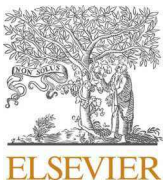
Journal of Experimental Social Psychology 108 (2023) 104497



Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

When the punk wishes you a great day, he still appears friendly: Stereotypes do not reliably guide spontaneous trait inferences from behavior

Jana Mangels *, Juliane Degner

Universität Hamburg, Germany

ARTICLE INFO

Editor: Shlomo Hareli

Keywords:
Spontaneous trait inferences
Stereotypes
Replication
Impression formation

ABSTRACT

One of the most robust effects in person perception research is the spontaneous trait inference (STI) effect, defined as the spontaneous tendency to draw dispositional inferences from actors' behaviors. Yet, research has suggested that stereotypes affect STIs by inhibiting stereotype incongruent or facilitating stereotype congruent STIs. These findings are remarkable considering (a) the robustness of STI effects and (b) the typical design of behavioral statements in this research as unambiguously indicative of traits. We present a series of four high-powered, preregistered experiments ($N = 1004$) that originally aimed at replicating stereotype effects on STIs as basis for investigating their underlying psychological mechanisms. We employed a probe recognition paradigm that has been used in prior research, pairing trait-implying behavioral statements with category labels implying either trait-congruent or -incongruent stereotypes. We additionally implemented several methodological improvements like a larger and extensively pretested stimulus set.

While we observed highly robust STI effects in all experiments, these were largely unaffected by actor stereotypes: Only one of the four experiments showed the hypothesized STI-stereotype interaction with a small effect size. We discuss how these findings add to the rather small number of existing publications on STIs and stereotypes and how the observed robustness of behavior-based impressions parallels prior research on intentional impression formation. We aim to instigate debate, further theorizing, and research that enhances our understanding of the boundary conditions of stereotype effects on spontaneous trait inferences in impression formation from unambiguous behaviors.

Imagine walking on the sidewalk minding your business, when suddenly a car slams on the brakes beside you, and you notice its young driver with multi-colored spiky hair and facial piercings. The stranger lowers his window, looks at you – then smiles blissfully, and says "Have a wonderful day!", before moving along. What is your resulting spontaneous impression of that person? Possibly, the stranger's appearance has activated a *social categorization* – like *punk*. The respective stereotype-based inferences may include characteristics such as rebellious, rude, or even dangerous. Inferences based on the stranger's *behavior*, however, may include characteristics such as friendly, jovial, and kind. How will such conflicting inferences affect your first impression of that person? Will one of them dominate what you think about this stranger?

When forming impressions of others, people regularly face a multitude of different, potentially contradicting, information. The various types of social inferences have received considerable attention in social cognitive research. For the current research, two long-standing lines of research are of interest. First, research on so-called *spontaneous trait inferences (STIs)* has demonstrated persistent and robust effects of people inferring traits from others' *behaviors* (Bott et al., 2022). Second, research on *social categorization and stereotyping* has documented similarly robust effects on perceivers impressions (Kunda & Spencer, 2003; Macrae & Bodenhausen, 2000; Quinn et al., 2007).

Although both behavior-based and stereotype-based impressions have long-standing traditions in social psychological research, both lines of research have developed largely independently from each other – theoretically and empirically (e.g., Chen et al., 2021). To our knowledge, only a handful of studies have investigated the interplay of spontaneous trait inferences (STIs) from behavior and categorical information about actors. For example, Wigboldus and colleagues (2003)

* Corresponding author at: Universität Hamburg, Department of Social Psychology, Von-Melle-Park 5, 20146 Hamburg, Germany.
E-mail address: jana.mangels@uni-hamburg.de (J. Mangels).

https://doi.org/10.1016/j.jesp.2023.104497
Received 29 July 2022; Received in revised form 10 April 2023; Accepted 28 May 2023
Available online 19 June 2023

manipulated the congruency between traits implied by an actor's behavior versus traits implied by their social category membership in a probe recognition paradigm (see below) and observed that stereotype incongruency reduced the occurrence of STIs from behavior. That is, for example, when learning that "the garbage man wins the science quiz", participants were less inclined to infer the trait "smart" as compared to when "the professor wins the science quiz" (Wigboldus et al., 2003; p. 473). Similar results were observed in a series of extended replication studies (Ramos et al., 2012), which further demonstrated that, when presented with counter-stereotypic behavior, participants were more likely to draw spontaneous inferences about an underlying *situation* rather than trait inferences (Ramos et al., 2012). Although these findings suggest a replicable effect of stereotype congruency on STIs, we argue that the available research entails some limitations and open questions: First, the small number of published research papers in this domain provides a rather weak basis for generalization ($k = 8$; as compared to $k = 97$ publications on STIs; see Bott et al., 2022). Second, as we will elaborate below, effects show a high level of inconsistency across the few published studies. Third, the available studies predominantly relied on very small stimulus sets further limiting their generalizability (Judd et al., 2012). Fourth, scholars have suggested multiple potential mechanisms driving the effects, but none of these have been systematically investigated. The initial aim of our current research was to investigate the underlying psychological mechanisms of stereotype congruency effects on trait inferences from behavior. We also aimed at increasing the generalizability of previous findings by employing a much larger stimulus set, including more behaviors, traits, and social categories, and by addressing some methodological and analytical limitations of previous studies.

To foreshadow results: While we observed large and robust STI effects (i.e., trait inferences from behavior) in all four experiments, our attempts to replicate a moderating effect of stereotypes on STIs yielded predominantly non-significant results and small effect sizes. In order to explain these unexpected findings, we shifted our attention towards potential systematic methodological variations between our studies and the published literature, which, however, yielded further inconclusive findings. Overall, we thus present results of four experiments questioning the ubiquity of stereotype effects on STIs from unambiguous behavior.

For the sake of transparency, we will first review theoretical assumptions and empirical evidence showing that both individual behaviors and stereotypes are powerful sources of person inferences affecting impression formation. We then present methods and results of our initial two experiments, which did not replicate the effect described in the literature. In Experiment 3 and 4, we then explore whether systematic methodological deviations from the published research might account for our unsuccessful replications.

1. Spontaneous inferences from behavior

When observing a stranger greeting someone, or solving a quiz, people tend to spontaneously infer that the person is friendly, or smart, even without intention or awareness of doing so. Numerous studies have investigated this effect (for reviews, see Moskowitz, 2005; Uleman et al., 2008; Uleman, Newman, et al., 1996; Uleman et al., 2012) and a recent systematic meta-analysis on STIs attested a moderate to large average effect size of $d_z = 0.59$ (Bott et al., 2022).

STI research mainly relies on indirect experimental paradigms: Typically, participants are presented with descriptions of others' behaviors and then complete word-based categorization, recall, or recognition tasks (but see Fiedler & Schenck, 2001; Kruse et al., 2023; for nonverbal behavior presentations). Depending on the paradigm, spontaneous trait inferences are then deduced from memory performances (e.g., in the false recognition paradigm; Todorov & Uleman, 2002) or response latencies (e.g., in the probe recognition paradigm; McKoon & Ratcliff, 1986). For example, in the so-called probe recognition paradigm, participants read a number of behavioral statements, such as "John gets an A for the test" (Ham & Vonk, 2003; p. 445). Immediately following each statement, various probe words are presented, for which participants indicate whether they had occurred in the previous statement or not. These probe words contain the trait implied by the behavioral statement (e.g., *smart*) or an unrelated control trait (e.g., *helpful*). Typically, responses are slower for rejecting the implied trait probes than the control trait probes (Ham & Vonk, 2003; Newman, 1991; Todd et al., 2011). This pattern of results has been interpreted as an indicator that the trait was inferred when encoding the trait-implying behavioral statement, that is, for spontaneous trait inferences.

A large number of studies has supported the assumption that these trait inferences are indeed *spontaneous* (Uleman et al., 2012; Uleman, Hon, et al., 1996) and can occur independent of intentional control (e.g., Krull & Erickson, 1995; Todorov & Uleman, 2002). These findings are in line with classic theorizing on the impression formation process (e.g., Gilbert et al., 1988; Trope, 1986) stating that initial categorization of behavior by implied traits occurs automatically and persistently (McCarthy & Skowronski, 2011; Okten & Moskowitz, 2020).

2. Spontaneous inferences from stereotypes

Stereotypes are defined as mental representations – knowledge structures about members of social groups that also contain associations of group memberships with typical traits (e.g., Devine, 1989; Dijksterhuis & Van Knippenberg, 1996). Classic models of person construal (Brewer, 1988; Fiske & Neuberg, 1990) propose that category-based stereotyping is the automatic first stage of impression formation. It has further been assumed that stereotypes are often prioritized over individuating attributes, and thus encourage category-consistent impression formation (Fiske & Neuberg, 1990; but see Monroe et al., 2018). Just like for STIs, research in the domain of stereotypes has relied on indirect measurement paradigms, which has demonstrated that stereotypes can be activated and applied to individuals in an automatic fashion according to different indicators for automaticity (for an overview, see Roth et al., 2019). For instance, perceivers activate stereotypes following category primes without their subjective awareness (Moskowitz et al., 2012) and with high processing efficiency (Payne, 2001; but see Spencer et al., 1998). Similarly, stereotypes can be applied to judgments instantly (Correll et al., 2002), again without awareness (Devine, 1989; Graham & Lowery, 2004).

3. The interplay of spontaneous trait inferences and stereotypes

As Chen et al. (2021) have elaborated in detail, research on STIs and on stereotypes have developed largely separately from each other, with only a few exceptions (see below). However, theoretical assumptions and experimental approaches of both research domains bear various resemblances. First, both view perceivers as active interpreters of information, whose impressions go beyond the given information (Chen et al., 2021). Second, both contain evidence from indirect experimental approaches, often relying on response latencies or memory performance (see Roth et al., 2019; Uleman et al., 2012). Third, perceivers' initial impression formation based on behaviors and on stereotypes are assumed to possess characteristics of automatic processes (Roth et al., 2019; Uleman et al., 2012).

Previous research has provided many different perspectives on the interplay between individual (behavior-based) and categorical (stereotype-based) impressions. Indeed, some of the first and most seminal studies in the domain of stereotype research revealed pervasive effects of stereotyping on behavioral interpretations, demonstrating, for example, that an ambiguous shove was interpreted as more violent behavior (and more likely attributed to dispositional causes) when performed by a Black as compared to a White actor (Duncan, 1976; see also, Dijksterhuis & Van Knippenberg, 1996; Dunning & Sherman, 1997; Otten & Stapel, 2007; Quadflieg & Macrae, 2011; Sagar & Schofield, 1980). However,

stereotypes seem less influential for impression formation when perceivers face more relevant or more diagnostic individual behavioral information, for instance, when observed behavior is unambiguous, extreme, or has clear trait implications (e.g., Beckett & Park, 1995; Bodenhausen et al., 1999; Kunda & Thagard, 1996).

Given that, in the domain of STI research, trait-implying behavior is typically carefully pretested and selected to be highly diagnostic and to have clear and unambiguous trait implications, one may argue that stereotypes may similarly be less influential for spontaneous impressions from unambiguous behavior. Remarkably, however, several studies report significant stereotype effects on STIs: Wigboldus et al. (2003), employed a probe recognition paradigm and provided participants with unambiguously trait- implying behavioral statements – like "wins the science quiz" implying the trait *smart* – which were either paired with a neutral actor label ("the human"), or a social category label that was stereotype congruent ("the professor") or incongruent ("the garbage man") with the implied trait. Throughout five studies, STI effects were observed to be significantly smaller for the stereotype incongruent actor-trait pairings as compared to the stereotype congruent pairings. Similar results have been documented with regard to stereotypes related to race (Stewart et al., 2003; Wigboldus et al., 2004), age (Wang & Yang, 2017), and gender (Wang et al., 2015; Yan et al., 2012). Ramos et al. (2012) extended these findings, demonstrating that reduced trait inferences for stereotype incongruent trials were accompanied by increased spontaneous situation inferences. These findings are remarkable considering that (a) behavioral statements in STI paradigms are typically pretested to be univocal and unambiguous, and (b) STIs have been shown to be highly robust (e.g., Bott et al., 2022). Consequently, the occurrence of stereotypes effects on Spontaneous Trait Inferences (as measured by classic STI paradigms) may indicate a high level of pervasiveness of stereotype effects in the impression formation process.

However, what appears to be a consistent research finding, reveals some open questions and inconsistencies at closer inspection: First, scholars have proposed different assumptions about mechanisms driving stereotype effects on STIs: While Wigboldus et al. (2003) speculated that stereotypes serve as *inhibition* tool, blocking the effect of incongruent STIs on person construal, Ramos et al. (2012) speculated that stereotypes serve as *facilitation* tool guiding inferential activity and encoding of congruent trait inferences. These competing assumptions were proposed as explanations for inconsistent empirical results regarding the exact experimental conditions driving the observed effects (e.g., Ramos et al., 2012; Wigboldus et al., 2003) but up to now, we lack systematic research investigating these underlying mechanisms. Second, stereotype effects on STIs seem to shift unpredictably between dependent variables: While some authors reported significant effects of stereotype congruency only for response latencies (Wang et al., 2015; Wang & Yang, 2017; Wigboldus et al., 2003, 2004; Yan et al., 2012), others reported stereotype effects only for error rates (Stewart et al., 2003), or even found effects shifting between response latencies and error rates (Ramos et al., 2012; with stereotype congruency effects on *trait* probes occurring in the error rates, and effects on *situational* probes occurring in the response latencies). While speed-accuracy-tradeoffs are not unusual in response-time based experimental paradigms (e.g., Ratcliff, 1993), we need to take this variance into account when evaluating the robustness of the published results, because it implies that the results of many studies actually contain several non-significant effects of stereotypes on STIs as well (see Table S.1 in the supplemental materials).

Third, while some published research reported significant stereotype effects on STIs for their entire samples (e.g., Ramos et al., 2012 and Wigboldus et al., 2003), others observed such effects only for specific sub-samples (e.g., participants under high cognitive load; Stewart et al., 2003; Wigboldus et al., 2004; participants after a negative mood manipulation; Wang et al., 2015; participants in a high power condition; Wang & Yang, 2017). Thus, research indicates only a partial replication of the original results that were obtained without capacity-limiting additional tasks (Wigboldus et al., 2003).

A fourth concern with regard to the generalizability of stereotype effects on STIs is related to the small number of stimuli typically employed, ranging from six (e.g., Wigboldus et al., 2003) to fourteen (Stewart et al., 2003) trait-implying behavioral statements in most studies (with the recent exception of 24 stimuli in Yang et al., 2022). Employing such small stimulus samples not only ignores potential variations between stimuli, limiting generalizability (Judd et al., 2012), but also reduces statistical power (Judd et al., 2017). A final concern addresses potential effects of analytical decisions with regard to the handling of response latency data, which usually requires corrections for outliers (i.e., singular trials with extremely slow responses due to inattention or distraction, which can largely bias aggregate response times; Ratcliff, 1993). There exist no conventions with regard to trimming of extreme values and/or transformations for the probe recognition paradigm, but researchers' decisions may affect the reliability and replicability of observed results. As trimming and transformation criteria vary across the published literature, it remains an open question if and to what extent these variations have affected the robustness of findings.

In summary, albeit published research has indicated that stereotypes may influence spontaneous trait inferences from behavior, we argue that the available empirical evidence is too scarce, with too few replications, too many methodological limitations, and with too many inconsistencies to warrant generalization to a general influence of stereotypes on trait inferences from non-ambiguous behavior.

4. The current research

We conducted four pre-registered experiments closely following the design of previous research investigating stereotype effects on STIs (e.g., Ramos et al., 2012; Wigboldus et al., 2003). All four studies employed the probe recognition paradigm (as adapted by Todd et al., 2011), in which we paired trait-implying behavioral statements with social category labels implying stereotypes about actors. For example, the behavioral statement "… picked up a stack of boxes like it was nothing" implying the trait "strong", was associated to the different actors: "the bodybuilder" (stereotype congruent), "the old man" (stereotype incongruent), or "Leslie" (stereotype neutral). Extending the published literature, we employed a considerably larger and thus more generalizable set of 33 carefully pretested behavioral statements paired with 66 stereotype (in)congruent actor labels. Given that the occurrence of STIs is typically inferred from slower response latencies when correctly rejecting implied than control traits (e.g., Ham & Vonk, 2003), we added control trait probes to the experimental procedure. These control probes allowed to identify effect sizes for potential inhibiting versus facilitating effects within each of the conditions. In our analyses, we employ the full variety of outlier correction methods that have been used in published research on this topic and explore the impact of the outlier correction methods on the effects of stereotype congruency on STIs. In *Experiment 1*, we aimed at replicating stereotype effects on STIs as observed by Wigboldus et al. (2003). Participants saw the full range of our larger stimulus set with actor labels counter-balanced between participants. Given that we did not observe significant stereotype effects in the pre-registered analyses of response times, we conducted *Experiment 2* to replicate this effect with preregistered analyses of error rates. Again, we did not observe the expected effects of stereotypes on STIs. Faced with these unsuccessful replications, we shifted our attention to methodological differences between our and the original studies, specifically on differences in number and repetition rate of stimuli within the probe recognition paradigm, which we discuss below in more detail. Finally, we conducted *Experiments 3* and *4*, in which we increased procedural similarity to the published research (i.e., employing fewer stimuli with higher repetition rates) in order to investigate whether this change would enable a replication of stereotype effects on STIs.

5. Experiment 1

The aim of Experiment 1 was to (a) replicate the finding that stereotypes influence spontaneous trait inferences with a larger stimulus set and appropriate control conditions, in order to (b) elucidate whether stereotypes lead to facilitation and/or inhibition of (in)congruent trait inferences. We adopted the probe recognition paradigm and manipulated stereotype congruency of actors regarding the respective implied trait by using stereotype congruent or incongruent social group labels; or a neutral first name, for the stereotype neutral trials. In line with the original findings by Wigboldus et al. (2003), we preregistered hypotheses with regard to response latencies expecting (a) that the employed stimulus materials would trigger general STI effects, as indicated by significantly slower response latencies for the implied trait probes than for the control trait probes in the stereotype neutral trials and (b) that stereotype congruency would modulate STI effects, thus expecting a significant interaction between stereotype congruency (congruent vs. neutral vs. incongruent) and probe type (implied trait vs. control trait) on response latencies. We expected this interaction to be driven by a larger STI effect (slower responses to implied vs. control probes) in stereotype congruent trials than in incongruent trials, but had no directional hypotheses whether STI effects in stereotype (in)congruent trials would be smaller, bigger, or the same as compared to the stereotype neutral condition (as there was evidence for all of the three possibilities; see Ramos et al., 2012; Wigboldus et al., 2003; Yan et al., 2012). We additionally ran exploratory analyses of error rates, reported in the supplemental materials.

5.1. Method

5.1.1. Participants

Analyses of Experiment 1 are based on valid data from $N = 230$ participants (90 male, 138 female, 2 diverse; age: 18 to 86 years, $M = 36.1$, $SD = 13.7$) recruited via the online recruitment platform Prolific (www.prolific.co; Palan & Schitter, 2018) for a financial reward of £1.75.

We had preregistered a required sample size of $N = 216$ valid data sets in a power analysis with a power of $1 - \beta = .80$ and $\alpha = .05$ to target a minimum effect size of interest, $\eta_p^2 = .022$, for the interaction effect in the central 2 (probe type: implied, control) x 3 (stereotype congruency: congruent, neutral, incongruent) repeated measures ANOVA. Given an expected exclusion rate of 5–10%, we collected data of $N = 241$ participants. Following preregistered criteria, we excluded data from $n = 11$ participants. For a detailed description of participant eligibility criteria, exclusion and a sensitivity power analysis of the final sample size, see supplemental materials.

5.1.2. Design

The full design of the probe recognition task followed a three (stimulus set assignment: 1, 2, 3, between) by two (statement: target, filler; within) mixed design, with the further three (stereotype congruency: congruent, neutral, incongruent) by four (probe type: implied trait, control trait, included noun, included verb) within factors nested into the target trials and the within factor filler type (1,2,3) nested into to the filler trials.

The critical trials used for analyses formed a 2 (probe type: implied trait, control trait) by 3 (stereotype congruency: congruent, neutral, incongruent) within-subjects design. Mean response latencies per cell of the target design served as the main, and error rates as auxiliary dependent variable.

5.1.3. Materials

We created a new, large, and extensively pretested stimulus pool of trait-implying behavioral statements paired with stereotype (in) congruent actor labels with a mix of qualitative and quantitative approaches in a total of $k = 7$ pretests and an aggregated pretest sample size of $N = 754$. The procedure and decision criteria of each pretest, as well as the final stimulus pool used in all four experiments, is described in the supplemental materials.

5.1.3.1. Target items. We selected 33 extensively pretested trait-implying behavior descriptions to imply traits without explicitly mentioning them. In the stereotype (in)congruent experimental conditions, statements were paired with the social category label pretested to be either typical or untypical for the implied trait. In the stereotype-neutral condition, statements were combined with a gender-neutral first name. Every target statement was assigned four probe words: (1) the implied trait (implied), and (2) a trait of same valence implied by another target sentence (control), both requiring a negative response, as well as (3) an included noun, and (4) an included verb, both requiring an affirmative response (see Table S.2).

The target statements were separated into three stimulus sets of eleven items, such that within participants, one third of the target items were presented in the stereotype congruent, neutral, and incongruent condition, respectively. Using a Latin-square design, stimulus sets were counterbalanced across participants such that each stimulus appeared only once for each participant but equally often in all three conditions throughout the study.

5.1.3.2. Filler items. In order to prevent the formation of response biases during completion of the probe recognition task, we employed 33 additional filler statements (see Table S.3 in the supplemental materials), thus balancing the number of required affirmative and negative responses (a) across all adjective probes and (b) across the entire probe recognition task. Filler statements explicitly included trait adjectives or adverbs that were also used as probes. For every filler sentence, we selected seven respective probes, namely (1) the included trait, (2) another included adjective/adverb, (3) a trait of same valence included in another sentence, (4) an included noun, (5) an included verb, (6) a new noun, and (7) a new verb. The filler statements were randomly assigned to be presented in one of three filler type conditions that differed regarding the selection of probes, with the restriction that for each participant, each filler condition appeared equally often. Eleven of the filler statements included a social category label, the remaining 22 included a first name as actor (such that, throughout the entire task, first names and labels appeared equally often).

5.1.4. Procedure

We implemented an adapted version of a Probe Recognition Paradigm (Todd et al., 2011), administered online using Inquisit web [Computer software] (Inquisit 4 (4.0.10.0), 2016; Millisecond Software LLC, http://www.millisecond.com). The experiment was displayed full-screen to prevent distractions. Participants were introduced to the task as a study on text comprehension. They were instructed to read a series of behavioral statements and to indicate whether the subsequently presented four words had been part of the previous statement. We provided participants with one example for a statement and respective probes and asked them to leave their index fingers on the response keys throughout the task and to focus on responding accurately (“please try to make as few mistakes as you can while still responding quickly”).

Each of the behavioral statements was presented for 4000 ms in the center of the computer screen, followed by a blank screen for 500 ms, a row of five fixation crosses for 1000 ms and series of four probe words in random order, of which each was presented until participants responded (with a post response pause of 100 ms). All of the stimuli were presented in blue letters in the center of the white screen and accompanied by a reminder of the response keys (“[A] No”, “[L] Yes” in the left and right lower corners, respectively) in black letters. In case of incorrect responses, a red X appeared in the middle of the screen for 500 ms before the next probe word appeared (no requirement of response correction). Participants first completed three practice trials and received feedback

about their performance before completing the experimental task consisting of 66 trials (including each 33 target and filler statements) presented in random order.

After completing the probe recognition task, participants received feedback about their aggregate performance, provided demographic information (age, gender, native language[s]), reported their task compliance ("How seriously did you work on the task?"; "How concentrated could you stay while working on the task?") on Likert-like scales ranging from 0 = *not seriously/concentrated at all* to 10 = *very seriously/concentrated*, and were asked for their assumptions about the study purpose ("What do you think this study was about?") in an open response format. Finally, they were fully debriefed about the purpose of the study and given the option to confirm or withdraw initial consent for data storage and analyses. The whole experiment lasted approximately 16 min.

5.2. Results

5.2.1. Outlier, data transformation and aggregation

We had preregistered various trimming and transformation criteria for response latencies and provide a complete overview of how applying each criterion affected the focal 2 (Probe type: implied trait vs. control trait) by 3 (stereotype congruency: congruent vs. neutral vs. incongruent) interaction effect in Table S.5 in the supplemental materials. In the main body of this article, we report results based on the outlier-correction used by Wigboldus et al. (2003), a general cut-off of responses ≥2000 ms (excluding 1.1% of trials).

For the main analyses, we computed mean response latencies of correct responses (overall M_{RT} = 776 ms, *SD* = 158 ms), separately for each cell of the design and participant.

5.2.2. Planned analyses

To verify that our stimulus materials and our adoption of the probe recognition paradigm were generally sensitive for the assessment of STI effects, we first conducted a one-tailed repeated measures *t*-test of responses in the stereotype neutral trials only. As predicted, participants' responses to implied trait probes were significantly slower than to neutral control trait probes, thus validating stimulus selection (see Table 1, for descriptive values and STI effect test statistics in each of the conditions).

The planned 2 (Probe type: implied trait vs. control trait) by 3 (stereotype congruency: congruent vs. neutral vs. incongruent) within-subjects ANOVA on mean response latencies for correct responses revealed a significant main effect of Probe type, $F(1, 229) = 274.41$, $p < .001$, $\eta_p^2 = .545$, 90% CI [.48; .60], but no significant main effect of stereotype congruency, $F(2, 458) = 0.85$, $p = .428$, $\eta_p^2 = .004$, 90% CI [.00; .02]), nor a significant interaction, $F(2, 458) = 1.43$, $p = .239$, $\eta_p^2 = .006$, 90% CI [.00; .02]. We additionally conducted three separate *t*-tests of implied versus control probes, which confirmed that STI effects were significant in all three conditions with moderate-to-large effect sizes (see Table 1).

5.3. Discussion

Results of Experiment 1 indicated robust spontaneous trait inferences from behavior, characterized by a moderate-to-large effect size that is typical for the probe recognition paradigm (see Bott et al., 2022). Thus, our newly developed stimulus materials and adopted procedure was generally sensitive for capturing STI effects. Contrary to our hypotheses, there was no interaction between stereotype congruency and probe type in response latencies, and STI effects were similarly moderate to large in all congruency conditions. This finding diverges from previous research where participants' response latencies to implied probes were affected by stereotype congruency (e.g., Wigboldus et al., 2003).

Our exploratory analyses reported in the supplemental materials indicated that the hypothesized effect instead might have shifted into

Table 1
Response latencies (in ms) for implied and control probes in the stereotype congruent, neutral, and incongruent condition of Experiments 1-4.

	Stereotype congruent							Stereotype neutral							Stereotype incongruent						
	Implied		Control		STI effect[b]			Implied		Control		STI effect[b]			Implied		Control		STI effect[b]		
Exp.	*M*	(*SD*)	*M*	(*SD*)	*df*	*t*	*p*	d_z	*M*	(*SD*)	*M*	(*SD*)	*df*	*t*	*p*	d_z	*M*	(*SD*)	*M*	(*SD*)	*df*	*t*	*p*	d_z
1	808	(167)	748	(147)	229	10.79	< .001	0.71	798	(162)	748	(149)	229	9.33	< .001	0.62	802	(158)	754	(153)	229	8.87	< .001	0.58
2[a]	648	(102)	627	(106)	76	2.68	.004	0.31	631	(93)	621	(90)	76	1.60	.057	0.18	645	(107)	625	(102)	76	2.28	.013	0.26
3	855	(179)	765	(168)	226	11.43	< .001	0.76	829	(183)	757	(153)	226	8.94	< .001	0.59	828	(182)	783	(163)	226	5.35	< .001	0.35
4a	842	(75)	774	(52)	235	11.54	< .001	0.75	840	(62)	771	(53)	235	13.17	< .001	0.86	832	(61)	780	(63)	235	9.37	< .001	0.61
4b	854	(103)	770	(76)	233	9.74	< .001	0.64	831	(91)	764	(79)	233	8.79	< .001	0.57	844	(106)	777	(82)	233	7.51	< .001	0.49

Note. [a] These analyses had been preregistered as exploratory. [b] STI-effects refer to the within-participants difference of response times in the implied vs. control probe conditions, respectively. Experiment 4a refers to the low, and 4b to the high repetition condition of Experiment 4, respectively.

the error rates: We observed a significant interaction effect between probe type and stereotype congruency on error rates with an intermediate effect size ($\eta_p^2 = .086$). This effect was driven by a significantly larger STI effect in the stereotype congruent condition compared to the stereotype neutral baseline and incongruent condition. However, given the procedural characteristics of the probe recognition task as we employed it (i.e., with an accuracy-focused procedure and instruction), we neither expected nor preregistered error rates as the main dependent variable. Correspondingly, participants in Experiment 1 produced a very low average error rate of $M = .06$ ($SD = .09$), as typical for this paradigm, and 34 participants (14.8%) did not commit any errors. We thus cannot rule out the possibility that the observed effect is a false positive effect. It remains an open question whether this unexpected result signaled that the hypothesized effect of stereotype congruency on STIs shifted into the error rates due to procedural characteristics of our experimental design, or whether we observed a false positive effect. To answer this question, we conducted a close replication of Experiment 1 in which we (a) adapted the experimental procedure to drive the effects of our manipulations into the error rates and (b) established error rates as the main dependent variable and preregistered our hypotheses accordingly.

6. Experiment 2

Experiment 2 was a close replication of Experiment 1, with the only exception that we instructed participants to focus on responding fast and implemented a response deadline procedure with feedback for slow responses. Parallel to Experiment 1, we expected general STI effects in the stereotype neutral trials, indicated by higher error rates following implied trait probes as compared to control trait probes, as well as a significant interaction between stereotype congruency (congruent vs. neutral vs. incongruent) and probe type (implied trait vs. control trait) on error rates in the target trials. The hypothesized direction of this interaction was also parallel to Experiment 1 (larger STI-effect in stereotype-congruent than in incongruent trials, but we had no directional hypotheses regarding STI-effects in stereotype (in)congruent versus stereotype-neutral trials).

6.1. Method

6.1.1. Participants

The analyses of Experiment 2 are based on data from $N = 77$ participants ($N = 76$ for error rate analyses; 36 male, 40 female, 1 diverse; age: 18 to 74 years, $M = 36.0$, $SD = 12.2$) recruited via Prolific with the same eligibility requirements and the same financial reward as in Experiment 1. Sample size was estimated for an effect size of $\eta_p^2 = .060$ based on the first Experiment's interaction effect in the exploratory 2 (probe type: implied, control) x 3 (stereotype congruency: congruent, neutral, incongruent) ANOVA on error rates, $\eta_p^2 = .086$, and its lower bound of the 60% CI [.06, .11] (leaving a 20% risk that the corresponding population effect might be lower than the confidence lower bound; Perugini et al., 2014). We had preregistered the required sample size of $N = 78$ based on a power analysis with $1 - \beta = .80$ and $\alpha = .05$. Although this required sample size was considerably smaller than in Experiment 1, we had deemed this approach reasonable, given that the exploratory result of Experiment 1 provided us with the most relevant estimate for this specific effect of interest (as compared to a general estimate for an effect size); as well as prudent, given that we relied on the lower bound of the 60% confidence interval.

Given an expected exclusion rate of 5–10%, we collected data of $N = 80$ participants. For the analyses of response latencies, we excluded data from $n = 3$ participants, and for the error rates, of $n = 4$ participants, respectively (see supplemental materials).

6.1.2. Design

The full design was identical to Experiment 1, with error rates serving as principal dependent variable.

6.1.3. Materials and procedure

We employed the same stimulus materials and identical procedure as in Experiment 1, with the exception that participants were instructed to focus on responding *fast* while also responding accurately, and that we implemented a response deadline procedure. That is, if participants' responses fell above 750 ms, the probe word was replaced by a "too slow" message written in bright blue letters in the center of the screen until they responded. This response deadline was chosen based on participants' mean response times of 756 ms to target trials in Experiment 1 to trigger subjective time pressure without severely hampering performance. Participants received no feedback on correct/incorrect responses. The whole study lasted approximately 16 min.

6.2. Results

6.2.1. Outlier, data transformation and aggregation

Data preparation and treatment was identical to Experiment 1 (see also Table S.5 for the multiverse approach). Table 2 reports descriptive values and test statistics of the STI effects of error rates in all conditions and experiments. Overall, participants responded faster ($M = 633$ ms, $SD = 100$ ms), but with higher error rates ($M = .15$, $SD = .14$) in Experiment 2 as compared to Experiment 1, indicating that the response deadline procedure was successful.

6.2.2. Planned analyses

We first verified that our adoption of the probe recognition paradigm using a response deadline and error rates as dependent variable was sensitive for the assessment of STI effects by conducting a one-tailed repeated measures *t*-test of responses in the stereotype-neutral trials. As predicted, participants' error rates were significantly higher for implied trait probes than for control trait probes (see Table 2).

The planned 2 (Probe type: implied trait vs. control trait) by 3 (stereotype congruency: congruent vs. neutral vs. incongruent) within-subjects ANOVA on error rates revealed a significant main effect of Probe type, $F(1, 75) = 38.73$, $p < .001$, $\eta_p^2 = .341$, 90% CI [.20; .46], but there was no significant main effect of stereotype congruency, $F(2, 150) = 1.19$, $p = .307$, $\eta_p^2 = .016$, 90% CI [.00; .05]), nor a significant interaction effect, $F(2, 150) = 0.40$, $p = .669$, $\eta_p^2 = .005$, 90% CI [.00; .03].

Results of exploratory analyses of response latencies are reported in the supplemental materials.

6.3. Discussion

Results of Experiment 2 again indicated significant and robust STI effects in both dependent variables. Importantly, we did not observe any significant interaction between stereotype congruency and probe type, neither on error rates nor on response latencies. STI effects in error rates were characterized by similar moderate-to-large effect sizes, and STI effects in response latencies were characterized by similar small-to-medium effect sizes throughout conditions. Thus, although our procedure should have fostered the effect of stereotype congruency on STIs in the error scores, we observed no such effect. Taken together, these findings suggest that effects of stereotypes on STIs are not replicable or unstable at best, and that the significant stereotype congruency effect on STIs observed in the error rates of Experiment 1 indeed may have been a false positive. The question remains, however, why the present studies did not consistently replicate stereotype effects on STIs as reported in the literature (e.g., Wigboldus et al., 2003).

One reason for this non-replication may be located in procedural characteristics that differed between our studies and the original research. Specifically, most available studies implemented a relatively small number of stimuli (between six and fourteen trait-implying behavioral statements; see supplemental materials). Importantly, most

Table 2
Error rates for implied and control probes in the stereotype congruent, neutral, and incongruent condition of Experiments 1-4.

	Stereotype congruent								Stereotype neutral								Stereotype incongruent							
	Implied		Control		STI effect[b]				Implied		Control		STI effect[b]				Implied		Control		STI effect[b]			
Exp.	*M*	(*SD*)	*M*	(*SD*)	*df*	*t*	*p*	d_z	*M*	(*SD*)	*M*	(*SD*)	*df*	*t*	*p*	d_z	*M*	(*SD*)	*M*	(*SD*)	*df*	*t*	*p*	d_z
1[a]	0.13	(0.13)	0.03	(0.06)	195	11.97	< .001	0.86	0.08	(0.09)	0.03	(0.06)	195	9.22	< .001	0.66	0.08	(0.10)	0.04	(0.06)	195	6.04	< .001	0.43
2	0.20	(0.15)	0.12	(0.13)	75	4.14	< .001	0.47	0.18	(0.15)	0.12	(0.11)	75	4.35	< .001	0.50	0.17	(0.16)	0.11	(0.13)	75	3.84	< .001	0.44
3[a]	0.14	(0.14)	0.06	(0.10)	164	5.93	< .001	0.46	0.12	(0.14)	0.05	(0.09)	164	6.77	< .001	0.53	0.13	(0.15)	0.04	(0.09)	164	7.26	< .001	0.56
4a[a]	0.11	(0.10)	0.02	(0.04)	210	13.13	< .001	0.90	0.08	(0.08)	0.02	(0.05)	210	8.46	< .001	0.58	0.08	(0.09)	0.02	(0.05)	210	7.94	< .001	0.55
4b[a]	0.16	(0.16)	0.04	(0.09)	166	9.11	< .001	0.71	0.11	(0.13)	0.05	(0.10)	166	5.56	< .001	0.43	0.13	(0.13)	0.06	(0.11)	166	5.88	< .001	0.46

Note. [a] These analyses had been preregistered as exploratory. [b] STI-effects refer to the within-participants difference of response times in the implied vs. control probe conditions, respectively. Experiment 4a refers to the low, and 4b to the high repetition condition of Experiment 4, respectively.

previous research manipulated stereotype congruency within statements *and* participants, leading to repetitions of statements, actor labels, and trait probes within participants. For example, in Wigboldus et al.'s (2003) studies, participants read repeated presentations of the same behavioral statement paired with different actors (e.g., "The skinhead hits the saleswoman", "The girl hits the saleswoman") as well as explicit actor-trait pairs ("The skinhead is aggressive", "The girl is aggressive") – and each of these statements again repeated with different probe words. We, in contrast, employed a much higher number of behavioral stimuli and ensured that each behavioral statement was presented only once to each participant. Notably, we had originally implemented these differences in order to reduce unwanted error variance by avoiding potential carry-over effects due to repeated stimulus presentations, such as stimulus-response bindings (e.g., Henson et al., 2014) or negative priming effects (e.g., Tipper, 2001). We had thus expected to *increase* the sensitivity of the probe recognition paradigm for any effects genuinely related to spontaneous impression formation from behavior and stereotypes. Given the unexpected results, however, we began to suspect that the procedural differences between our and the original studies might be responsible for the non-replication of stereotype effects in our experiments. For example, the high repetition rate of the same stimuli with different actor labels may have increased the salience of these actor labels. Further, the comparison between congruent and incongruent labels for each behavior may have induced contrast effects that signaled the relevance of these actor labels, which in turn may have increased the accessibility of related stereotype contents and/or increased the tendency to apply activated stereotypes to the behavioral inference. Indeed, the assumption that such stimulus characteristics affect person inferences is not new: There is ample evidence from impression formation research demonstrating that categorical information impacts impressions only when it is made salient (Beckett & Park, 1995), accessible (Köpetz & Kruglanski, 2008), and relevant (Gawronski et al., 2003; Köpetz & Kruglanski, 2008; for a theoretical model with similar assumptions, see Kunda & Thagard, 1996).

We thus conducted a third experiment, in which we employed our carefully pretested stimulus materials from Experiment 1 and 2 but designed the procedure of the probe recognition task to be as similar as possible to the procedure employed by previous research (e.g., Wigboldus et al., 2003).

7. Experiment 3

In Experiment 3, we implemented an experimental procedure as similar as possible to the original research by Wigboldus et al. (2003) and Ramos et al. (2012). Therefore, we presented each participant with only a subset of six different behavioral statements. Overall, participants saw each statement altogether twelve times throughout the task: six times as a behavioral statement, and six times as trait-version of that statement. Both the behavioral and trait statements appeared as target and as filler trials (with different sets of probes) and in each of the three stereotype congruency conditions (with different actors). For example, the same group of participants were presented twice (with different probes) with the behavioral statements "The bodybuilder picked up a stack of boxes like it was nothing" (stereotype congruent), "The old man picked up a stack of boxes like it was nothing" (stereotype incongruent), and "Leslie picked up a stack of boxes like it was nothing" (stereotype neutral); and twice with each of the trait statements (e.g., "The bodybuilder was strong", "The old man was strong", "Leslie was strong"). We also modified probes of the filler trials (but not of target trials) to better resemble previous research in that the social group labels also appeared as probe words.

Assuming that these procedural characteristics were relevant in producing an effect of stereotype congruency on STIs, we expected to observe a significant interaction between stereotype congruency (congruent vs. neutral vs. incongruent) and probe type (implied trait vs. control trait) on response latencies in the target trials. Like in the

previous studies, we expected this interaction to be driven by a larger STI effect in stereotype congruent than in incongruent trials and had no directional hypotheses regarding STI-effects in stereotype (in)congruent versus stereotype-neutral trials.

7.1. Method

7.1.1. Participants

The analyses of Experiment 3 are based on data from $N = 227$ participants (88 male, 134 female, 4 diverse; age: 18 to 74 years, $M = 33.2$, $SD = 13.5$) recruited via Prolific with the same eligibility requirements and financial reward as in Experiment 1 and 2. We had preregistered a required sample size of $N = 216$ valid data sets based on the same power analysis as in Experiment 1. We collected data of $N = 237$ participants (including an expected exclusion of 5–10%). Following preregistered criteria, we excluded data from $n = 10$ participants (see supplemental materials).

7.1.2. Design

The design of the probe recognition task diverged from Experiment 1 and 2 in that each participant only saw a subset of our larger stimulus sample whilst still completing a similarly long probe recognition task. The task thus followed a 5 (Stimulus set assignment: 1–5; between) by 2 (Statement: target, filler; within) by 3 (stereotype congruency: congruent, neutral, incongruent; within) mixed design, with the further 4 level within-factor probe type (implied trait, control trait, included noun, included verb) nested into the target trials and the 3 level within-factor filler type (1, 2, 3; see *Materials*) nested into to the filler trials.

The critical trials formed the same 2 (probe type: implied, control) x 3 (stereotype congruency: congruent, neutral, incongruent) within-subjects design as in Experiment 1 and 2. Like in Experiment 1, mean response latencies per cell of the target design served as the main, and error rates as auxiliary dependent variable.

7.1.3. Materials

7.1.3.1. Target items. We removed three statements from the initial statement pool (see Table S.2) and split the remaining 30 statements into stimulus sets of 6 statements each. The target statements were paired with the same four probe words as in Experiment 1 and 2, except for the control trait probes, where we ensured that the control traits were implied by another statement within the same stimulus subset.

7.1.3.2. Filler items. Filler stimuli differed in various aspects from our previous two experiments. We developed three types of filler stimuli with the same demands as in Experiment 1 and 2 for balancing correct affirmative and negative responses, and the additional aim to resemble Wigboldus et al. (2003) more closely (see Table S.4 in the supplemental materials). *Filler type 1* consisted of the target statements (with stereotype congruent, incongruent, and neutral actors), each paired with four filler probes: (1) a noun included in another statement, requiring a negative response, and (2) an included verb, (3) an included actor, and (4) an included adjective/adverb (or another included word, in case no adjective/adverb was available, such as a pronoun, determiner, preposition, or noun); all three requiring an affirmative response. For *filler type 2* and *3*, following Wigboldus et al. (2003), we developed one adjective-version of each target statement consisting of the trait implied in the behavioral statement and the respective actor. For example, from the target statement "The bodybuilder picked up a stack of boxes like it was nothing", we formed the filler statement "The bodybuilder was strong". For *filler type 2*, the probes were the same as used for the respective target statement (e.g., strong – caring – boxes – picked). Because the adjective implied in the target trial is included in these filler statements, this probe required an affirmative response, whereas the remaining three probes required a negative response. Probes for *filler type 3* consisted of (1) the included trait and (2) the actor of the statement, both requiring an affirmative response, as well as (3) the control trait and (4) a verb included in another target statement of the same set, both requiring a negative response. In sum, this resulted in 54 filler statements per stimulus set.

7.1.4. Procedure

The procedure of the probe recognition paradigm was close to Experiment 1, with the difference that participants were required to correct their response after error feedback in order to continue with the task. Participants were randomly assigned to one of the five experimental stimulus sets consisting of 72 trials presented in individual random order. Of the 72 trials, 18 were target statements (six statements presented three times, with a stereotype congruent, neutral, and incongruent actor); the remaining 54 trials were filler trials. After completing the probe recognition task, participants responded to the same questions as in Experiment 1 and provided informed consent. The whole study lasted approximately 17 min.

7.2. Results

7.2.1. Outlier, data transformation and aggregation

Like in Experiment 1, our main dependent variable was computed by aggregating mean response latencies of correct responses (overall $M =$ 803 ms, $SD = 175$ ms) per participant and cell of the design. Likewise, while we report analyses based on the general cut-off criterion of responses $\geq$2000 ms, resulting in exclusion of 1.3% of trials, we report results of analyses using different trimming and transformation methods in the supplemental materials (Table S.5). Note that results were not as homogeneous as in the previous studies.

7.2.2. Planned analyses

A one-tailed repeated measures *t*-test of response latencies in the stereotype-neutral trials showed that participants' responses to implied trait probes were significantly slower than responses to control trait probes, indicating a sensitivity of the modified procedure for assessing STI effects. Tables 1 and 2 report descriptive statistics of response latencies and error rates, as well as test statistics of STI effects in each of the conditions.

The planned 2 (probe type: implied trait vs. control trait) by 3 (stereotype congruency: congruent vs. neutral vs. incongruent) within-subjects ANOVA on mean response latencies for correct responses yielded a significant main effect of probe type, $F(1, 226) = 184.74$, $p < .001$, $\eta_p^2 = .450$, 90% CI [.37; .51], and also a significant main effect of stereotype congruency, $F(2, 452) = 3.57$, p[GG] $= .030$, $\eta_p^2 = .016$, 90% CI [.00; .04]. Importantly, the expected 2 (probe type: implied trait vs. control trait) by 3 (stereotype congruency: congruent vs. neutral vs. incongruent) interaction effect was significant, $F(2, 452) = 8.80$, $p < .001$, $\eta_p^2 = .037$, 90% CI [.01; .07] (see Fig. 1; Note that the interaction was non-significant for 6 out of 15 trimming / transformation methods; see Table S.5).

Effect sizes for Spontaneous Trait Inferences (i.e., differences between implied and control traits) were significant in all stereotype congruency conditions, but descriptively highest for stereotype congruent trials, followed by the neutral control trials, and the stereotype incongruent trials (see Table 1). For further inspection of this interaction, we submitted individual STI-indices (difference scores of response latencies to implied minus control traits) to three planned follow-up *t*-tests between the congruency conditions (applying Bonferroni correction for multiple comparisons, thus $p < .017$). STI difference scores did not differ significantly between the stereotype congruent ($M_{diff} = 90$ ms, $SD_{diff} = 119$ ms) and neutral ($M_{diff} = 72$, $SD_{diff} = 121$) condition, $t(226) = 1.77$, $p = .078$, $d_z = 0.12$, 95% CI [−0.01; 0.25]; but did differ between the stereotype neutral and the stereotype incongruent ($M_{diff} = 45$ ms, $SD_{diff} = 126$ ms) condition, $t(226) = 2.45$, $p = .015$, $d_z =$ 0.16, 95% CI [0.03; 0.29], and between the stereotype congruent and

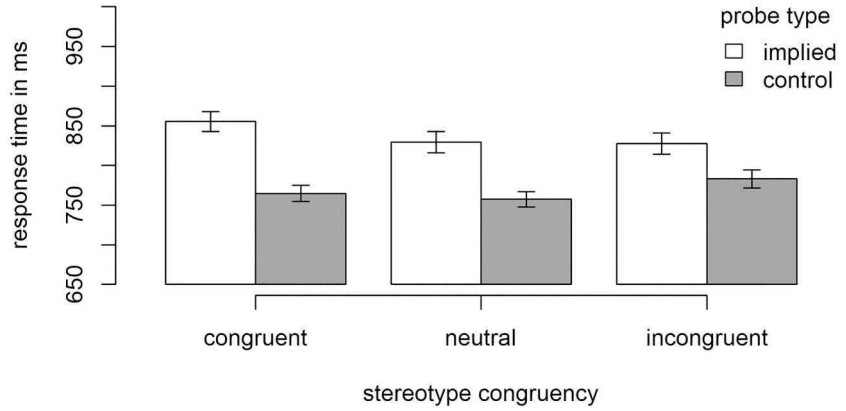


Fig. 1. Response latencies in Experiment 3 for implied and control probes in the stereotype congruent, neutral, and incongruent condition. The central tendency is the mean, error bars represent 95% confidence intervals for means in within-subjects designs (see Morey, 2008).

incongruent condition; $t_{\text{one-tailed}}(226) = 4.05$, $p < .001$, $d_z = 0.27$, 95% CI [0.13; 0.41] (see Fig. 1).[1]

7.3. Discussion

We had conducted Experiment 3 as close replication of published research (e.g., Wigboldus et al., 2003) and the altered procedure seemed indeed more sensitive for an effect of stereotype congruency on STIs: We observed a significant interaction between probe type and stereotype congruency, albeit characterized by a small effect size. Pairwise comparisons further indicated that the STI effect was *reduced* for stereotype *incongruent* actors as compared to neutral actors, but there was no *increase* of STI effects for stereotype *congruent* actors, replicating some of the prior research (Stewart et al., 2003; Wigboldus et al., 2003, 2004). Note, however, that this interaction was not entirely robust for variations of trimming and transformation methods of response times (see Table S.5) and thus should be interpreted with caution. The comparison of Experiment 3 to Experiments 1 and 2 lends support to our initial suspicion that stimulus repetitions may be responsible for the observed interaction effect between stereotype congruency of actor-based and behavior-based inferences. However, in order to systematically investigate this suspicion, it seems essential to conduct a direct comparison between the two employed procedures.

8. Experiment 4

In order to allow for a direct comparison of the different experimental procedures, and to replicate our previous pattern of results, we conducted Experiment 4, in which participants were randomly assigned to either a low-repetition procedure (parallel to Experiment 1) or a high repetition procedure (parallel to Experiment 3). We hypothesized a significant three-way interaction between procedure (high vs. low repetition; between subjects), stereotype congruency (congruent vs. neutral vs. incongruent; within subjects) and probe type (implied trait vs. control trait; within subjects) on response latencies. We expected this interaction to be driven by a larger 2 (probe type) x 3 (stereotype congruency) interaction effect in the high as compared to the low repetition condition.

8.1. Method

8.1.1. Participants

The analyses of Experiment 4 are based on data from $N = 470$ participants (261 male, 202 female, 5 diverse; age: 18 to 87 years, $M = 40.7$, $SD = 14.1$) recruited via Prolific with the same eligibility requirements as the prior experiments, for a reward of £2.60. A subgroup of $n = 236$ completed the low repetition procedure of the probe recognition paradigm (as in Experiment 1 and 2), $n = 234$ completed the high repetition procedure (as in Experiment 3).

We had preregistered a target sample size of $N = 480$ (determined based on financial constraints). Allowing for an expected exclusion of 4.6%, we had collected $N = 507$ valid data sets (preregistered: $N = 503$). Following preregistered criteria, we excluded data from $n = 37$ participants (see supplemental materials).

The final sample size of $N = 470$ was sensitive to detect an interaction effect of $\eta_p^2 = .010$ ($\alpha = .05$, $1 - \beta = .80$) in the central 2 (procedure: high vs. low repetition; between) x 2 (probe type: implied, control; within) x 3 (stereotype congruency: congruent, neutral, incongruent; within) mixed ANOVA (MorePower 6.0; Campbell & Thompson, 2012).

8.1.2. Design, materials, and procedure

In Experiment 4, participants were randomly assigned to complete either an identical replication of Experiment 1 ("low repetition procedure") or Experiment 3 ("high repetition procedure"). Therefore, the critical design of the target trials followed a 2 (Procedure: high vs. low repetition; between) by 3 (Stereotype congruency: congruent, neutral, incongruent; within) by 2 (Probe type: implied trait, control trait; within) mixed design, with the main DV response latency, and the auxiliary DV error rate.

8.2. Results

8.2.1. Outlier, data transformation and aggregation

We employed the same preregistered trimming and transformation methods as in the previous experiments (see Tables S.5 and S.6). Note that results depended on the choice of trimming and transformation methods more than in the previous experiments. For sake of cohesiveness, we adhere to reporting full results based on the general cut-off criterion of responses $\geq$2000 ms in the main body of this article but inform readers about divergent results for the central interaction effects to be found in the supplemental materials (Tables S.5 and S.6).

Our main dependent variable was computed by aggregating mean response latencies of correct responses (overall $M = 807$ ms, $SD = 170$ ms) per participant and cell of the design.

[1] Parallel to Experiment 1 and 2, the follow-up *t*-test between the stereotype congruent and incongruent condition had been preregistered as one-tailed, and the tests including the stereotype neutral condition had been preregistered as two-tailed.

8.2.2. Planned analyses

Replicating prior results, the preregistered one-tailed repeated measures *t*-tests of response latencies in the stereotype neutral trials showed that participants' reactions to implied trait probes were significantly slower than to control trait probes in both the low repetition procedure ($d_z = 0.86$) and the high repetition procedure ($d_z = 0.57$; see Table 1), confirming a sensitivity for assessing STI effects.

8.2.2.1. Joint analysis for low and high repetition procedures. The planned 2 (Procedure: high vs. low repetition; between subjects) by 2 (Probe type: implied trait, control trait; within subjects) by 3 (Stereotype congruency: congruent, neutral, incongruent; within subjects) mixed ANOVA on mean response latencies did not yield a main effect of Procedure, $F(1, 468) = 1.46$, $p = .228$, $\eta_p^2 = .003$, 90% CI [.00; .02], but a significant main effect of Probe Type, $F(1, 468) = 434.84$, $p < .001$, $\eta_p^2 = .482$, 90% CI [.43; .53], as well as a significant main effect of Stereotype Congruency, $F(2, 936) = 3.15$, $p = .043$, $\eta_p^2 = .007$, 90% CI [.00; .02].

The central 2 (procedure) by 2 (probe type) by 3 (stereotype congruency) interaction effect was not significant, $F(2, 936) = 1.15$, $p = .318$, $\eta_p^2 = .002$, 90% CI [.00; .01] (see Fig. 2). We observed a marginally significant 2 (probe type) by 3 (stereotype congruency) interaction effect, $F(2, 936) = 3.06$, $p = .047$, $\eta_p^2 = .007$, 90% CI [.00; .02]. Note, however, that this 2 × 3 interaction effect was non-significant for 12 out of 15 trimming / transformation methods; see Table S.6.

8.2.2.2. Separate analyses for low and high repetition procedures. As preregistered, we further conducted two separate 2 (Probe type) by 3 (Stereotype congruency) within-subjects ANOVAs for the high and the low repetition procedures, respectively.

8.2.2.2.1. Low repetition procedure. In the low repetition procedure, this analysis yielded a significant main effect of probe type, $F(1, 235) = 297.30$, $p < .001$, $\eta_p^2 = .559$, 90% CI [.49; .61], no significant main effect of stereotype congruency, $F(2, 470) = 0.25$, $p = .781$, $\eta_p^2 = .001$, 90% CI [.00; .01], but a marginally significant interaction effect, $F(2, 470) = 3.11$, p[GG] $= .048$, $\eta_p^2 = .013$, 90% CI [.00; .03] (interaction non-significant for 10 out of 15 trimming / transformation methods; see Table S.5).

We computed follow-up t-tests, applying Bonferroni correction for multiple comparisons (thus $p < .017$), comparing STI effects (i.e., differences between implied and control traits) between the stereotype congruency conditions. These t-tests revealed that, unexpectedly, STI effects in the stereotype congruent ($M_{diff} = 68$ ms, $SD_{diff} = 90$ ms) and stereotype incongruent ($M_{diff} = 53$, $SD_{diff} = 86$) condition did not differ significantly, $t(235) = 2.01$, $p = .023$, $d_z = 0.131$ (one-tailed); and neither did the congruent and the neutral ($M_{diff} = 69$ ms, $SD_{diff} = 81$ ms) condition; $t(235) = -0.17$, $p = .869$, $d_z = -0.011$ (two-tailed). However, STI effects were significantly smaller in the incongruent as compared to the neutral condition; $t(235) = 2.54$, $p = .012$, $d_z = 0.165$ (two-tailed).[2]

8.2.2.2.2. High repetition procedure. In the high repetition procedure, the 2 × 3 ANOVA yielded a significant main effect of probe type, $F(1,233) = 181.08$, $p < .001$, $\eta_p^2 = .437$, 90% CI [.36; .50], as well as a significant main effect of stereotype congruency, $F(2, 466) = 3.72$, $p = .025$, $\eta_p^2 = .016$, 90% CI [.00; .04], whereas the 2 (Probe type) by 3 (Stereotype congruency) interaction effect was non-significant, $F(2, 466) = 1.66$, $p = .191$, $\eta_p^2 = .007$, 90% CI [.00; .02], thus not replicating results of Experiment 3.

8.3. Discussion

In Experiment 4, we had aimed at systematically comparing the different experimental procedures employed in our first two studies (i.e., low repetition procedure) versus our Experiment 3 and most published research (i.e., high repetition procedure). Spontaneous Trait Inference effects (i.e., differences between implied and control traits) were significant in both procedures and all stereotype congruency conditions, mirroring results from our previous experiments. Contrary to our hypotheses, however, we did not observe a significant moderation of stereotype effects on STIs by procedure. Descriptively, the pattern of results was even reversed to the results obtained in Experiment 1 and 3, with a significant stereotype congruency effect on STIs in the low repetition condition (for 5 out of 15 preregistered trimming and transformation methods), but no such effect in the high repetition condition. Therefore, results neither replicated our initial results, nor the results of published research (i.e., an effect of stereotype congruency on STIs with a high repetition procedure; e.g., Wigboldus et al., 2003). Furthermore, although the probe type by stereotype congruency interaction was significant in the low repetition condition, it was driven by higher STI effects in the stereotype neutral as compared to the congruent and incongruent conditions. The direction of the interaction effect is thus also contrary to prior research. Finally, our multiverse analysis approach (i.e., our comparison of different plausible trimming and transformation methods to control for outliers in response latencies) indicated a relatively low level of robustness of the stereotype effect on STIs, which was more often non-significant than significant. In sum, the results of Experiment 4 contradict the hypothesis that a procedure with higher repetitiveness produces effects of stereotypes on STIs. Instead, we again observed large and highly robust STI effects that do not appear to be reliably moderated by stereotype congruency.

9. General discussion

The present research had originally aimed at investigating the underlying mechanisms of stereotype effects on spontaneous trait inferences from behavior (STIs). Our four high-powered pre-registered experiments, however, mostly obtained null results or small effects of stereotypes on STIs that were not robust to different trimming and transformation methods. Eventually, only one of our four high-powered and preregistered studies showed the hypothesized moderating effect of stereotypes on STIs, and that with a small effect size. Further, results yielded inconclusive results regarding the notion that the experimental procedure (low versus high repetitiveness) may represent a systematic moderator for the occurrence of such stereotype effects – with one affirmatory and one adversarial study. These results seem remarkable given that we had deemed our experiments methodologically superior to the prior research: We included a higher number of carefully pretested stimuli, control trait probes for every behavioral statement to compare actual STI effects between stereotype congruency conditions, larger sample sizes with high test power to detect small effect sizes, and meticulously preregistered hypotheses, methods, and analyses.

In sum, our experiments paint the picture of highly robust spontaneous trait inferences in all experiments, conditions, and dependent variables but signal that stereotype effects on STIs may be less robust than one may presume from the previously published research.

9.1. Determinants and generalizability of stereotype effects on STIs

Up to date, effects of stereotypes on STIs are typically postulated as a certain empirical finding without reference to potential limitations or boundary conditions (e.g., Chen et al., 2021; Uleman et al., 2012). Our own research, however, does not support this postulate and we would like to argue that the published research is also not as conclusive as warranted to be interpreted as a robust and generalizable research finding. At closer inspection, published research findings yield small effect sizes that shift *unpredictably* between different dependent variables (between and within lines of research; e.g., Ramos et al., 2012), that are sometimes absent for subsamples (e.g., Wigboldus et al., 2004),

[2] Parallel to Experiments 1 to 3, the follow-up *t*-test between the stereotype congruent and incongruent condition had been preregistered as one-tailed, and the tests including the stereotype neutral condition had been preregistered as two-tailed.

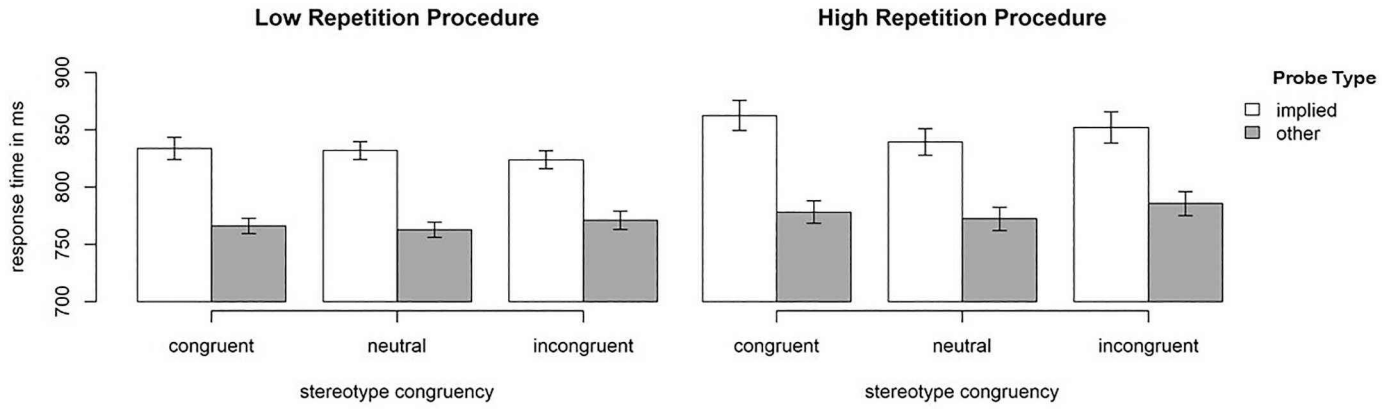


Fig. 2. Response latencies in the low vs. high repetition procedure in Experiment 4, for implied and control probes in the stereotype congruent, neutral, and incongruent condition. The central tendency is the mean, error bars represent 95% confidence intervals for means in within-subjects designs (see Morey, 2008).

and that are partly based on rather small sample sizes ($N = 1681$ in $k = 8$ publications, with $k_{studies} = 24$ experiments; compared to $N = 1004$ in the current research), and a very restricted number of stimuli (between 6 and 14 statements; e.g., Stewart et al., 2003; Wigboldus et al., 2003; with the recent exception of 24 stimuli in Yang et al., 2022). Furthermore, given the general research culture of the time when most studies on stereotype effects on STIs were published, we cannot rule out that some existing non-significant research findings sucummed to the file-drawer effect and have not yet been published.

Taken together, we cannot rule out that the effect of stereotypes on STIs is either non-systematic or so small – especially compared to the large and robust effect of STIs – that its occurrence appears to some extent random and/or determined by yet unknown moderating factors. Such moderating factors may boost or hinder stereotype effects by (a) amplifying the frequency or strength of stereotype activation, (b) amplifying the likelihood of application of stereotype-based inferences to the spontaneous impression of that individual, and/or (c) enhancing the ambiguity of trait-implying behavior. We had explored one potential moderating factor in two experiments, namely high repetitiveness of stereotype labels and behavioral statements within the experimental procedure,. We had we hypothesized that it may influence the relative salience, activation and/or application of stereotypes, which in turnmay influence stereotype effects on STIs. However, we did not observe consistent findings. Similarly, other proposed moderators such as cognitive load (Wigboldus et al., 2004), negative mood (Wang et al., 2015), or high power (Wang & Yang, 2017) have not yet been replicated. Possibly, yet unknown moderators may appear as unsystematic variance and thus contribute to the low robustness of stereotype effects, potentially because they may depend on participant-stimulus interactions (e.g., level of individual accessibility and/or personal endorsement of each single stereotype-based inference) – which cannot be adequately detected with the current methods. Further investigating moderators of stereotype effects on STIs thus seems, in our view, a necessary and fruitful endeavor for future research.

9.2. *Implications for stereotype effects in impression formation*

We had introduced the current research by referring to seminal published literature documenting that stereotypes can and do guide perceivers' impressions in some instances, even when perceivers have access to individuating information about others (e.g., Duncan, 1976; Dunning & Sherman, 1997; Sagar & Schofield, 1980); a notion that is also reflected in classical models of impression formation (e.g., Brewer, 1988; Fiske & Neuberg, 1990). In light of our current results, it is, however, important to note that these seminal findings have predominantly been shown for ambiguous behaviors and situations – in which stereotypes can be used to disambiguate incoming information. This does not imply that social categorization and stereotype activation inevitably or universally affect impression formation (Roth et al., 2019). On the contrary, when perceivers have access to individuating information about others, such as the individual behavior described in our stimulus materials, it appears more functional to use these as basis of impression formation rather than category labels and associated stereotype traits. Indeed, research on the integration of individuating and categorizing information into person impressions documents that perceivers' reliance on stereotypes can be small to non-existent (e.g., Beckett & Park, 1995; Köpetz & Kruglanski, 2008; Monroe et al., 2018; Rubinstein et al., 2018). For example, target gender has been shown to influence perceivers' judgments *only* when made more salient than the individuating information (Beckett & Park, 1995) or when judged as subjectively relevant by perceivers (with strong gender-stereotypic associations; Gawronski et al., 2003).

We argue that our current research is in line with these findings because in typical STI studies, participants are provided with individuating information about actors exhibiting highly trait-diagnostic and unambiguous behaviors, which gives little room for more general stereotypic information to influence the inference process in person perception.

Considering our results and our inspection of the variability of effects in the published literature, we thus agree with Kunda and Thagard's (1996) conclusion that we can "not assume that stereotypes dominate impressions, or that they are used earlier and more automatically than are other types of information" (p. 302). Instead, they may be unreliable or small, and may be highly dependent on specific characteristics of the situation or experimental procedure – both of which remain to be systematically investigated in future research.

9.3. *Conclusion*

Our research demonstrates that stereotype effects on spontaneous impressions from unambiguous behavior may be less robust than previously assumed. To be clear, we do not deny that stereotypes play a crucial role in people's making sense of others and can have influential and pervasive effects. The exact mechanisms of when and under which conditions stereotypes are activated and applied to an individual person of which more or less diagnostic information is available, however, need to be further investigated. We thus argue that it is crucial to formulate and investigate small-scale theories about the determinants and moderators for effects of stereotypes on STIs in order to advance large-scale theories about the role of categories and behavior in the impression formation process (see Degner et al., 2006).

We believe that we need far more systematic research to understand if and under which circumstances social categorization and stereotype activation influence the impression formation process. The two research

traditions of category-based and behavior-based person inferences have, so far, developed largely separately from each other (Chen et al., 2021) – and it appears high time that we engage in cross-domain integrative theorizing and research.

Open practices

Preregistrations of methods, hypotheses and analyses plans, as well as the experimental files and analyses codes for all experiments can be found at https://osf.io/bjndf/?view_only=8f95c66579024c0fb6b44fdadb899cd6. All measures, manipulations, and exclusions in all four experiments are disclosed.

Author note

This research was supported by a Ph.D. scholarship from the Hans-Böckler-Foundation to the first author. We thank Nicoleta Mihailova and Ricardo Bolaños González for their valuable help in analyzing the qualitative data of the pretests.

Declaration of Competing Interest

The authors report no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2023.104497.

References

Beckett, N. E., & Park, B. (1995). Use of category versus individuating information: Making base rates salient. *Personality and Social Psychology Bulletin, 21*(1), 21–31. https://doi.org/10.1177/0146167295211004

Bodenhausen, G. V., Macrae, C. N., & Sherman, J. W. (1999). On the dialectics of discrimination: Dual processes in social stereotyping. In S. Chaiken, & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 271–290). Guilford Press. https://doi.org/10.1037/h0091141.

Bott, A., Brockmann, L., Denneberg, I., Henken, E., Kuper, N., Kruse, F., & Degner, J. (2022). Spontaneous trait inferences from behavior: A systematic meta-analysis. *Personality and Social Psychology Bulletin*, 1–25. https://doi.org/10.1177/01461672221100336

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

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods, 44*(4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0

Chen, J. M., Quinn, K. A., & Maddox, K. B. (2021). *Bridging the gap between spontaneous behavior- and stereotype-based impressions* (pp. 1–40). https://doi.org/10.31234/osf.io/2nbde

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*(6), 1314–1329. https://doi.org/10.1037//0022-3514.83.6.1314

Degner, J., Wentura, D., & Rothermund, K. (2006). Indirect assessment of attitudes with response-time-based measures: Chances and problems. *Zeitschrift Fur Sozialpsychologie, 37*(3), 131–139. https://doi.org/10.1024/0044-3514.37.3.131

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. doi: 10.1037%2F0022-3514.56.1.5.

Dijksterhuis, A., & Van Knippenberg, A. (1996). The knife that cuts both ways: Facilitated and inhibited access to traits as a result of stereotype activation. *Journal of Experimental Social Psychology, 32*(3), 271–288. https://doi.org/10.1006/jesp.1996.0013

Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology, 34*(4), 590–598. https://doi.org/10.1037/0022-3514.34.4.590

Dunning, D., & Sherman, D. A. (1997). Stereotypes and tacit inference. *Journal of Personality and Social Psychology, 73*(3), 459–471. https://doi.org/10.1037/0022-3514.73.3.459

Fiedler, K., & Schenck, W. (2001). Spontaneous inferences from pictorially presented behaviors. *Personality and Social Psychology Bulletin, 27*(11), 1533–1546. https://doi.org/10.1177/01461672012711013

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. https://doi.org/10.1016/s0065-2601(08)60317-2

Gawronski, B., Ehrenberg, K., Banse, R., Zukova, J., & Klauer, K. C. (2003). It's in the mind of the beholder: The impact of stereotypic associations on category-based and individuating impression formation. *Journal of Experimental Social Psychology, 39*(1), 16–30. https://doi.org/10.1016/S0022-1031(02)00517-6

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54*(5), 733–740. https://doi.org/10.1037/0022-3514.54.5.733

Graham, S., & Lowery, B. S. (2004). Priming unconscious racial stereotypes about adolescent offenders. *Law and Human Behavior, 28*(5), 483–504. https://doi.org/10.1023/b:lahu.0000046430.65485.1f

Ham, J., & Vonk, R. (2003). Smart and easy: Co-occurring activation of spontaneous trait inferences and spontaneous situational inferences. *Journal of Experimental Social Psychology, 39*(5), 434–447. https://doi.org/10.1016/S0022-1031(03)00033-7

Henson, R. N., Eckstein, D., Waszak, F., Frings, C., & Horner, A. J. (2014). Stimulus-response bindings in priming. *Trends in Cognitive Sciences, 18*(7), 376–384. https://doi.org/10.1016/j.tics.2014.03.004

Inquisit 4 (4.0.10.0). (2016). https://www.millisecond.com.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. https://doi.org/10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*(October 2016), 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

Köpetz, C., & Kruglanski, A. W. (2008). Effects of accessibility and subjective relevance on the use of piecemeal and category information in impression formation. *Personality and Social Psychology Bulletin, 34*(5), 692–705. https://doi.org/10.1177/0146167207313730

Krull, D. S., & Erickson, D. J. (1995). Inferential hopscotch: How people draw social inferences from behavior. *Current Directions in Psychological Science, 4*(2), 35–38. https://doi.org/10.1111/1467-8721.ep10770986

Kruse, F., Sprecht, E., & Degner, J. (2023). Comparing person inferences from behavior observations and behavior descriptions. In *[Manuscript in preparation]*.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin, 129*(4), 522–544. https://doi.org/10.1037/0033-2909.129.4.522

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*(2), 284–308. https://doi.org/10.1037/0033-295X.103.2.284

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*, 93–120. https://doi.org/10.1146/annurev.psych.51.1.93

McCarthy, R. J., & Skowronski, J. J. (2011). What will Phil do next? Spontaneously inferred traits influence predictions of behavior. *Journal of Experimental Social Psychology, 47*(2), 321–332. https://doi.org/10.1016/j.jesp.2010.10.015

McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(1), 82–91. https://doi.org/10.1037/0278-7393.12.1.82

Monroe, B. M., Koenig, B. L., Wan, K. S., Laine, T., Gupta, S., & Ortony, A. (2018). Re-examining dominance of categories in impression formation: A test of dual-process models. *Journal of Personality and Social Psychology, 115*(1), 1–30. https://doi.org/10.1037/pspa0000119

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology, 4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Moskowitz, G. B. (2005). Correspondence bias and spontaneous trait inferences. In *Social cognition: Understanding self and others* (pp. 267–309).

Moskowitz, G. B., Stone, J., & Childs, A. (2012). Implicit stereotyping and medical decisions: Unconscious stereotype activation in practitioners' thoughts about African Americans. *American Journal of Public Health, 102*(5), 996–1001. https://doi.org/10.2105/ajph.2011.300591

Newman, L. S. (1991). Why are traits inferred spontaneously? A developmental approach. *Social Cognition, 9*(3), 221–253. https://doi.org/10.1521/soco.1991.9.3.221

Okten, I. O., & Moskowitz, G. B. (2020). Easy to make, hard to revise: Updating spontaneous trait inferences in the presence of trait-inconsistent information. *Social Cognition, 38*(6), 571–625. https://doi.org/10.1521/soco.2020.38.6.571

Otten, S., & Stapel, D. A. (2007). Who is this Donald? How social categorization affects aggression-priming effects. *European Journal of Social Psychology, 37*, 1000–1015. https://doi.org/10.1002/ejsp

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*(2), 181–192. https://doi.org/10.1037//0022-3514.81.2.181

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power Estimates. *Perspectives on Psychological Science, 9*(3), 319–332. https://doi.org/10.1177/1745691614528519.

Quadflieg, S., & Macrae, C. N. (2011). Stereotypes and stereotyping: What's the brain got to do with it? *European Review of Social Psychology, 22*(1), 215–273. https://doi.org/10.1080/10463283.2011.627998

Quinn, K. A., Macrae, C. N., & Bodenhausen, G. V. (2007). Stereotyping and impression formation: How categorical thinking shapes person perception. In M. A. Hogg, & J. Cooper (Eds.), *The SAGE handbook of social psychology: Concise student edition* (pp. 68–92). SAGE Publications Ltd.. https://doi.org/10.4135/9781848608221.n4

Ramos, T., Garcia-Marques, L., Hamilton, D. L., Ferreira, M., & Van Acker, K. (2012). What I infer depends on who you are: The influence of stereotypes on trait and situational spontaneous inferences. *Journal of Experimental Social Psychology, 48*(6), 1247–1256. https://doi.org/10.1016/j.jesp.2012.05.009

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Roth, J., Deutsch, R., & Sherman, J. W. (2019). Automatic antecedents of discrimination. *European Psychologist, 24*(3), 219–230. https://doi.org/10.1027/1016-9040/a000321

Rubinstein, R. S., Jussim, L., & Stevens, S. T. (2018). Reliance on individuating information and stereotypes in implicit and explicit person perception. *Journal of Experimental Social Psychology, 75*(July 2017), 54–70. https://doi.org/10.1016/j.jesp.2017.11.009

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39*(4), 590–598. https://doi.org/10.1037/0022-3514.39.4.590

Spencer, S., Fein, S., Wolfe, C. T., Fong, C., & Dunn, M. A. (1998). Automatic activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin, 24*(11), 1139–1152. https://doi.org/10.1177/01461672982411001

Stewart, T. L., Weeks, M., & Lupfer, M. B. (2003). Spontaneous stereotyping: A matter of prejudice? *Social Cognition, 21*(4), 263–298. https://doi.org/10.1521/soco.21.4.263.27003

Tipper, S. P. (2001). Does negative priming reflect inhibitory mechanisms? A review and integration of conflicting views. *The Quarterly Journal of Experimental Psychology, 54*(2), 321–343. https://doi.org/10.1080/713755969

Todd, A. R., Molden, D. C., Ham, J., & Vonk, R. (2011). The automatic and co-occurring activation of multiple social inferences. *Journal of Experimental Social Psychology, 47*, 37–49. https://doi.org/10.1016/j.jesp.2010.08.006

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*(5), 1051–1065. https://doi.org/10.1037//0022-3514.83.5.1051

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93*(3), 239–257. https://doi.org/10.1037/0033-295X.93.3.239

Uleman, J. S., Hon, A., Roman, R. J., & Moskowitz, G. B. (1996). On-line evidence for spontaneous trait inferences at encoding. *Personality and Social Psychology Bulletin, 22*(4), 377–394. https://doi.org/10.1177/0146167296224005

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. *Advances in Experimental Social Psychology, 28*(C), 211–279. https://doi.org/10.1016/S0065-2601(08)60239-7

Uleman, J. S., Rim, S., Saribay, S. A., & Kressel, L. M. (2012). Controversies, questions, and prospects for spontaneous social inferences. *Social and Personality Psychology Compass, 6*(9), 657–673. https://doi.org/10.1111/j.1751-9004.2012.00452.x

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology, 59*, 329–360. https://doi.org/10.1146/annurev.psych.59.103006.093707

Wang, M., Xia, J., & Yang, F. (2015). Flexibility of spontaneous trait inferences: The interactive effects of mood and gender stereotypes. *Social Cognition, 33*(4), 345–358. https://doi.org/10.1521/soco.2015.33.4.1

Wang, M., & Yang, F. (2017). The malleability of stereotype effects on spontaneous trait inferences: The moderating role of perceivers' power. *Social Psychology, 48*(1), 3–18. https://doi.org/10.1027/1864-9335/a000288

Wigboldus, D. H. J., Dijksterhuis, A., & van Knippenberg, A. (2003). When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. *Journal of Personality and Social Psychology, 84*(3), 470–484. https://doi.org/10.1037/0022-3514.84.3.470

Wigboldus, D. H. J., Sherman, J. W., Franzese, H. L., & van Knippenberg, A. (2004). Capacity and comprehension: Spontaneous stereotyping under cognitive load. *Social Cognition, 22*(3), 292–309. https://doi.org/10.1063/1.3033202

Yan, X., Wang, M., & Zhang, Q. (2012). Effects of gender stereotypes on spontaneous trait inferences and the moderating role of gender schematicity: Evidence from Chinese undergraduates. *Social Cognition, 30*(2), 220–231. https://doi.org/10.1521/soco.2012.30.2.220

Yang, F., Li, M., Han, Y., Fan, X., & Zhang, Q. (2022). My general manager is warmer than department manager: Stereotypes about senior and junior high-power individuals and their influences on spontaneous trait inference. *Frontiers in Psychology, 13*(December), 1–19. https://doi.org/10.3389/fpsyg.2022.1015736