), tended to provide effect sizes that were smaller than those obtained from the use of targets that were culled from printed material (e.g., yearbooks). Moreover, the people who were recruited for stimulus creation in a laboratory, who presumably had less control over the impression they conveyed, could have "leaked" more information while being videotaped or photographed, explaining why accuracy was highest for targets falling into this group. Because we had no ability to control for self-presentation, future empirical work could attempt to explore and address these questions. The other moderator that was examined in regard to targets was the number of stimuli used in any given empirical investigation that was included in the current sample. The number of targets did not have a significant effect on the accuracy of identification. Researchers, however, should still be concerned about the number of targets that they use in their studies: Although accuracy did not change as a function of the number of targets, smaller samples of targets could yield insufficient power to detect significant effects when the photographs, video, and audio clips are treated as units of analysis (Ambady et al., 1999).

Finally, we examined whether the publication year and relative impact (i.e., number of citations) for each individual article was related to the magnitude of the effect size. This analysis showed no relationship between the effect sizes in a given article and the number of times that the article was cited when controlling for the year of publication. The effect sizes did become smaller, however, as time progressed. This decrease in effect size might be due to more rigorous methodological procedures (e.g., standardization of stimuli, use of more conservative analytic techniques) and the increased control over extraneous variables that occurred as the field of experimental psychology matured throughout the latter half of the 20th century and on until the present.

Naturally, this meta-analysis has certain limitations. Although it has been identified in previous research that the categorization of ambiguous groups is an automatic and rather spontaneous process (e.g., Rule et al., 2009), this conclusion cannot be inferred from this study because there was no opportunity to examine automaticity or the spontaneity of categorization as a moderator variable. Rather, the effect sizes studied here typically consisted of judgments made on a conscious level. Future empirical research should be inclined to explore the implicit processes that may underlie categorizations to provide a more complete sense of the ecological validity or real-world presence of these categorizations. Furthermore, this meta-analysis provided only a limited set of moderators that could be identified from the research articles. Future studies should explore other possible moderators of the effects that may not be available in the current corpus of studies. Researchers, for example, may wish to explore whether ambiguous group members intentionally or subconsciously "leak" information to help affiliate with other ingroup members. This should be especially relevant to groups that tend to be stigmatized in a given society and may be in danger if their true identity is revealed (e.g., Jews in Nazi Germany). This information "leakage" could facilitate categorization accuracy to levels that exceed chance guessing to increase opportunities for affiliation with other group members, while still permitting the targets to remain ambiguous enough to avoid negative outcomes. This area of research is particularly interesting because there may be cognitive manifestations of identity (e.g., body, sway; Johnson et al., 2007) that "leak" the information about a person to perceivers even without a target's awareness. Furthermore, researchers should attempt to explore the possibility of interactions between ambiguous and obvious identities, as it is possible that targets of different races, sexes, and ages are perceived with different levels of accuracy. Unfortunately, the present analysis does not possess enough information to answer these questions and, although a few studies have begun to explore this possibility (e.g., Johnson & Ghavami, 2011; Remedios et al., 2011; Rule, 2011; Rule, Ishii, Ambady, Rosen, & Hallett, 2011), more research is needed to address these questions. In addition, it is important to mention that many null effects found in this meta-analytic review could be considered to rule out the many different factors that might influence the ability to accurately categorize targets into groups, which in turn will inform future research in this area.

The present study provides important implications for future research. The findings suggest that the average person can distinguish ambiguous group members better than chance guessing. This is not only relevant for the understanding of perceptually ambiguous groups but for the understanding of perceptually obvious groups as well. Specifically, the ability of the human mind and perceptual system to make correct categorizations of individuals' social group memberships from such minimal information suggests a robust cognitive-perceptual capacity for extracting information about others from their appearance and other subtle nonverbal cues. It may be difficult to highlight such processes with obvious groups because the judgments are made rapidly, automatically, and with high consensus (Macrae & Bodenhausen, 2000). Redundancies from multiple, obvious features (e.g., the overlap between skin tone and Afrocentric features in the perception of race; Maddox, 2004) might preclude an in-depth investigation of the component processes involved in perceiving and categorizing others. The cognitive and perceptual mechanisms underlying person perception may therefore be best pursued by studying groups that are less identifiable, which may allow for deeper testing through experimental manipulations.

The practical implications of identifying ambiguous group members are important as well. For example, the current results provide additional scientific support for the recent repeal of the U.S. military's "don't ask, don't tell" policy targeting gay individuals (Policy Concerning Homosexuality in the Armed Forces, 1993). Because the data show that the accuracy for indentifying the sexual orientation of random targets is greater than chance, the policy would seem futile as it rests on the assumption that explicit disclosure is needed to accurately infer a person's sexual orientation. Furthermore, the results could provide important considerations for other legislative and business practices in which discrimination based on creed, sexual orientation, ideology, or religious preference may be relevant (Employment Non-Discrimination Act of 2007, 2007). Indeed, knowing that subtly-distinguished and presumably invisible social group memberships can be actively ascertained from minimal cues holds implications for numerous domains and is of great potential value to policy makers, researchers, and the lay public.

Conclusion

People constantly make rapid, automatic, and accurate inferences about members of obvious groups (Macrae & Bodenhausen, 2000). The present study found that the accuracy of identifying members of ambiguous groups was significantly above-chance levels. Indeed, while modest in absolute size, the mean effect found in this analysis for the categorization of perceptually ambiguous group members (r = .29) was larger than recent estimates of the average study in social and personality psychology (r = .23; Vazire, 2010). Moreover, the study provides an overall effect that takes into account both the studies that reported positive and negative effects for the perception and categorization of ambiguous individuals, and evidence from moderation analyses suggests that this overall effect may be an underestimate of the true effect. The present research also demonstrated that the stimulus modality, analytic approach, publication year, and source of stimuli exert significant influences on the accuracy of perceivers' categorizations. The study of perceptually ambiguous groups is therefore valuable for increasing common social understanding of the legibility of social distinctions often assumed to be invisible, as well as for providing insights into the processes involved in person perception that are often inaccessible when studying perceptually obvious groups.

Acknowledgments

The authors would like to express their gratitude to Judy Hall and Nalini Ambady for their instruction, insight, and support.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded in part by the Social Sciences and Humanities Research Council of Canada.

Notes

1. Although these studies suggest that people are accurate in making these judgments, no information has been provided about the physical cues that may facilitate this accuracy.
2. Of 24 video clips, 5 were accompanied by sound. The clips containing auditory information and those that were silent did not yield statistically different effect sizes, $t(22) < 1$, $p = .97$, $r = .00$.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Adams, W. L. (2004). *Don't ask, don't tell: Nonverbal cues to sexual orientation* (Undergraduate honors thesis). Harvard College, Cambridge, MA.

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

*Allport, G. W., & Kramer, B. M. (1946). Some roots of prejudice. *Journal of Psychology*, *22*, 9-39.

*Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology*, *77*, 538-547.

*Andrzejewski, S. A., Hall, J. A., & Salib, E. R. (2009). Anti-Semitism and identification of Jewish group membership from photographs. *Journal of Nonverbal Behavior*, *33*, 47-58.

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, *323*, p. 1183.

Babel, M. (2007). *Social judgments and their acoustic cues in read speech* (UC Berkeley Phonological Lab Annual Report, pp. 226-264). Berkeley, CA: UC Berkeley Phonological Lab.

*Benjamin, D. J., & Shapiro, J. M. (2009). Thin-slice forecasts of gubernatorial elections. *Review of Economics and Statistics*, *91*, 523-536.

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

Berry, D. S., & McArthur, L. Z. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology*, *48*, 312-323.

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

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

*Carter, L. F. (1948). The identification of ''racial'' membership. *Journal of Abnormal and Social Psychology*, *43*, 279-286.

Church of Jesus Christ of Latter-Day Saints. (2004). *True to the faith: A gospel reference*. Salt Lake City, UT: Author.

*Dorfman, D. D., Keeve, S., & Saslow, C. (1971). Ethnic identification: A signal detection analysis. *Journal of Personality and Social Psychology*, *18*, 373-379.

*Elliott, D. N., & Wittenberg, B. H. (1955). Accuracy of identification of Jewish and non-Jewish photographs. *Journal of Abnormal and Social Psychology*, *57*, 339-341.

Enstrom, J. E. (1989). Health practices and cancer mortality among active California Mormons. *Journal of the National Cancer Institute*, *81*, 1807-1814.

Enstrom, J. E., & Breslow, L. (2008). Lifestyle and reduced mortality among active California Mormons, 1980-2004. *Preventive Medicine*, *46*, 133-136.

Fiske, S. T., & Taylor, S. E. (1984). *Social cognition* (1st ed.). Reading, MA: Addison-Wesley.

*Freeman, J. B., Johnson, K. L., Ambady, N., & Rule, N. O. (2010). Sexual orientation perception involves gendered facial cues. *Personality and Social Psychology Bulletin*, *36*, 1318-1331.

*Gaudio, R. P. (1994). Sounding gay: Pitch properties in the speech of gay and straight men. *American Speech*, *69*, 30-57.

Hallahan, M. (1998). *Reanalysis of Berger, Hank, Rauzi, & Simkins, 1987*. Unpublished manuscript, Clemson University, Clemson, South Carolina.

Herek, G. M. (2004). Beyond "homophobia": Thinking about sexual stigma and prejudice in the twenty-first century. *Sexuality Research and Social Policy*, *1*, 6-24.

Herek, G. M., Cogan, J. C., & Gillis, J. R. (2002). Victim experiences in hate crimes based on sexual orientation. *Journal of Social Issues*, *58*, 319-339.

*Himmelfarb, S. (1966). Studies in the perception of ethnic group members: Accuracy, response bias, and anti-semitism. *Journal of Personality and Social Psychology*, *4*, 347-355.

H.R. 3685—110th Congress: Employment Non-Discrimination Act of 2007. (2007). *GovTrack.us* [database of federal legislation]. Retrieved from http://www.govtrack.us/congress/bill.xpd?bill=h110-3685

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: SAGE.

*Jahoda, G. (1954). Political attitudes and judgments of other people. *Journal of Abnormal and Social Psychology*, *49*, 330-334.

*Johnson, K. L., & Ghavami, N. (2011). At the crossroads of conspicuous and concealable: What race categories communicate about sexual orientation. *PLoS ONE*, *6*, e18025.

*Johnson, K. L., Gill, S., Reichman, V., & Tassinary, L. G. (2007). Swagger, sway, and sexuality: Judging sexual orientation from body motion and morphology. *Journal of Personality and Social Psychology*, *93*, 321-334.

*Kendig, C. J., & Maresca, N. (2004). Guessing sexual orientation: Heterosexuals ability to accurately estimate their "gaydar." *The Id: Graduate Faculty, Psychology Social Bulletin*, *2*, 71-78.

*Lindzey, G., & Rogolsky, S. (1950). Prejudice and identification of minority group membership. *Journal of Abnormal and Social Psychology*, *45*, 37-53.

*Linville, S. E. (1998). Acoustic correlates of perceived versus actual sexual orientation in men's speech. *Folia Phoniatrica et Logopaedica*, *50*, 35-48.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE.

*Lund, F. H., & Berg, W. C. (1946). Identifiability of nationality characteristics. *Journal of Social Psychology*, *24*, 77-83.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*, 93-120.

Macrae, C. N., & Martin, D. (2007). A boy primed Sue: Feature-based processing and person construal. *European Journal of Social Psychology*, *37*, 793-805.

Macrae, C. N., & Quadflieg, S. (2010). Perceiving people. In S. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *The handbook of social psychology* (5th ed., pp. 428-463). New York, NY: McGraw-Hill.

Maddox, K. B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review*, *8*, 383-401.

Munson, B., & Babel, M. (2007). Loose lips and silver tongues, or, projecting sexual orientation through speech. *Language and Linguistics Compass*, *1*, 416-449.

*Olivola, C. Y., Sussman, A. B., Tsetsos, K., Kang, O. E., & Todorov, A. (2012). Republicans prefer republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological & Personality Science*. Advance online publication. doi:10.1177/1948550611432770

*Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*, 315-324.

*Piccolo, F. (2008a). Perceived sexual orientation and attitudes towards sounding gay or straight. *University of Pennsylvania Working Papers in Linguistics*, 14, 129-138.

*Piccolo, F. (2008b). *Listener-identified phonetic correlates of gay-, lesbian-, and straight sounding voices* (Doctoral dissertation). Retrieved from dissertation and theses database. (UMI no. 3311895).

Policy Concerning Homosexuality in the Armed Forces, 10 U.S.C. § 654 (1993).

Poutvaara, P., Jordahl, H., & Berggren, N. (2009). Faces of politicians: Babyfacedness predicts inferred competence but not electoral success. *Journal of Experimental Social Psychology*, *45*, 1132-1135.

*Pulos, L., & Spilka, B. (1961). Perceptual selectivity, memory, and anti-Semitism. *Journal of Abnormal and Social Psychology*, *62*, 690-692.

Quanty, M. B., Keats, J. A., & Harkins, S. G. (1975). Prejudice and criteria for identification of ethnic photographs. *Journal of Personality and Social Psychology*, *32*, 449-454.

Remedios, J. D., Chasteen, A. L., Rule, N. O., & Plaks, J. E. (2011). Impressions at the intersection of ambiguous and obvious social categories: Does gay + black = likable? *Journal of Experimental Social Psychology*, *47*, 1312-1315.

Rice, D. R., & Mullen, B. (2003). Isaac, Ishmael, and Janus: Past and future lessons regarding the ethnic categorization of faces. *Applied Cognitive Psychology*, *17*, 1129-1147.

Richeson, J. A., & Trawalter, S. (2005). On the categorization of admired and disliked exemplars of admired and disliked racial groups. *Journal of Personality and Social Psychology*, *89*, 517-530.

*Rieger, G., Linsenmeier, J. A. W., Gygax, L., Garcia, S., & Bailey, J. M. (2010). Dissecting "gaydar": Accuracy and the role of masculinity-femininity. *Archives of Sexual Behavior*, *39*, 124-140.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: SAGE.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature review. *Annual Review of Psychology*, *52*, 59-82.

Rosenthal, R., & Rosnow, R. L. (2007). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York, NY: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*, *106*, 332-337.

*Rule, N. O. (2011). The influence of target and perceived race in the categorisation of male sexual orientation. *Perception*, *40*, 830-839.

*Rule, N. O., & Ambady, N. (2008). Brief exposures: Male sexual orientation is accurately perceived at 50 ms. *Journal of Experimental Social Psychology*, *44*, 1100-1105.

*Rule, N. O., & Ambady, N. (2010). Democrats and Republicans can be differentiated from their faces. *PLoS ONE*, *5*, e8733.

*Rule, N. O., Ambady, N., Adams, R. B., Jr., & Macrae, C. N. (2007). Us and them: Memory advantages in perceptually ambiguous groups. *Psychonomic Bulletin & Review*, *14*, 687-692.

*Rule, N. O., Ambady, N., Adams, R. B., Jr., & Macrae, C. N. (2008). Accuracy and awareness in the perception and categorization of male sexual orientation. *Journal of Personality and Social Psychology*, *95*, 1019-1028.

Rule, N. O., Ambady, N., Adams, R. B., Jr., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology*, *98*, 1-15.

*Rule, N. O., Ambady, N., & Hallett, K. C. (2009). Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal of Experimental Social Psychology*, *45*, 1245-1251.

*Rule, N. O., & Andrzejewski, S. A. (2011). *Bias-colored glasses: The influence of homophobia on social perception*. Unpublished manuscript.

*Rule, N. O., Garrett, J. V., & Ambady, N. (2010a). Places and faces: Geographic environment influences the in group memory advantage. *Journal of Personality and Social Psychology, 98*, 343-355.

*Rule, N. O., Garrett, J. V., & Ambady, N. (2010b). On the perception of religious group membership from faces. *PLoS ONE, 5*, e14241.

*Rule, N. O., Ishii, K., Ambady, N., Rosen, K. S., & Hallett, K. C. (2011). Found in translation: Cross-cultural consensus in the accurate categorization of male sexual orientation. *Personality and Social Psychology Bulletin*, *37*, 1499-1507.

*Rule, N. O., Macrae, C. N., & Ambady, N. (2009). Ambiguous group membership is extracted automatically from faces. *Psychological Science*, *20*, 441-443.

*Rule, N. O., Rosen, K. S., Slepian, M. L., & Ambady, N. (2011). Mating interest improves women's accuracy in judging male sexual orientation. *Psychological Science*, *22*, 843-848.

*Samochowiec, J., Wänke, M., & Fiedler, K. (2010). Political ideology at face value. *Social Psychological & Personality Science*, *1*, 206-213.

*Savitz, L. D., & Tomasson, R. F. (1959). The identifiability of Jews. *American Journal of Sociology*, *64*, 468-475.

*Scodel, A., & Austrin, H. (1957). The perception of Jewish photographs by non-Jews and Jews. *Journal of Abnormal and Social Psychology*, *54*, 278-280.

*Shelp, S. (2002). Gaydar: Visual detection of sexual orientation among gay and straight men. *Journal of Homosexuality*, *44*, 1-14.

*Smyth, R., Jacobs, G., & Rogers, H. (2003). Males voices and perceived sexual orientation: An experimental and theoretical approach. *Language in Society*, *32*, 329-350.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy and Law*, *7*, 36-97.

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*, 1046-1055.

*Sylva, D., Rieger, G., Linsenmeier, J. A. W., & Bailey, J. M. (2010). Concealment of sexual orientation. *Archives of Sexual Behavior*, *39*, 141-152.

*Tabak, J. A. (2009). *The Gaydar effect: Automatic perception of sexual orientation from faces presented upright and upside-down* (Undergraduate honors thesis). Cornell University, Ithaca, NY.

*Tabak, J. A., & Zayas, V. (2011). *Deconstructing the gaydar effect: The roles of featural and configural face processing in snap judgments of sexual orientation*. Unpublished manuscript.

*Valentova, J., Rieger, G., Havlicek, J., Linsenmeier, J. A. W., & Bailey, J. M. (2011). Judgments of sexual orientation and masculinity-femininity based on thin slices of behavior: A cross-cultural comparison. *Archives of Sexual Behavior*. doi:10.1007/s10508-011-9818-1

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281-300.

Wright, D. B., & Stroud, J. N. (2002). Age differences in lineup identification accuracy: People are better with their own age. *Law and Human Behavior*, *26*, 641-654.

Zebrowitz, L. A. (1997). *Reading faces: Window to the soul*? Boulder, CO: Westview.

*Zimman, L. (2010). Female-to-male transsexuals and gay-sounding voices: A pilot study. *Colorado Research in Linguistics*, *22*, 1-21.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 2-60). San Diego, CA: Academic Press.