

TELL ME YOUR BIASES.
A FRAMEWORK OF REPORTING ON AUTOMATIC
COGNITIONS AND BEHAVIORS.



Inauguraldissertation
zur Erlangung des Doktorgrades
der Humanwissenschaftlichen Fakultät
der Universität zu Köln
nach der Promotionsordnung vom 18.12.2018

vorgelegt von

Alexandra Gödderz

aus Aachen

Januar 2023

Erstgutachter: Prof. Dr. Andreas Glöckner

Zweitgutachter: Dr. Adam Hahn

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität Köln im
Mai 2023 angenommen.

Datum der mündlichen Prüfung: 23.05.2023

Erklärung

Chapter 2 beruht auf folgendem Manuskript:

Goedderz, A., Rahmani-Azad, Z., & Hahn, A. (2023). Awareness of implicit attitudes revisited: Meta-analysis on replications across samples and settings. [Manuscript in preparation].

Der dritte Autor hat das Studiendesign entwickelt und einen Teil der Studien programmiert, erhoben, analysiert und publiziert. Ich habe 9 der 17 inkludierten Studien programmiert, erhoben, analysiert und für die Meta-Analyse aufbereitet. Die Idee für diese Meta-Analyse, sowie die Entscheidungen darüber, welche Analysen verglichen und präregistriert werden habe ich gemeinsam mit dem dritten Autor entwickelt. Die zweite Autorin hat bei der Aufbereitung und Analyse von einigen Studien geholfen. Ich habe die Meta-Analyse durchgeführt und das Manuskript geschrieben. Die zweite Autorin und der dritte Autor haben wertvolle Vorschläge für das finale Manuskript eingebracht.

Chapter 3 beruht auf folgendem Manuskript:

Goedderz, A., & Hahn, A. (2022). Biases left unattended: People are surprised at racial bias feedback until they pay attention to their biased reactions. *Journal of Experimental Social Psychology*, 102, 104374.
<https://doi.org/10.1016/j.jesp.2022.104374>

Der zweite Autor hat die zentrale Idee entwickelt. Ich habe die Idee weiterentwickelt und zusammen mit dem zweiten Autor die Studiendesigns entworfen. Ich habe die Studien programmiert, die Daten erhoben, die Daten analysiert und das Manuskript geschrieben. Der zweite Autor hat in jedem dieser Schritte aktiv gewinnbringende Vorschläge eingebracht.

Chapter 4 beruht auf folgendem Manuskript:

Goedderz, A., & Hahn, A. (2023). Predicting implicit preferences towards social groups vs. food items: The role of normativity and social desirability in accurate IAT score predictions. [Manuscript submitted for publication at the Personality and Social Psychology Bulletin].

Ich habe die Idee für die Untersuchung von Backwaren eigenständig entwickelt. Ich habe das Studiendesign entworfen, die Studie programmiert und die Daten erhoben. Die Daten wurden in enger Zusammenarbeit mit dem zweiten Autor analysiert und für das Manuskript final von mir aufbereitet. Der zweite Autor und ich haben das Manuskript in enger Zusammenarbeit geschrieben.

Abstract

A central debate in social psychology concerns whether implicit measures capture conscious or unconscious mental content. Supporting the notion that cognitions reflected in implicit evaluations are consciously accessible, recent research has documented that people are able to accurately predict the pattern of their results on Implicit Association Tests (IATs). At the same time, the same participants systematically underestimated and mislabeled their test results and research has documented that people often react with surprise to their IAT results. This suggests that there is a lot that people do not know about their automatic cognitions. To reconcile these competing findings, this dissertation presents a framework which proposes that when and how people gain awareness of their automatic cognitions is determined by different concepts of awareness. Specifically, the framework distinguishes between *introspective awareness* (the ability to sense and report on one's own automatic cognitions) and *social calibration* (the act of labeling those cognitions in accordance with conventions in the reference sample). I present three lines of research that are in line with the propositions of the framework and establish the main premises by (1) replicating the fundamental findings on which the framework is based, (2) examining why people report surprise at IAT results, even though they are able to report on them, and (3) applying the proposed concepts of awareness to a new attitudinal domain. I conclude that the cognitions reflected in implicit evaluations are neither conscious nor unconscious but that they often reside in a preconscious state until people pay attention to them, and that people often lack knowledge about the societal meaning of their automatic cognitions. This new nuanced perspective has important implications for theories of implicit social cognition and can ultimately help gain a better understanding of how much people know about themselves.

Keywords: implicit cognition, unconscious, automaticity, introspection

Deutsche Zusammenfassung

Eine zentrale Debatte in der Sozialpsychologie ist, ob implizite Messungen bewusste oder unbewusste mentale Inhalte erfassen. Jüngste Forschungsergebnisse zeigen, dass Menschen in der Lage sind, ihr eigenes Ergebnismuster auf mehreren impliziten Assoziationstests (IATs) akkurat vorherzusagen. Dies stützt die Annahme, dass Kognitionen, die sich in impliziten Bewertungen widerspiegeln, bewusst zugänglich sind. Gleichzeitig haben dieselben Versuchspersonen ihre Testergebnisse systematisch unterschätzt und falsch eingeordnet und weitere Forschung zeigt, dass Menschen oft überrascht auf ihre IAT-Ergebnisse reagieren. Dies deutet darauf hin, dass es vieles gibt, was Menschen nicht über ihre automatischen Kognitionen wissen. Um diese widersprüchlichen Ergebnisse in Einklang zu bringen, wird in dieser Dissertation ein Rahmenkonzept vorgestellt, das davon ausgeht, dass wann und wie sich Menschen ihrer automatischen Kognitionen bewusst werden, davon abhängt wie Bewusstsein konzeptualisiert wird. Hierbei wird zwischen *introspektivem Bewusstsein* (der Fähigkeit, die eigenen automatischen Kognitionen zu erkennen und zu berichten) und *sozialer Kalibrierung* (dem Akt der Benennung dieser Kognitionen in Übereinstimmung mit Konventionen der Stichprobe) unterschieden. Ich stelle drei Forschungslinien vor, die mit dieser Konzeptualisierung übereinstimmen und die Hauptprämissen des Rahmenkonzeptes stützen, indem sie (1) die originalen Befunde, auf denen das Konzept basiert, replizieren (2) untersuchen, warum Menschen über IAT-Ergebnisse überrascht sind, obwohl sie in der Lage sind, diese vorherzusagen, und (3) die vorgeschlagenen Bewusstseinskonzepte auf eine neue Einstellungsdomäne anwenden. Ich schlussfolgere, dass die Kognitionen, die sich in impliziten Bewertungen widerspiegeln, weder bewusst noch unbewusst sind, sondern dass sie sich oft in einem vorbewussten Zustand befinden, bis Aufmerksamkeit auf sie gerichtet wird, und dass Menschen oft das Wissen über die gesellschaftliche Bedeutung ihrer automatischen Kognitionen fehlt. Diese neue, nuancierte Perspektive hat wichtige theoretische Implikationen und kann letztendlich ein besseres Verständnis darüber befördern, wie viel Menschen über sich selber wissen.

Acknowledgements

This dissertation would not have been possible without the continuous support of many people and I want to take a moment to thank them.

First and foremost, I would like to thank my supervisor Adam Hahn. You sparked my interest in social cognitive research in your outstanding social cognition course, and gave me the opportunity to work with you first as an intern, then as a research assistant, and finally as a doctoral student. You made me believe in my scientific abilities when I most doubted them and gave me a sense of belonging in a place I often felt I did not belong to. Thank you for giving me a place to grow intellectually and personally.

I also want to thank Andreas Glöckner for generously providing shelter when Adam was appointed to a new position, and two of my former chairs no longer existed. I greatly benefitted from the diverse and inclusive work environment you created.

I thank Angela Dorrough for being my academic mentor, an admirable female role model, and a good friend in moments of doubt. You helped me through some of the toughest parts of this process, provided guidance when I felt lost, and encouraged me to pursue and believe in my own ideas.

Along the way, I was fortunate to have been surrounded by many great colleagues who not only offered invaluable feedback on my work but also provided guidance in navigating myself through the academic world. Thank you for making my time at the University of Cologne a time to remember, Kathi Diel, Mike Schreiber, Felix Speckmann, Lea Sperlich, Mareike Westfal, and Tobias Wingen.

Thank you also to Jimmy Calanchini and his group, Kayla Chaplin, Emily Esposito, Deja Simon, and Liz Wilson, for making my international research stay at the University of California, Riverside one of the most fun times of my life, and for giving valuable feedback that substantially improved my academic writing and thinking.

Finally, I want to thank my family, who always encouraged me to pursue my dreams and supported me unconditionally. Thank you, Mum & Dad, for getting excited just by the thought of being able to once call me “Doctor.” Thank you for helping me escape the academic bubble and refining my thinking through endless discussions.

Thank you, Martin, for staying by my side through all ups and downs, for being my rock, for always believing in me, and for making me a better person. Without you, I would probably not even have made it through my bachelor’s degree.

Thank you, Luna, for making my life so full of love and laughter and for reminding me every day of what really matters.

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION.....	1
1.1. Implicit Social Cognition	3
1.2. Awareness and Implicit Evaluations	7
1.3. A Framework of Reporting on Automatic Cognitions and Behaviors.....	12
1.3.1. Introspective Awareness.....	14
1.3.2. Social Calibration	16
1.4. The Current Research.....	18
CHAPTER 2. AWARENESS OF IMPLICIT ATTITUDES REVISITED: A META-ANALYSIS ON REPLICATIONS ACROSS SAMPLES AND SETTINGS	20
Abstract	21
2.1. Introduction	22
2.2. Awareness and Implicit Attitudes – Theoretical Framework and Specific Questions.....	23
2.2.1. Are People Able To Predict Their IAT Scores?	25
2.2.2. Do People Consider Other Information for Traditional Explicit Reports Than What Is Reflected on Their Implicit Measures?	26
2.2.3. Do People Predict Their Own Evaluations or A Normative Pattern? ..	27
2.2.4. Do People Know Where Their IAT Scores Rank in Comparison To Other People?	28
2.2.5. Summary of the Original Findings	30
2.3. The Need for Replications.....	30
2.4. The Current Meta-Analysis	31
2.5. Method	32
2.5.1. Data Inclusion.....	32
2.5.2. Materials	35
2.5.3. Procedure	38
2.5.4. Analyses.....	38
2.6. Results	41
2.6.1. Prediction Accuracy	41
2.6.2. Implicit-Explicit Relationship	44
2.6.3. Predictions vs. Explicit Ratings.....	44

2.6.4. Simultaneously Predicting the Predictions from IAT Scores and Explicit Ratings	47
2.6.5. Prediction Accuracy Beyond Normative Patterns Based on Other Participants' Predictions.....	50
2.6.6. Between-Subjects Analysis	53
2.7. Discussion	56
2.7.1. Subgroup Analyses	59
2.7.2. Limitations.....	62
2.8. Conclusion.....	63

CHAPTER 3. BIASES LEFT UNATTENDED: PEOPLE ARE SURPRISED AT RACIAL BIAS FEEDBACK UNTIL THEY PAY ATTENTION TO THEIR BIASED REACTIONS.	65
Abstract	66
3.1. Introduction	67
3.2. Previous Research on Reactions to IAT Racial Bias Feedback.....	68
3.3. Implicit Evaluations as Unconscious Attitudes	70
3.4. Potential Reasons for Surprise	72
3.4.1. Surprise and Harshness of Feedback: The Feedback Wording Hypothesis	72
3.4.2. Implicit Evaluations as Preconscious Attitudes: The Attention Hypothesis	72
3.4.3. Real Surprise? The Social Desirability Hypothesis	74
3.4.4. Feedback Wording, Attention, or Social Desirability?	74
3.5. The Present Research	76
3.6. Pilot Study.....	78
3.6.1. Method	78
3.6.2. Results and Discussion.....	79
3.7. Study 1	80
3.7.1. Method	80
3.7.2. Results	83
3.7.3. Discussion	84
3.8. Study 2	85
3.8.1. Method	86

3.8.2. Results	88
3.8.3. Discussion	92
3.9. Study 3	93
3.9.1. Method	95
3.9.2. Results	96
3.9.3. Discussion	101
3.10. Study 4a.....	102
3.11. Study 4b	104
3.11.1. Method	105
3.11.2. Results	106
3.11.3. Discussion	108
3.12. Additional Analyses	109
3.12.1. Meta-Analysis	109
3.12.2. Non-White Participants and Participants with Pro-Black Biases.....	110
3.13. General Discussion.....	111
3.13.1. Limitations	116
3.13.2. Why Are Biases Left Unattended?.....	118
3.13.3. Theoretical Implications for Implicit Bias Research.....	119
3.13.4. Practical Implications	121
3.14. Conclusion	122
3.15. Open Practices.....	123

CHAPTER 4. PREDICTING IMPLICIT PREFERENCES TOWARDS SOCIAL GROUPS VS. FOOD ITEMS: IMPLICATIONS FOR THE ROLE OF	
NORMATIVITY AND SOCIAL DESIRABILITY IN ACCURATE IAT SCORE	
PREDICTIONS.....	124
Abstract	125
4.1. Introduction	126
4.2. Previous Research on Awareness and Implicit Evaluations	127
4.2.1. Normative Patterns: Knowing One’s Evaluations or One’s Cultural	
Norms?	128
4.2.2. Social Sensitivity: Awareness vs. Calibration.....	129
4.3. Implicit Evaluations of Food Targets.....	130
4.4. Overview of the Study.....	131

4.5.	Method	132
4.5.1.	Baked Goods Sample.....	132
4.5.2.	Comparison Social Group Sample	135
4.6.	Results	136
4.6.1.	Testing Assumptions About Evaluations Toward Baked Goods and Social Groups	136
4.6.2.	Awareness.....	140
4.6.3.	Predictions Beyond Normative Patterns?	142
4.6.4.	Calibration	144
4.6.5.	Awareness vs. Calibration in Different Domains	145
4.7.	Discussion	146
4.7.1.	Predicting IAT Scores Beyond Normative Patterns	147
4.7.2.	Predicting IAT Score Patterns in Socially Less Sensitive Domains – Awareness vs. Calibration.....	148
4.7.3.	Implications for Dual-process Models	150
4.7.4.	Limitations.....	151
4.7.5.	Conclusion.....	152
<hr/> CHAPTER 5. GENERAL DISCUSSION		154
5.1.	Introspective Awareness or Inferences from External Information?	157
5.2.	What Do People Introspect Upon?.....	159
5.3.	Traditional Explicit Measures and Predictions	162
5.4.	Social Calibration or Socially Desirable Responding?	164
5.5.	Implications for Theories on Implicit and Explicit Evaluations	165
5.6.	Introspective Awareness and Social Calibration of Other Automatic Cognitions	168
5.7.	Generalizability to Other Implicit Measures	169
5.8.	Practical Implications	170
5.9.	Conclusion.....	171
<hr/> REFERENCES		173

Chapter 1. General Introduction

Most western societies today advocate egalitarian values, and most people would probably agree that individuals should be treated equally regardless of their racial background, gender, sexual orientation, or religious belief. Nonetheless, inequalities and discrimination remain a widespread issue across the world. One of social psychology's most important contributions to understanding this mismatch in beliefs and observable behavior is the concept of implicit social cognition (Greenwald & Banaji, 1995). Some researchers have argued that people harbor implicit biases that they are either unwilling or unable to report when explicitly asked, and that such implicit biases may automatically influence their behavior toward different social groups (Greenwald & Banaji, 1995; Kelly & Roedder, 2008; Kurdi, Seitchik et al., 2019). Measurement tools aimed at capturing such implicit biases are typically only weakly related to people's explicitly endorsed attitudes (Hofmann, Gawronski et al., 2005). In consequence, implicit biases are often conceptualized as capturing unconscious mental contents that are inaccessible to introspection (Greenwald & Banaji, 1995; Nosek et al., 2002; Nosek, 2005). This idea has been empirically challenged by research showing that people are able to predict the pattern of their results on the Implicit Association Test (IAT; Greenwald et al., 1998) - a measurement tool aimed at capturing implicit biases - while at the same time reporting other attitudes on traditional explicit measurement scales (Hahn et al., 2014). While this suggests that implicit biases may not be completely inaccessible to introspection, other research challenges the idea that implicit biases are consciously accessible at all times (Hahn & Goedderz, 2020). For instance, people are often surprised at implicit bias feedback revealing that they harbor racial biases which suggests that such feedback captures information they did not expect (Goedderz & Hahn, 2022; Howell et al., 2013). Further, even though people may know that they harbor biases, they often believe that they are less biased than everyone else (Hahn et al., 2014; Howell &

Ratliff, 2017). These observations suggest that people may not be constantly aware of their automatic cognitions and that there may still be a lot that people do not know about their biases.

The purpose of the present dissertation is to reconcile these competing findings in the literature. Consequently, I introduce a framework that proposes that when and how people gain awareness of their automatic cognitions is determined by different concepts of awareness. Specifically, the framework distinguishes between the concept of *introspective awareness* – referring to the ability of a person to know their own automatic cognitions – and *social calibration* – referring to the ability of a person to accurately label their automatic cognitions in accordance with external conventions (Hahn & Goedderz, 2020). Furthermore, the framework suggests that both concepts of awareness rely on different processes, require different empirical designs, and require different analytical approaches. In the present thesis, I present three lines of research that support the propositions of the framework. On the basis of the framework and findings of my research, I propose that people are generally able to access and report their automatic cognitions consciously but that these cognitions often reside in a preconscious state until they are paid attention to (Chapter 3). Moreover, even if people pay attention to their automatic cognitions, especially in socially sensitive domains, they may often appear unaware of them because they may lack knowledge about how to label their automatic cognitions and may be unwilling to label their own cognitions in undesirable ways (Chapter 2 and 4).

These propositions have important theoretical implications for the field of implicit social cognitions by moving the discussion beyond a dichotomy of unconscious or conscious cognitions and providing a more nuanced understanding of when and how people are able to report on their automatic cognitions. Ultimately, it offers new avenues for raising awareness about implicit biases, leading to practical implications for implicit bias interventions.

1.1. Implicit Social Cognition

For over two decades now the field of implicit social cognition has been of central interest to social psychologists. Conventionally, implicit social cognition refers to research that uses indirect measurement tools that infer people's thoughts and evaluations from automatic behavior on computerized reaction tasks without directly asking them about these (Hahn & Gawronski, 2018). These measurement tools started to gain traction with the introduction of the evaluative priming task (EPT, Fazio et al., 1986) and the development of the Implicit Association Test (IAT, Greenwald et al., 1998). Since then, implicit measures have been applied to a wide range of psychological research fields such as prejudice and stereotypes (e.g., Kurdi, Mann et al., 2019)(Kurdi, Mann et al., 2019), clinical psychology (e.g., Nock et al., 2010), personality psychology (e.g., Fatfouta & Schröder-Abé, 2018), and consumer choices (e.g., Friese et al., 2006). Though the field of implicit social cognition now spans across a wide range of topics, the central aim of implicit measures was initially to overcome difficulties in explicitly asking people about their cognitions, especially in potentially socially sensitive topics such as racial attitudes (Fazio et al., 1995; Greenwald & Banaji, 1995). The argument has been made that people may not explicitly report their true attitudes in these domains because of social desirability concerns (Edwards, 1957) or because it may be difficult for them to do so due to a lack of introspection into their own attitudes (Nisbett & Wilson, 1977). Indirect measures sought to address these issues by limiting strategic control over people's responses and by inferring people's attitudes from their response patterns such that introspection is not required. This had led to the prominent summary of indirect measures as capturing cognitions that people are "unwilling or unable to report" (e.g., Nosek, 2005, p. 566).

These two conceptualizations of the cognitions captured by indirect measures can be traced back to two historical lines of research that started the field of implicit social cognition

through the development of the EPT and the IAT (for a summary of the history of implicit social cognition see Payne & Gawronski, 2010). One line of research was based on the assumption that attitudes are stored in memory as associations of varying strength between objects and evaluations and that these associations are automatically activated during the encounter of an attitude object (Fazio, 2007). For example, if a person has a stronger associative link between the social category Black and a negative evaluation, the encounter with a Black person would automatically make negative concepts more readily accessible. From this perspective, an indirect measure was thought to capture this automatically activated link between an object and its evaluation when people do not have time to control their responses. In contrast, it was assumed that explicit measures will show different results from implicit measures whenever people do have time to control their responses and are motivated to present themselves in a positive light. This idea is most prominently summarized in the MODE model (Motivation and Opportunity as DEterminants; Fazio, 1990, 2007). Thus, this line of research assumes that indirect measures reflect evaluations people are unwilling to report and may intentionally hide.

The second line of research in turn was more concerned with consciousness and developed from findings in implicit memory showing that past experiences can influence current behavior without conscious recall of the past experience (Jacoby & Witherspoon, 1982; Schacter, 1987). This principle was picked up on by Greenwald and Banaji (1995) and adapted to the attitude literature resulting in the prominent definition of implicit attitudes as “introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects” (p. 8). One could argue that this definition refers to the source of the mediated response (the experience) as being introspectively unidentified; however it has often been misinterpreted to mean that the cognitions which result from past experiences are unavailable to introspection

(Gawronski et al., 2006). As such, this definition has contributed to the idea that implicit measures capture unconscious cognitions that people are unable to introspect upon. Consequently, the term “implicit” started to get used interchangeably with “unconscious” leading many researchers and the general public to conclude that implicit measures capture “unconscious biases” or even “unconscious racism” (Akram, 2018; Basu, 2018; Cole, 2018; Devlin, 2018; Haider et al., 2011; Haider et al., 2014; Lai et al., 2013; Nosek et al., 2002; Quillian, 2008).

The observation that implicit and explicit evaluations are only weakly correlated (Hofmann, Gawronski et al., 2005) is often interpreted as evidence that people are unaware of the cognitions reflected on their implicit evaluations (Nosek et al., 2002; Nosek, 2005). In contrast, more recent theories such as the Associative-Propositional Evaluation model (APE model, Gawronski & Bodenhausen, 2006, 2011) propose a different explanation for dissociations between implicit and explicit measures. The model assumes that responses on implicit measures reflect the outcome of automatic associative processes whereas responses on explicit measures reflect the outcome of controlled, propositional processes. Consistency between the associative and propositional processes determines whether implicit and explicit evaluations overlap. That is, according to the APE model (Gawronski & Bodenhausen, 2006, 2011), the encounter of an attitudinal object (e.g., a Black person) elicits an automatic associative response (e.g., a negative affective reaction) which is then validated through a propositional process (e.g., egalitarian beliefs, knowledge about discrimination). If the propositional response is inconsistent with the associative response, a person would discard their associative response as invalid and base their explicit report on the propositional thoughts (e.g., report positive attitudes toward Black people on an explicit measure, although they show negative reactions toward Black people on an implicit measure). Consequently, in contrast to the assumption that implicit measures capture unconscious cognitions, the APE

model assumes that people are fully aware of the cognitions reflected on their implicit evaluations and that they consciously reject these when considering other inconsistent propositional information.

These different theoretical positions highlight that there is an ongoing debate in social cognition research regarding what the term *implicit* means (Gawronski et al., 2020). In line with the historical perspective of automatically activated attitudes, some researchers use the term *implicit* to refer to a specific kind of *measurement tool* that captures information about a psychological attribute by limiting people's control over their responses (Fazio & Olson, 2003). In line with the idea of unconscious attitudes influencing people's responses, other researchers use the term *implicit* to refer to the *mental content* that is captured by indirect measurement tools, often using the term *implicit* to refer to *unconscious* mental representations (Greenwald et al., 2002). Due to conceptual unclarities of both of these definitions, in the present dissertation I adopt a framework proposed by De Houwer et al. (2009) and use the term *implicit* to refer to the *outcome* of an *indirect* measurement instrument, and the term *explicit* to refer to the *outcome* of a *direct* measurement instrument. As such, I make no assumptions about the underlying cognitions reflected on these measures. Instead, I hope to inform the understanding of the underlying cognitions with the present research.

In summary, the term implicit social cognition refers to research that uses indirect (implicit) measures which were developed to overcome shortcomings of explicit measures by limiting people's control over their responses. The two most known implicit measures were based on two distinct lines of research that either assumed that implicit measures capture cognitions that people are unwilling or unable to report. An influential definition by Greenwald and Banaji (1995), together with observations of low correlations between implicit and explicit measures led to the assumption that implicit measures capture

unconscious cognitions. In contrast, the APE model (Gawronski & Bodenhausen, 2006, 2011) proposes that people are generally able to consciously access the cognitions reflected on their implicit measures but consciously reject these cognitions because they believe other information to be more valid bases for their explicit reports. Resting on these opposing theoretical considerations, a central debate in implicit social cognition research revolves around whether the cognitions reflected on implicit measures reflect conscious or unconscious mental contents (Gawronski et al., 2006; Hahn & Gawronski, 2018; Hahn & Goedderz, 2020).

1.2. Awareness and Implicit Evaluations

When empirically investigating the extent to which implicit measures capture unconscious cognitions, Gawronski et al. (2006) argued that unconsciousness can refer to three different aspects of a cognition. Specifically, a person can be unaware of (1) the experience that shaped the cognition (*source unawareness*), (2) the cognition itself (*content unawareness*), and (3) the influence the cognition has on subsequent behavior (*impact awareness*). As pointed out earlier, a widespread assumption in research on implicit social cognition is that people are unaware of the cognitions reflected on implicit measures altogether. Thus, in the present dissertation I focus on *content unawareness* of the cognitions reflected on implicit measures. While implicit measures have been applied to many different topics, the historical roots of implicit social cognition lie within the domain of attitudes, and the question of awareness also primarily revolves around research on implicit evaluations. As such, in this section, I focus on the evidence of awareness in research on implicit evaluations.

The claim that the cognitions reflected on implicit evaluations are consciously inaccessible to introspection is often based on the observation that implicit and explicit evaluations are only weakly correlated (Hofmann, Gawronski et al., 2005). Even though it is true that if people were unaware of the cognitions reflected on their implicit evaluations,

correlations between implicit and explicit evaluations would necessarily be low, the reverse conclusion is not warranted (Hahn & Gawronski, 2018). Indeed, a lot of empirical evidence and theoretical considerations speak against the idea that responses on implicit and explicit measures diverge because of unawareness (Gawronski et al., 2006). Instead, motivational aspects, deliberate processing during self-report, conceptual correspondence between measurements, and measurement error can contribute to low correlations between implicit and explicit evaluations (for as summary of the available evidence see Gawronski et al., 2006). In support of motivational reasons for low convergences between implicit and explicit measures, research has documented that the motivation to control prejudice moderates the relationship between implicit and explicit measures toward racial minorities (Dunton & Fazio, 1997; Gawronski et al., 2003; Hofmann, Gschwendner et al., 2005). Further, the announcement of a test in a bogus pipeline paradigm increased correlations between the two measures (Nier, 2005). Showing that deliberate processing during self-report impacts implicit-explicit convergence, a meta-analysis by Hofmann, Gschwendner et al. (2005) has established that implicit and explicit measures correlate more strongly when the explicit measure captures a more spontaneous reaction. Further, implicit and explicit measures have been shown to correlate more strongly when the measured concepts are more aligned, e.g., when the explicit measure asks a question about an affective reaction rather than a deliberate opinion (Banse et al., 2001; Payne et al., 2008). Finally, studies show that the correlations between implicit and explicit measures increase when using latent variable models suggesting that measurement error is another important component of low implicit-explicit convergences (Carpenter et al., 2022; Cunningham et al., 2001). Taken together, low correlations between implicit and explicit measures can have various other reasons than unawareness of the cognitions reflected on implicit evaluations. In line with this, dual-process theories such as the MODE model (Fazio, 2007) or the APE model (Gawronski & Bodenhausen, 2006, 2011)

suggest that the cognitions reflected on implicit measures are fully consciously accessible but rejected for explicit reports due to some of the aforementioned reasons. However, just as unawareness is inferred from low correlations between implicit and explicit measures, awareness is also often inferred from the fact that correlations between implicit and explicit measures are malleable and respond to experimental manipulations (Gawronski et al., 2006). The reasoning is, that if the cognitions reflected on implicit evaluation were unconscious, experimental manipulations should not have a systematic impact on their correlation with explicit measures, because people would be simply unable to introspect upon these cognitions. While this is certainly true, as we have seen, implicit-explicit correlations are determined by many other factors beyond awareness, which limits the informative value of such correlations for examining awareness.

To overcome these difficulties and provide a more direct test of whether people are aware of the cognitions reflected in their implicit evaluations and able to report on them, Hahn et al. (2014) introduced a new paradigm. In their studies, the researchers asked participants to predict how they would score on five upcoming IATs, measuring their implicit evaluations toward Black people, Asian people, Latin American people, celebrities, and children in comparison to White people, regular people (non-celebrities), and adults, respectively. This paradigm differed in two important ways from previous research. First, instead of inferring awareness from the extent to which implicit and explicit measures correlated, they directly asked participants to indicate what they believed an implicit measure would show while also asking them about their explicitly endorsed feelings on explicit measures. This procedure allowed the researchers to delineate whether people would be able to predict their IAT results even though they reported other explicit feelings. Second, by asking participants to predict their results toward five different target-pairs the researchers were able to compute a within-subject correlation per participant between participants'

predictions and their IAT results indicating how accurately participants were able to report on their own *pattern* of IAT results. The authors argued that such within-subject correlations would be a better indicator of introspective awareness because they reflect a participant's ability to report on their own reactions toward several attitude objects at a time. In other words, it is an indicator of how well participants know e.g., that they have more positive reactions toward White than Black people, and yet different reactions toward Asian people. In turn, they argued that between-subject correlations between participants' predictions and their IAT scores across participants calculated separately for each target pair – e.g., the correlation between the participants' predictions and the participants' actual scores on a Black-White IAT – do not only rely on a person's introspective awareness but also on their knowledge about labeling conventions in the sample.

Results were in line with previous findings and theoretical considerations proposing that the cognitions reflected on implicit evaluations should be available for introspection. Across four studies, Hahn et al. (2014) found that participants were quite accurate in predicting the pattern of their IAT results, as indicated by high average within-subject correlations between predictions and IAT scores ($r = .54$). At the same time, classical explicit ratings were less strongly related to participants' IAT score patterns ($r = .20$), and the relationship between explicit ratings and IAT score patterns was entirely explained by participants' predictions. In line with theorizing by the APE model (Gawronski & Bodenhausen, 2006, 2011), this suggests that different information is reflected in implicit and explicit measures and even though the participants in Hahn et al.'s (2014) studies were able to report on the information reflected on their implicit evaluations, they decided to provide other information on their explicit reports.

Even though these findings suggest that people can consciously access the cognitions reflected in their implicit evaluations and to accurately report on them, results also showed

that participants often used inadequate labels to refer to their own cognitions (Hahn et al., 2014). That is, while participants were able to accurately predict the *pattern* of their IAT results as evidenced in high within-subject correlations, the between-subject correlations between participants' predictions and IAT scores averaged across targets were much lower ($r = .31$). A closer examination of participants' predictions suggested that the reason for these differences in within- and between-subject correlations may have been due to participants using weaker labels to refer to their own cognitions than what their IAT results would conventionally be labeled. That is, if their IAT results suggested that they had a "strong" preference for White people over Black people, and a "moderate" preference for White people over Asian people, they would predict that they had a "mild" preference for White over Black people and "little to no" preference for White over Asian people. Note, that in this example participants correctly sensed that their reactions were more favorable toward Asian than Black people (in comparison to White people), but they used different labels to refer to these reactions. This finding corroborates Hahn et al.'s (2014) idea that within- and between-subject analyses may tap into two different concepts when people report on their cognitions reflected on implicit evaluations. It further suggests that even though people may have insight into their own automatic cognitions, there may still be things about these cognitions they do not know. Thus, an open question is which parts of their cognitions people are aware of and which parts may reside outside conscious awareness.

Another puzzling observation in light of Hahn et al.'s (2014) findings is that people often report surprise when they are confronted with feedback from implicit bias tests, such as the IAT, suggesting that they harbor biases (Goedderz & Hahn, 2022; Hillard et al., 2013; Howell et al., 2013). Such surprise reactions are frequently taken as evidence that people are unaware of their biases (Gawronski, 2019; Krickel, 2018). Indeed, it is hard to reconcile with Hahn et al.'s (2014) findings, such that if people evidently know what their IAT results will

show when asked prior to IAT completion, why would they react with surprise at feedback telling them this result? This challenges the idea that people are constantly aware of their biases and poses the question of when people are able to introspect upon their automatic cognitions and when these cognitions may remain unconscious.

Taken together, a considerable amount of research suggests that people are able to gain introspective access to the cognitions reflected on their implicit evaluations. However, research showing people's surprise when confronted with implicit bias feedback raises the question of whether people are constantly aware of their biases. Furthermore, findings by Hahn et al. (2014) show that participants seemed to be less aware of where their biases rank in comparison to other participants in the sample, further questioning whether people are aware of all parts of their cognitions. In the present dissertation I present a framework aimed at reconciling these competing findings and answering the question of when and how people become aware of their automatic cognitions.

1.3. A Framework of Reporting on Automatic Cognitions and Behaviors

The basic ideas of the framework I present next have been articulated in a paper by Adam Hahn and myself published in 2020 in *Social Cognition* (Hahn & Goedderz, 2020). In this dissertation, I provide a formalized version of the framework and discuss the framework's implications in light of the existing literature.

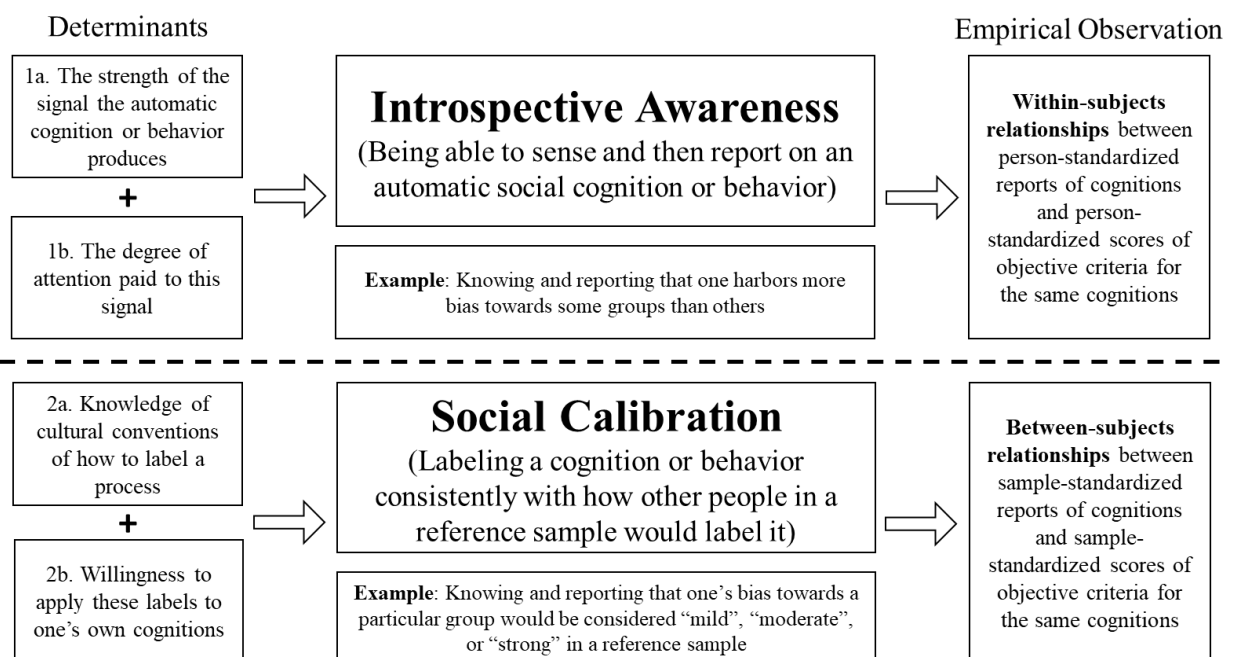
Based on Hahn et al.'s (2014) findings, the main assumption of the framework is that people are, in principle, able to gain awareness of their automatic cognitions and behaviors and are able to report on those. However, in contrast to propositions by dual-process models such as the APE (Gawronski & Bodenhausen, 2006, 2011) or the MODE model (Fazio, 2007), the present framework does not assume constant accessibility of these cognitions as in a trait-definition of consciousness (Hahn & Goedderz, 2020). Resting on theories of consciousness (Dehaene et al., 2006; Hofmann & Wilson, 2010), we rather assume that

automatic cognitions are often *preconscious* and enter a reportable state of consciousness when certain conditions are met. In this context, the term “preconscious” is used to describe the pre-attentive state of a cognition that is known to be consciously accessible but remains unconscious if not attended to (for definitions of trait-unconsciousness, state-unconsciousness, and preconsciousness in the context of implicit evaluations see Hahn & Goedderz, 2020).

Figure 1

Framework of Reporting on Automatic Cognitions and Behaviors: Determinants of Introspective Awareness and Social Calibration and Empirical Strategies to Study Them

Reporting on Automatic Cognitions and Behaviors



The framework further takes into consideration that participants in Hahn et al.’s (2014) studies were quite accurate in predicting their own pattern of IAT results (showed high within-subject correlations) but were less accurate in labeling their IAT results in accordance with conventions (lower between-subject correlations). As such, it proposes that within- and between-subject correlations tap into two distinct constructs that are involved

when people report on their automatic cognitions that give insight into different aspects of awareness: Introspective awareness and social calibration (Goedderz & Hahn, 2023; Hahn et al., 2014; Hahn & Goedderz, 2020). The framework is depicted in Figure 1 and summarizes the different processes that introspective awareness and social calibration depend on, along with the different empirical strategies and analytical methods required to study both constructs.

1.3.1. Introspective Awareness

The concept of *introspective awareness* refers to the ability to sense and report on one's own automatic cognitions and behaviors toward different attitudinal objects. In the domain of racial biases, a person would for instance show high introspective awareness if they know and are able to report that they have a more positive reaction toward White people than Black people and yet a different reaction toward Asian people or Latin-American people (Hahn et al., 2014; Hahn & Goedderz, 2020). Thus, to be able to empirically study introspective awareness, one must ask participants to report on their automatic cognitions toward several targets at a time and then assess the objective criteria of these automatic cognitions. Analytically, the within-subject correlation between person-standardized reports and person-standardized test results across all targets serves as an indicator of how accurately people are able to report their own *pattern* of automatic cognitions when comparison standards or labeling conventions are not taken into consideration.

Integrating theories of consciousness (Dehaene et al., 2006; Hahn & Goedderz, 2020; Hofmann & Wilson, 2010) with research by Hahn and Gawronski (2019) and Goedderz and Hahn (2022, see Chapter 3) suggests that an individual's introspective awareness of their automatic cognitions should depend on (1a) the strength of the signal the cognition produces and (1b) the attention that is paid to this signal. That is, Dehaene et al. (2006) propose in their framework of the global workspace that several processes are constantly competing for the

limited access to active representation in a global workspace state. A process' access to the global workspace and momentary representation in the global workspace state is determined by its bottom-up activation and its top-down attentional amplification. Applied to the topic of implicit social cognition, this means that an automatic cognition may often remain preconscious because the signal the cognition produces is not sufficiently strong (process 1a), or the signal is not attended to (process 1b).

Regarding process 1a, Hofmann and Wilson (2010) postulate that a cognition that produces a strong signal should more easily reach conscious awareness compared to a cognition that produces a weak signal. This signal can potentially be a variety of different experiences such as gut reactions, affective reactions, fluency perceptions, familiarity, or confidence (Hofmann & Wilson, 2010). Applied to the domain of implicit evaluations past research suggests that implicit measures capture spontaneous affective reactions (Gawronski & Bodenhausen, 2006; Gawronski & LeBel, 2008; Hofmann, Gawronski et al., 2005; Hofmann & Wilson, 2010). As such, the extent to which a person has insight into their own cognitions, as reflected on implicit evaluations, should depend on the strength of the affective reactions the cognitions elicit. On the one hand, this suggests that stronger affective reactions, for instance reflected in higher IAT scores, should be easier to consciously access than weaker affective reactions reflected in lower IAT scores. On the other hand, this suggests that experimentally altering the signal a cognition produces should influence people's awareness of their automatic cognitions.

Regarding process 1b, research by Hahn and Gawronski (2019) provides first evidence that for a cognition to be consciously accessible, people need to pay attention to the signal produced by the cognition. Hahn and Gawronski (2019) found in the domain of evaluations of social groups that participants' explicit reports were more aligned with their implicit evaluations and that they acknowledged their biases more after IAT score

predictions. These findings suggest that people learn something new about their own automatic cognitions when they first pay attention to the affective signal the cognition produced. In turn, if people are not actively asked to pay attention to their automatic cognitions, these may often reside outside people's awareness, leaving the cognitions momentarily unavailable to introspection but not generally inaccessible to introspection (Hahn & Goedderz, 2020).

1.3.2. Social Calibration

The concept of social calibration refers to the act of labeling an automatic cognition in accordance with labeling conventions in the reference sample. For instance, in the domain of racial biases, being socially calibrated would mean that a person knows and reports that they have a stronger bias toward Black people than another person in the sample and that their own bias would be called "strong" on a test, while the other person's bias would be called "mild". Empirically, social calibration can be measured by asking participants to report on an automatic cognition toward one target at a time and assess an objective criterion of this cognition. Analytically, the between-subject correlation between sample-standardized reports and sample-standardized test results toward one target serves as an indicator of how accurately people report on the relative strength of their cognition in comparison to other people in accordance with labeling conventions.

We propose that the extent to which a person is socially calibrated in reporting on their automatic cognitions should depend on (2a) the person's knowledge of labeling conventions and (2b) the person's willingness to apply a label to their own cognition. That is, for a between-subject correlation between participants' reports and a criterion to be high, participants need to be able to report the rank of their own criterion. In the domain of implicit racial biases for instance, the most biased person would have to report that they have a "strong bias" while the least biased person would have to agree that their bias should be

called “mild”. However, many people may find it inappropriate to talk about their preferences for one group over another, such that people may lack experience in labeling their own evaluations in comparison to others. Additionally, conventions used to describe implicit evaluations reflected on IAT scores may be known to researchers familiar with the scoring algorithm of the IAT, but it is unlikely that outside this circle people have an idea of the effect sizes the IAT produces, and the labels used to describe these effects because such labels are usually arbitrarily set (Gawronski, 2019; Kruglanski, 1989). Complicating the case of accurate social calibration further, motivational aspects may additionally distort people’s reports on their automatic cognition. People are motivated to maintain a positive view of themselves and expect themselves to score better than the average person in many dimensions (Alicke et al., 1995; Alicke & Sedikides, 2009). In line with this, Howell and Ratliff (2017) have documented that across 10 different topics, including racial bias, weight bias, and gender bias, people generally believed that they were less biased than other people. Even though participants in Hahn et al.’s (2014) studies accurately predicted the pattern of their IAT results, they also thought that they harbored fewer biases than the other participants on average, a statistically impossible result.

These considerations suggest that people should be better calibrated in areas where they have more knowledge about appropriate qualifiers for their cognitions (process 2a), for instance because they more openly discuss their preferences in some areas or because they have been educated about labeling conventions. Further, people should be more willing to apply appropriate labels to their cognitions (process 2b) in domains where they are less concerned with self-presentational issues or for instance if labeling conventions are adapted to be less threatening.

1.4. The Current Research

In the present research, I formalized a framework of reporting on automatic cognitions and behaviors based on findings by Hahn et al. (2014) and theories of consciousness and introspection by Dehaene et al. (2006) and Hofmann and Wilson (2010). In the upcoming chapters I establish the main premises of the framework by examining the robustness of Hahn et al.'s (2014) findings and I present research in line with the proposed processes involved when people report on their automatic cognitions.

To this end, Chapter 2 presents a meta-analysis of 17 published and unpublished studies replicating the original prediction paradigm used by Hahn et al. (2014) in which participants predict and complete five IATs toward 5 different social groups. In line with Hahn et al.'s (2014) findings and in line with the framework, I hypothesized that (1) participants across studies would be able to accurately predict the pattern of their IAT results even though they reported different evaluations on traditional measures (introspective awareness), and (2) that participants would be less accurate in labeling their cognitions in accordance with conventions in the reference sample (social calibration).

Chapter 3 then addresses the question of why people often react with surprise at their IAT results even though they are ostensibly able to accurately predict the pattern of their IAT results. Across four preregistered studies and a mini-meta analysis, I tested the hypotheses that (1) people react with surprise because people rarely pay attention to their biased cognitions, (2) people are surprised at the labels used to describe their cognitions, or (3) people merely pretend to be surprised due to social desirability concerns. Results in favor of the attention-hypothesis would be in line with the proposition of the framework that attention (process 1b) is crucial for gaining introspective awareness of one's automatic cognitions.

In Chapter 4, I examined whether the proposed processes determining introspective awareness and social calibration would be applicable to another domain beyond social

groups. To this end, we asked participants to predict their IAT scores and complete five IATs toward baked goods. In addition, we compared the results to a comparable sample of participants that completed the prediction paradigm toward social groups. In line with the proposed framework, I hypothesized that people should show similar levels of introspective awareness in both domains because signal strength (process 1a) and attention to this signal (process 1b) were held constant across domains. In contrast I expected participants to be better socially calibrated in the domain of baked goods compared to the domain of social groups. This is because people more often talk about their preferences for food and should therefore have more knowledge about labeling conventions (process 2a) and they should be more willing to apply these labels to their preferences for food because the topic is less socially sensitive (process 2b).

It is important to note that Chapters 2,3, and 4 are based on published manuscripts, manuscripts under review, or manuscripts in preparation such that each chapter includes a separate introduction and discussion section. Hence, parts of the introduction or the general discussion of this dissertation may show redundancy with the manuscripts. Further, the chapters were written before the formalization of the framework in this dissertation. As such, the research presented here is influenced by some ideas integrated into the framework but does not constitute a direct or exhaustive test of the complete framework. Instead, the current research has influenced the development of the framework as it is presented in this dissertation and presents findings in line with the propositions of the framework.

In Chapter 5 I provide a general summary of the findings across the Chapters and discuss these in light of the proposed framework. I further debate implications for theories of implicit and explicit evaluations, the generalizability of the present research, general limitations, and practical implications. I end with a final conclusion.

Chapter 2. Awareness of Implicit Attitudes Revisited: A Meta-Analysis on Replications Across Samples and Settings

This chapter is based on the following manuscript:

Goedderz, A., Rahmani-Azad, Z., & Hahn, A. (2023). Awareness of implicit attitudes revisited: Meta-analysis on replications across samples and settings. [Manuscript in preparation].

Please note that headings, citation style, and formatting were changed to fit the layout of this dissertation. The content of the article was not changed.

Abstract

A long-standing debate in social psychology is whether the cognitions reflected on implicit measures are unconscious. Research by Hahn et al. (2014) has documented that people are able to predict the patterns of their results on Implicit Association Tests (IATs) towards five pairs of social groups prospectively. The present article presents a meta-analysis of 17 published and unpublished exact replication studies conducted by or in close supervision of the original author. Replicating Hahn et al., participants in all 17 studies were able to accurately predict the patterns of their IAT results (meta-analytical effect: $b = .44$, equivalent to an average within-subjects correlation). This prediction-accuracy effect was smaller for online ($b = .28$) than lab ($b = .48$) studies, as well as for general-public ($b = .27$) as opposed to student samples ($b = .47$). Moreover, predictions fully explained implicit-explicit relations, and they seemed to reflect unique insights into participants' own cognitions beyond knowledge about normatively expected patterns of implicit responses. This pattern of results remained the same across samples, settings, countries (Canada, US, and Germany), and languages (English vs. German). Further analyses suggested that lower prediction accuracy in online samples seems to partly reflect a suppression effect from higher consistency between traditional explicit evaluations and predictions. Once explicit evaluations were controlled for (exerting a negative unique effect on IAT scores beyond IAT score predictions), online prediction accuracy rose, rendering the online-lab difference non-significant. Together, the results strengthen the hypothesis that cognitions reflected on implicit evaluations are accessible to conscious awareness.

Keywords: Implicit attitudes, consciousness, introspection, racial bias, meta-analysis

2.1. Introduction

In 2014, Hahn et al. published a paper that showed that participants were able to predict the patterns of their results on five IATs toward different social groups. These findings challenged more traditional views conceptualizing the cognitions reflected on implicit¹ evaluations as revealing unconscious attitudes to which people have no introspective access (Greenwald & Banaji, 1995; Lai et al., 2013; McConnell et al., 2011; Nosek et al., 2002). Instead, they favored interpretations by other dual-process models that assume that implicit evaluations are in principle introspectively accessible (Fazio, 2007; Gawronski et al., 2006; Gawronski & Bodenhausen, 2006, 2011). According to these models, dissociations between implicit and explicit evaluations can be explained such that people tend to consider different information for their answers to explicit questions than the spontaneous cognitions that show on implicit measures, which are often rejected (Fazio, 2007; Gawronski et al., 2006; Gawronski & Bodenhausen, 2006, 2011). Since the publication of these studies, there have been some successful conceptual replications showing the generalizability of Hahn et al.'s (2014) findings to other attitudinal domains (Goedderz & Hahn, 2023; Morris & Kurdi, 2022; Rahmani Azad et al., 2022). At the same time, our lab has conducted several direct replications of the original paradigm in the domain of social groups of which some are published and others remain unpublished. Some of these are pilot studies that tested whether the effects would hold in different settings (e.g. online), different languages (i.e., German vs. English), and with culturally different samples (e.g. in the US vs. Canada vs. Germany), which may not ever get published individually. Such unpublished studies, however, may lead

¹ With the exception of the title, we use the terms *implicit* and *explicit* to refer to measurement outcomes. Accordingly we use the term “implicit evaluation” when we refer to an evaluation that is inferred from an indirect computerized measurement instrument, and the term “explicit evaluation” when we refer to an evaluation that is stated on a direct self-report measure (De Houwer et al., 2009; Hahn & Gawronski, 2018). This terminology differs from Hahn et al. (2014), who used the term “implicit” to describe the underlying attitude instead of the measurement outcome. Hence, we made an exception to our measurement-focused usage of the terms in the title because we wanted to reference the original article.

to biased estimations of effect sizes and in the worst case perpetuate false positive findings (Murayama et al., 2014; Nuijten et al., 2015).

In the present research, we address this issue with a preregistered meta-analysis of all published and unpublished studies that directly replicated the prediction paradigm introduced by Hahn et al. (2014) in the domain of social groups. In doing so, we pursue three main goals. First, by including all published and unpublished studies, we want to provide a less biased estimate of an average effect size of the prediction accuracy effect. Second, we test the generalizability of the original findings to different samples and settings by examining differences in effect sizes between different study characteristics. Third, and last, we replicate and meta-analyze additional analyses proposed by Hahn et al. (2014) aimed at answering theoretically relevant questions. Specifically, we examine the extent to which predictions explain unique variance in IAT score patterns over and above traditional explicit measures (i.e., thermometer ratings). Further, we investigate whether participants have unique insights into their own patterns of IAT results or whether predictions are culturally normative and interchangeable across participants from the same sample. Finally, we analyze whether participants are better in predicting their own relative evaluations of different social groups than communicating the relative strengths of their evaluations of one social group in comparison to other participants in the sample, a process we call “social calibration” (Goedderz & Hahn, 2023; Hahn & Goedderz, 2020). We will explain each of these points and the initial results found by Hahn et al. (2014) in more detail next, after a quick summary of the different theoretical models that they address.

2.2. Awareness and Implicit Attitudes – Theoretical Framework and Specific Questions

Different theoretical models make different predictions upon whether the cognitions reflected on implicit measures are consciously accessible or not. On the one hand, there are

models based on Greenwald and Banaji's (1995) conceptualization of implicit social cognition as "introspectively unidentified traces of experience" (Greenwald & Banaji, 1995, p. 8; Lai et al., 2013; McConnell et al., 2011; Nosek et al., 2002). Specifically, when indirect attitude measures were first developed, some researchers argued that explicit measures tap into consciously accessible attitudes while implicit measures capture unconscious evaluations. This idea was met with enthusiasm by other researchers and the public media who started to use the terms "implicit attitudes", "unconscious biases" or even "unconscious racism" interchangeably (Basu, 2018; BBC News, 2017; Devlin, 2018; Haider et al., 2011; Haider et al., 2014; Quillian, 2008). From this perspective, the fact that implicit and explicit measures are often only weakly correlated ($r = .24$ in a meta-analysis by Hofmann, Gawronski et al., 2005) is interpreted as evidence that people are unable to report on the cognitions reflected on implicit measures (Nosek et al., 2002; Nosek, 2005).

On the other hand, several dual-process models propose that the cognitions reflected in implicit measures differ from explicit reports because people consider other information when they have time and resources to think about a deliberate answer. For instance, the MODE model (Fazio, 1990; Fazio & Olson, 2003) proposes that if people are motivated and have the opportunity, they will report different evaluations on explicit measures than they will show on implicit measures. The Associative-Propositional Evaluations Model (APE; Gawronski & Bodenhausen, 2006, 2011) hypothesizes that implicit and explicit measures will diverge when people do not hold their spontaneous reactions to be valid bases for their deliberate evaluations. For instance, a person may feel that they have a more negative spontaneous feeling toward a Black person than toward a White person. However, when asked directly and with time to think about an honest answer that person may think of different information. For example, they may think about Black friends they have, that they genuinely believe that all men are created equal, and that they have egalitarian worldviews. In

this scenario, the person would probably show biases against Black people on an implicit measure but would not report such biases on an explicit measure. Importantly, however, the reason for this discrepancy would not be rooted in a lack of awareness of the spontaneous reactions. Instead, it would reflect the fact that they do not consider their spontaneous reactions to be the only valid bases for their general evaluation of Black people.

Based on these opposing theoretical considerations, Hahn et al. (2014) empirically tested whether people are generally aware of their (biased) evaluations of social groups or not. In these studies, participants were asked to first explicitly rate how they felt toward different social groups on “feeling thermometer” scales. They then went on to predict how they would score on five IATs measuring their reactions toward the social categories Black, Latino, Asian, Child, and Celebrity in contrast to Regular (non-Celebrity) White Adults. Afterwards, they completed all five respective IATs. This study design enabled the researchers to investigate several questions regarding participants’ awareness of the cognitions captured on implicit evaluations. We discuss these questions and the evidence from the original studies next.

2.2.1. Are People Able To Predict Their IAT Scores?

The main focus of Hahn et al.’s (2014) studies was to investigate whether people were generally aware of the cognitions reflected on their implicit evaluations. To test this, they used a within-subject design, letting participants predict and complete five IATs, and examined whether they were able to predict the patterns of their IAT results prospectively. Their reasoning for this particular design was as follows: To investigate whether people know about their own evaluative reactions toward different targets, participants would have to predict how their reactions toward one attitude object differs from their reaction toward another attitude object. This can only be analyzed using within-subject correlations between predictions and implicit evaluation scores for several attitude objects per participant (see also

Hahn & Goedderz, 2020). Results showed that participants predicted the patterns of their IAT scores with significant accuracy, with an average within-subject correlation of $r = .54$ across four studies. Results further showed that this prediction accuracy was independent of (1) whether implicit attitudes were described as true attitudes or culturally learned associations (Studies 1 and 2), (2) whether participants were told to specifically predict their behavior on an IAT (e.g., “which block would be easier for you?”, Study 1) or their “implicit attitudes” (Studies 2-4), or (3) how much explanations they received about the IAT or how much experience they had with the task (Study 4). Overall, these findings are first evidence that people are able to report on the cognitions reflected in their implicit evaluations, suggesting that these cognitions are generally consciously accessible.

2.2.2. Do People Consider Other Information for Traditional Explicit Reports Than What Is Reflected on Their Implicit Measures?

Another goal of Hahn et al.’s (2014) studies was to empirically investigate the theoretical considerations of models such as the APE model (Gawronski & Bodenhausen, 2006, 2011; Fazio, 2007), which postulate that different information factors into explicit and implicit measures. These models hypothesize that the degree to which implicit and explicit measures are correlated depends on how much people rely on their spontaneous gut reaction for their explicit reports. The studies supported this idea. First, correlations between IAT scores and explicit thermometer ratings tended to be low, while correlations between participants’ predictions and their IAT score patterns were always considerably higher. This supports the notion that participants can generally have insight into the cognitions reflected on their IAT scores but nonetheless often report other information on traditional explicit measures. Second, the relationship between explicit thermometer ratings and IAT scores was entirely explained by participants’ predictions in all studies. In line with the APE model (Gawronski & Bodenhausen, 2006, 2011), this shows that beyond a first spontaneous

reaction, people consider additional information for explicit reports that are not captured in implicit measures. Together, these findings constitute supporting evidence for the hypothesis that the reason why implicit-explicit correlations often tend to be low is not that people are unaware of their implicit evaluations. Instead, the data are more compatible with the notion that people do have access to the cognitions reflected on implicit evaluations, but that they rely on additional information for their explicit reports.

2.2.3. Do People Predict Their Own Evaluations or A Normative Pattern?

One explanation for the fact that participants in Hahn et al.'s (2014) studies accurately predicted their IAT score patterns is that they had unique insight into the cognitions reflected on their implicit evaluations. Another possibility is that the IAT score patterns toward the social groups followed a normative pattern and participants were accurate because they predicted what they assumed made most sense in their cultural context. For example, an American citizen may assume that the average other American citizen will have negative reactions toward Black people and Latinos, neutral to somewhat negative reactions toward Asians, and somewhat positive reactions toward Children and Celebrities. If participants' own patterns of IAT results are in line with these assumptions, the participants in Hahn et al.'s (2014) studies would not have predicted what they believed to be their *own* evaluative pattern toward the social groups but rather what they believe to be the culturally shared normative evaluation of the social groups in their context.

To investigate this idea, Hahn et al. (2014) used two approaches. First, they reexamined data of their studies by pairing a random other participant's prediction of the same sample with participants' own IAT responses and vice versa. They argued that if participants only predicted a normative pattern, then any other participant's predictions should be as good as a predictor for their IAT results as their own predictions. In contrast, if participants predicted their own patterns of evaluations, their own predictions should predict

unique variance of their IAT results over and above the randomly paired other participants' predictions. Results supported the latter explanation: The random other participants' predictions showed lower correlations with participants' own patterns of IAT results than their own predictions. Additionally, participants' own predictions predicted unique variance in their own patterns of IAT results over and above the random other participants' predictions.

Second, they tackled the question experimentally. In one study, they additionally asked participants to predict how the average student at their university would score on the respective IATs. Their reasoning was that if participants had unique insight into their own cognitions reflected on implicit evaluations, their own predictions should explain the patterns of their own IAT results over and above what they believed the IAT score results for the average student at their institution would be. In line with this reasoning, results indeed showed that participants' own predictions explained variance in the pattern of their own IAT results over and above their assumed pattern of results for the average student.

These results suggest that participants in Hahn et al.'s (2014) studies reported their own evaluations rather than what they believed to be cultural consensus, at least to some degree. Nonetheless, both approaches also showed that a significant proportion of every participant's unique IAT score pattern was also predicted by random others and by their idea of what an average participant would show.

Taken together, the studies by Hahn et al. (2014) suggest that accurate predictions may be a combination of unique insight and cultural knowledge of normative responses, with the former playing a somewhat larger role.

2.2.4. Do People Know Where Their IAT Scores Rank in Comparison To Other People?

Thus far, the main analyses in Hahn et al.'s (2014) studies focused on within-subject correlations. Using within-subject analyses in a multilevel design allowed the researchers to

estimate whether participants are able to accurately say how their own reactions on the IATs would differ for, e.g., a Black/White IAT compared to a Latino/White IAT and a Child/Adult IAT. However, Hahn et al. (2014) pointed out that most previous research investigating whether people know the cognitions reflected on their implicit evaluations looked at between-subject analyses. However, between-subject implicit-explicit correlations tend to be low (Hofmann, Gawronski et al., 2005), which could be interpreted such that the participants do not seem to know their implicit evaluations. Opposing this interpretation, Hahn et al. (2014), argued that this level of analysis answers a different question: Namely, whether participants know where their results on the IAT rank in comparison to other participants in the same sample. As such, a low correlation in a between-subject analysis could show that participants do not know whether they have more or less biases than other people in the sample, or that all participants use the prediction scale labels differently. Hahn and Goedderz (2020) summarized these two perspectives of awareness that are connected to the two ways of analyses as “introspective awareness” (within-subject analyses) vs. “social calibration” (between-subject analyses). They argue that both types of analyses and thinking about people’s knowledge about their implicit evaluations can tell us different things about what kind of awareness people have of the cognitions reflected on their implicit evaluations.

Following this reasoning, as an additional analysis, Hahn et al. (2014) looked at the between-subject correlations between predictions and IATs computed per target-pair IAT and averaged across the five IATs per study. These averaged between-subject correlations for all 4 studies were still significantly different from zero. However, they were lower than the within-subject correlations. That is, participants seemed to be more accurate in predicting their own pattern of IAT results than estimating whether their biases were “slight”, “moderate”, or “strong” in comparison to other participants in the sample. At the same time,

correlations between explicit thermometer ratings and the IATs did not seem to differ in size for the within-subject or between-subject analyses.

2.2.5. *Summary of the Original Findings*

The studies by Hahn et al. (2014) provided first evidence that people may be aware of the cognitions reflected on their implicit evaluations. Participants in these studies were able to accurately predict the patterns of their IAT scores even beyond normatively expected evaluative patterns and even though they reported other evaluations on traditional explicit scales. These findings speak against older conceptualizations of implicit evaluations capturing unconscious mental contents. Instead, they favor theoretical models that assume that people are generally aware of the cognitions reflected on their implicit evaluations but that people consider other information when they have time to think about a deliberate answer. Lastly, at the same time as participants were able to accurately sense their own biases toward different social groups, they seemed to be less accurate in sensing where their biases ranked in the sample distribution. An open question is whether these different findings pertaining to important theoretical considerations are reliable and robust.

2.3. The Need for Replications

Recent developments in scientific rigorousness highlight the importance of replications for scientific progress (Nosek et al., 2022). First, replications ensure that the original study is not based on a random false positive by showing that the result is reproducible when directly following the original design using a similar sample and setting (Murayama et al., 2014; Simmons et al., 2011). Secondly, a direct replication using a different sample (e.g., users of a survey platform vs. university students, participants in different countries) or a different setting (online vs. laboratory) can speak to the generalizability of the finding by the original studies (Henrich et al., 2010). The original paper by Hahn et al. (2014) already contained four studies, thus the authors already replicated

their initial finding three times while at the same time showing that slight changes in the design did not significantly change the prediction effect. However, these studies were all conducted with undergraduate students at the same US university in the same laboratory. This poses the question whether the findings replicate in other samples and settings. That is, there could be something specific about the students at the specific US university that make them more aware of their implicit biases. For instance, it could be that the topic of implicit biases is very present in the United States such that people are already paying more attention to their biased reactions, or undergraduate students may be a very specific population that is highly sensitive to the topic of implicit biases. Additionally, a laboratory setting may enhance pressure on participants to “admit” to biases in the specific set-up of the study and thus exacerbate actual levels of awareness. Oppositely, the laboratory setting could also underestimate awareness when the presence of an experimenter may make them unwilling to admit to biases of which they are aware.

To ensure that the effects reported by Hahn et al. (2014) are not a random false positive or a specific effect of the investigated group of undergraduates at a US university in a laboratory setting, replications with different samples and in different settings are needed.

2.4. The Current Meta-Analysis

The current meta-analysis reviews published and unpublished replications in different samples and settings all conducted by, or in close supervision of, the original author of the Hahn et al. (2014) studies. As such, the present meta-analysis has three main goals. First, by including published and unpublished studies with varying effect sizes, we aim to inform future research that wishes to replicate Hahn et al.’s (2014) paradigm with a more accurate effect size estimation of the original prediction accuracy. Second, we aim to systematically investigate whether the original findings replicate in different samples and settings by running subgroup analyses for different study characteristics to investigate the

generalizability of the previous findings. Third, and finally, we systematically investigate whether the different results and theoretical considerations suggested by Hahn et al. (2014) hold across all studies. Specifically, beyond the meta-analytical effect of the prediction accuracies across studies, we also examine (1) whether predictions explain variance in IAT score patterns beyond explicit thermometer ratings, (2) whether participants have unique insight into the cognitions reflected on their implicit evaluations beyond normative patterns, and (3) whether participants are better in predicting the patterns of their IAT scores than placing their evaluations accurately in the sample distribution.

2.5. Method

2.5.1. Data Inclusion

All published and unpublished studies that used the prediction procedure put forward in Hahn et al.'s (2014) studies were considered for the present meta-analysis. The considered studies were all conducted or supervised by the original first author of the Hahn et al. (2014) article. We preregistered a list of criteria for the inclusion or exclusion of studies for the present meta-analysis (<https://osf.io/mejzp/>). These criteria aimed to ensure that the examined studies were as comparable as possible in their procedural aspects while allowing other characteristics to vary between studies (e.g. setting and sample characteristics). In total, we collected data from 26 studies² with an overall sample size of 5180 participants. We retained only studies that used the original five social group-pairs used in Hahn et al. (2014) that were Black/White, Asian/White, Latino/White, Child/White Adult, Celebrity/White Regular Adult. As such, we excluded five studies that used other target pairs – for instance baked-goods or occupational groups (e.g., Goedderz & Hahn, 2023). We further excluded three studies

² We included the four original studies reported in the Hahn et al. (2014) in this meta-analysis. Because effects did not differ significantly between different manipulations, we collapsed the data across the four studies and treated them as one study (Study 17). All meta-analytical effects hold when the original studies are not included in the meta-analyses (see supplemental materials).

because participants did not see any or the same pictures as used in the upcoming IATs on their prediction slides and the predictions and IATs in these studies contained only five (instead of ten) pictures and words per category. Another two studies were dropped because the procedure of the implemented IATs differed slightly due to programming errors. We thus kept 17 studies with a total sample size of 3201³. Nine of these studies contained one or more experimental conditions that altered the prediction procedure or procedural aspects of the studies diverged from our preregistered inclusion criteria. As preregistered, we retained these studies but included only conditions for analyses that followed our preregistered inclusion criteria and exclude conditions that differed from these criteria. We further excluded participants with missing data on any of the central variables for our main analyses (predictions, thermometer ratings, IAT scores), participants who did not finish the study, or who failed attention checks or seriousness checks where applicable (e.g. in online studies). Following recommendations by Greenwald et al. (2003) we deleted trials ≥ 10.000 ms before calculating IAT scores, and excluded participants that responded ≤ 300 ms in over 10% of the trials in any of the five IATs. In line with the original publication by Hahn et al. (2014) and as a final step, we dropped participants that had participated in a study with the prediction paradigm before, but kept their data in their first participation. The final sample size thus consisted of 17 studies with a total of 1734 participants. An overview of the final set of studies, their initial sample size and the retained sample size after the data cleaning process, as well as the central demographic characteristics can be examined in Table 1.

³ A table including a list of all considered studies and the respective exclusion criterion for the present meta-analysis can be found in the supplemental materials.

Table 1*Overview of All Studies, Sample Sizes, Sample Characteristics, and Study Characteristics*

Study ID	Study Code	year	Total N	Final N	Status	Setting	Sample Group	Language	Mean Age	Age SD	% Female	% White	dominant citizenship
1	UOBSGQG2020	2020	65	65	Unpublished	Online	Students	German	30.02	11.39	61.5	83.1	Germany (94%)
2	UOBSGQU2020	2020	79	57	Unpublished	Online	General population	English	34.18	8.10	45.6	68.4	USA (100%)
3	UOBSGQO2020	2020	61	59	Unpublished	Online	General population	English	32.88	13.24	64.4	64.4	UK (71 %)
4	ULBSGDG2019	2019	81	72	Unpublished	Lab	Students	German	24.40	7.05	69.4	76.4	Germany (82%)
5	ULESGDG2019	2019	290	66	Unpublished	Lab	Students	German	22.18	3.90	72.7	78.8	Germany (88 %)
6	UOBSGIU2019	2019	126	94	Unpublished	Online	General population	English	38.26	11.68	50.0	77.7	USA (93%)
7	ULESGIG2019	2019	220	71	Unpublished	Lab	Students	German	23.79	6.58	84.5	83.1	Germany (94 %)
8	ULESGIG2018_a	2018	248	118	Unpublished	Lab	Students	German	22.53	3.88	78.8	84.7	Germany (92%)
9	ULESGIG2018_b	2018	318	95	Unpublished	Lab	Students	German	22.85	3.34	80.0	87.4	Germany (96%)
10	ULESGDG2018	2018	256	74	Unpublished	Lab	Students	German	23.23	4.64	79.7	85.1	Germany (96%)
11	PLESGIG2016	2016	243	125	Published	Lab	Students	German	23.50	6.02	78.4	84.0	Germany (94 %)
12	ULBSGDG2015	2015	65	65	Unpublished	Lab	Students	German	25.14	7.93	84.6	N/A	Germany (95%)
13	PLESGDG2015	2015	205	95	Published	Lab	Students	German	23.26	4.00	77.9	89.5	Germany (96%)
14	ULESGDC2014	2014	253	65	Unpublished	Lab	Students	English	18.48	1.25	63.1	61.5	Canada (72 %)
15	PLESGDC2013	2013	150	75	Published	Lab	Students	English	22.40	5.21	65.3	40.0	Canada (49%)
16	ULBSGPU2012	2012	111	110	Unpublished	Lab	Students	English	19.25	1.58	50.0	78.2	USA (88 %)
17	PLESGPU2011	2009-2012	430	428	Published	Lab	Students	English	19.16	1.61	60.5	79.9	USA (N/A*)

Note. Study Codes were created to capture important information about the studies. The abbreviations are as follows: U/P = Unpublished/Published, L/O = Laboratory/Online, B/E = Basic Paradigm/Experimental manipulations in some conditions, SG/OG = Targets are Social Groups/Other Groups (in the present meta-analysis only studies with social groups were included), D/I/Q/P = Study was programmed in DirectRT/Inquisit/Qualtrics/Python, C/G/U/O = Data was collected in Canada/Germany/USA/Other Country, all Study Codes end with the year of data collection, if all else criteria resulted in the same Study Code “_a” or “_b” was added to distinguish the studies. N/A indicates that data on this topic was not collected and was hence not available. *Study 17 was run in the United States at the University of Colorado. Citizenship was not specifically assessed, such that precise percentages are missing.

2.5.2. *Materials*

All materials are almost identical to the materials used in Hahn et al. (2014) and were created by the first author of Hahn et al. (2014) or the first author of the current paper. All materials are openly accessible at the OSF repository (<https://osf.io/mejzp/>).

2.5.2.1. Explicit ratings. To assess explicit evaluations toward the five social group pairs, participants rated their feelings toward each group on standard thermometer scales. The scales ranged from 0 (very cold feelings) to 100 (very warm feelings) and were combined with a depiction of a thermometer colored in green or blue on one end (cold feelings) and red on the other end (warm feelings). Participants were shown each social group separately and asked to indicate how warmly or coolly they feel toward this social group. The social groups were “Black people”, “Latinos/Latinas”, “Asian people”, “White people”, “Celebrities”, “Regular people (non-celebrities)”, “Children”, and “Adults”. For better comparison to the predictions and IAT scores the final explicit rating was computed subtracting participants’ rating for the target groups from their rating for the respective comparison group. Positive scores thus indicate more positive explicit evaluations toward the social categories White, Regular, and Adult than toward Black, Latino, Asian, Celebrity, or Child ⁴.

2.5.2.2. Predictions. Participants were asked to predict how they would score on the five upcoming IATs. Before doing so, they read an introduction briefly explaining the concept of implicit evaluations and introducing the IAT as a method developed to measure such implicit evaluations. Participants received procedural details about the IAT in three of the four studies run by Hahn et al. (2014), but in none of the other studies. The introductions differed slightly between studies and the exact wordings of each study can be found on OSF. After this, participants completed – depending on the study - one or two trial predictions toward

⁴ Study 1 in Hahn et al. (2014) didn’t ask participants about “regular people” and “adults” separately, such that “White people” were always used as the comparison group. This mistake was fixed starting with Study 2.

Dogs vs. Cats and/or Insects vs. Flowers to get familiar with the prediction procedure. They went on to complete the critical predictions toward the five upcoming IAT social group pairs (Black/White, Latino/White, Asian/White, Celebrity/Regular, Child/Adult). The prediction slide was structured as follows: In the top part, all pictures that were used in the upcoming IAT were depicted. They were sorted such that all pictures of the target groups were depicted on one side, and all pictures of the comparison group on the other. There was a prompt in the center asking participants to indicate what they think their implicit attitude toward these social categories is. The bottom part showed the prediction scale (depending on the study this was a 7-point-scale or a slider) ranging from “a lot more positive toward [TARGET GROUP]” to “a lot more positive toward [COMPARISON GROUP]”. The direction of the scale was in line with the depiction of the pictures above, such that if the target groups were presented on the left-hand side positive reactions toward the target group were also indicated on the left end of the scale and vice versa.

2.5.2.3. Implicit evaluations. Implicit evaluations were assessed using evaluative Implicit Association Tests (IATs, Greenwald et al., 1998) following the procedure used in the studies by Hahn et al. (2014). The studies used different software for implementing the IATs. They were either programmed using Inquisit, DirectRT or an adapted version of a JavaScript based program in Qualtrics developed by Carpenter et al. (2019). In every study, participants completed five evaluative IATs in individually randomized order toward five different social groups with the same comparison group (non-celebrity White Adults). The IATs used the labels “Black vs. White”, “Latino vs. White”, “Asian vs. White”, “Celebrity vs. Regular”, and “Child vs. Adult”. Participants were instructed to position their fingers on a left and a right key on their keyboard (depending on the study, the left key was “A” or “E” and the right key was “5” (on the number pad) or “I”) and to sort pictures and words according to the assignments on the top of the screen. The categories “Bad” and “Good” were represented by 10 positive and negative words each. The specific words differed slightly between studies and

can be found on OSF. The social categories were represented by 10 pictures (five male, five female) per target category (Black, Latino, Asian, Celebrity, Child). The comparison category (White, Regular, Adult) was represented by 10 pictures (five male, five female) per IAT (50 different pictures in total). The pictures differed slightly between studies and can be clustered in three sets of pictures which can be found in the materials section on OSF. Which set of pictures was used per study can be found in the study overview in the supplemental materials.

Participants were instructed to respond as fast as possible while making as few mistakes as possible. If participants pressed the wrong key, they saw a red “X” and could only proceed by pressing the correct key. The latency for wrong trials was taken from trial onset until the correct response key was pressed. All IATs used a 250ms interstimulus interval. Before completing the five target IATs, every participant completed an initial 20-trial word-sorting block in which they sorted positive and negative words to the left and right side. After that, every IAT consisted of 4 blocks. Block 1 consisted of 20 trials in which participants sorted pictures of the target and the comparison group. In Block 2, participants had to sort both pictures and words for the duration of 40 trials. Block 3 consisted of 40 trials in which pictures had to be sorted on reversed sides. In Block 4 participants again were asked to sort both pictures and words while the pictures had changed sides such that pictures that had to be sorted on one side with e.g. negative words in Block 2 now had to be sorted with e.g. positive words on that side and vice versa.

Following recommendations by Greenwald et al. (2003) we calculated an IAT *D* score for each IAT by subtracting the mean reaction time for the compatible block (in which positive words are paired with the comparison groups) from the mean reaction time for the incompatible block (in which positive words are paired with the target groups) divided by their pooled standard deviation. We proceeded like this for the first half and the second half of the compatible and incompatible blocks (Blocks 2 and 4) such that we derived two *D* scores which we averaged to obtain a final *D* score. Higher *D* scores indicate faster reactions when

positive words were paired with the comparison groups (White, Regular, or Adult) and negative words were paired with the respective target group (Black, Asian, Latino, Celebrity, or Child).

2.5.3. Procedure

All included studies followed the basic prediction paradigm introduced by Hahn et al. (2014). Participants first completed the thermometer ratings which were block-randomized to avoid confusion. That is, participants completed three blocks of thermometer ratings in randomized order: (1) ethnic groups in randomized order, (2) celebrities followed by regular people, and (3) children followed by adults. Next, participants completed the prediction procedure with the prediction slides presented in random order for each participant. Lastly, participants completed the five IATs toward the social group pairs in randomized order. Participants resumed by answering demographic questions.

2.5.4. Analyses

2.5.4.1. Prediction Accuracy. To calculate how accurately participants predicted the patterns of their IAT results in each study, we ran a multi-level model predicting participants' person-standardized IAT scores from their person-standardized predictions on level 1. The random slopes in this analysis are equivalent to a correlation coefficient per participant between their IAT scores and predictions. To examine the mean prediction accuracy in each sample we examine the slope on level-2 (fixed effect) which equals the average size of the random slopes. To estimate the average effect size of the prediction accuracy across studies, we ran a random-effects model weighing the estimates of the fixed effects with the inverse-variance method.

2.5.4.2. Implicit-Explicit Relationship. To analyze how strongly participants' explicit evaluations were related to their implicit evaluations, we ran a multi-level model per study regressing person-standardized IAT scores onto person-standardized thermometer difference scores on level-1 and examined the level-2 fixed effect. To further investigate

whether the relation between explicit and implicit evaluations could be explained by participants' predictions, we regressed participants person-standardized IAT scores simultaneously onto participants' person-standardized predictions and thermometer difference scores on level-1 and examined the level-2 fixed effects. We meta-analyzed the fixed effect estimates from both analyses (fixed effects for thermometer ratings and predictions) in a random-effects model using inverse-variance weighting to investigate their average sizes across studies.

2.5.4.3. Simultaneously Predicting the Predictions from IAT Scores and Explicit Ratings.

We preregistered an additional analysis that is not explained above that Hahn et al. (2014) only conducted in their Study 4. This analysis looks at the unique within-subjects relationships between participants' predictions and their explicit and implicit evaluations, controlling for the respective other type of evaluation. The purpose was to see to what degree participants' predictions were based on the same information that went into their explicit evaluations beyond the spontaneous reactions reflect on implicit evaluations; and vice versa, to what degree their predictions uniquely incorporated the spontaneous reactions reflected on implicit measures in ways that is not reflected on explicit measures. To this end, we regressed participants' person-standardized predictions simultaneously onto their person-standardized IAT scores and thermometer difference scores on level-1 and examined the fixed effects on level-2. We meta-analyzed the two fixed effects from this analysis (fixed effect for IAT scores and thermometer ratings) with two random-effects models using inverse-variance weighting to investigate their average effect sizes across studies.

2.5.4.4. Prediction Accuracy Beyond Normative Prediction Patterns That Are Shared with Other Participants.

To investigate whether participants' predictions are related to their patterns of IAT results beyond a normative tendency, we adopted the analytic approach described earlier by Hahn et al. (2014) in which they predicted participants' IAT

scores from another persons' predictions in the same sample. In their study, they paired one person with one other person from the same sample. To ensure that the obtained results are not bound to the specific random pairing for this one model, we iterated this procedure 1000 times (see also Rahmani Azad et al., 2022 for similar analyses). Specifically, we ran a multi-level model in which every participant's person-standardized IAT scores were predicted by another participant's person-standardized predictions on level 1 and examined the level-2 fixed effect indicating how accurately on average another person in the sample predicted the participants' pattern of IAT scores. We repeated this analysis process 1000 times such that every participant's IAT scores were predicted 1000 times by another person's predictions and averaged the 1000 fixed effects. In a second step, we took the same 1000 random pairings and entered them into a multi-level model in which participants' IAT scores were predicted simultaneously by both their own predictions and the random other persons' predictions. We again averaged the fixed effects on level 2 to estimate whether overall participants' own predictions explained variance in their patterns of IAT scores over and above the other persons' predictions.

To estimate the meta-analytical effect of the described analyses across studies, we ran three random effects models on the obtained averaged fixed effects from both analyses weighing them using the inverse-variance method.⁵

2.5.4.5. Between-Subjects Analysis. To examine the degree to which participants knew how much bias they would show in comparison to others, we also examined between-subjects correlations per social category. To this end, we standardized predictions and IAT scores by social category and ran a multilevel model predicting IAT scores from predictions by IAT type. We examine the fixed effect which is equivalent to the average correlation

⁵ To our knowledge, there are no conventions on how to run a meta-analysis on effects derived from bootstrapping analyses. We thus decided to apply the same method to the average effects of the bootstrapping analyses as we used for the effects from the standard multi-level analyses.

between predictions and IAT scores across IAT types. To estimate the meta-analytical effect of this analysis we ran a random-effects model weighting the fixed effects derived from this analysis using the inverse-variance method.

2.6. Results

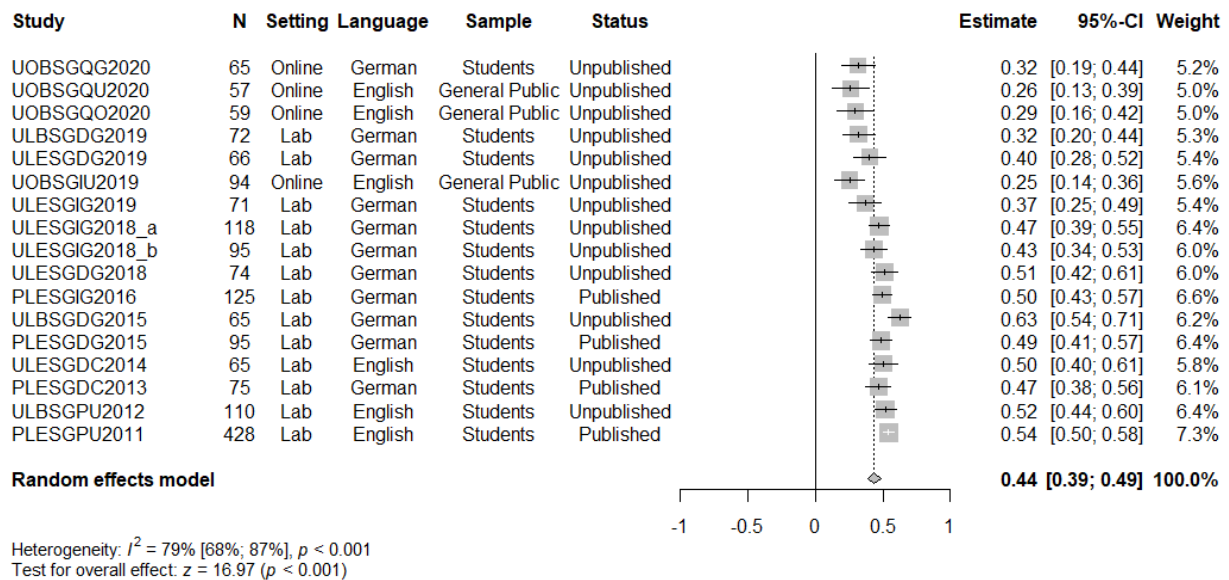
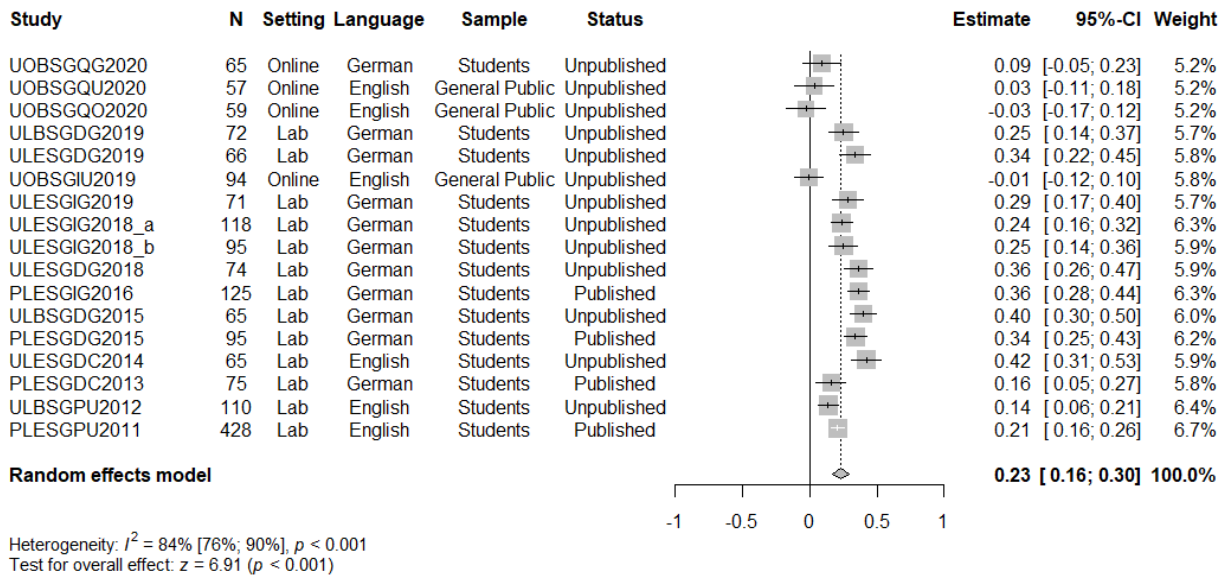
2.6.1. Prediction Accuracy

In all 17 studies participants were able to accurately predict the pattern of their IAT results with effects ranging from a minimum of $b = 0.26$, 95% CI [0.13, 0.39], $t(56) = 3.87$, $p < .001$ up to a maximum of $b = 0.63$, 95% CI [0.54, 0.71], $t(324) = 14.41$, $p < .001$. The meta-analytical effect in a random-effects model was $b = 0.44$, 95% CI [0.39, 0.49] and was significantly different from zero, $Z = 16.97$, $p < .001$ (see Figure 1). Effect sizes seemed to vary systematically between studies as indicated by the significant Cochran's Q statistic for heterogeneity, $Q(16) = 77.35$, $p < .001$ with $I^2 = 79\%$, 95% CI [68%, 87%].

To examine this heterogeneity in effect sizes, we ran subgroup analyses for the study setting (Online vs. Lab), the sample (General Public vs. Students), the study language (English vs. German), and the current status of publication (Published vs. Unpublished). The meta-analytical effects for each subgroup can be found in Figure 2.

Figure 1

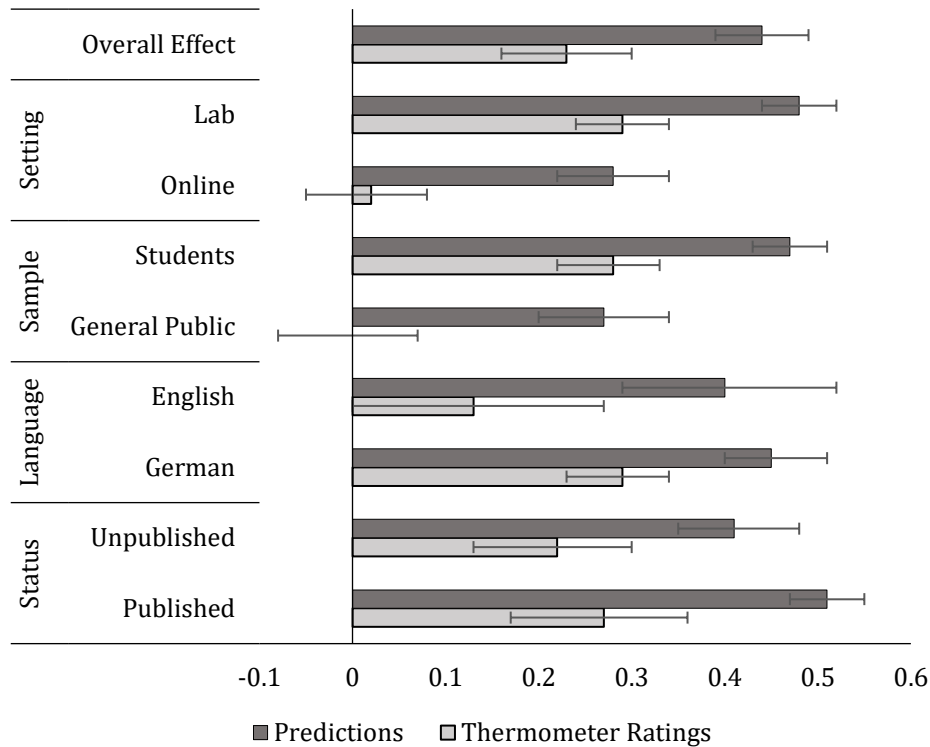
Overview of Fixed-effect Estimates Across Studies and the Meta-analytical Effect Based on Multi-level Models Separately Predicting IAT Scores From IAT Score Predictions or Thermometer Ratings

a. Predictions**b. Thermometer Ratings**

Note. Estimates in each study are calculated on standardized scores within-subjects, once per participant, aggregated across participants in a multi-level analysis regressing IAT Scores on IAT Score predictions (Panel A), and IAT scores on thermometer ratings (Panel B). The meta-analytical effect weighs the estimates of the fixed effects with the inverse-variance method. Note that the confidence intervals in these figures may differ slightly from those reported in the multi-level analyses because in the meta-analysis confidence intervals are calculated using degrees of freedom based on the sample size while in the multi-level model confidence intervals were based on the satterthwaite approximation of degrees of freedom.

Figure 2

Overview of Meta-analytical Effects by Subgroups Based on Multi-level Models Separately Predicting IAT Scores From IAT Score Predictions or Thermometer Ratings



Note. The effects are based on two separate multi-level models per study predicting IAT scores from IAT score predictions or thermometer ratings. All scores are standardized within-subjects per participant and aggregated across participants in the multi-level analysis. The resulting fixed effect was imputed in a meta-analysis using the inverse-variance method for weighing.

Results showed that the prediction accuracy effect was significantly higher in studies that were conducted in the laboratory ($b = 0.48$, 95% CI [0.44, 0.52]) than in studies that were conducted online ($b = 0.28$, 95% CI [0.22, 0.34]), $Q(1) = 29.45$, $p < .001$. Further, studies with student samples showed higher prediction accuracy effects ($b = 0.47$, 95% CI [0.43, 0.51]) than studies conducted on the general public ($b = 0.27$, 95% CI [0.20, 0.34]), $Q(1) = 23.46$, $p < .001$. Though differences were less pronounced for the publication status of the studies, prediction accuracy effects were larger for published studies ($b = 0.51$, 95% CI [0.47, 0.55]) than for unpublished studies ($b = 0.41$, 95% CI [0.35, 0.48]), $Q(1) = 6.57$, $p = .010$. Effects did not significantly differ for studies that were conducted in English ($b = 0.40$, 95%

CI [0.29, 0.52]) as opposed to studies that were conducted in German ($b = 0.45$, 95% CI [0.40, 0.51]), $Q(1) = 0.62$, $p = .430$.

2.6.2. *Implicit-Explicit Relationship*

Thermometer ratings were inconsistently related to the pattern of IAT results with effect sizes ranging from $b = -0.03$, 95% CI [-0.17, 0.12], $t(58) = -0.39$, $p = .697$ to $b = 0.42$, 95% CI [0.31, 0.53], $t(64) = 7.73$, $p < .001$. The meta-analytical effect in a random-effects model was $b = 0.23$, 95% CI [0.16, 0.30] and was significantly different from zero, $Z = 6.91$, $p < .001$ (see Figure 1). Effect sizes varied systematically between studies as indicated by the significant Cochran's Q statistic for heterogeneity, $Q(16) = 100.71$, $p < .001$ with $I^2 = 84\%$, 95% CI [76%, 90%].

Subgroup analyses showed that thermometer ratings were more strongly related to IAT patterns in laboratory settings ($b = 0.23$, 95% CI [0.16, 0.30]) than in online settings ($b = 0.02$, 95% CI [-0.05, 0.08]), $Q(1) = 40.71$, $p < .001$. Effects were also stronger for student samples ($b = 0.28$, 95% CI [0.22, 0.33]) than for the general public ($b = 0.00$, 95% CI [-0.08, 0.07]), $Q(1) = 36.45$, $p < .001$. Studies conducted in German did also show stronger effects ($b = 0.29$, 95% CI [0.23, 0.34]) than studies conducted in English ($b = 0.13$, 95% CI [0.00, 0.27]), $Q(1) = 4.42$, $p = .036$. Effects did not significantly differ between published ($b = 0.27$, 95% CI [0.17, 0.36]) and unpublished studies ($b = 0.22$, 95% CI [0.13, 0.30]), $Q(1) = 0.65$, $p = .419$.

2.6.3. *Predictions vs. Explicit Ratings*

In all 17 studies, predictions were more strongly related to participants' pattern of IAT results than thermometer ratings. A pattern that was even more strongly pronounced in the simultaneous model predicting IAT score patterns from predictions and thermometer ratings. While in the simultaneous model predictions remained a significant predictor in all 17 studies (Effects ranged from $b = 0.27$, 95% CI [0.14, 0.41], $t(86) = 4.06$, $p < .001$ to $b = 0.58$, 95% CI [0.48, 0.68], $t(318) = 11.33$, $p < .001$) thermometer ratings only remained significantly

(positively) related to IAT score patterns in two studies (Effects ranged from $b = -0.19$, 95% CI [-0.35, -0.03], $t(62) = -2.39$, $p < .020$ to $b = 0.16$, 95% CI [0.02, 0.29], $t(115) = 2.33$, $p < .022$; for an overview of all effects see Table 2).

Table 2.

Prediction Accuracy and Relationship Between Thermometer Ratings and IAT Scores for Each Study in Simple Regression Models and a Simultaneous Regression Model

Study ID	Study	Predictor	Prediction model estimates	Implicit-explicit model estimates	Simultaneous model estimates
1	UOBSGQG2020	IAT score predictions	.35***		.45***
		Explicit ratings		.11	-.17*
2	UOBSGQU2020	IAT score predictions	.26***		.35***
		Explicit ratings		.03	-.14
3	UOBSGQO2020	IAT score predictions	.29***		.41***
		Explicit ratings		-.03	-.19*
4	ULBSGDG2019	IAT score predictions	.32***		.27***
		Explicit ratings		.25***	.11
5	ULESGDG2019	IAT score predictions	.40***		.32***
		Explicit ratings		.34***	.13
6	UOBSGIU2019	IAT score predictions	.25***		.38***
		Explicit ratings		-.01	-.21**
7	ULESGIG2019	IAT score predictions	.47***		.51***
		Explicit ratings		.24***	-.05
8	ULESGIG2018_a	IAT score predictions	.43***		.46***
		Explicit ratings		.25***	-.04
9	ULESGIG2018_b	IAT score predictions	.51***		.49***
		Explicit ratings		.36***	.07
10	ULESGDG2018	IAT score predictions	.50***		.45***
		Explicit ratings		.36***	.10*
11	PLESGIG2016	IAT score predictions	.63***		.58***
		Explicit ratings		.40***	.10
12	ULBSGDG2015	IAT score predictions	.49***		.45***
		Explicit ratings		.34***	.06
13	PLESGDG2015	IAT score predictions	.47***		.51***
		Explicit ratings		.16**	-.07
14	ULESGDC2014	IAT score predictions	.37***		.33***
		Explicit ratings		.29***	.08
15	PLESGDC2013	IAT score predictions	.50***		.41***
		Explicit ratings		.42***	.16*
16	ULBSGPU2012	IAT score predictions	.52***		.53***
		Explicit ratings		.14***	-.02
17	PLESGPU2011	IAT score predictions	.54***		.56***
		Explicit ratings		.21***	-.02

Note. Relationships are calculated on standardized scores within-subjects, once per participant, and then aggregated across participants in a multi-level analysis.

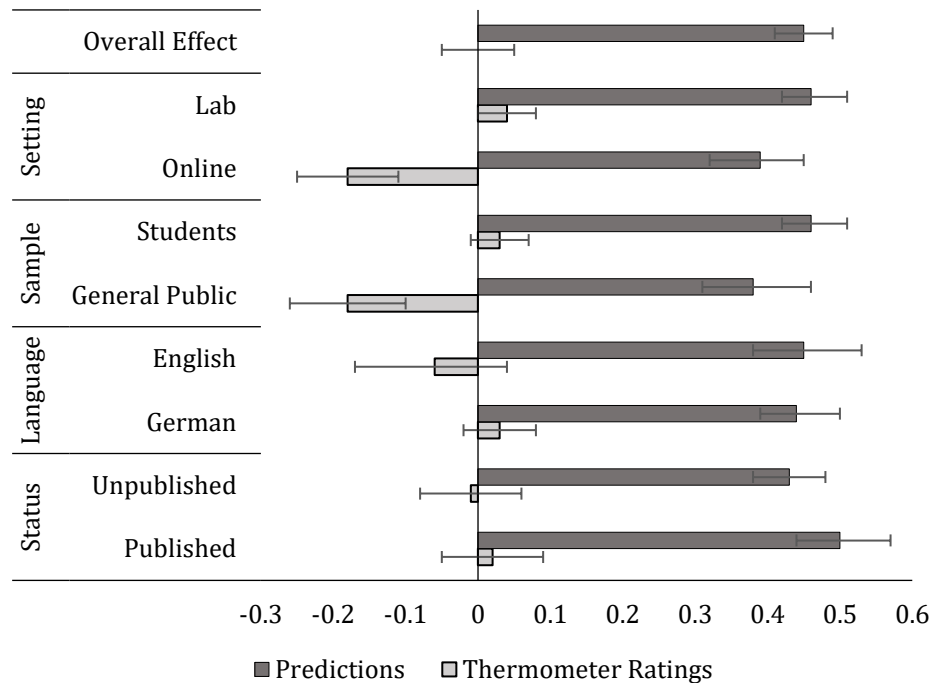
* indicates significance at the $p < .05$ level, ** at the $p < .01$ level, and *** at the $p < .001$ level.

The meta-analyses on both effects in the simultaneous models corroborated these results and showed that the meta-analytical effect of the prediction accuracy remained significant, $b = 0.45$, 95% CI [0.41, 0.49], $Z = 21.02$, $p < .001$, while the meta-analytical effect of the thermometer ratings did not significantly differ from zero, $b = 0.00$, 95% CI [-0.05, 0.05], $Z = -0.10$, $p = .924$. Both effects showed significant Cochran's Q statistics for heterogeneity, $Q_{\text{predictions}}(16) = 49.06$, $p < .001$, $I^2_{\text{Predictions}} = 67\%$, 95% CI [46%, 80%], $Q_{\text{Thermometer}}(16) = 53.80$, $p < .001$, $I^2_{\text{Thermometer}} = 70\%$, 95% CI [51%, 82%].

Interestingly, no subgroup analyses for the prediction accuracy-beyond-explicit-ratings effect in the simultaneous model revealed significant differences between groups (Setting: $Q(1) = 3.23$, $p = .072$; Sample: $Q(1) = 3.20$, $p = .074$; Language: $Q(1) = 0.05$, $p = .827$; Publication status: $Q(1) = 3.41$, $p = .065$). In contrast, the meta-analytical effect for thermometer ratings beyond predictions differed in the subgroup analyses depending on the study setting ($Q(1) = 27.47$, $p < .001$), and sample ($Q(1) = 20.75$, $p < .001$). Thermometer ratings were negative unique predictors of IAT score patterns when studies were conducted online ($b = -0.18$, 95% CI [-0.25, 0.11]), or on the general public ($b = -0.18$, 95% CI [-0.26, 0.10]). In contrast, there was simply no (negative or positive) effect of Thermometer ratings beyond predictions in the lab ($b = 0.04$, 95% CI [-0.00, 0.08]) and student ($b = 0.03$, 95% CI [-0.01, 0.07]) samples. The thermometer effects did not differ for the subgroups on language ($Q(1) = 2.63$, $p = .105$), and publication status ($Q(1) = 0.29$, $p = .591$). That is, while raw prediction accuracy was lower in online samples than in lab samples (and general-population as opposed to student samples), this difference disappeared when controlling for explicit ratings. In other words explicit ratings showed a suppression effect on prediction accuracy in the online and general-public samples, but not the lab and student samples. Once controlling for this suppression effect, prediction accuracy did not differ between the lab and online settings. An overview of all meta-analytical effects of the prediction accuracy and thermometer ratings for all subgroups can be found in Figure 3.

Figure 3

Overview of Meta-analytical Effects by Subgroups Based on Multi-level Models Simultaneously Predicting IAT Scores From IAT Score Predictions and Thermometer Ratings



Note. The effects are based on a multi-level analysis per study in which IAT scores were simultaneously predicted from IAT score predictions and thermometer ratings. All scores were standardized within-subjects per participant and aggregated across participants in the multi-level analysis. The resulting fixed effects were imputed in a meta-analysis using the inverse-variance method for weighing.

2.6.4. Simultaneously Predicting the Predictions from IAT Scores and Explicit Ratings

In all 17 studies, predictions were significantly predicted by both, participants' IAT score patterns and patterns of thermometer ratings in a simultaneous model. The effects for IAT scores ranged from $b = 0.21$, 95% CI [0.10, 0.32], $t(71) = 3.82$, $p < .001$ to $b = 0.49$, 95% CI [0.40, 0.58], $t(72) = 11.24$, $p < .001$, with a meta-analytical effect of $b = 0.34$, 95% CI [0.30, 0.38], $Z = 15.24$, $p < .001$. The effects for thermometer ratings ranged from $b = 0.16$, 95% CI [0.02, 0.29], $t(115) = 2.33$, $p = .022$ to $b = 0.57$, 95% CI [0.45, 0.68], $t(60) = 9.61$, $p < .001$, with a meta-analytical effect of $b = 0.43$, 95% CI [0.37, 0.49], $Z = 15.15$, $p < .001$. Both effects showed significant Cochran's Q statistics for heterogeneity, $Q_{\text{IAT}}(16) = 87.87$, $p < .001$, $I^2_{\text{IAT}} = 82\%$, 95% CI [72%, 88%], $Q_{\text{Thermometer}}(16) = 110.60$, $p < .001$, $I^2_{\text{Thermometer}} = 86\%$,

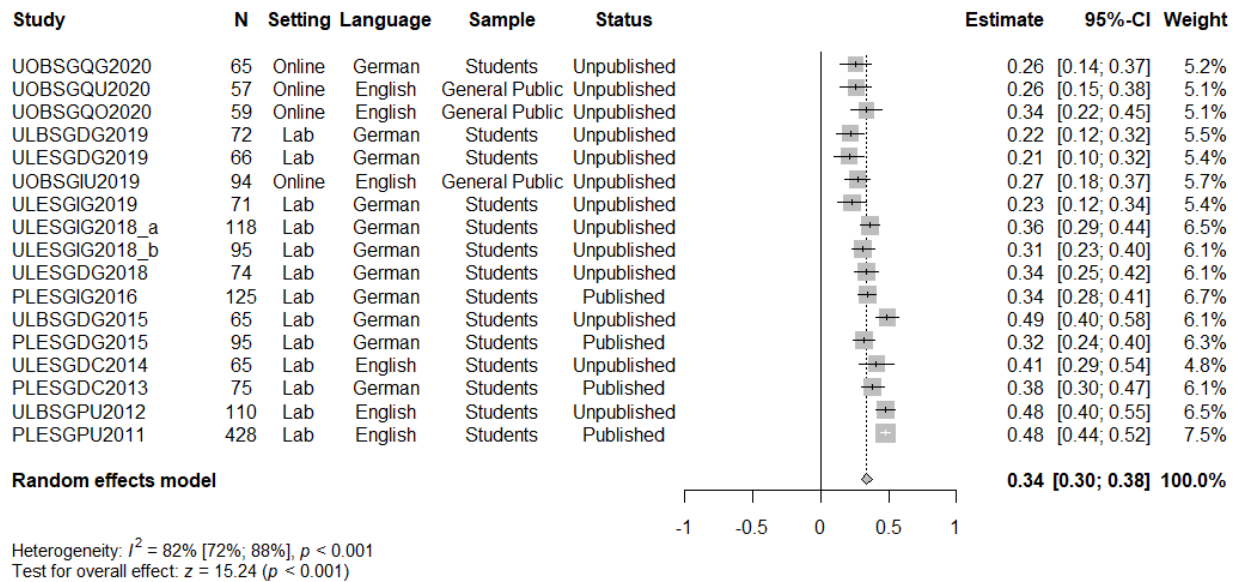
95% CI [78%, 90%]. An overview of the fixed-effect estimates across studies and the meta-analytical effects can be found in Figure 4.

Subgroup analyses showed that prediction patterns were less strongly related to participants' IAT score patterns when studies were conducted online ($b = 0.28$, 95% CI [0.23, 0.33]), than when they were conducted in the laboratory ($b = 0.36$, 95% CI [0.30, 0.41]), $Q(1) = 4.03$, $p = .045$. All other subgroup analyses did not show significant differences between groups (Sample: $Q(1) = 2.43$, $p = .119$; Language: $Q(1) = 1.51$, $p = .220$; Publication status: $Q(1) = 1.69$, $p = .194$). Thermometer ratings were less strongly related to participants' prediction patterns in studies conducted in the laboratory ($b = 0.41$, 95% CI [0.35, 0.48]) than in studies conducted online ($b = 0.51$, 95% CI [0.45, 0.57]), $Q(1) = 4.21$, $p = 0.40$. This result mirrors the suppression effect above in that it shows more consistency between IAT score predictions and thermometer ratings in online as opposed to lab settings. Effects were also smaller when studies were conducted in English ($b = 0.35$, 95% CI [0.23, 0.46]) than when they were conducted in German ($b = 0.48$, 95% CI [0.44, 0.52]), $Q(1) = 4.41$, $p = .036$. Subgroup analyses did not show significant differences for the type of sample ($Q(1) = 1.61$, $p = .205$) or the publication status ($Q(1) = 0.11$, $p = .739$). All meta-analytical effects can be found in Figure 5.

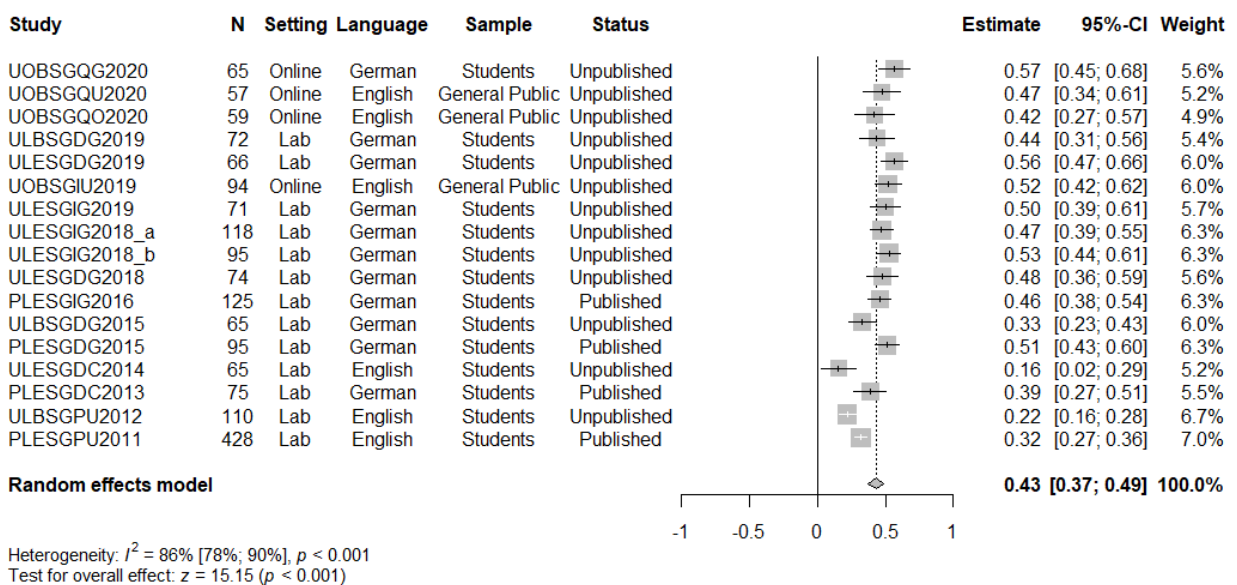
Figure 4

Overview of Fixed-effect Estimates Across Studies and the Meta-analytical Effect Based on a Multi-level Model Predicting IAT Score Predictions From IAT Scores and Thermometer Ratings

a. IAT Scores Predicting IAT Score Predictions Beyond Thermometer Ratings



b. Thermometer ratings Predicting IAT Score Predictions Beyond IAT Scores

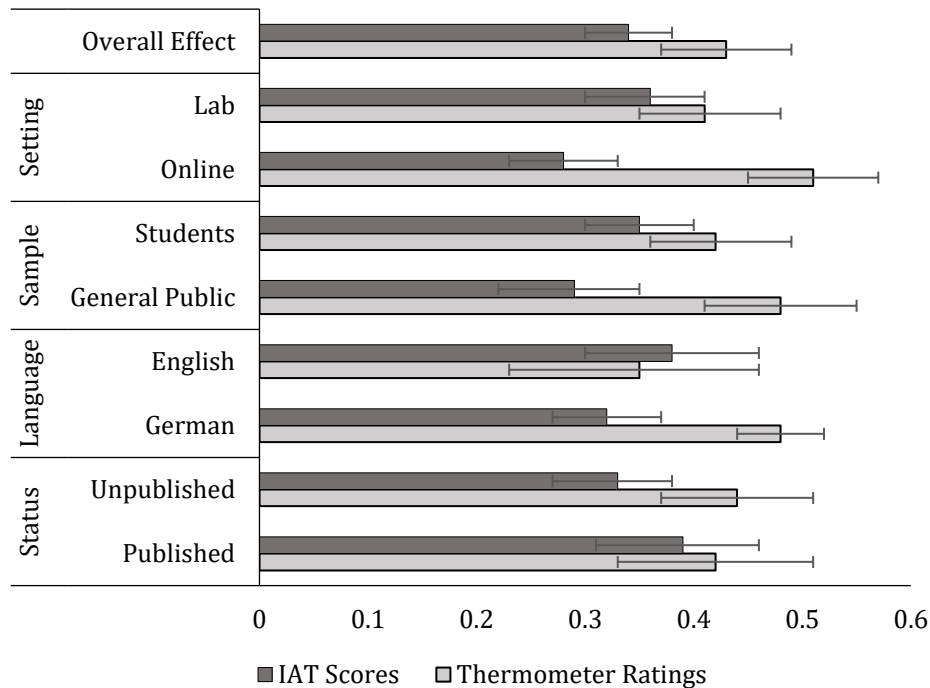


Note. Estimates in each study are calculated on standardized scores within-subjects, once per participant, aggregated across participants in a multi-level analysis regressing IAT Scores on IAT Score predictions (Panel A), and IAT scores on thermometer ratings (Panel B). The meta-analytical effect weighs the estimates of the fixed effects with the inverse-variance method.⁶

⁶ Note that the confidence intervals in these figures may differ slightly from those reported in the multi-level analyses because in the meta-analysis confidence intervals are calculated using degrees of freedom based on the sample size while in the multi-level model confidence intervals were based on the satterthwaite approximation of degrees of freedom.

Figure 5

Overview of Meta-analytical Effects by Subgroups Based on Multi-level Models Simultaneously Predicting IAT Score Predictions From IAT Scores and Thermometer Ratings



Note. The effects are based on a multi-level analysis per study in which IAT score predictions were simultaneously predicted from IAT scores and thermometer ratings. All scores were standardized within-subjects per participant and aggregated across participants in the multi-level analysis. The resulting fixed effects were imputed in a meta-analysis using the inverse-variance method for weighing.

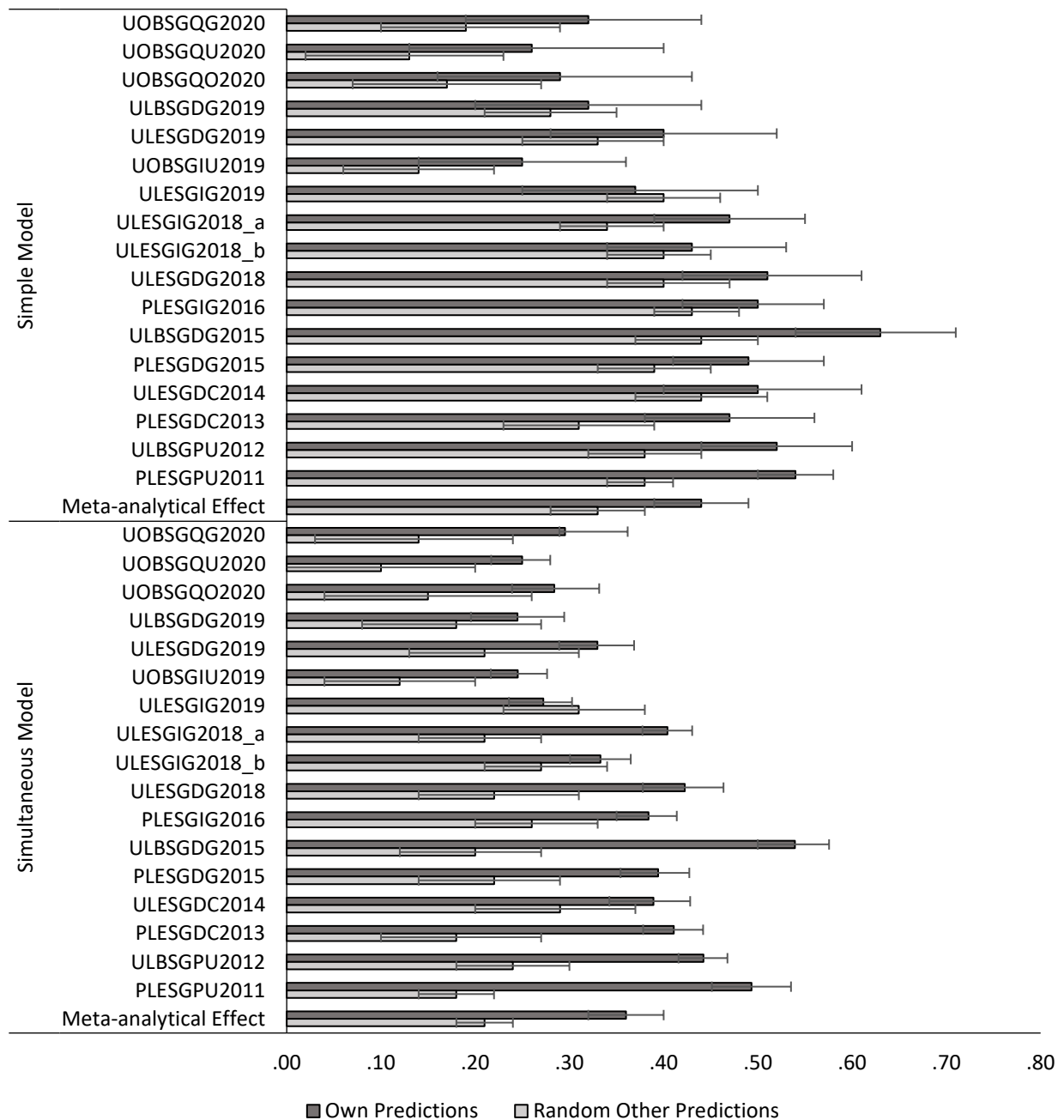
2.6.5. Prediction Accuracy Beyond Normative Patterns Based on Other Participants'

Predictions.

The average prediction accuracy of the randomly paired other participants ranged from $b = 0.08$, 95% CI [0.02, 0.24] to $b = 0.44$, 95% CI [0.38, 0.50] indicating that in all 17 studies the randomly paired other participants' prediction patterns were related to participants' own pattern of IAT results above zero on average. In 16 out of these 17 studies, participants' own predictions descriptively showed higher accuracies than the random other participants' predictions. This difference was significant in 12 out of the 17 studies as indicated by the 95% CI of the random other prediction accuracies that did not include the average participants' own prediction accuracy in these studies (see Figure 6).

Figure 6

Overview of Fixed-effect Estimates Across Studies and the Meta-analytical Effects Based on Multi-level Models Separately or Simultaneously Predicting Participants' Own IAT Scores From Their Own IAT Score Prediction or From Randomly Paired Other Participants' Predictions



Note. The effects of participants' own predictions are based on a multi-level analysis per study in which participants' IAT scores were predicted from their own IAT score predictions and are the same as the effects reported in the prediction accuracy subsection. The effects for random other participants' predictions are the average of 1000 iterations of a multi-level analysis predicting participants' IAT scores from a randomly paired other participants' predictions. All scores were standardized within-subjects per participant and aggregated across participants in the multi-level analysis. The resulting fixed effects were imputed in a meta-analysis using the inverse-variance method for weighing.

The meta-analytical effect corroborated this finding and showed that across all studies, participants' own prediction patterns were significantly more related to their pattern of IAT scores ($b = 0.44$, 95% CI [0.39, 0.49], $Z = 17.38$, $p < .001$) than the average other participants' prediction patterns ($b = 0.33$, 95% CI [0.28, 0.38], $Z = 13.51$, $p < .001$).

These effects replicated when running the simultaneous model in which we predicted participants' pattern of IAT scores simultaneously from their own prediction patterns and the randomly paired others' prediction patterns. In 16 out of the 17 studies the randomly paired other participants' prediction patterns explained variance in participants own IAT score patterns above participants' own prediction patterns (effects ranged from $b = 0.10$, 95% CI [-0.01, 0.21] to $b = 0.31$, 95% CI [0.23, 0.38]). However, again, in 16 out of the 17 studies participants' own prediction accuracies descriptively outperformed the random other participants' prediction accuracies (effects ranged from $b = 0.24$, 95% CI [0.20, 0.29] to $b = 0.54$, 95% CI [0.50, 0.58]). In line with this, the meta-analytical effects again supported this pattern of findings, showing that across all studies, participants' own prediction patterns significantly explained variance in their own pattern of IAT results ($b = 0.36$, 95% CI [0.32, 0.40], $Z = 16.85$, $p < .001$) over and above the randomly paired others' prediction patterns ($b = 0.21$, 95% CI [0.18, 0.24], $Z = 15.95$, $p < .001$).

Subgroup analyses showed that the average random other participants' prediction accuracies in the simple model were higher when the studies were conducted in the laboratory ($b = 0.38$, 95% CI [0.36, 0.41]) than when the studies were conducted online ($b = 0.16$, 95% CI [0.11, 0.20]), $Q(1) = 69.60$, $p < .001$, and higher for student samples ($b = 0.37$, 95% CI [0.34, 0.40]) than for the general public ($b = 0.15$, 95% CI [0.09, 0.20]), $Q(1) = 49.80$, $p < .001$. There were no significant differences for the studies' language ($Q(1) = 1.85$, $p = .173$) or publication status ($Q(1) = 3.46$, $p = .063$). This pattern of results replicated in the simultaneous model. Effects were larger for studies conducted in the laboratory ($b = 0.23$, 95% CI [0.20, 0.25]) than when the studies were conducted online ($b = 0.13$, 95% CI [0.08,

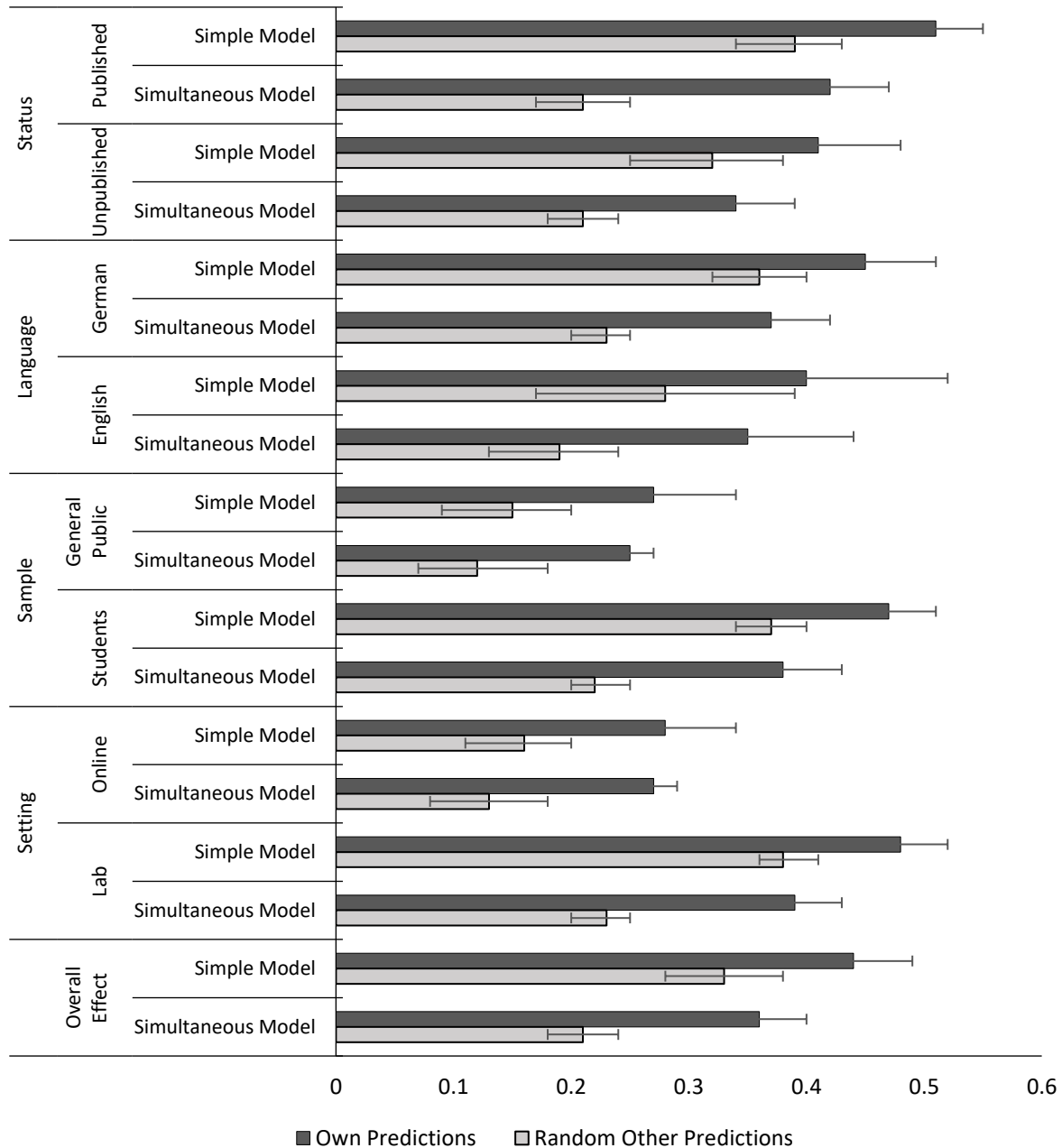
0.18]), $Q(1) = 13.02, p < .001$, and higher for student samples ($b = 0.22, 95\% \text{ CI } [0.20, 0.25]$) than for the general public ($b = 0.12, 95\% \text{ CI } [0.07, 0.18]$), $Q(1) = 10.69, p < .001$. Effects did not differ significantly depending on the studies' language ($Q(1) = 1.64, p = .200$) or the studies' status ($Q(1) = 0.01, p = .921$). The effect of participants' own prediction patterns on their IAT score pattern in the simultaneous model was also larger in laboratory studies ($b = 0.39, 95\% \text{ CI } [0.35, 0.43]$) than in online studies ($b = 0.27, 95\% \text{ CI } [0.24, 0.29]$), $Q(1) = 23.28, p < .001$, and larger for student samples ($b = 0.38, 95\% \text{ CI } [0.34, 0.43]$) than for the general public ($b = 0.25, 95\% \text{ CI } [0.23, 0.27]$), $Q(1) = 28.96, p < .001$. This effect was also larger for published studies ($b = 0.42, 95\% \text{ CI } [0.37, 0.47]$) than for unpublished studies ($b = 0.34, 95\% \text{ CI } [0.29, 0.39]$), $Q(1) = 4.84, p = .028$, but did not differ significantly depending on the studies' language, $Q(1) = 0.11, p = .744$. Overall, across all subgroups both participants' own predictions and the random other participants' predictions predicted participants' own IAT score patterns but participants' own predictions were consistently more strongly related to their own patterns of IAT results than a randomly-paired other participant-s predictions (see Figure 7).

2.6.6. Between-Subjects Analysis

The average between-subject correlations ranged from $b = 0.08, 95\% \text{ CI } [-0.03, 0.25]$, $t(4) = 1.39, p = .238$ to $b = 0.34, 95\% \text{ CI } [0.14, 0.54]$, $t(4) = 4.65, p = .010$, with a meta-analytical effect of $b = 0.22, 95\% \text{ CI } [0.19, 0.26]$, $Z = 12.06, p < .001$. An overview of the between-subject effects across studies and the meta-analytical between-subject effect can be found in Figure 8.

Figure 7

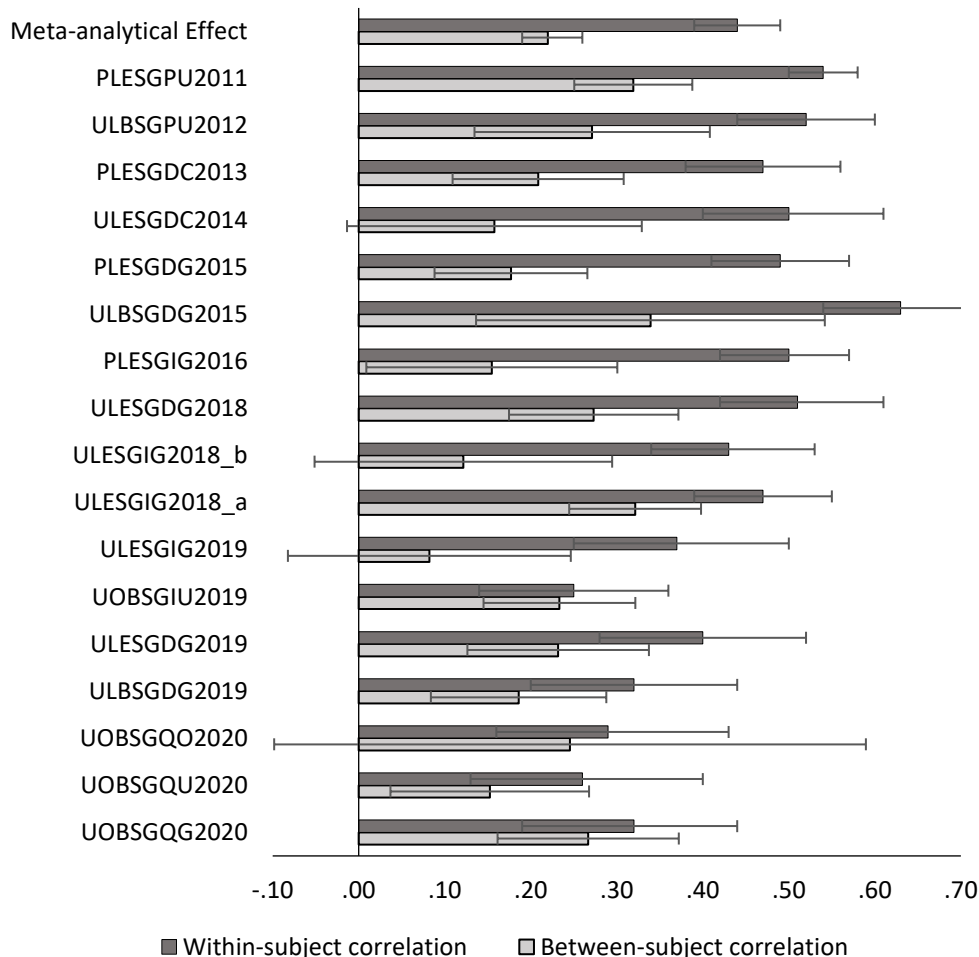
Overview of Meta-analytical Effects by Subgroups Based on Multi-level Models Separately or Simultaneously Predicting Participants' Own IAT Scores From Their Own IAT Score Predictions or From Randomly Paired Other Participants' Predictions



Note. The simple model effects are based on of based on a multi-level analysis per study in which participants' IAT scores were separately predicted from their own IAT score predictions or 1000 iterations of randomly paired other participants' predictions. In the simultaneous model effects represent the average of 1000 iterations of a multi-level analysis simultaneously predicting participants' IAT scores from their own predictions and a randomly paired other participants' prediction. All scores were standardized within-subjects per participant and aggregated across participants in the multi-level analyses. The resulting average effects were imputed in a meta-analysis using the inverse-variance method for weighing.

Figure 8

Overview of Fixed-effect Estimates Across Studies and the Meta-analytical Effects Based on Within-subject or Between-subject Correlations Between IAT Score Predictions and IAT Scores



Note. The within-subject effects are based on multi-level analysis predicting IAT scores from IAT score predictions. In this analysis all scores were standardized within-subjects per participant and aggregated across participants in the multi-level analysis. The between-subject effects are based on a multi-level analysis predicting IAT scores from IAT score predictions with scores standardized between-subjects per target group aggregated across target groups in the multi-level analysis. The resulting fixed effects were imputed in a meta-analysis using the inverse-variance method for weighing.

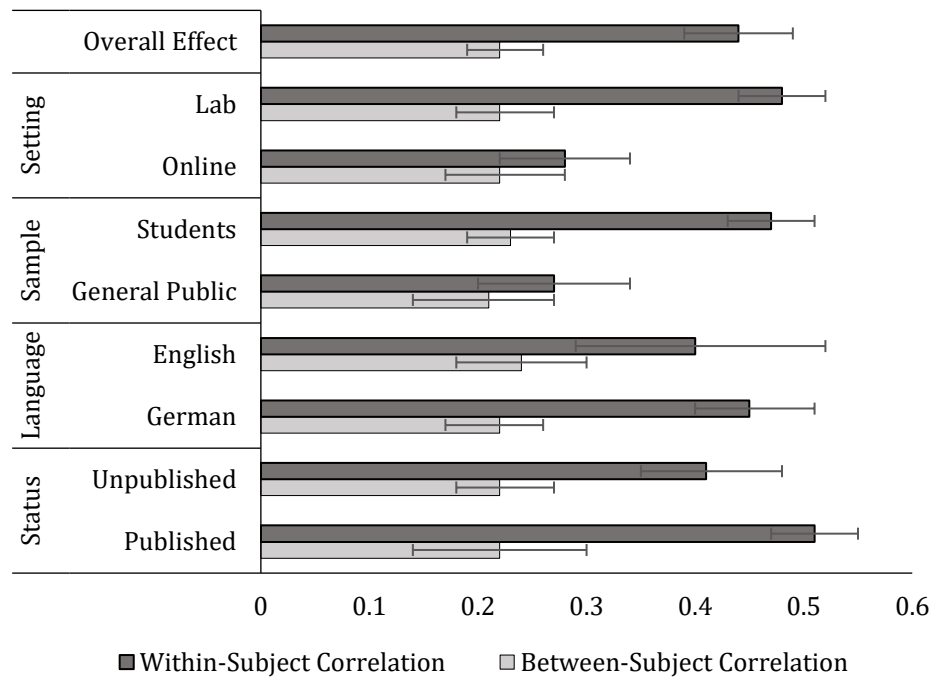
Effects varied substantially between studies as indicated by a significant Cochran's Q statistics for heterogeneity, $Q_{IAT}(16) = 38.34$, $p = .001$, $I^2_{IAT} = 58\%$, 95% CI [29%, 76%].

Despite this heterogeneity in effect sizes, none of the subgroup analyses showed systematic differences between the defined groups (Setting: $Q(1) = 0.00$, $p = .999$; Sample: $Q(1) = 0.27$,

$p = .604$; Language: $Q(1) = 0.39, p = .533$; Publication status: $Q(1) = 0.00, p = .978$). The meta-analytical between-subject effects for each subgroup can be found in Figure 9.

Figure 9

Overview of Meta-analytical Effects by Subgroups Based on Within-subject or Between-subject Correlations Between IAT Score Predictions and IAT Scores



Note. The within-subject effects are based on multi-level analysis predicting IAT scores from IAT score predictions. In this analysis all scores were standardized within-subjects per participant and aggregated across participants in the multi-level analysis. The between-subject effects are based on a multi-level analysis predicting IAT scores from IAT score predictions with scores standardized between-subjects per target group aggregated across target groups in the multi-level analysis. The resulting fixed effects were imputed in a meta-analysis using the inverse-variance method for weighing.

2.7. Discussion

The main goal of the present meta-analysis was to examine whether the findings by Hahn et al. (2014) that people are able to accurately predict the patterns of their IAT results would replicate across different samples and settings and to provide a more accurate estimate of the average effect size of the prediction accuracy. To this end, we reanalyzed 17 published and unpublished studies that followed the original prediction paradigm closely. Results replicated Hahn et al.'s (2014) findings and showed that in all 17 studies, participants were

able to accurately predict the patterns of their IAT results with an average within-subject correlation across studies of $b = 0.44$. This finding further strengthens Hahn et al.'s (2014) claim that the cognitions reflected on implicit evaluations are consciously accessible and reportable. Further, in all studies predictions were more strongly related to IAT score patterns than explicit thermometer ratings with an average within-subject correlation between thermometer ratings and IAT scores of $b = 0.23$. This highlights that people are willing and able to report on the cognitions reflected on their implicit evaluations even though they may report different evaluations on traditional explicit measures. Finally, predictions remained a significant predictor explaining variance in IAT score patterns over and above thermometer ratings in a simultaneous model with an average meta-analytical effect size of $b = 0.45$. At the same time, thermometer ratings only remained a significant (positive) predictor of IAT score patterns in 1 out of the 17 studies and the meta-analytical effect of thermometer ratings in this model dropped to $b = 0.00$. This finding is in line with theorizing by dual-process models such as the APE (Gawronski & Bodenhausen, 2006) or the MODE model (Fazio, 1990, 2007), which propose that people may be well aware of their automatic cognitions and these may partially inform their explicit evaluations, but additional (propositional) information is considered for their final deliberate answer.

An often-voiced concern with the original findings is whether participants are indeed introspectively aware of their biases or whether they merely infer their own pattern of biases from normatively expected patterns (Morris & Kurdi, 2022). In line with the idea that IAT score patterns are partially normatively shared, in all 17 studies randomly paired other participants' prediction patterns were significantly related to participants' own IAT score patterns. Importantly, however, in 16 out of the 17 studies participants' own prediction patterns were more strongly related to their own IAT scores, and participants' own prediction patterns explained variance in their own IAT score patterns over and above the randomly paired other participants' prediction patterns in 16 out of the 17 studies. The meta-analytical

effects supported these findings and showed that across all studies, even though the randomly paired other participants' prediction patterns partially explained variance in participants' own IAT score patterns with an effect-size of $b = 0.33$ ($b = 0.21$ in the simultaneous model), participants' own prediction patterns were more strongly related to their own IAT score patterns, $b = 0.44$ ($b = 0.36$ in the simultaneous model). These results suggest that participants' predictions may be a combination of introspective insight and cultural knowledge, with unique insight playing a slightly larger role.

As pointed out earlier, these findings all rely on within-subject correlations between participants' person-standardized predictions and their person-standardized IAT results and thus indicate the degree to which participants are aware of their own automatic reactions toward different target groups. As such, the findings indicate how much people know their own reactions if comparison and labeling standards are not taken into consideration. Hahn et al. (2014) showed that between-subject correlations between predictions and IAT scores standardized per target group were considerably smaller than those within-subject correlations. We replicated these findings in all 17 studies and found that descriptively the average between-subject correlation was always smaller than the average within-subject correlation with a meta-analytical between-subject effect of $b = 0.22$ and a meta-analytical within-subject effect of $b = 0.44$ (see Figure 8). In line with theorizing by Hahn and Goedderz (2020), this further highlights the importance of distinguishing between the concept of *introspective awareness* and *social calibration*. Hahn and Goedderz (2020) have proposed to use the term *introspective awareness* to refer to a person's ability to sense and report on their own cognitions toward different targets, while the term *social calibration* may be used to describe a person's ability and willingness to apply labels to their own cognitions in accordance with culturally shared conventions. The present study is not in the position to make further claims on different processes involved in *introspective awareness* and *social calibration*, but it suggests that it is worthwhile to distinguish between the two concepts. If

researchers continue to primarily inspect between-subject correlations to assess awareness, they may assume that their participants lack awareness when really they are just poorly calibrated (Goedderz & Hahn, 2023).

2.7.1. Subgroup Analyses

The overall pattern of results as they pertain to the different theoretical considerations replicated in all examined subgroups. Nonetheless, it is noteworthy that effect sizes differed substantially between some of the subgroups. Participants were overall less accurate in online studies than in the laboratory (average within-subjects correlations $b = .28$ vs. $b = .48$). This may have several reasons. First, studies conducted online could potentially show lower effect sizes than studies conducted in the laboratory due to less motivated and/or concentrated participants and an overall less-controlled environment. Second, in the specific case of predicting reactions on socially sensitive topics, a laboratory environment may lead participants to feel more encouraged to report on their implicit biases because they more strongly believe those will be found out either way by the researcher. However, the most plausible explanation in our opinion is proposed by the present results. In addition to lower prediction accuracies, explicit thermometer ratings were also substantially less related to IAT scores in online settings, and when controlling for predictions they even showed negative relationships with IAT scores. In contrast, once for thermometer ratings were controlled for, the relationship between predictions and IAT scores no longer differed between online and laboratory studies. This suggests that higher consistency between explicit ratings and predictions in online samples might exert a suppression effect on prediction accuracy. In other words, online participants' explicit ratings diverged more strongly from their IAT scores compared with lab participants. Because participants' explicit ratings seem to partially influence participants' predictions, online predictions are consequently also less related to their IAT score patterns. Once we control for these explicit evaluations, predictions explain

more variance in their IAT score patterns in online samples, closer to the prediction accuracy found in the other samples.

We found the same pattern of results for student samples as opposed to the general public. It is important to note that most online studies were also conducted on the general public such that there is only one study in the present meta-analysis that was conducted online on a student sample. As such, it may be difficult to distinguish whether differences in effect sizes are due to differences in the setting or in the sample. Both factors may play an important role but thus far we believe the data suggests that the setting is more important. That is, results for the only online study conducted on a student sample (Study 1) are descriptively more in line with the other online study results than with the other student sample results. Note, for instance, that thermometer ratings are only unrelated to IAT scores in the online studies, and the relationship between explicit ratings and the IATs turns negative in the simultaneous model only in the online studies (see Table 2). More studies in different settings using more diverse samples are needed to support this speculation.

Beyond this notable difference between online and lab samples, no other meaningful difference emerged between subsamples. Perhaps most strikingly, German participants did not differ from Canadian and US-American participants on any measure with the exception of higher correspondence between implicit and traditional explicit evaluations. In contrast to the English-speaking world, where “implicit bias” is a matter of continuous public debate, this construct has barely reached public discourse outside the academy in Germany (although general discussions about diversity beyond “implicit bias” are equally prevalent). Additionally, the biggest immigrant groups in German society come from Central-Eastern and Eastern Europe, as well as the middle-East and Turkey; while the proportion of the population that identifies as Black, East-Asian, and Latino/-a has historically been much lower than in the

US, Canada, and the UK.⁷ This lower experience with the groups in question for this paradigm may partly explain higher implicit-explicit correlations. German participants seem to have more readily based their explicit evaluations on spontaneously activated knowledge, while considering fewer pieces of additional propositional information, perhaps because less other information (from experience or public discourse) was available to them. Importantly, however, the lack of exposure to the construct of “implicit bias” or the groups does not seem to have made it harder for them to accurately predict their IAT scores. This further contradicts the notion that accurate prediction of IAT scores merely reflects cultural knowledge in American participants. If this were true, then participants with less exposure to cultural information about “implicit bias” (i.e., German participants) should be worse at predicting their IAT scores. These interpretations remain exploratory and speculative and need to be corroborated by more targeted research and analyses.

Lastly, unpublished studies showed somewhat lower effect sizes than published ones. A notable proportion of these unpublished studies were run online as pilot studies, to see whether the paradigm could be moved online to save resources, trying different recruitment platforms, programming languages, countries, and samples. Hence, the lower effect sizes in unpublished studies can in large part be explained through the lower effects in online samples discussed above. While we have so-far concluded that the present paradigm cannot be run online without significant sacrifices in data precision, we hope that our decision to publish these hitherto unpublished samples in the current format will help future researchers gauge what effect sizes to expect if they try to replicate or extend the present findings in different modalities.

⁷ E.g., freely available data from the Federal Statistical Office of Germany (*Statistisches Bundesamt*, available at <https://www-genesis.destatis.de/genesis/online>) as well as the Robert Koch Institute (www.rki.de) for 2021 suggest that less than 1% of the German population each have a sub-Saharan African, South-American, and East-Asian background. The Federal Statistical Office of Germany does not collect data on racial identification.

In sum, the present subgroup analyses suggest that Hahn et al.'s (2014) patterns of results replicate across modalities, two languages and three countries, as well as general-public and student populations; with some notable differences in patterns when the paradigm is administered online. Future research with even more languages and cultures, conducted by independent researchers, is needed to corroborate these effects.

2.7.2. Limitations

There are several limitations to the present meta-analysis. First, this meta-analysis only includes studies conducted by or in close supervision of the original first author. While this ensured that the procedures were maximally comparable across studies, internal meta-analysis may provide a biased estimation of effect sizes (Vosgerau et al., 2019). To minimize this risk, we preregistered all criteria on inclusion and exclusion of studies and participants within studies, and preregistered all analytical strategies. Though this may reduce the risk of biased estimates due to selective reporting, specifics about the procedure may also impact the size of the effect (Simons, 2014). In order to enable other researchers to closely replicate effects reported in this meta-analysis, all materials are openly accessible on OSF. First evidence by Morris and Kurdi (2022) show that the overall prediction effect replicates when other researchers follow the prediction paradigm such that we are confident that our overall conclusions will hold when studies are conducted by other researchers. Future studies are necessary to delineate whether specific aspects of the predictions help participants gain awareness of the cognitions reflected on implicit evaluations.

Second, the majority of our participants self-identified as White (between 40% – 89.5%), all studies were conducted in Western countries (Germany, USA, Canada, United Kingdom), and almost all studies were conducted with student samples. One aim of the present study was to examine the extent to which Hahn et al.'s (2014) findings hold in different samples and settings, and we have found that effect sizes differ between study characteristics, but that the overall pattern of results remained the same. This makes us

believe that the overall theoretical ideas formulated here and in Hahn et al.'s (2014) studies may at least generalize across WEIRD (Western, Educated, Industrialized, Rich, and Democratic, Henrich et al., 2010) samples, where debates around prejudice and stereotyping, and egalitarian norms are highly salient. More research on more diverse samples is needed to see whether these findings generalize to other non-WEIRD populations.

Third, and finally, to ensure that the theoretical considerations of Hahn et al. (2014) and the present paper hold beyond the specifics of the prediction paradigm using IATs toward social groups, it is important to conceptually replicate the present findings in different attitudinal domains and to other implicit measures. First evidence that people may be generally aware of the cognitions reflected on their implicit evaluations regardless of the attitudinal domain has already been provided by a few studies. For instance, Rahmani Azad et al. (2022) showed that participants were able to accurately predict their patterns of implicit gender stereotyping, and Goedderz and Hahn (2023) found that participants accurately predicted their pattern of implicit preferences for food items. Further, Morris and Kurdi (2022) extended Hahn et al.'s (2014) paradigm to a wide range of other attitudinal domains and showed that participants were also able to predict their results on the Affect Misattribution Procedure (Payne et al., 2005). Considering all this additional evidence, we are optimistic that the theoretical considerations of this paper will generalize to other attitudinal domains and other implicit measures, but more research is needed.

2.8. Conclusion

A long-standing debate in research on implicit evaluations is whether they capture unconscious mental content (Greenwald & Banaji, 1995). In contrast to this conceptualization, Hahn et al. (2014) found that participants were able to accurately predict the patterns of their IAT results suggesting that participants were aware of the cognitions reflected on their IAT scores. The present meta-analysis reanalyzed 17 published and unpublished studies replicating Hahn et al.'s (2014) prediction paradigm. All patterns of

results replicated across the examined studies and showed that participants were able to accurately predict the patterns of their IAT results even though they often report different evaluations on traditional explicit measures. Results further indicated that participants had unique insight into their own automatic cognitions beyond knowledge about normatively shared patterns. While participants were quite accurate in reporting their patterns of IAT results toward different target-groups, they were less accurate in labeling their cognitions in accordance with conventions in the sample. While effect sizes were smaller for online studies conducted on the general public than for lab studies run on university students, the overall pattern of results remained unchanged throughout the examined subgroups. Together, these findings provide further evidence that the cognitions reflected in implicit evaluations are consciously accessible. We hope this meta-analysis can guide researchers willing to study awareness of implicit evaluations in two important ways: First by providing meta-analytical effect size estimations, and second, by contributing to a better theoretical understanding of studying awareness in research on implicit evaluations.

Chapter 3. Biases left unattended: People are surprised at racial bias feedback until they pay attention to their biased reactions.

This chapter is based on the following publication:

Goedderz, A., & Hahn, A. (2022). Biases left unattended: People are surprised at racial bias feedback until they pay attention to their biased reactions. *Journal of Experimental Social Psychology*, 102, 104374.
<https://doi.org/10.1016/j.jesp.2022.104374>

Please note that headings, citation style, and formatting were changed to fit the layout of this dissertation. The content of the article was not changed.

Abstract

Why are people surprised at racial bias feedback, such as test results from Implicit Association Tests (IATs), even though they can predict their IAT racial bias scores prospectively? The present research tested three hypotheses: People are surprised at racial bias feedback due to (1) the feedback wording, (2) implicit evaluations often being preconscious and unattended, or because (3) pretending to be surprised at racial bias feedback is socially desirable. One pilot, four preregistered studies, and a mini-meta-analysis supported hypothesis (2): Although racial biases such as those reflected on IAT scores are observable, people rarely pay attention to them. Specifically, predicting IAT results (Studies 2-4b) and encouragement to pay attention to one's biased reactions before IAT completion (Study 3) reduced surprise, independent of explanation of "implicit bias" (Study 4b). Contradicting the social-desirability hypothesis (3), neither encouragement to admit to bias in the form of abstract predictions (Study 3), nor non-threatening explanations of implicit bias (Study 4b), reduced surprise in the absence of encouragement to pay attention to one's own biases. Speaking against hypothesis (1), surprise was independent of feedback severity (Studies 1-3); and the prediction effect was mediated by recognition of bias, but not correspondence of predictions and feedback (Study 3). These studies suggest that surprise is a consequence of the preconscious nature of automatic social cognitions: People may be motivated to keep consciously accessible racial biases out of awareness. Implications for theories of implicit social cognition and the generality of these effects beyond research on implicit bias are discussed.

Keywords: Attitudes, IAT, implicit bias, preconscious, unconscious

3.1. Introduction

“I was really surprised at the [IAT] results as I never thought of myself as having any biases against Black people.” (Study participant).

Recently, the idea that racial biases are widespread even among egalitarian-minded people has increasingly gained traction in public discourse. For instance, Gallup (2021) reports that agreement with the observation that minorities are treated poorly was at an all-time high among Americans of all backgrounds in 2021. One of social psychology’s most prominent contributions to this debate – and simultaneously one of its most criticized constructs – has been the concept of implicit bias (BBC News, 2017; Devlin, 2018; Green & Hagiwara, 2020; Grinberg, 2015; Robson, 2021). Implicit bias research has shown widespread implicit racial biases across most Western countries among all strata of society (Nosek et al., 2002; Redford, 2018), and in response, implicit bias trainings, in which informing people about their biases is often an important feature, have been on the rise across the world (Chamorro-Premuzic, 2020; Wen, 2020).

In contrast to the observation that acknowledgement of the widespread nature of racial biases is on the rise, however, and in line with the quote in the beginning, research has documented that people respond defensively and with surprise to IAT feedback that communicates that they themselves might harbor racial biases (Howell et al., 2013; Howell et al., 2017; Vitriol & Moskowitz, 2021). Two common explanations for surprise responding to bias feedback have been that racial biases are either purposefully hidden (such that surprise may be a reaction to the disclosure of a hidden response), or that they must be entirely “unconscious” (such that participants could not have known, Haider et al., 2014; Nosek et al., 2002; Quillian, 2008). However, both ideas are at odds with findings that people can predict the patterns of their IAT scores (Hahn et al., 2014) and that people can easily be brought to acknowledge their own (racial) biases (Hahn & Gawronski, 2019). The present research

addresses this apparent contradiction: Why are people ostensibly surprised when they learn about their own implicit racial biases (Gawronski, 2019; Howell et al., 2013), even though research suggests that people may be able to predict the patterns of racial bias IAT scores accurately (Hahn et al., 2014)?

Focusing on the IAT as the most widely used measure of implicit bias (Gawronski & De Houwer, 2014), we investigated three explanations: People are surprised because they (1) disagree with the labeling of the racial bias feedback, (2) rarely pay attention to their racial biases, or (3) pretend to be surprised because they believe this is the socially desirable reaction to racial bias feedback. To this end, we gave people feedback to Black-White IATs, measured their surprise reaction, and investigated whether this surprise would decrease in response to (1) different labeling of racial bias feedback, (2) making people pay attention to their racial biases ahead of taking a test, (3a) making people admit to their racial biases, or (3b) describing IAT racial bias scores in more socially acceptable ways. As such, this paper aims to provide evidence about what aspects of IAT racial bias feedback are surprising to lay people and what this can tell us about the – purportedly “unconscious” or “conscious” – nature of the cognitions reflected on implicit bias scores. On a more general level, we suggest that many racial biases – including those reflected on implicit measures such as the IAT – are often “preconscious” (Dehaene et al., 2006; Hahn & Goedderz, 2020): They are rarely attended to, even though, in principle, they are observable. This would not only advance our theoretical understanding of implicit bias, but also point towards simple interventions: It would suggest that people should be encouraged to pay attention to their own reactions to notice their own automatic biases (Hahn & Gawronski, 2019).

3.2. Previous Research on Reactions to IAT Racial Bias Feedback

People prefer positive to negative feedback, they want to see themselves in a positive light (Sedikides et al., 2003), and they expect to score better than the average other person on most tests and dimensions (Alicke et al., 1995). From this perspective, it is unsurprising that

most people, including people who complete IATs, generally think they are less biased than others (Howell & Ratliff, 2017). Indeed, even as Hahn et al.'s (2014) participants predicted the patterns of their IAT scores accurately, they thought other participants in the same study would show a lot more bias on average – a statistically impossible result. Different from these findings, however, the beginning statement to this paper – if we believe it to be honest – indicates that many people do not just think they are less biased than others, they appear to think that they are not biased at all.

Although there is a thriving research field that investigates defensive reactions to IAT feedback, their causes and consequences, and ways to overcome them (Howell et al., 2013; Howell et al., 2017; Vitriol & Moskowitz, 2021), the specific reaction of surprise has not received similar attention, even though it is often mentioned anecdotally or implied (Gawronski, 2019; Howell et al., 2013). For instance, research has found that people are defensive to the degree that their IAT feedback deviates from their explicit evaluations (Howell et al., 2013; Howell et al., 2015; Howell et al., 2017; Howell & Ratliff, 2017). Such defensiveness reactions could indicate that participants are also surprised at being told that they are biased. However, defensiveness and surprise are independent and distinct reactions. For instance, it is possible to be defensive about being told one is biased without being surprised about it (e.g., by expecting a test to be biased); and a person might be surprised without becoming defensive (e.g., new and unexpected information can be considered interesting). As such, defensive responding to IAT feedback is limited in terms of clarifying whether people are surprised about their bias scores or not. Additionally, surprise is different to defensiveness as it focuses primarily on the feeling of unexpectedness (Stiensmeier-pelster et al., 1995) and may thus be especially fruitful when examining the conscious or unconscious nature of the cognitions reflected on implicit bias scores.

Looking at research that has investigated surprise reactions to IAT feedback specifically, a classroom study by Hillard et al. (2013) showed that more bias feedback on the

IAT was associated with higher levels of surprise. However, levels of surprise in this study were rather low (below 2.0 on a 5-point scale with 1 indicating “very slightly or not at all” and 5 indicating “extremely”, Hillard et al., 2013, p. 506). Hence, these results question the assumption that people are generally surprised at IAT feedback, but they support the notion that people will tend to be more surprised the more bias the feedback communicates. In a one-item measure Howell et al. (2013) found that participants were more surprised at their IAT feedback the more it deviated from their explicitly reported attitudes. A qualitative analysis by Schlachter and Rolf (2017) looking at comments about IAT feedback on the internet showed somewhat mixed results. Some participants said that they were surprised at their feedback and others that they were not, suggesting that surprise might vary considerably across people and circumstances. In line with this, Perry et al. (2015) found that how their participants reacted to bias feedback was a function of participants’ individual differences of bias awareness.

Taken together, there is some evidence that people might be surprised at IAT bias feedback, especially when it deviates from explicit attitudes, but this might differ largely between people. Whether people are surprised at learning that they might harbor any biases (as opposed to no biases) remains an open question awaiting further empirical evidence.

3.3. Implicit Evaluations as Unconscious Attitudes

Surprise reactions at IAT feedback are often cited as evidence that the cognitions reflected in implicit measures must be unconscious (Gawronski, 2019; Krickel, 2018; Lane et al., 2007). After all, if people were aware of their biases, they should not be surprised to learn about them. In early debates around implicit bias, the claim that implicit evaluations reflect unconscious attitudes additionally used to often appear in discussions around low correlations between implicit and explicit measures⁸ of the targets (Hofmann et al., 2009; Hofmann,

⁸ We use the term “implicit” to refer to evaluations inferred from indirect computerized reaction time measurements instruments such as the IAT, and “explicit” to refer to self-reported evaluations (Hahn and Gawronski, 2018; De Houwer et al., 2009). As such, the usage of this terminology makes no assumptions about

Gawronski et al., 2005; Hofmann, Gschwendner et al., 2005; Nosek, 2005, 2007; Nosek & Hansen, 2008). However, low correlations between implicit and explicit measures do not per se speak to the inaccessibility of the cognitions reflected in implicit measures (Gawronski et al., 2006; Hahn et al., 2014; Hahn & Gawronski, 2014). Various prominent dual-process models provide different explanations for why implicit and explicit measures diverge (Fazio, 2007; Gawronski & Bodenhausen, 2006, 2011).

For instance, the MODE model (*Motivation and Opportunity as DEterminants*) suggests that explicit and implicit evaluations diverge as a function of motivation and opportunity (Fazio, 2007). The main claim here is that people report different evaluations on explicit measures because they are motivated to present themselves in socially desirable ways (Dunton & Fazio, 1997). Gawronski and Bodenhausen's (2011) Associative-Propositional Evaluations (APE) model proposes that people do not always consider the reactions reflected on implicit evaluations to be valid bases for their explicit judgements. According to this model, people are aware of their negative associations with ethnic minorities. However, they might nevertheless report positive attitudes toward them on explicit ratings because they consider other propositional information, e.g. their egalitarian values or specific exemplars of minority members they admire, to be more valid bases for their reported attitudes. As such, both the MODE and APE models argue that the cognitions reflected on implicit measures are generally consciously accessible, but often rejected (Fazio & Olson, 2003; Gawronski & Bodenhausen, 2011). And indeed, across several studies Hahn et al. (2014) and Hahn and Gawronski (2019) found that their participants were able to predict the patterns of their implicit evaluations when asked directly. The authors asked their participants to predict how they will score on five different IATs measuring their spontaneous reactions toward five social groups (Black, Asian, Latino, Children, Celebrities) compared to non-celebrity White

the underlying cognitions reflected on these measures. We hope to contribute to understanding the underlying cognitions with this paper.

adults. Their participants were generally good at predicting the patterns of their IAT scores, even though they reported different explicit evaluations. These findings challenge the unconsciousness hypothesis. People can predict the patterns of their implicit evaluations, and there are other explanations for why they report different evaluations when asked explicitly.

3.4. Potential Reasons for Surprise

How can the observation that people are surprised at racial bias feedback be reconciled with findings that they can predict the patterns of their IAT scores prospectively? Integrating different theories and empirical evidence with respect to implicit social cognition led us to three different hypotheses.

3.4.1. Surprise and Harshness of Feedback: The Feedback Wording Hypothesis

One hypothesis is that people generally know that they are biased, but they might be surprised at the specific wording that is used to describe their biases (Gawronski, 2019). That is, Hahn et al. (2014) found that participants knew *that* they harbored biases, but they didn't seem to know how biased they were compared to other people; and – consistent with the better-than-average effect (Alicke et al., 1995; Howell & Ratliff, 2017) – they suspected that they were less biased than others.

From this perspective, participants may be surprised at any IAT bias feedback that goes beyond “a slight preference” for one group over another. If this is true, then surprise should be a specific reaction to the (arbitrarily set) conventions and language for IAT feedback, rather than the bias feedback per se (Gawronski, 2019), and people should be less surprised at mild than strong bias feedback. However, as the beginning statement indicates, many people do not only seem to reject the strength of their bias feedback, but the fact that they may harbor any biases at all.

3.4.2. Implicit Evaluations as Preconscious Attitudes: The Attention Hypothesis

If it is in fact true that many people are surprised at harboring any biases at all, then the question remains: How is such surprise compatible with the fact that people can predict

the patterns of their IAT scores accurately? Integrating research by Hahn and Gawronski (2019) with theories of consciousness (Dehaene et al., 2006; Hofmann & Wilson, 2010; Hahn & Goedderz, 2020) suggests the following explanation: The cognitions reflected on implicit measures might not generally be unconscious, but often “preconscious” – people rarely pay attention to them unless they are encouraged to do so. In line with the need to view oneself positively, they will hence tend to believe that they are unbiased until they are encouraged to face their biases.

Specifically, Hahn and Gawronski (2019) found that participants aligned their explicit evaluations with their implicit evaluations and acknowledged being biased after they predicted their IAT scores. This indicates that people may learn something new about themselves when they predict their IAT scores. Merely completing IATs (announced as tests of implicit racial attitudes) without predictions changed neither explicit evaluations nor acknowledgment of bias compared to control conditions and pre-test ratings. This last point is important, because it emphasizes that the prediction procedure did not just make participants more honest about cognitions that they knew all along. If that were the case, then knowledge of measurement, and hence completion of IATs, should have had similar effects. Instead, it seems that predicting IAT scores led participants to discover new information about themselves, and this changed their explicit evaluations and their perceptions of how biased they are. Models of consciousness may help clarify this point.

That is, in line with Hofmann and Wilson (2010) and others (Dehaene et al., 2006; Dehaene & Naccache, 2001), we propose that a cognitive process reaches awareness when (1a) it produces a signal that is strong enough, and (1b) attention is paid to this signal. Moreover, a process that produces a detectable signal (1a is present) but remains outside of conscious awareness because it is left unattended (1b is absent), may be called “preconscious” (Dehaene et al., 2006). A lot of research and theorizing suggest that the signal produced by the cognitions reflected on implicit evaluations is a spontaneous affective reaction (Gawronski &

Bodenhausen, 2006; Hahn & Gawronski, 2019; Ranganath et al., 2008; Smith & Nosek, 2011). Integrating these thoughts, people might be surprised at their IAT results because they rarely pay attention to their spontaneous affective reactions to people with different backgrounds. From this perspective, surprise after IAT feedback would indeed be a reaction to learning that one is biased. However, the reason for this surprise is not that the cognitions reflected on implicit measures are generally unconscious. Much rather, surprise would demonstrate that these cognitions are preconscious – people rarely pay attention to them. If this hypothesis is true, then drawing people’s attention to their spontaneous affective reactions before receiving IAT feedback should lower their surprise at this feedback.

3.4.3. Real Surprise? The Social Desirability Hypothesis

One last explanation for why people indicate surprise at bias feedback may be social desirability (Crowne & Marlowe, 1960). That is, people may be aware that they harbor biases, but report surprise as an act of self-presentation, because pretending that racial biases are unexpected might be the most desirable answer to give. If this explanation is true and the participant quoted at the beginning of this article was not actually surprised but simply dishonest, then people should always indicate surprise even at “slight” preference feedback because any level of bias is undesirable. However, this surprise should still be a function of the strength of the feedback. That is, showing “strong” racial preferences is less desirable than showing “slight” preferences, such that reported surprise would have to be a function of the desirability of the specific feedback participants get. Furthermore, presenting the IAT procedure in a non-offensive way should lower participants’ surprise because they may perceive their IAT results as less of a threat to their values and beliefs.

3.4.4. Feedback Wording, Attention, or Social Desirability?

To test these three hypotheses, the four studies presented in this paper measured surprise reactions in response to IAT racial bias feedback. Although the three hypotheses are compatible in some instances, we designed our studies such that they would answer three

empirical questions to which the three hypotheses make opposing predictions. The first is whether or not participants report more surprise in response to all levels of racial bias feedback - even low levels of bias - when compared to no-bias feedback. The second is whether this surprise is a function of the degree of bias the feedback communicates, such that feedback of a “strong” bias would lead to more surprise than feedback of “mild” bias. The last is whether making people pay attention to their biased reactions before test completion reduces surprise. The three hypotheses’ predicted answers to the three questions are summarized in Table 1 and described next.

Table 1

Empirical Questions in the Present Research and Their Predicted Outcomes According to the Three Hypotheses

	Hypotheses		
	Feedback wording hypothesis	Attention hypothesis	Social-desirability hypothesis
Reason for surprise at IAT feedback	Arbitrary labels: People know their biases but disagree with the labels.	Preconscious attitudes: People rarely pay attention to their biases.	Pretend surprise: People know their biases but admitting this is undesirable.
Empirical Questions	1. Do people generally report surprise at any racial bias feedback compared to no-bias feedback, including low levels of bias?	No	Yes
	2. Is the level of surprise a function of the degree of bias communicated in the feedback?	Yes	Both answers compatible/no prediction
	3. Does paying attention to one’s biases before IAT completion reduce surprise at IAT feedback?	No	Yes

The feedback wording hypothesis predicts no surprise at bias feedback that is clearly at the low end of the scale, but increased surprise the more comparative bias the feedback indicates. This surprise reaction should further not change when people are asked to pay attention to their biases before IAT completion. According to the feedback wording

hypothesis, participants already know that they harbor biases, and hence asking them to pay attention to those biases should not lead to any new insights.

The attention hypothesis predicts surprise at any feedback indicating bias unless the person is encouraged to first pay attention to their biased reactions. The attention hypothesis makes no predictions regarding reactions to severity of feedback.

Lastly, the social desirability hypothesis predicts both surprise at all bias feedback and increased surprise the less socially desirable said feedback sounds. Making participants pay attention to their biases before IAT completion may reduce “pretend surprise”, but only to the degree that it either induces participants to admit to biases they would otherwise hide, or changes their perception of what a socially desirable response is. Hence, it should not require any specific *attention-to-bias* manipulation.

Instead, any request to admit to biases before IAT completion should suffice to both induce a person into admitting bias and shift their perception of a desirable response. We explain these last points in more detail in Study 3 when we test them directly.

In sum, observing the patterns of results to the three questions we investigated allowed us to see which hypothesis explains best why people report surprise at IAT bias feedback.

3.5. The Present Research

The aim of the present studies was to investigate three potential explanations for why people react with surprise at racial bias feedback even though they can predict the patterns of their IAT scores prospectively: (1) disagreement with the feedback wording, (2) the preconscious nature of implicit attitudes (attention hypothesis), or (3) pretend surprise due to social desirability concerns.

We started this research project with a pilot study in which we asked participants to imagine hypothetical feedback to test the surprise scale we developed for subsequent studies. In Study 1, we tested whether people in fact react with surprise to performance-based bias feedback compared to feedback that declared “no meaningful bias” (Question 1, see Table 1).

To test the feedback wording and the social desirability hypotheses, Studies 1-4 further investigated whether surprise was a function of the degree of bias communicated in the feedback (Question 2, see Table 1). Additionally, we altered the feedback to be less socially undesirable in Study 2. Addressing the attention hypothesis, Studies 2-4 tested whether participants would be less surprised at IAT feedback after encouragement to pay attention to their spontaneous affective reactions to stimuli of the targets in question (Question 3, see Table 1).

We first operationalized attention to reactions as predicting IAT scores before completing IATs (Studies 2-4, Hahn & Gawronski, 2019). In Studies 3 and 4b, we then disentangled whether the effect of predictions on surprise could be better explained by attention, social desirability concerns, and/or the wording of the feedback. Specifically, Study 3 tested whether people simply pretend to be biased unless they are induced to admit to biases ahead of time (social desirability hypothesis), or if they are instead truly surprised at IAT feedback as long as they are not encouraged to pay attention to their biased reactions (attention hypothesis). Study 4b investigated whether non-threatening information about implicit evaluations and the IAT could explain the prediction effect in the absence of attention to one's affective reactions (social desirability hypothesis). Finally, to compare the attention and feedback wording hypotheses directly, a mediation analysis in Study 3 also tested whether the effect of prediction on surprise could be better explained by correspondence of feedback with expectations (feedback wording hypothesis) or by recognition of bias (attention hypothesis).

We preregistered all studies (except for the pilot study) and report all data, measures, manipulations, and exclusions in each study to allow for increased transparency, replicability, and trustworthiness of our findings (Lindsay et al., 2016). Data analyses were only conducted once the full samples reported here were collected and preregistered data exclusions were completed. We report all preregistered analyses and indicate where we conducted non-

preregistered analyses. All materials, data sets, preregistrations, and analysis files can be found at <https://osf.io/bezqx/>.⁹

3.6. Pilot Study

This study was aimed at piloting a scale developed to measure surprise at IAT feedback in the present line of research. The study also tested whether people would be more surprised when imagining bias as opposed to no-bias feedback. It was not preregistered.

3.6.1. Method

3.6.1.1. Participants. One-hundred and twenty-two participants were recruited on Amazon's Mechanical Turk platform (MTurk) in exchange for US-\$ 0.10 basic payment and US-\$ 0.10 possible bonus payment. After excluding eight participants who failed at least one of two attention check items, the final sample consisted of 114 participants¹⁰ (52.6% female; median age = 36, age range = 18-64 years). All participants were American citizens and most (76.3%) identified as White (8,8% Black/African-American, 7,9% (East-) Asian, 0,9% Latino/Hispanic, 6,1% more than one ethnic category).

3.6.1.2. Materials and Procedure. Participants were asked to imagine they were to receive feedback on a Black-White IAT that either reveals that they have “...*a strong automatic preference for WHITE over BLACK*” (strong-bias condition) or “...*NO statistically detectable preference for either BLACK or WHITE*” (no-bias condition), as well as to repeat this feedback on the next page (attention check). Seeing their feedback on top of the screen, participants then completed a ten-item surprise scale aimed at self-reported surprise and unexpectedness of IAT feedback (see Table 2 for final six items and their psychometric

⁹ We chose the preregistration template from “as predicted”, which is not stored by name on the Open Science Framework (OSF). Hence, the preregistrations can only be identified by their dates within the project, which we indicate in each study. We also provide direct links to each registration.

¹⁰ We conducted power sensitivity analyses using G*Power (Faul et al., 2007) for all studies. With the final sample size of 114 (53 participants in the bias condition, and 61 in the no-bias condition; assuming alpha = 0.05; two-tailed; power = 80%), this pilot study had to have a minimum effect size of $d = 0.53$ to show up as significant, which could be reached at a critical t value of 1.98.

properties, and OSF repository for all ten initial items), on 7-point Likert scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*), and in individually randomized orders.

The scale included another attention check item (“Attention-Check: Please select ‘3’”). The study concluded on demographic information and a debriefing.

Table 2

Item Factor Loadings on The First Component of a Principal Components Analysis, and Variance Explained by This Component, across all 4 studies and the pilot.

Item	Factor Loadings per Study				
	Pilot	1	2	3	4b
1. I was surprised at my IAT result.	.96	.91	.88	.89	.85
2. I expected a different IAT result.	.97	.91	.90	.87	.87
3. I did not expect to get the IAT result that I got.	.96	.89	.92	.92	.88
4. My IAT result confirmed what I expected to find. (rev.)	.98	.95	.92	.94	.89
5. My IAT result supported my initial expectation. (rev.)	.88	.94	.91	.93	.90
6. I expected the IAT to show the result that it showed. (rev.)	.95	.95	.94	.91	.90
% of Variance explained	89.78	85.83	83.04	82.79	78.03
Cronbach’s alpha	.98	.97	.96	.96	.94

6.6.2. Results and Discussion

A *t*-test on the average of the ten items of the surprise scale showed that participants who imagined receiving strong-bias feedback scored higher ($M = 4.94$, $SD = 1.76$) than participants who imagined getting no-bias feedback ($M = 2.39$, $SD = 1.60$), $t(112) = 8.11$, $p < .001$, $d = 1.49$, 95% CI [1.08, 1.91]. This result also replicated on every item individually (all t s > 6.83 , all p s $< .001$)¹¹, such that we selected a smaller number of three forward and three backward-phrased items for the remaining studies (see Table 2, Cronbach’s $\alpha = .98$).

¹¹ There were no noticeable differences in psychometric quality between the 10 items. All ten items loaded on the same first component in a principal components analysis (PCA, 89.8% of variance explained) and the original ten-item reliability (Cronbach’s $\alpha = .99$) never dropped below .98 from the exclusion of any particular item.

Taken together, the pilot established our surprise scale and provided first evidence for Question 1: participants reacted with more surprise when they imagined receiving feedback of harboring strong racial biases as opposed to feedback that they harbor no biases.

3.7. Study 1

Study 1 aimed at providing empirical evidence concerning whether people are surprised at IAT feedback indicating racial biases.¹² Specifically, we tested whether surprise at standard feedback as typically given at www.projectimplicit.com would be higher than performance-independent no-bias feedback (Question 1, see Table 1), as well as whether surprise would be a function of the degree of bias the feedback indicates (Question 2, see Table 1). The preregistration for this study can be found at <https://osf.io/8uev5/> and was registered on January 10th, 2019.

3.7.1. Method

3.7.1.1. Participants and Design. Study 1 featured a two-condition between-subjects design (standard feedback vs. no-bias feedback). A G*Power analysis (Faul et al., 2007) for an independent samples *t*-test with an allocation ratio of two-to-one (allocating twice as many participants to the standard feedback condition as to the no-bias feedback condition) revealed that, to find a medium sized effect of $d = .50$ with 80% Power, we would need at least 144 participants. To account for our preregistered exclusion criteria, we aimed at recruiting 180 participants via the service TurkPrime on Amazon's Mechanical Turk (MTurk). Participants were first informed that the study contained a computerized reaction time task, and that they could only proceed to the end if they followed instructions and would not click buttons randomly on this task. Participants who responded with ≤ 300 ms in 10% or more of the trials

¹² We report the second attempt at running the exact same study. Our first study revealed a minor programming mistake. When participants completed the surprise scale in that study, the title instructions encouraged them to "imagine" they received feedback on an IAT, even though they had just in fact received feedback on IATs, so we decided to run the study again. The main effect largely remained the same so that we decided to only report the second study (see supplemental materials). The preregistration for the first study can be found at <https://osf.io/rnf3j/> and was registered on December 12th, 2018. All materials, data, and analyses are available online.

were excluded from the study after IAT completion (Greenwald et al., 2003). All participants who completed the study received a basic payment of US \$0.70. Another US-\$0.30 were paid if the two attention check items also used in the pilot were answered correctly. In total, 198 participants started the study on MTurk, of which 180 completed all tasks and passed the preregistered exclusion criteria^{13,14} (50% Female; median age = 33, age range = 18-65 years, 98.9% US-American citizens). 72.8% of the participants self-identified as White (7.8% Black/African American, 3.3% Latino/Hispanic, 6.1% East-Asian, 3.9% South-Asian, 6.2% more than one ethnic category or another ethnicity).

3.7.1.2. Black-White IAT. To create a seven-block IAT (Greenwald et al., 1998) in Qualtrics, we used the IATgen tool (<https://iatgen.wordpress.com>, Carpenter et al., 2019). As targets, we used pictures of 10 Black and 10 White individuals (five male and five female individuals per target category) adapted from Hahn et al. (2014) who used pictures from the productive aging lab database (Minear & Park, 2004). The attributes consisted of 10 positive and 10 negative words (see all stimuli on the OSF repository at <https://osf.io/yxrpb/>). In the first block, participants completed 20 practice trials categorizing the attribute words to the left or right side by using the E or I key on their computer keyboards. The second block consisted of another 20 practice trials categorizing the target pictures to the left or right side as “White” or “Black”. Blocks 3 (20 trials) and 4 (40 trials) were combined blocks where participants had to either react with one key to pictures of Black people and negative words, and with the other key to White people and positive words (prejudice-compatible), or the other way around (prejudice-incompatible). In the fifth block, the target pictures switched sides and participants

¹³ With the final sample size of 180 (56 participants in the no-bias feedback condition, and 126 in the standard-feedback condition; assuming alpha = 0.05; two-tailed; power = 80%), Study 1 had to have a minimum effect size of $d = 0.46$ to show a significant effect, which could be reached at a critical t value of 1.97.

¹⁴ As preregistered, although 25 participants failed to answer the attention check regarding their feedback correctly, they were kept in the final sample. In the previous study, 20 participants were excluded due to this exclusion criterion. We decided to not honor this exclusion criterion again because (1) it excluded participants unequally from conditions, (2) results remained largely the same with or without these participants, and (3) participants were reminded of their actual IAT feedback before IAT completion such that it was ensured without this attention check that they knew their IAT result.

spent 40 trials practicing the reversed categorization. Block 6 and 7 were structurally similar to Blocks 3 and 4, but with a changed combination. Participants who completed the prejudice-compatible blocks first now completed the prejudice-incompatible blocks and those who first completed the prejudice-incompatible blocks now completed the prejudice-compatible blocks. The order of blocks (prejudice-compatible or prejudice-incompatible first) as well as the key-assignments (good-left, bad-right or bad-right, good-left) were randomly assigned between subjects. As such, participants completed one of four possible IAT combinations. When participants made an error, they were shown a red “X” and asked to correct their response by pressing the other button. Reaction times were measured from the stimulus onset until participants indicated the correct response (Greenwald et al., 2003). A *D*-Score was computed according to Greenwald et al. (2003), dividing the reaction time differences for Block 3 and 6 (incompatible – compatible) by the pooled standard deviation of both blocks. The same was done for Blocks 4 and 7. The final *D*-Score was derived from the mean of these two scores with a positive value indicating faster reaction times in the compatible blocks compared to the incompatible blocks, which is interpreted as a pro-White bias. Negative scores indicate a pro-Black bias. The IAT showed satisfactory reliability (Cronbach’s alpha = .77, calculated from the two *D*-scores).

3.7.1.3. IAT Feedback. The original Javascript Code by Carpenter et al. (2019) was altered to calculate a *D*-Score within Qualtrics that was used to present a feedback statement to participants: “*Your data suggest [...] automatic preference for [Group A] over [Group B]*”. Two-thirds of the participants received a feedback statement with qualifiers based on conventions used on <http://www.implicit.harvard.edu>: *little to no* for $|D| \leq .15$, *a slight* for $.15 < |D| \leq .35$, *a moderate* for $.35 < |D| \leq .65$, and *a strong* for $|D| \geq .65$. The groups were imputed depending on the sign of the *D*-Score (i.e., “...preference for WHITE” or “...BLACK”). The remaining third of the participants received performance-independent

feedback: “Your data suggest *NO* meaningful automatic preference for either *BLACK* or *WHITE*”.

3.7.1.4. Procedure. All participants first provided informed consent and were informed about possible bonus payments and exclusions. Afterwards, participants were randomly assigned to either the “standard feedback” (2/3rd of participants) or the “no-bias feedback” (1/3rd of participants) condition. In both conditions, participants received a brief introduction to the Black-White IAT, were told that they would receive feedback on their IAT, and then completed the IAT. Next, all participants received their respective feedback, followed by an attention check item that asked them to indicate which of several feedback options they had just received. Finally, all participants completed the surprise scale (Cronbach’s alpha = .97, see Table 2), demographic information, and were given the chance to comment on the study. At the end, all participants were debriefed about the purpose of the study, including a detailed explanation pertaining to feedback scoring conventions of the IAT.

3.7.2. Results

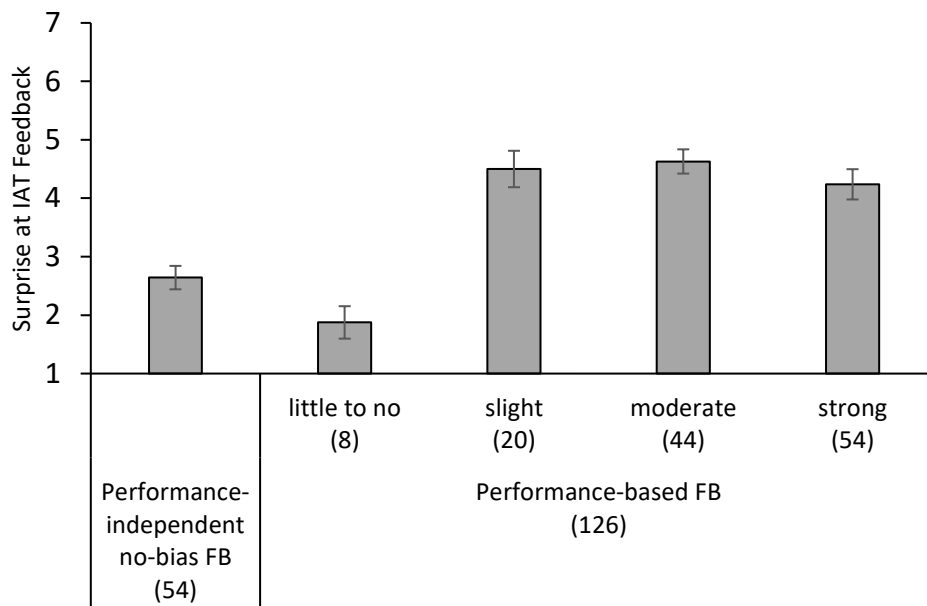
An independent-samples *t*-test on the average of the six surprise items showed that participants who received standard IAT feedback ($M = 4.27$, $SD = 1.71$) were more surprised than those who received no-bias feedback independent of performance ($M = 2.64$, $SD = 1.47$), $t(178) = 6.07$, $p < .001$, $d = 0.99$, 95% CI [0.65, 1.32].

To investigate whether participants were more surprised the stronger the wording of their bias feedback (Question 2, see Table 1), we looked at the correlation between surprise and the absolute feedback on the IAT (“Slight” = 2, “Moderate” = 3, “Strong” = 4) for the group who received bias feedback based on their performance. As explained in Table 1, we consider the effects of the strength of the wording of the feedback on surprise (i.e., Question 2) a different question than whether *any* bias feedback leads to more surprise than no-bias feedback (Question 1). Hence, participants who received “little to no” bias feedback were excluded from this analysis on the effects of feedback wording (See Figure 1). There was no

significant correlation between surprise and the feedback categories, $r(118) = -.08, p = .374$, and neither were there any other significant differences in surprise between the three bias feedback conditions, all $ps > .24$ (see no rise in surprise between the bars that indicate “slight”, “moderate”, or “strong” bias in Figure 1).

Figure 1

Study 1: Level of Surprise as a Function of IAT Feedback (FB) on Racial Preferences



Note. Error bars represent population-estimated standard errors. Numbers in parentheses represent number of participants who received said feedback.

3.7.3. Discussion

The results of Study 1 confirmed the so-far anecdotal observation that people react with surprise to IAT feedback indicating bias as opposed to no-bias feedback, independent of the severity and wording of the feedback. Even participants who were told that they had a “slight preference” were more surprised than participants who were told that they were not biased, but no less surprised than participants who were told they had a “strong preference”.

These findings are in line with the attention hypothesis, which states that, when not encouraged to pay attention to their own biases, people’s need to see themselves positively leads them to believe and expect to be unbiased, independent of the severity of the feedback.

At the same time, they provide first evidence against the feedback wording hypothesis, which would not predict surprise at low-bias feedback, and increased surprise the more severe the feedback. Although the social-desirability hypothesis predicts that participants will pretend to be surprised at any IAT bias feedback including low-bias feedback, it would also predict a relationship between harshness (= undesirability) of feedback and surprise (see Table 1). Hence, the results of Study 1 so far favor the attention hypothesis as an explanation for surprise reactions to IAT bias feedback.

3.8. Study 2

The aim of Study 2 was to experimentally test whether the attention or the feedback wording hypothesis is better suited to explain why people report surprise at IAT bias feedback (as Study 1 showed). To manipulate the attention people pay to their spontaneous biased reactions, participants either predicted their IAT scores or not. Hahn and Gawronski (2019) found that people tend to discover previously unattended biases when they predict IATs. Building on these findings, we reasoned that, if the attention hypothesis is true, participants who complete predictions should be less surprised at their IAT results than participants who do not complete predictions.

To test the feedback wording hypothesis, which states that the strength of bias communicated in the feedback predicts the level of surprise people report, we either gave participants standard IAT feedback as used on www.implicit.harvard.edu, or they received reduced feedback; thus complementing the correlational findings of Study 1 with an experimental manipulation. The reduced feedback simply said that the IAT indicated “an automatic preference” for one group over the other, without adding qualifications based on the degree of bias. If the feedback wording hypothesis is true, we reasoned, then participants who receive reduced feedback should be less surprised at their IAT feedback than those who receive standard feedback, independent of whether they predicted their IAT scores or not. The

preregistration for this study can be found at <https://osf.io/fh542/> and was registered on December 18th, 2018.

3.8.1. Method

3.8.1.1. Participants and Design. The study featured a 2 (prediction vs. no prediction) by 2 (standard feedback vs. reduced feedback) between-subjects design. A power analysis using G*Power (Faul et al., 2007) indicated that to find a small to medium effect size $f = 0.20$ with at least 90% power we would need a total sample size of $N = 265$. We rounded this number and aimed at recruiting 300 participants via TurkPrime. Participants who completed all parts of the study received a basic payment of US\$ 0.80 for participation and another US\$ 0.40 if they correctly answered two attention check items embedded in the study. Overall, 349 participants started the study of which 28 instantly opted out. As preregistered, 17 participants who responded too fast on the IAT (Greenwald et al., 2003) were dropped from the study, and another two participants were excluded from data analysis because they failed the attention check items. The final dataset consisted of 302 participants¹⁵ (55.6% Female; median age = 32, age range = 19-72 years). 97.4% of the participants reported being US-American citizens and 91.7% were born in the USA. 70.5% self-identified as White/Caucasian (9.9% Black/African American, 5.6% Latino/Hispanic, 5.0% East-Asian, 2.6% South-Asian, 6.3% more than one of the ethnic categories or another ethnicity).

3.8.1.2. Materials.

Prediction Task. Participants in the prediction condition first received an introductory text explaining the concept of implicit attitudes and the IAT as “spontaneous affective reactions” that often differ from what people would express when asked directly. Then they completed a trial prediction towards cats and dogs before they completed the actual prediction

¹⁵ With the final sample size of $N = 302$ (randomly assigned to one of four conditions; assuming alpha = 0.05; power = 80%), Study 2 had to reach a minimum effect size of $f = 0.16$ to show a significant effect, which could be reached at a critical F value of 3.87.

towards Black and White people. In it, they were asked to look at the pictures that represent the social categories Black and White (that were also used in the IAT), and to listen to their gut reactions to predict what an IAT on these social categories would show. The prediction question said “I predict that the IAT comparing my reactions to BLACK vs WHITE will show that my implicit attitude is...” with a scale ranging from -3 (“...a lot more positive toward BLACK”) to 3 (“...a lot more positive toward WHITE”). Participants in the no-prediction condition completed four similarly-formatted filler items on consumer preferences (casual vs. formal cloths, junk food vs. vegetables, texting vs. talking on cellphone, outdoor vs. indoor activities), but the topic of race was not mentioned.

IAT Feedback. All participants received feedback based on their performance on the IAT. While participants in the standard-feedback condition received feedback according to the same conventions as described in Study 1, participants in the reduced-feedback condition were only told that they showed either “little to no automatic preference” ($|D| < .15$) or “an automatic preference” ($|D| \geq .15$).

3.8.1.3. Procedure. Participants were first informed about all conditions of payment and participation, including possible bonus payments and exclusions, provided standard informed consent, and were randomly assigned to one of the four conditions. Participants then completed the predictions as explained above, or the filler items. Afterwards, all participants read the introduction to the IAT informing them that they will receive feedback, completed the same Black-White IAT as described in Study 1 (Cronbach’s $\alpha = .73$), and received feedback on their performance on the IAT depending on condition. Next, participants were asked to recall and indicate the feedback they received to increase attention to it. Finally, all participants completed the surprise scale (Cronbach’s $\alpha = .96$, see Table 1) with their feedback repeated on top. They ended the study with demographic information, a chance to comment on the study, and finally, a debriefing. This debriefing included an explanation and discussion on the different feedback conditions, similar to Study 1.

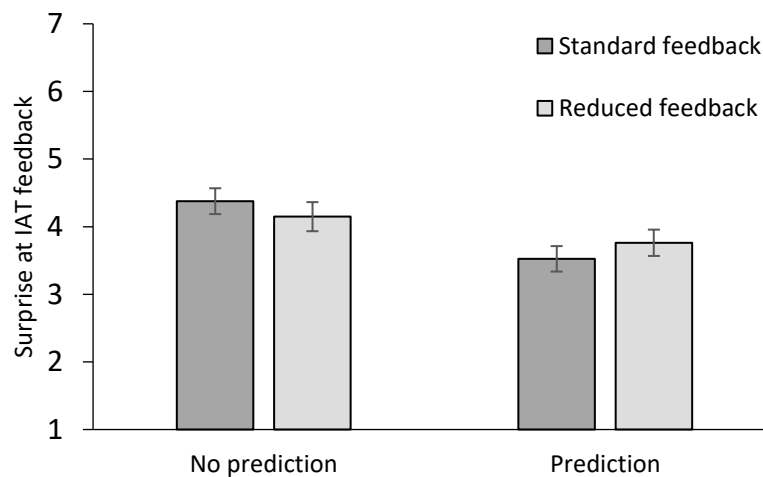
3.8.2. Results

3.8.2.1. Level of Bias. To explore whether the prediction manipulation prior to completing the IAT influenced participants' bias scores, we ran an independent-samples *t*-test on average *D*-scores (not preregistered). There was no statistically significant difference in *D*-scores between participants who predicted their IAT results ($M = 0.47$, $SD = 0.40$) and participants who did not ($M = 0.54$, $SD = 0.39$), $t(300) = 1.66$, $p = .098$, $d = 0.19$, 95% CI[-0.04; 0.42].

3.8.2.2. Surprise at IAT Feedback. We conducted a 2 (prediction vs. no prediction) x 2 (standard feedback vs. reduced feedback) between-subjects ANOVA on average responses on the surprise scale. Results showed a significant main effect of prediction, $F(1, 298) = 9.83$, $p = .002$, $\eta_p^2 = .032$, indicating that participants who completed predictions were less surprised at their IAT feedback than those who did not complete predictions (see Figure 2).

Figure 2

Study 2: Level of Surprise by Experimental Condition



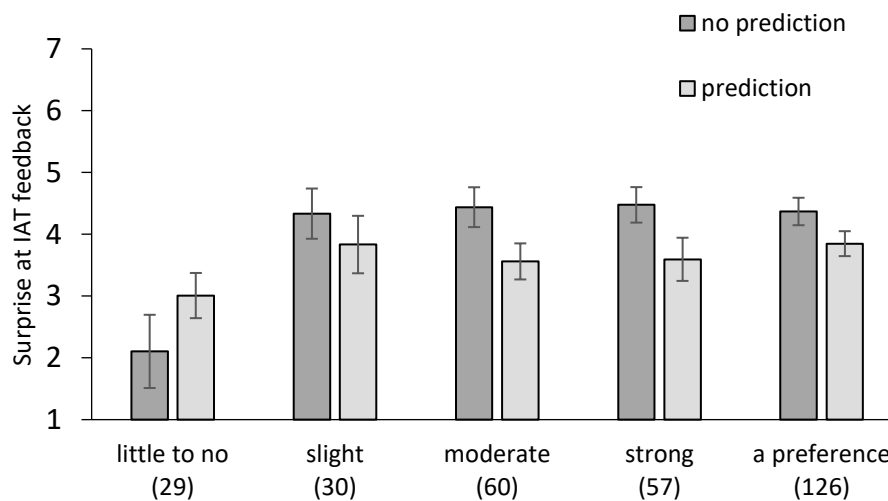
Note. Level of surprise as a function of IAT score prediction and type of IAT feedback received. Errors bars depict standard errors of estimated marginal means from a 2 (IAT score prediction vs. no prediction) x 2 (standard IAT feedback vs. reduced IAT feedback) ANOVA. $N = 302$, randomly assigned to condition.

A non-significant main effect of the type of feedback, $F(1, 298) < 0.01$, $p = .983$, $\eta_p^2 < .001$, further showed that there was no significant difference in surprise between participants who received feedback with potentially threatening and undesirable labels compared to bias feedback without specific qualifiers. Additionally, there was no significant interaction, $F(1, 298) = 1.39$, $p = .239$, $\eta_p^2 = .005$, confirming that independent of the type of feedback, completing predictions made participants less surprised at their IAT feedback.

3.8.2.3. Surprise and Strength of Bias Feedback. We again looked at the relationship between surprise and IAT feedback labels (see Study 1), including only participants who received standard feedback and excluding participants who received “little to no” bias feedback.

Figure 3

Study 2: Level of Surprise as a Function of IAT Score Prediction for the Five Different Performance-based Feedback Labels



Note. Level of surprise as a function of IAT score prediction for the five different performance-based feedback labels participants received. Errors bars depict standard errors of estimated marginal means from a 2 (IAT score prediction vs. no prediction) x 5 (levels of feedback) ANOVA. Numbers in parenthesis represent number of participants who received said feedback.

As can be seen in Figure 3, degree of surprise was again independent of the level of bias indicated by the feedback, $r(147) = .01, p = .896$ (see no rise in surprise for the three pairs of bars in the center of the figure). The same result emerged when we included participants in the reduced feedback condition who were told that they have “an automatic preference” (= 2; equivalent to the “slight” feedback category), $r(273) < -.01, p = .980$. Hence, this analysis continued to find no evidence that the strength of the bias communicated in the wording of the feedback was the reason for the surprise reactions (Gawronski, 2019).

3.8.2.4. Prediction Accuracy. Although not the primary purpose of the present study, we decided to examine the correlation between IAT score predictions and actual IAT scores. Hahn et al. (2014, see also Hahn & Goedderz, 2020) have argued that awareness of the cognitions reflected on IAT scores can only be analyzed in within-subjects correlations across several IATs and several predictions. This is because between-subjects correlations between predictions of one IAT and IAT scores require that people calibrate their attitudes consistently: A more biased person needs to use a stronger bias label than a less-biased person. Notwithstanding these limitations (we only administered one IAT, precluding within-subjects analyses), the between-subjects prediction accuracy in the current sample was significant, $r(159) = .42, p < .001$, indicating both awareness of bias and consistent social calibration across participants.

3.8.2.5. Surprise and Deviation of Predictions From Feedback. As a last step, we conducted several non-preregistered analyses, investigating whether participants’ reported surprise was a function of the deviation between their predictions and feedback. To this end, we coded all responses in the standard feedback condition to create a variable with three levels. They indicated whether participants (1) predicted IAT scores that exactly matched their specific IAT feedback (e.g., predicting a “slight preference for WHITE over BLACK” and receiving feedback indicating “a slight preference for WHITE over BLACK”, 22.0% of participants); (2) predicted their IAT scores to be in the same *direction* as the feedback (e.g.,

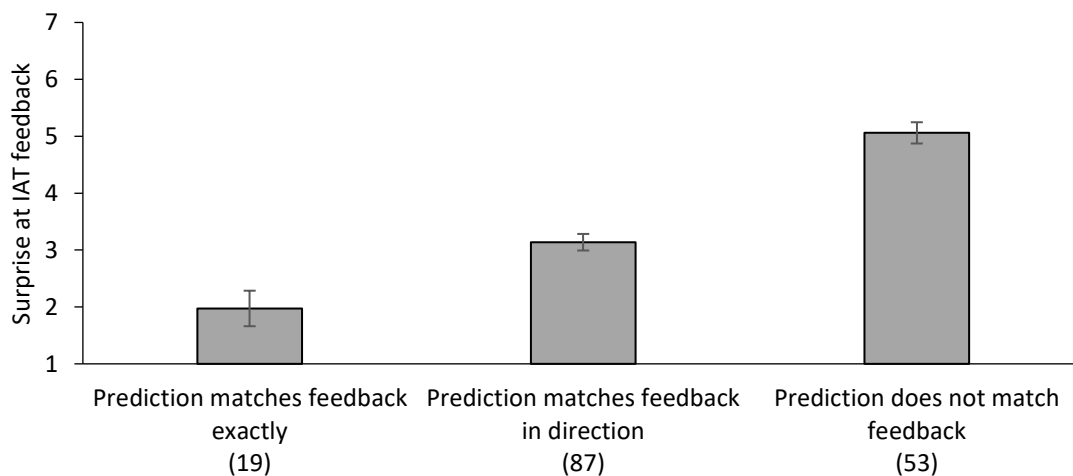
correctly predicting a preference for one group over the other, independent of the degree of bias, 51.2%); or (3) did not predict the same direction of bias as the feedback (e.g., predicting no preference, but receiving feedback of a preference for one group over the other, 26.8%).

For participants in the reduced feedback condition, 58.4% predicted the same direction of bias as the feedback, and 40.3% predicted their scores to be opposite to the IAT feedback or predicted no bias. One participant predicted showing “little to no bias” and actually received that feedback.

As can be seen in Figure 4, participants’ level of surprise significantly differed between the three categories, $F(2, 156) = 49.11$, $p < .001$, $\eta_p^2 = .386$.

Figure 4

Study 2: Level of Surprise by Correspondence of Predictions and Feedback for Participants in the Prediction Conditions



Note. Level of surprise as a function of correspondence between predictions and feedback for participants in the prediction conditions ($N = 159$). Error bars depict standard errors of estimated marginal means from a one-way ANOVA testing differences between the three conditions. Numbers in parenthesis represent number of participants in each category.

Examining contrasts further revealed that participants who received feedback that exactly matched their predictions were less surprised ($M = 1.97$, $SD = 0.89$) than those who predicted their IAT scores with different labels ($M = 3.14$, $SD = 1.39$), $F(1, 156) = 40.58$, $p < .001$, $\eta_p^2 = .206$. However, those participants whose feedback indicated bias in the same

direction as they predicted were still less surprised than the midpoint of the scale (4), $t(86) = 5.77, p < .001$, and than those who received feedback that was entirely inconsistent with their predictions ($M = 5.06, SD = 1.44$), $F(1, 156) = 97.21, p < .001, \eta_p^2 = .384$.

Consistent with Howell et al. (2013), we also found a strong relation between surprise and the absolute deviation of participants' feedback from their predictions, $r(159) = .55, p < .001$. Hence, in contrast to the analyses on continuous feedback, these analyses suggest that feedback wording does contribute to surprise at IAT feedback, at least to the degree that it deviates from participants' stated expectation (Gawronski, 2019; Howell et al., 2013).

3.8.3. Discussion

Study 2 tested whether attention or the strength of the feedback wording could explain why people react with surprise to IAT feedback indicating bias. The attention hypothesis states that people react with surprise because they rarely pay attention to their biases. The feedback wording hypothesis states that people react with surprise at the wording used to describe their biases. The data supported the attention hypothesis. Participants who were asked to pay attention to their spontaneous affective reactions by predicting their IAT results were less surprised at their IAT feedback than participants who did not predict their IAT results.

Concerning the feedback wording hypothesis, results were mixed. Participants were equally surprised when they received standard feedback as when they received reduced feedback, and strength of feedback was again uncorrelated with surprise. At the same time, people were more surprised the more their feedback deviated from their predictions. One interpretation of these findings is that, even though participants often choose different labels for their biases, it is not the *harshness* of the feedback that drives this effect. Another interpretation is that omitting all qualifiers in the feedback was not perceived as less harsh, such that our manipulation did not work as intended. We will return to this point in the general discussion.

Although the effect of IAT score predictions on surprise is a direct prediction from the attention hypothesis, it is also compatible with the social-desirability hypothesis if we assume that surprise reports are dishonest. Prediction might lower people's surprise not because it made them pay attention to their biases, but because it made them admit to biases ahead of time. This would make a pretend-surprise reaction superfluous, and it may shift perception such that admitting to biases would now become a socially desirable response. Additionally, the prediction manipulation may have presented the IAT in more socially desirable ways, such that the feedback became less threatening. Studies 3 and 4 aimed at investigating these alternative explanations.

3.9. Study 3

The purpose of Study 3 was to investigate whether the effect of IAT score prediction on surprise could be explained by attention to biased reactions, or by social desirability concerns. Specifically, the prediction procedure consists of three steps (see Table 3). Relevant for this study, it includes a manipulation to pay attention to one's biased reactions (explained in Step 1, executed in Step 2), and a request to predict these biased reactions on 1-7 scales (Step 3). The attention hypothesis states that Steps 1 and 2, the process of paying attention to one's biased reaction, is responsible for the reduced surprise reactions. However, according to the social desirability hypothesis, Step 3, the overt rating of this reaction on a scale, may be responsible for the effect. Specifically, concrete predictions may induce participants to "admit" to biases of which they are always fully aware, but which they would otherwise hide. It might also shift participants' perception such that admitting to biases becomes the more desirable response. If this hypothesis is true, we reasoned, then any question that induces participants to admit to harboring biases ahead of time should reduce reported surprise, even if it does not include any encouragement to pay attention to one's biased reactions.

To test these competing explanations, we designed a 2-by-2 design in which we independently manipulated whether participants were asked to predict their results on the

prediction scale or not (see Table 3, Step 3, prediction manipulation), and whether they were asked to pay attention to their spontaneous affective reactions or not (see Table 3, Steps 1 and 2, attention manipulation). If the completion of the prediction scale reduced surprise in Study 2 because it induced participants to admit to biases they knew all along and thus shifted perceptions of social desirability, then this should result in a main effect of the prediction manipulation in this study. On the other hand, if the effect is driven by attention to one's spontaneous reactions, then this should result in a main effect of the attention manipulation. The preregistration can be read at <https://osf.io/ze3pg/> and was registered on February 21st, 2019.

Table 3

Components of the Prediction Procedure Included in Each Condition in Study 3. The Exact Wording and All Materials Can Be Found on OSF.

		Conditions			
		Attention		No Attention	
		Prediction	No Prediction	Prediction	No Prediction
		Standard prediction		Control	
Prediction procedure					
Attention Manipulation	Step 1: Explanation of IAT as measure of spontaneous affective reactions	Yes	Yes	-	-
	Step 2: Pay attention to reactions to pictures used in the IAT				
Prediction Manipulation	Step 3: Completion of prediction scale	Yes	-	Yes	-

Note. The explanation in Step 1 included a non-threatening explanation of implicit attitudes as spontaneous reactions that may be different from explicitly endorsed attitudes. The no-explanation version simply introduced the IAT as a test of “implicit attitudes” that would reveal a preference for BLACK or WHITE. The control condition included filler items about consumer preferences without mentioning race or bias. Steps 2 and 3 were always completed twice, once hypothetically towards cats and dogs, and then towards BLACK and WHITE.

3.9.1. Method

3.9.1.1. Participants and Design. Study 3 consisted of a 2 (attention vs. no attention) by 2 (prediction vs. no prediction) between-subjects design. A power analysis using G*Power (Faul et al., 2007) based on the effect found in Study 2 ($\eta_p^2 = .032$) indicated that we would need at least 396 participants to find the attention effect on surprise with a power of 95%. Accordingly, we set TurkPrime to collect data from 400 participants, 100 per condition. The exclusion criteria were the same as described in Study 2. In total 461 participants started the experiment of which 29 immediately opted out. Another 32 participants met our preregistered fast-rate criteria and were excluded (Greenwald et al., 2003). Six participants failed the attention check embedded in the surprise scale and were also dropped from the final analyses. One person participated twice, and we only included their first set of data, leaving a final sample of 393 participants¹⁶ (48.6% Female; 0.3% non-binary; median age = 34, age range = 18-70 years). Most participants indicated having US-American citizenship (98.7%) and having been born in the USA (94.9%). The majority self-identified as White/Caucasian (74.8% 6.6% Black/African American, 4.3% Latino/Hispanic, 5.9% East-Asian, 2.0% South-Asian, 0.5% Middle-Eastern, 5.9% more than one ethnic category or another ethnicity).

3.9.1.2. Materials. The materials and procedure were based on Study 2, and we manipulated paying attention and predicting IAT scores by omitting individual steps of the prediction procedure. Participants in the attention-prediction condition completed the study exactly as participants in Study 2. They first saw a short explanation of the IAT as a measure of spontaneous affective reactions and were asked to reflect on biases in their reactions while looking at the pictures of Black and White people used on the upcoming IAT. Next, they were asked to rate their reaction by predicting how they would score on a 7-point scale.

¹⁶ With the final sample size of $N = 393$ (randomly assigned to one of four conditions; assuming alpha = 0.05; power = 80%), Study 3 had to reach a minimum effect size of $f = 0.14$ to show a significant effect, which could be reached at a critical F value of 3.87.

In the attention-no-prediction condition, the first part was identical, but the last part was missing. Specifically, the rating scale under the pictures (Step 3) was omitted and replaced by the sentence “Press ‘>>’ when you have thought about what an IAT will show about your spontaneous reactions towards BLACK and WHITE.” Hence, these participants were explained that the IAT measures biased affective reactions and then encouraged to pay attention to their biased reactions, but they were never asked to predict the feedback they expected on a scale.

Reversely, participants in the no-attention-prediction condition were only asked to predict how they would score on a test measuring their “implicit attitudes” towards Black and White people, but without explanation of IAT scores reflecting biased affective reactions to pictures of Black and White people, and without encouragement to pay attention to their gut reactions (Steps 1 and 2 were omitted). Participants in the control no-attention-no-prediction condition completed neither of the two steps. Instead, they completed the same filler items as described in Study 2.

3.9.1.3. Procedure. After providing informed consent and being informed about potential bonus payments, participants were randomly assigned to one of the four conditions and completed the respective tasks. Participants then completed the Black-White IAT (Cronbach’s alpha =.75), received feedback based on their performance, and completed the surprise scale (Cronbach’s alpha =.96, see Table 1). At the end, all participants filled out demographic information, were given the opportunity to provide feedback on the study, and saw a short debriefing.

3.9.2. Results

3.9.2.1. Level of Bias. A non-preregistered 2 (attention vs. no-attention) by 2 (prediction scale completion vs. no completion) between-subjects ANOVA on average *D*-scores showed that the manipulations prior to completing the IATs did not significantly influence participants’ level of bias. Neither the attention manipulation, $F(1, 389) < 0.01$, $p <$

.942, $\eta_p^2 < .001$, nor completing the prediction scale, $F(1, 389) = 0.03$, $p < .853$, $\eta_p^2 < .001$, nor the interaction, $F(1, 389) = 0.49$, $p < .487$, $\eta_p^2 = .001$ showed significant effects on participants' bias scores.

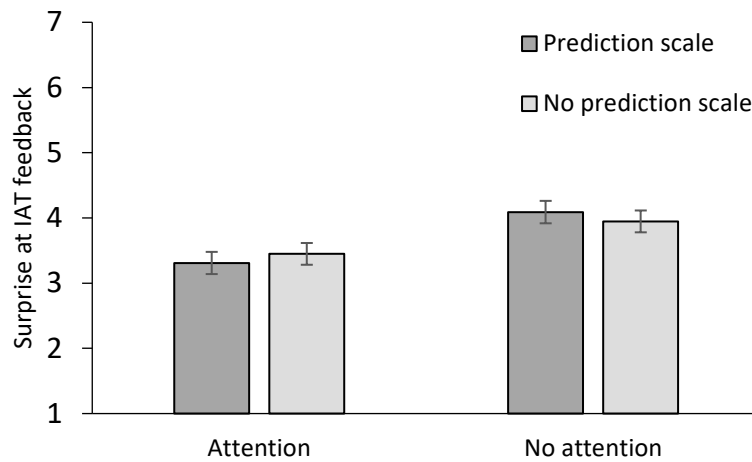
3.9.2.2. Surprise at IAT Feedback. A 2 (attention vs. no-attention) by 2 (prediction scale completion vs. no completion) between-subjects ANOVA on average responses on the surprise scale revealed a significant main effect of the attention manipulation, $F(1, 389) = 14.28$, $p < .001$, $\eta_p^2 = .035$ (see Figure 5). The presence of the prediction scale did not have a significant effect on reported surprise, $F(1, 389) < 0.01$, $p = .992$, $\eta_p^2 < .001$. Neither did the data reveal a significant interaction between attention and the presence of the prediction scale, $F(1, 389) = 0.697$, $p = .404$, $\eta_p^2 = .002$.

In line with the attention hypothesis, participants who were asked to pay attention to their own affective reactions while looking at pictures used in the IAT reported less surprise at their IAT feedback, independent of whether or not they completed the prediction scale. In contrast, induction to predict (and hence to “admit”) one's level of bias on the prediction scale ahead of time had no significant effect on participants' surprise in the absence of encouragement to pay attention to their biases. This last point contradicts the idea that completing the prediction scale shifted participants' perception such that admitting to biases became the more socially desirable response. If that were true, then prediction without attention should have reduced surprise, but this condition showed the most surprise.

3.9.2.3. Surprise and Strength of Bias Feedback. Replicating findings from Studies 1 and 2, surprise was again uncorrelated with strength of bias feedback, $r(354) = .07$, $p = .174$, nor were there any other significant differences in surprise depending on whether participants received feedback of having “a slight”, “a moderate” or “a strong” bias (all pairwise comparisons, $ps > .14$). However, replicating findings from Study 2 and Howell et al. (2013), participants were again more surprised the more their feedback deviated from their predictions, $r(192) = .65$, $p < .001$.

Figure 5

Study 3: Level of Surprise by Experimental Condition



Note. Level of surprise as a function of paying attention to pictures in IAT score prediction (attention vs. no attention) and using an IAT prediction scale (prediction scale vs. no prediction scale). Error bars depict standard errors of estimated marginal means from a 2 (attention vs. no attention) x 2 (prediction scale vs. no prediction scale) ANOVA. $N = 393$, randomly assigned to condition.

3.9.2.4. Prediction Accuracy. Within the two prediction conditions, we also looked at the between-subjects correlations between participants' predictions and their IAT scores. In line with Hahn and Goedderz (2023) and as preregistered, people in the present study were more accurate at predicting their level of bias in the attention condition where they saw pictures, $r(97) = .33$, $p = .001$, compared to the no-attention condition without pictures, $r(95) = .17$, $p = .096$. However, in line with observations that differences between correlations require much more power than finding differences between means (Judd et al., 2017), this difference was statistically not significant, $Z = 1.14$, $p = .128$. There was also a non-significant trend for people to predict more bias when they saw pictures ($M = 0.93$, $SD = 1.17$) than when they did not see pictures ($M = 0.60$, $SD = 1.27$), $t(190) = 1.86$, $p = .064$, $d = 0.27$, 95% CI[-0.02; 0.55].

As previously argued (Hahn et al., 2014), one of the problems with between-subjects correlations in discussions around awareness is that such correlations confound awareness

with accurate calibration. That is, the degree of a between-subjects correlation does not only depend on whether people are aware of their biases. It also depends on whether they know how biased they are compared with other people, such that more biased people would have to predict more bias than less biased people. Importantly, the attention hypothesis states that attention makes people recognize their (otherwise preconscious) biases, not that it helps them calibrate them accurately. Following this reasoning, in an additional non-preregistered analysis, we compared whether people more often recognized the direction of their biases (e.g., White over Black, no preference, or Black over White) that showed on the IAT (= 1) or not (= -1)¹⁷, independent of the specific comparative label they chose on the prediction scale. Supporting the attention hypothesis, significantly more people recognized their biases in the attention condition (80.4%) as compared to the no-attention condition (65.3%), $X^2(1, 192) = 5.58, p = .018$.

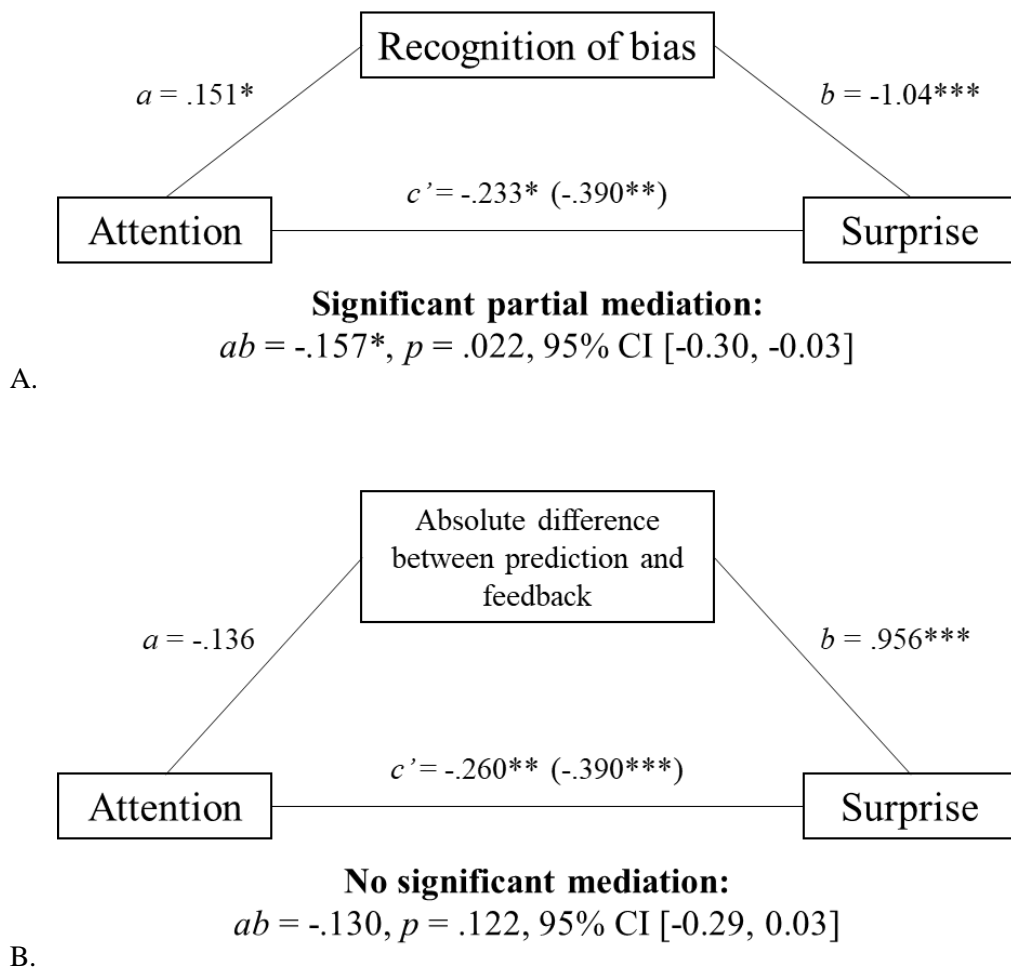
3.9.2.5. Mediation: Deviation from Expectations or Recognition of Bias?. The analyses so far suggest two possible explanations for why the attention manipulation reduces surprise. First, in line with the feedback wording hypothesis, attention may have lowered participants' surprise only to the degree that their predictions were consistent with the wording used in their IAT feedback. This interpretation is supported by the observation that surprise was a function of deviation of feedback from predictions. However, it stands at odds with the observation that predictions were not significantly more accurate in the attention condition (although this may be a power problem, Judd et al., 2017). A second explanation is in line with the attention hypothesis. It states that attention made people recognize their (formerly unattended) biases, and it was this recognition, independent of the accuracy of their specific predictions, that primarily lowered surprise. To examine these competing

¹⁷ Participants were also coded as recognizing their biases (1) when they predicted no preference and showed none, and as not recognizing their biases (-1) when they predicted bias but ended up showing no bias (i.e., $|D| < .15$).

explanations, we ran two non-preregistered mediational analyses on participants who predicted their IAT results (Attention-prediction and no-attention-prediction condition; $N = 192$). Results favored the attention over the feedback wording hypothesis.

Figure 6

Study 3: Mediation Analyses with Recognition of Bias (Panel A) or Deviation of Predictions from Feedback (Panel B) Mediating the Effect of Attention on Surprise



Note. Mediation analyses with recognition of bias (Panel A, recognition of bias = 1, no recognition of bias = -1), or the absolute difference between prediction and feedback (Panel B) mediating the effect of attention (attention = 1, no attention = -1) on surprise (scale from 1-7). Bootstrapping analysis indicated significant partial mediation for recognition of bias (Panel A) $ab = -.157, p = .022, 95\% \text{ CI } [-0.30, -0.03]$, and no significant mediation for the absolute difference between prediction and feedback (Panel B), $ab = -.130, p = .122, 95\% \text{ CI } [-0.29, 0.03]$. Values represent unstandardized path coefficients. The total effect is presented in parentheses.

* indicates significance at the $p < .05$ level, ** at the $p < .01$ level, and *** at the $p < .001$ level.

Figure 6, Panel A, shows that an analysis treating recognition of bias as the mediator showed significant mediation.

Participants in the attention condition were significantly more likely to recognize their biases than participants in the no-attention condition ($a = .151, p = .018$), and participants who recognized their biases were significantly less surprised than participants who did not recognize their biases ($b = -1.04, p < .001$). A bootstrapping analysis for the indirect effect based on 1000 bootstrap samples using R 3.6.1 (R Core Team, 2019) and the *mediation* package (Tingley et al., 2014) revealed a significant partial mediation effect, $ab = -.157, p = .022, 95\% \text{ CI } [-0.30, -0.03]$.

A mediation model treating the absolute difference between predictions and feedback as the mediator did not show evidence for significant mediation (see Figure 6 Panel B). In line with the above reasoning, participants' predictions did not deviate significantly more from their feedback in the attention condition than in the no-attention condition ($a = -.135, p = .115$), even though deviation of feedback from predictions was related to more surprise ($b = .956, p < .001$). The bootstrapping analysis using 1000 bootstrap samples revealed no significant mediation, $ab = -.130, p = .122, 95\% \text{ CI } [-0.29, 0.03]$.

Together, these two mediation analyses support the idea that making people pay attention to their spontaneous affective reactions lowers surprise at IAT feedback because it helps them discover their biases. Whether they specifically choose the same labels for their biases as the feedback indicates does not seem to explain lowered surprise in response to predictions.

3.9.3. Discussion

In Study 3, we tested all three competing hypotheses for why IAT score prediction lowers surprise at IAT feedback against each other. Both the attention and the social-desirability hypotheses can explain why prediction lowers surprise, but they predict different mechanisms for this effect. According to the attention hypothesis, predicting IAT scores

makes participants pay attention to their biases, which they otherwise rarely consider. In contrast, the social desirability hypothesis states that completing a prediction scale may induce participants to admit to biases they always know. Because revelation of bias is announced during IAT score prediction, and because there is a motivation to be accurate and consistent in one's report (Jussim et al., 1995), prediction may have diminished any value in acting surprised, and shifted participants' perception such that admitting to biases became the socially desirable response.

Results supported the attention hypothesis. People reported less surprise at their IAT feedback when asked to pay attention to their biases prior to IAT completion, even when they never completed the prediction scale. In contrast, merely seeing and completing the prediction scale without an induction to pay attention to one's biases did not make people report less surprise at their IAT feedback. These findings speak against the idea that people are generally aware of their biases but act surprised because such a reaction sounds desirable. Instead, it supports the hypothesis that people discover their biases when they are asked to pay attention to their gut reactions; and this recognition then leads them to report less surprise.

Two non-preregistered mediation analyses additionally showed that recognition of bias was a better explanation for the effect of attention on surprise than lower deviations of predictions from feedback. These analyses provide additional support against the feedback wording hypothesis and in favor of the attention hypothesis. It was recognition of bias, not acceptance of the feedback wording, that explained why attention to biased reactions reduces surprise.

3.10. Study 4a

Studies 4a and 4b aimed at examining whether surprise was reduced after predicting IAT results in Studies 2 and 3 because people were encouraged to pay attention to their own reactions (attention hypothesis), or because they read an explanation that the IAT measures spontaneous reactions (Step 1, see Tables 3 and 4, social desirability hypothesis). Specifically,

the explanation read as follows: “[...] In addition to the things you say when you are asked about your attitudes, you may have spontaneous reactions toward people at first that you wouldn't always express.[...] For instance, you may have a more positive affective reaction toward a picture of a skinny top model than toward a picture of a regular woman, even though you may not think or say that skinny top models are better people than regular women.[...] [Your implicit attitudes] may be different from the explicit attitudes you would report when you have had time to think about them.”.

Reading this information could potentially lead people to be less surprised at their IAT feedback for at least two reasons. First, it may lead them to expect their IAT biases to differ from their explicit attitudes because the explanation says so. Second, the explanation might present IAT results as less of a threat to participants' values and beliefs. Both of those effects may make it less socially undesirable to admit to bias and report lowered surprise.

To investigate this potential alternative explanation in Study 4a, participants either completed the standard prediction procedure as implemented before; only read the explanation but never predicted their IAT results; or neither read an explanation nor predicted their IAT results, resulting in 3 conditions (explanation and prediction, only explanation, control). Unexpectedly, this study failed to replicate the original prediction effect shown in both Studies 2 and 3. There were no significant differences in reported surprise between conditions, $F(2, 380) = 0.11, p = .900, \eta_p^2 = .001$.

Because this null-result could be a random false-negative (Lakens & Etz, 2017), and the effect replicated once before, we decided to rerun a slightly altered version of Study 4a, but to include Study 4a in a final mini-meta-analysis (Goh et al., 2016) to investigate whether the main prediction effect still holds when this failed replication is included (see Figure 8). The preregistration for this study can be found at <https://osf.io/t5wdn/> and was registered on February 26th, 2019. A more detailed description of the sample and the results can be found

in the supplemental materials section. All data, analysis, materials and details on the sample for Study 4a are available on OSF.

3.11. Study 4b

As our second attempt to test whether predictions reduced surprise in Studies 2 and 3 due to attention or due to a non-offensive explanation, Study 4b independently manipulated receiving an explanation about implicit attitudes (Step 1) and paying attention to one's reactions by predicting IAT results (Steps 2 and 3) in a 2-by-2 between-subjects design. Hence, Study 4b featured both a condition where people only paid attention to their reactions, but never read any explanation (No-explanation-attention condition), and a condition where they only read an explanation, but never paid attention to their own reactions (Explanation-no-attention condition). See Table 4 for a complete dissociation of the two effects.

Table 4

Components of the Prediction Procedure Included in Each Condition in Study 4b. The Exact Wording and All Materials Can Be Found on OSF.

		Conditions			
		Explanation		No Explanation	
		Attention	No Attention	Attention	No Attention
		Standard Prediction		Control	
	Prediction procedure				
Explanation Manipulation	Step 1: Explanation of IAT as measure of spontaneous affective reactions	Yes	Yes	-	-
Attention Manipulation	Step 2: Pay attention to reaction to pictures used in the IAT Step 3: Completion of prediction scale	Yes	-	Yes	-

Note. The explanation in Step 1 included a non-threatening explanation of implicit attitudes as spontaneous reactions that may be different from explicitly endorsed attitudes. The no-explanation version simply introduced the IAT as a test of “implicit attitudes”.

If the attention hypothesis holds true, then this should result in a main effect of the attention manipulation, independent of reading an explanation or not. In contrast, if the social desirability hypothesis is true then this should result in a main effect of the explanation manipulation. In this case, participants who receive the non-threatening introductory text should report less surprise at their IAT results, independent of whether they are asked to pay attention to their biases and predict their IAT scores or not. The preregistration for this study can be found at <https://osf.io/h9p6j/> and was registered on April 2nd, 2019.

3.11.1. Method

3.11.1.1. Participants and Design. Study 4b featured a 2 (explanation vs. no explanation) by 2 (attention vs. no attention) between-subjects design. The attention condition always included predictions of IAT scores based on reactions to pictures. We again aimed at recruiting 400 participants (100 per condition) corresponding to a 95% chance of finding the attention effect from our prior studies. Four-hundred and forty-nine participants started the study, of which 27 instantly opted out. Following the same preregistered exclusion procedure as in Studies 2 and 3, 22 participants were removed due to exceeding speed on the IAT (Greenwald et al., 2003). Another four participants failed to answer the attention check item embedded in our surprise scale, leading to a final sample of 396 participants¹⁸ (54.3% Female; 0.5% non-binary; median age = 35, age range = 18-84 years). Most participants held US-American citizenship (98.2%) and were born in the United States (93.7%). 72.2% of our participants self-identified as White/Caucasian (8.8% Black/African American, 8.1% Latino/Hispanic, 4.3% East-Asian, 1.3% South-Asian, 0.3% Middle-Eastern, 5.1% more than one or none ethnic category).

¹⁸ With the final sample size of $N = 396$ (randomly assigned to one of four conditions; assuming $\alpha = 0.05$; power = 80%), Study 4b had to reach a minimum effect size of $f = 0.14$ to find a significant effect, which could be reached at a critical F value of 3.87.

3.11.1.2. Materials and Procedure. After being explained the potential bonus payments, participants provided informed consent and were randomly assigned to one of four conditions (see Table 4). Participants in the explanation-attention condition completed the full prediction procedure including all three steps. Participants in the explanation-no-attention condition read the non-threatening intro cited in the introduction to Study 4a, but did not go on to observe their own reactions and predict their scores. Participants in the no-explanation-attention condition only read that they would complete a test that measures their “implicit attitudes” without further information and were then encouraged to observe their reactions towards pictures of Black and White people to predict its results. Participants in the control condition went straight to the IAT without an explanation or predictions.

After finishing the respective tasks, all participants completed a Black-White IAT (Cronbach’s alpha = .68), received feedback based on their performance, and were asked to fill out the surprise scale (Cronbach’s alpha = .94). Finally, all participants completed a questionnaire on demographic information, were given the chance to provide feedback on the study, and were debriefed.

3.11.2. Results

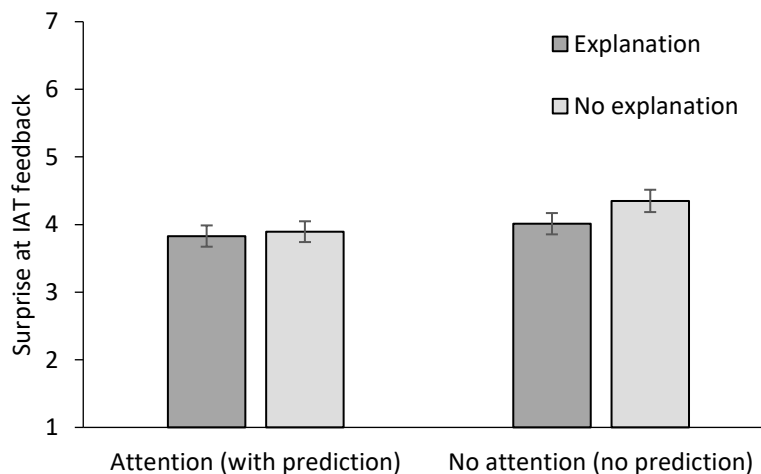
3.11.2.1. Level of Bias. A non-preregistered analysis on level of bias as a function of conditions showed that participants’ bias scores were not significantly affected by reading the explanation, $F(1, 392) = 1.17$, $p = .279$, $\eta_p^2 = .003$, or paying attention by predicting IAT results, $F(1, 392) = 0.38$, $p < .540$, $\eta_p^2 = .001$. There was no significant interaction between explanation and attention, either, $F(1, 392) < 0.01$, $p = .997$, $\eta_p^2 < .001$.

3.11.2.2. Surprise at IAT Feedback. We conducted a 2 (explanation vs. no explanation) by 2 (attention vs. no attention) between-subjects ANOVA on average responses on the surprise scale. This analysis supported the attention hypothesis with a significant main effect of attention, $F(1, 392) = 4.05$, $p = .045$, $\eta_p^2 = .010$. Participants who were asked to pay attention to their spontaneous affective reactions toward stimuli by predicting their IAT scores

were less surprised at their IAT feedback than participants who did not predict their IAT scores (See Figure 7). Overall, reading the explanation did not significantly lower surprise, $F(1, 392) = 1.61$, $p = .205$, $\eta_p^2 = .004$, and we did not find a significant interaction either, $F(1, 392) = 0.74$, $p = .391$, $\eta_p^2 = .002$. However, descriptively, participants who only read the explanation were somewhat less surprised than those in the control condition, yet more surprised than those who only paid attention to their reactions but never read any explanation (see Figure 7).

Figure 7

Study 4b: Level of Surprise by Experimental Condition



Note. Level of surprise as a function of attention to reactions and IAT explanation. Error bars depict standard errors of estimated marginal means from a 2 (attention vs. no attention) x 2 (explanation vs. no explanation) ANOVA. $N = 396$, randomly assigned to condition.

Follow-up simple effects revealed that the attention-only condition differed significantly from the control condition, $F(1, 392) = 3.97$, $p = .047$, $\eta_p^2 = .010$, while the explanation-only condition did not differ significantly from the control condition, $F(1, 392) = 2.23$, $p = .136$, $\eta_p^2 = .006$. At the same time, participants in the explanation-only condition were not significantly more surprised than participants who also paid attention to their reactions, $F(1, 392) = 0.69$, $p = .407$, $\eta_p^2 = .002$. Hence, although the explanation

descriptively lowered surprise, the resulting level of surprise fell non-significantly between all other conditions, whereas attention to reactions without explanation did lead to a significant reduction of surprise compared to control.

3.11.2.3. Surprise and Strength of Bias Feedback. Following the same procedure as in Studies 1-3 we looked at the level of surprise as a function of strength of bias feedback, excluding participants who received “little to no” bias feedback. Contrary to our prior findings we found a small but significant correlation between surprise and degree of bias, $r(353) = .12$, $p = .021$. In this study, participants were somewhat more surprised at their feedback the more bias it suggested.

3.11.2.4. Prediction accuracy. Participants in this study were again able to predict their IAT results, $r(200) = .40$, $p < .001$. Prediction accuracy did not differ as a function of whether participants read an explanation, $r(100) = .39$, $p < .001$, or not, $r(100) = .43$, $p < .001$, $Z = -0.33$ $p = .369$.

3.11.3. Discussion

The purpose of Study 4b was to investigate whether the prediction effect observed in Studies 2 and 3 was due to the fact that participants were encouraged to pay attention to their spontaneous affective reactions (attention hypothesis), or because of the non-threatening and more socially desirable explanation of implicit attitudes as different from explicit attitudes (social desirability hypothesis). To test these two hypotheses against each other, we independently manipulated whether participants read an explanatory text or not, and observed their reactions towards sets of pictures by predicting their IAT results or not. Results were again in line with the attention hypothesis. Participants who were asked to pay attention to their spontaneous affective reactions were less surprised at their IAT feedback independently of whether they read the non-threatening explanation or not. Only reading the explanation did not significantly lower people’s surprise compared to the control condition. However, it is noteworthy that, descriptively, participants reported somewhat less surprise at their feedback

when provided with a non-threatening explanation; but this explanation alone was neither necessary nor sufficient to lower surprise. And importantly, observing one's reaction without any non-threatening explanation *was* sufficient to reduce surprise.

Unexpectedly, and in contrast to Studies 1-3, participants reported more surprise the harsher the feedback they received. However, this correlation remained low and did not show in any of the other studies, so we treat it as a potential false-positive.

Overall, these findings again support the attention hypothesis and challenge a social desirability explanation. If people in fact reported less surprise in the previous studies because the non-threatening explanation encouraged them to think that admitting to biases is now desirable, simply reading this explanation should have been sufficient to lower surprise.

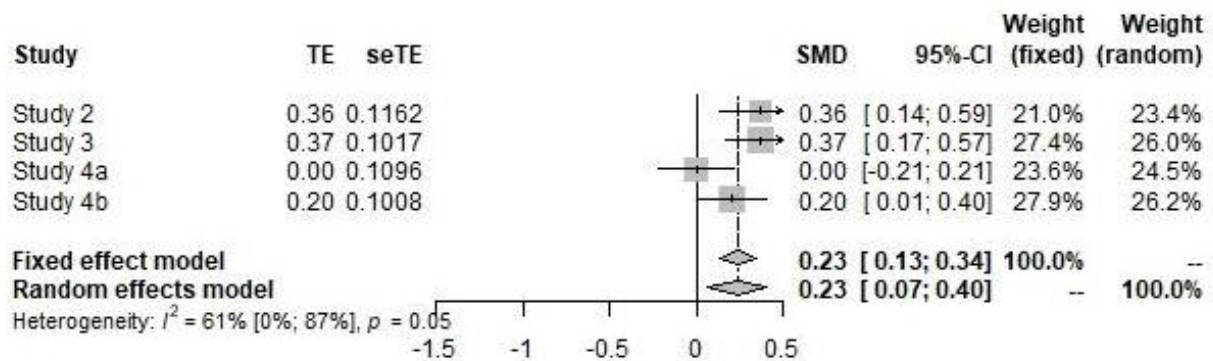
3.12. Additional Analyses

3.12.1. Meta-Analysis

Given the null results of Study 4a, we additionally wanted to investigate whether the overall attention effect would hold across all conducted studies when accounting for the failed replication (Lakens & Etz, 2017). To this end, we meta-analyzed the four studies that included the attention effect (Studies 2-4b, Goh et al., 2016) using R 3.6.1 (R Core Team, 2019) and the *meta* package (Balduzzi et al., 2019). We used a fixed and a random effects model in which the mean effect size (Cohen's *d*) for the main effect of attention in each study was weighted by sample size. Supporting the attention hypothesis, both the fixed and random effects model showed a significant effect for the attention manipulations on surprise including the failed replication (fixed effects model: $M d = .23$, 95% CI [.13, .34]; random effects model: $M d = .23$, 95% CI [.07, .40]; see Figure 8).

Figure 8

Forest Plot for the Meta-analysis of the Attention Effect over All Conducted Studies



3.12.2. Non-White Participants and Participants with Pro-Black Biases

The studies presented in this paper are supposed to present a general effect about intergroup bias, and we hence sampled our participants and preregistered our analyses independent of the majority or minority status of either the participants or the feedback. However, the interested reader may wonder whether White and non-White participants would react similarly in response to receiving pro-Black or pro-White bias feedback, as previous research has shown variations in defensive reactions (e.g., Howell et al., 2015). To investigate this point, we conducted a series of non-preregistered, exploratory analyses across all studies. All results are described in detail in the supplemental materials section. They showed that none of the central results reported in Studies 1-4 interacted with self-reported ethnicity (White vs. non-White) of the participants (all F s < 1.2, all p s > .29). Additionally, all main results from Studies 1-4 replicated independently on the non-White samples. Furthermore, when combing all relevant data from all studies to yield a more powerful sample, the answers to the three questions from Table 1 showed largely similar patterns on White and non-White participants who showed either pro-White or pro-Black bias separately. Specifically, independently of whether participants were White or non-White, or showed pro-White or pro-Black bias, they always reported more surprise when their feedback indicated any bias

(including low levels of bias) than when it indicated no bias (Question 1). In line with findings from Studies 1-3, none of the subsamples showed significant correlations between the reported level of surprise and the strength of the bias feedback (Question 2). Finally, regarding Question 3, whether paying attention to one's biases reduces surprise at IAT feedback, results followed the overall pattern for all subsamples, except for White participants who received pro-Black bias ($N = 59$). This may suggest that this small sample expected to show pro-White biases after predicting IAT results.

All statistical details on all analyses can be found in the supplemental materials section. Although these results can at best be considered suggestive, given their exploratory nature and the combination of all non-White participants into one sample, they do suggest that many intergroup biases, not just those that indicate undesirable pro-majority bias, but also those that seem less undesirable and may cause less defensive responding (Howell et al., 2015), may be preconscious and hence surprising for many majority and minority group members. More specific and targeted research is needed to investigate and complement these findings.

3.13. General Discussion

Awareness that racial biases are widespread across society has been growing (Gallup, 2021). Echoing these developments, social-cognitive research, too, has demonstrated that acknowledgement of implicit bias is possible with simple attention manipulations (Hahn & Gawronski, 2019). At the same time, however, many people seem to react to racial bias feedback on IATs defensively (Howell et al., 2015; Howell et al., 2017; Vitriol & Moskowitz, 2021) and with surprise (Hillard et al., 2013; Schlachter & Rolf, 2017). This can be read as indicating that people do not know that they harbor any racial biases (Gawronski, 2019; Krickel, 2018; Lane et al., 2007), hence supporting the notion that implicit racial bias scores reflect “unconscious” attitudes. The purpose of the current set of studies was to shed light on this apparent contradiction. We proceeded in three steps. First, Study 1 demonstrated the so-

far anecdotal observation that participants were in fact more surprised when their feedback indicated bias than when it indicated no bias. Second, Studies 2-4b showed that participants were less surprised at their bias feedback when they predicted their results on an IAT racial bias test prospectively, by looking at the stimuli used on the test and observing their own (biased) reactions. Third and last, Studies 3 and 4 dissociated whether attention, social desirability, or the wording of the feedback were better suited to explain this prediction effect. Results favored the attention hypothesis. Across the four studies, we show that people are surprised at racial bias feedback as long as they are not encouraged to pay attention to their biased reactions. This suggests that the cognitions reflected on implicit measures may often be “preconscious”: They are generally accessible, but people rarely pay attention to them (Hahn & Goedderz, 2020).

Alternatively, according to the social desirability hypothesis, people think surprise is the socially desirable response to bias feedback. Completing the prediction slides in the present studies might have reduced participants’ reported surprise (but not actual surprise) because the prediction procedure induced them to admit to biases they would otherwise hide. This may have caused them to be prepared for their biases (which they knew all along) to be revealed, and shifted their perception of what a socially desirable response to this revelation would be.

Voicing surprise at anti-Black bias feedback is likely more socially desirable than admitting to harboring biases, and this tendency most certainly contributed to the observed effects. However, our data did not confirm this to be the *main* explanation for the surprise effect. First, surprise was uncorrelated with the social desirability of the feedback. Participants were always more surprised when they received bias feedback as opposed to no-bias feedback, but how much bias the feedback communicated was unrelated to surprise. Second, Study 3 independently manipulated inducing participants to state their biases before the IAT (prediction) and paying attention to spontaneous reactions (attention). Our reasoning was that

if the prediction procedure were simply a method to induce participants to “admit” to biases before IAT completion and make this admission seem socially desirable, then prediction without attention should suffice to reduce surprise, whereas attention without prediction should not suffice to reduce surprise. Results did not support this reasoning and instead favored the attention hypothesis. Participants reported less surprise even when they only thought about their affective reactions toward specific stimuli, but never saw or completed a prediction scale. In contrast, merely predicting IAT results on a scale without paying attention to reactions toward pictures did not significantly reduce participants’ surprise. Finally, Study 4b showed that the prediction effect was independent of the social desirability of the explanation of the construct of implicit bias. Only paying attention to spontaneous reactions without reading any prior explanation already reduced participants’ surprise. Conversely, when participants only read a non-threatening explanation of the IAT as a measure of spontaneous affective reactions, surprise was not significantly different from either the control condition or the prediction condition. This last finding indicates that non-threatening explanations may help reduce surprise, but not as successfully as active attention to one’s reactions.

In sum, our studies did not confirm the idea that participants act surprised at IAT feedback because this answer seems more socially desirable. Instead, it favors the attention explanation: A simple manipulation encouraging people to notice their spontaneous affective reactions made people less surprised at bias feedback. This suggests that people are surprised at IAT feedback because they do not pay attention to their biases chronically – suggesting that these biases are often preconscious (Dehaene et al., 2006).

Yet another explanation for why people might react with surprise at IAT feedback is the feedback wording hypothesis. It says that people disagree with the labels chosen to describe their biases (Gawronski, 2019). Participants in the studies by Hahn et al. (2014) tended to be socially miscalibrated in their results. They disagreed on which biases should be

called “mild” or “strong”, even as they predicted the patterns of their individual IAT scores accurately. This lack of consensus on what to call a specific reaction might explain why people are surprised at the feedback that they get. Additionally, people are motivated to see themselves as above average on desirable traits and below average on undesirable traits (Alicke et al., 1995). This motivation may lead to surprise at any feedback that suggests that people may have less socially desirable preferences and biases than others.

Empirically investigating this claim led to mixed results. On the one hand, Studies 1-3 did not find any relationship between surprise and the strength of the labels used in the feedback, and Study 4b showed a significant but very small correlation. On the other hand, participants were always more surprised when their feedback included a different qualifier than they had predicted (e.g., “moderate” vs. “strong”). These results do support the notion that surprise is partly a response to the fact that participants have different ideas of what to call their biases than standard IAT feedback communicates. However, it contradicts the notion that the strength or harshness of the feedback is specifically responsible for the surprise reaction. In line with this, a mediation analysis further showed that consistency between predictions and feedback labels did not significantly explain the effect of the attention manipulation on surprise. Instead, it supported the attention hypothesis by showing that the attention manipulation reduced participants’ surprise because participants more often recognized their biases compared to the no-attention condition. Lastly, experimentally altering the IAT feedback by omitting all qualifiers and telling all participants with D scores above $|.15|$ that they have “an automatic preference” for one group over the other did not significantly lower surprise compared to standard feedback. In this case, hearing that the IAT suggests “a preference” may have been a bad operationalization for non-threatening feedback; “an automatic preference” might be perceived as more offensive than having a “mild” or “moderate” bias.

Given the remaining ambiguity of these results, we find it important to note that how IAT feedback is communicated may still be an important factor that influences how people react to IATs. For instance Vitriol and Moskowitz (2021) show in their studies that if IAT results are communicated such that participants feel less blamed and perceive more control over their biases, it reduces their defensiveness and increases bias awareness. In the present studies, we only wanted to test the effect of the feedback qualifiers (“slight”, “moderate”, and “strong”) specifically, and found only mixed to negative results. Additional research is needed to further investigate whether communicating feedback in an entirely different way might make people react with less surprise and defensiveness. For instance, instead of using qualifiers that are chosen arbitrarily, it might be helpful to put the feedback in context by telling people where their IAT scores rank in comparison to other people’s scores. Additionally, instead of presenting IAT effects as “preferences” for one group over another, they could be framed as “automatic reactions”. Such changes in feedback communication might be a more meaningful interpretation of the available data, it could help people understand the meaning of their biases in comparison to others, and it could make them react less defensively to bias feedback.

Yet another explanation for why people react with surprise at IAT results could lie within problems of the IAT as a measure. That is, task-specific variance of the IAT may have led to distortion of bias scores and thus inaccurate bias feedback. Whether the IAT should be used as a measure of individual bias and thus a basis for individual feedback remains a contentious debate (Kurdi et al., 2020; Schimmack, 2019). However, we believe that task-specific variance cannot explain the attention effect on surprise observed in our studies. First, the fact that participants were able to predict their IAT results suggests that at least in the present studies, the IAT showed a certain degree of validity despite its methodological constraints. Second, if surprise was simply a reaction to “invalid” feedback from the IAT, a

prediction manipulation shouldn't have reduced it, since the feedback would have remained just as invalid either way.

In sum, our studies indicate that people are surprised at IAT feedback because they often remain inattentive to their biased reactions towards people. Once they are encouraged to pay attention to these reactions, they discover their biases, and surprise decreases. This suggests that racial biases, such as those reflected on implicit evaluations, are often preconscious. Although they are generally accessible and reportable, people often fail to pay attention to them until they are encouraged to do so.

3.13.1. Limitations

The present studies have several limitations. First, we focused on surprise and awareness of racial biases reflected on IAT scores. While we acknowledge the many criticisms surrounding the IAT and the construct of “implicit bias” (Blanton et al., 2007; Hahn & Gawronski, 2018), however, we believe that our findings may speak to a more general phenomenon. Whether or not the IAT measures them accurately, most humans are likely to hold stereotypic and prejudicial associations with different racial groups that are activated automatically. And independent of whether those automatic biases translate directly to discriminatory behavior (Kurdi, Seitchik et al., 2019), they are likely to show themselves in some responses and reactions. From this perspective, we believe our findings that people tend not to pay attention to their own biases has implications beyond the specific reactions toward IAT feedback we studied here. Although general awareness of racial biases in society may be on the rise (Gallup, 2021), many people may still be resistant to confront their own racial biases, resulting in surprise or even shock at feedback of having shown a bias, even when these biases are easily observable. And this phenomenon might apply to biases beyond those captured by IATs as well. Importantly, as the present studies demonstrate, however, a simple encouragement to pay attention to a biased reaction can change this effect, reduce surprise, and lead to acknowledgement of bias (Hahn & Gawronski, 2019). As we explain below, this

insight could prove useful in designing bias intervention and education programs. Future research will have to show how widely our findings that people tend to leave their biases unattended generalizes to other instantiations of racially biased behavior and thoughts.

Second, we only examined racial biases toward Black and White people which may question the generalizability of our findings to other attitudinal domains. We postulate that the specific surprise effect will primarily emerge whenever people's automatic cognitions conflict with their personal standards, such that paying attention to them might be threatening to a person's self-concept. In those cases, encouraging people to pay attention to their reactions before a test should reduce surprise. In contrast, people should be unsurprised at IAT feedback when their explicit evaluations are already based on their spontaneous reactions, such that they match. For instance, Nosek (2007) found high implicit-explicit correspondence in political attitudes, or attitudes towards Coke vs. Pepsi. Because of this correspondence, we would predict little surprise in response to IAT feedback in these domains. Which kinds of attitudes in which domains are subject to such divergences and concerns will likely show variation across cultures, countries, and individuals.

Third, it is important to note that the samples in the reported studies were majority White American participants (71-75%) and the majority of participants received feedback of having a pro-White bias (88- 90%). This limits the generalizability of our findings to other populations. Exploratory analyses across these categories showed that the patterns of results presented in each study and response to our three main questions from Table 1 were similar across White and non-White (American) participants who showed pro-White or pro-Black bias. While many reactions to intergroup bias feedback – from defensive rejection to emotional responses – will very likely differ as a function of a person's own racial status and the bias in question (e.g., Howell et al., 2015), these results suggest that the specific reaction of surprise may be more general. That is, people may pay little attention to the types of intergroup biases that are reflected on IATs generally. As a result, feedback about them may

be surprising independent of whether such feedback is at odds with one's beliefs and values or social desirability, unless one has been encouraged to pay attention. Future research is needed to investigate these questions more directly.

On a more general level, we have reason to believe that our general conclusion – that implicit evaluations often reflect preconscious attitudes to which people may or may not pay attention – will hold across different targets, instruments, and populations. Future research is needed to investigate the generalizability of our findings and conclusions.

3.13.2. Why Are Biases Left Unattended?

One question the present findings pose is how people manage to keep their biases out of awareness (Hahn & Goedderz, 2020). That is, most of the people who participated in these studies have likely met people with different backgrounds in their lives. Why, then, are they discovering new information when asked to observe their reactions? We see several possibilities. The most direct application of the idea of “unattended biases” is that people simply direct away their attention from their biases. Many real-life situations where people may show biases (and thus have a chance to observe them) are ambiguous with respect to the source of the bias, such that identifying a racial bias might need specific motivation. Another, compatible, possibility is that they misattribute their biases to other aspects in the situation. For instance, personal experience with IAT studies suggest that many participants attribute their IAT biases to the order of the blocks in which the IAT is completed. This suggests that they do initially notice their biased reactions, but attribute them to other aspects of the situation than race. This misattribution process might be even more common in real-life situations where there are many more aspects of the situation to which one could attribute one's bias. The current study cannot distinguish between these different interpretations. Importantly, however, regardless of why participants did not pay attention by themselves, once encouraged, participants in this study did notice their biases and reported less surprise at

IAT scores. We hope this research contributes to more research on why people are so often blind to biases in their reactions and behavior.

3.13.3. Theoretical Implications for Implicit Bias Research

We believe our studies question the common portrayal of implicit measures as capturing “attitudes people may be unable or unwilling to report” (<https://implicit.harvard.edu>, 2020). This description summarizes two theoretical perspectives proposed by dual-process models which differ in terms of the role they attribute to consciousness in the dissociation of implicit and explicit measures. On the one hand, some older conceptualizations have often claimed that implicit evaluations reflect unconscious attitudes people are unable to report (e.g., Greenwald & Banaji, 1995; Lai et al., 2013; Nosek et al., 2002). On the other hand, several dual-process models propose that people might be well-aware of the cognitions reflected on implicit evaluations, but consciously decide not to report those on explicit measures (Fazio, 2007; Fazio & Olson, 2003; Gawronski & Bodenhausen, 2006, 2011).

The present research challenges both perspectives. Assuming that the cognitions reflected on implicit measures are unconscious stands at odds with the present data for at least two reasons. First, people’s surprise at IAT feedback was reduced when they paid attention to their biases by predicting IAT scores. If the cognitions reflected in implicit measures were indeed completely unconscious, informing participants that their IAT results might diverge from their explicit attitudes would be the only way to reduce their surprise. However, as Study 4b showed, providing participants with such an explanation was neither sufficient nor necessary to lower people’s surprise, while making them pay attention to their biases was both. Second, participants who were asked to pay attention to their biases were overall accurate at predicting their IAT results. Together, these findings speak against the idea that implicit measures are completely inaccessible to introspection.

However, conceptualizing implicit evaluations as cognitions that are consciously accessible but rejected is also contradicted by the present data. The fact that participants reacted with surprise to IAT feedback in the first place is already hard to reconcile with the idea that people are aware of their biases at all times. Additionally, participants were less surprised after paying attention to their biases. This indicates that they learned new information about their cognitions that they had not considered previously from these predictions. If people were chronically aware of the biases reflected in their implicit evaluations, reduced surprise reactions should be explainable by other factors than paying attention and learning new information, such as the wording of the feedback or social desirability. However, as stated earlier, these explanations cannot fully account for the data.

In sum, the present research speaks against a simple dichotomy of implicit evaluations reflecting either entirely unconscious or entirely conscious cognitions, and thus also against the portrayal of these cognitions as “attitudes people may be unwilling or unable to report” (<https://implicit.harvard.edu>, 2020). We propose that the concept of “preconsciousness” is better suited to explain the present data and other contradicting findings in the literature on implicit evaluations (Hahn & Goedderz, 2020). A “preconscious” cognition is one that is generally accessible, but to which a person is not paying attention (Dehaene et al., 2006). The present studies suggest that people are surprised at IAT results because, if unattended, the biases captured on implicit evaluations reside outside of conscious awareness. However, those biases are easily accessible when people pay attention to their spontaneous affective reactions, as indicated by reduced surprise reactions and accurate IAT predictions. Hence, although implicit evaluations do not seem to capture “unconscious attitudes” per se, they may well capture cognitions that are preconscious for a lot of people a lot of the time – until they pay attention.

3.13.4. Practical Implications

In addition to these implications for theory, these findings might also have implications for societal debates around implicit bias. There has been a recent debate about the effectiveness of so-called “implicit bias trainings” (e.g., Basu, 2018; Chamorro-Premuzic, 2020; Powell, 2016; Wen, 2020). In light of this trend, we believe refining how implicit evaluations are defined and communicated to the public can ultimately help inform better interventions aimed at reducing biases. While most implicit bias trainings involve many different components and facets that likely need to be approached from different angles, describing implicit biases as “preconscious” and unattended constitutes a much more direct invitation to observe one’s own biases than a presentation of bias as “unconscious”. And giving people the opportunity to observe their own biases may lead to less defensiveness and hence more openness to the idea of harboring racially biased reactions and behavior.

Additionally, we believe a better and more parsimonious understanding of what implicit biases reflect may benefit societal debates around implicit bias more generally. Our research indicates that implicit biases most likely reflect spontaneous reactions that are often preconscious. Applied to discussions about the meaning of implicit bias for behavior, every person may ask themselves whether they believe their behavior may sometimes be guided by spontaneous, unintentional reactions rather than deliberate attitudes (Hahn & Gawronski, 2018). In contrast to this, discussions on whether implicit evaluations predict behavior have often been presented as discussion around whether behavior is guided entirely by “unconscious forces,” or not (e.g., Oswald et al., 2013) (e.g., Oswald et al., 2013). This presentation implies that, if the IAT were to predict behavior, we would be helpless victims of mysterious undetectable processes of our mind. Such presentations are not only unlikely to be true, they also appear less likely to lead to solution-oriented discussions than asking people to reflect on their biased impulses. For instance, assumptions about “unconscious forces” guiding our behavior may lead to the conclusion that the only way to correct biased behavior

is to reduce implicit biases. However, few interventions aimed at reducing implicit biases have led to long-term changes and may thus not be suited to reduce discriminatory behavior (Lai et al., 2014; Lai et al., 2016). Hence, in addition to adding possible interventions, we hope that our demonstration that the biased cognitions reflected on IAT scores are often preconscious will contribute to more informed discussions around the possible meaning of implicit biases for society.

3.14. Conclusion

Both societal trends and social-psychological research have recently shown that people acknowledge the prevalence of racial biases in society and are able to sense their own racial biases (Gallup, 2021; Hahn et al., 2014; Hahn & Gawronski, 2019). At the same time, researchers report that people who receive feedback of implicit racial bias scores often react defensively and with surprise (Gawronski, 2019; Howell & Ratliff, 2017). To reconcile these seemingly incompatible findings, we proposed the concept of “preconsciousness” (Dehaene et al., 2006). Specifically, we argued that people may often be surprised at their IAT feedback because they rarely pay attention to their racial biases even if in principle, they are accessible. In line with this, the present set of studies first confirmed the so far anecdotal observation that people are surprised at IAT feedback indicating bias. Second, it showed that surprise was reduced when participants were instructed to pay attention to their spontaneous affective reaction before completing a Black-White IAT. Together, these findings show that people often fail to pay attention to their biased reactions. Beyond contributing to our understanding of surprise reactions to racial bias feedback, this also illustrates that going beyond a simple dichotomy of conscious or unconscious attitudes can explain current inconsistencies in research on implicit evaluations, help improve implicit bias interventions, and refine our understanding of implicit social cognition in general.

Many societies around the globe show widespread discrimination and disadvantages for racial minority groups. While there are many reasons for these disparities, one of them

might lie in the fact that many well-intentioned people harbor racial biases. While the meaning and origin of these biases will require more research, our studies suggest that simple attention manipulations may be an effective first step towards preventing those biases from being left unattended.

3.15. Open Practices

All materials, data sets, and analysis files can be found at <https://osf.io/bezqx/>.

Preregistrations can be found at the following links:

Study 1: <https://osf.io/8uev5/>

Study 2: <https://osf.io/fh542/>

Study 3: <https://osf.io/ze3pg/>

Study 4a: <https://osf.io/t5wdn/>

Study 4b: <https://osf.io/h9p6j/>

**Chapter 4. Predicting Implicit Preferences Towards Social Groups vs. Food Items:
Implications for the Role of Normativity and Social Desirability in Accurate IAT Score
Predictions**

This chapter is based on the following manuscript:

Goedderz, A., & Hahn, A. (2023). Predicting implicit preferences towards social groups vs. food items: The role of normativity and social desirability in accurate IAT score predictions. *Manuscript submitted for publication to the Personality and Social Psychology Bulletin.*

Please note that headings, citation style, and formatting were changed to fit the layout of this dissertation. The content of the article was not changed.

Abstract

Research has shown that people are able to predict the patterns of their results on Implicit Association Tests (IATs) toward different social groups. An open question is whether these findings generalize to other attitudinal domains that follow less normative patterns and are less socially sensitive. We let participants predict and complete IATs toward five different food item pairs and compared the results to the social-groups domain. Participants showed similar levels of awareness in both domains (evidenced by comparable levels of within-subjects correlations), even though food evaluations followed less normative patterns. However, they were less calibrated in communicating their evaluations in the domain of social groups than food (evidenced by higher between-subjects correlations). This can be partly explained by participants abstaining from using harsh labels when reporting on their biases toward social groups due to social desirability concerns, and it emphasizes the importance of distinguishing between awareness and calibration.

Keywords: Awareness, attitudes, implicit measures, automaticity, consciousness

4.1. Introduction

Recent research has documented that people are able to predict the patterns of their scores on several Implicit Association Tests (IATs, Greenwald et al., 1998) towards social groups prospectively (Hahn et al., 2014), contradicting the long-standing hypothesis that the cognitions reflected on implicit measures¹⁹ are unconscious (Greenwald & Banaji, 1995). Evaluations toward social groups, however, often follow (1) normative patterns, and (2) are known to prompt social desirability concerns. This may question if people have true insight into the cognitions reflected on their IAT scores or simply infer them from knowledge about cultural norms, and whether patterns of awareness replicate in domains that are less socially sensitive.

To investigate these factors, we asked participants to predict how they would score on five IATs toward baked goods. We then compared our findings to a comparable sample that completed the paradigm with social groups. If people base their predictions on cultural norms, they should not be able to predict their pattern of IAT scores in the less-normative domain of baked goods. However, if people have true insight into the cognitions reflected on implicit measures, they should be equally good at predicting the pattern of their IAT scores in both domains. The reduced social desirability of the food domain may lead participants to be more willing to use extreme labels in their predictions. Importantly, however, this difference in labeling should not impact participants' accuracy in predicting the *patterns* of their IAT scores. Instead, it should only impact the accuracy of predicting the strengths of one's preference compared to other participants. We refer to this difference as a difference between awareness (knowing and reporting your relative preferences toward different targets) and

¹⁹ In line with De Houwer et al. (2009), we use the term "implicit measures" to refer to measurement outcomes (in this case evaluations) that are inferred from instruments that limit a participant's ability to control the measurement outcome. They are contrasted from "explicit measures" that we use to describe measures that are (explicitly) reported in self-report measures under full control. Hence, we do not mean to imply that the underlying cognitions reflected on implicit vs. explicit measures are necessarily different. We are trying to contribute to understanding what those differences, if any, may be.

calibration (decisions on what to call these preferences, Hahn & Goedderz, 2020). We explain this difference in detail after a review on previous research and theorizing on awareness in the domain of implicit measures.

4.2. Previous Research on Awareness and Implicit Evaluations

Based in part on low correlations between implicit and explicit measures of the same targets (Hofmann, Gschwendner et al., 2005), many researchers used to claim that explicit measures assess conscious evaluations while implicit measures tap into unconscious cognitions (Basu, 2018; Cunningham et al., 2004; Devlin, 2018; Greenwald & Banaji, 1995; Jost et al., 2002; Nosek, 2007; Phelps et al., 2000; Quillian, 2008; Rudman et al., 1999). However, various dual-process models question this conceptualization and argue that there are other reasons why outcomes on implicit and explicit measures differ (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006). Based on such dual-process models, Hahn et al. (2014) put the unconsciousness hypothesis to an empirical test. They asked participants to predict how they would score on five IATs toward different social groups (Asian vs. White, Black vs. White, Latino vs. White, Celebrity vs. Regular person (non-celebrity), and Child vs. Adult), which they afterwards completed. Results showed that participants were able to predict their patterns of results on these five IATs accurately (median within-subject correlation between prediction and IAT scores, $r = .65$), indicating that they had conscious awareness of the evaluations reflected on these IATs (see also Hahn & Gawronski, 2019, for a replication, and Rahmani Azad et al. (2022), for an extension to gender stereotyping).

Thus far, these studies have only used social groups and highly socially sensitive topics that are matters of continuous public debates: prejudices against minorities and gender stereotypes. There are at least two characteristics of these domains that may limit their generalizability: Normative patterns and social sensitivity.

4.2.1. Normative Patterns: Knowing One's Evaluations or One's Cultural Norms?

Intergroup biases are often discussed in public discourse (e.g., biases toward minority groups, gender stereotypes) and they tend to be culturally shared (Fiske, 2017; Payne et al., 2017). Applied to the topic at hand, people's ability to predict the patterns of their IAT scores may be less a result of true insight into their own evaluative responses, but rather a demonstration of this cultural knowledge (Morris & Kurdi, 2022). That is, participants in Hahn et al.'s (2014) studies may have simply parroted back that they expected to show biases against minority groups and in favor of celebrities and children. To counteract this interpretation, Hahn et al. (2014) showed that their participants could explain their own pattern of preferences better than the preferences of a random other participant, and beyond a binary predictor that contrasted anti-minority bias from pro-celebrity and pro-child bias (see Rahmani Azad et al. (2022), for similar analyses in the domain of gender stereotyping). In the present study, we aimed to examine this alternative interpretation more directly by having participants predict non-social preferences that show less consensus, and hence less normative patterns, than evaluations of social groups. If it is true that this domain shows less consensus, then there should be more between-subjects variation in individual IAT scores (i.e., more variation in how different people evaluate the targets). However, if people can truly feel the reactions reflected on implicit measures, then participants should still be able to predict the patterns of their IAT scores with similar accuracy, despite the fact that they cannot draw upon cultural norms to the same degree. Expanding theorizing on the role of cultural norms for accurate predictions, we further predicted that participants' own predictions should predict their own IAT results over and above the predictions of other participants in the sample in both domains (Hahn et al., 2014; Rahmani Azad et al., 2022). Importantly, however, if evaluations in the non-social domain follow less normative patterns, other participants' predictions should be less related to participant's own pattern of IAT results in the non-social domain than in the domain of social groups.

Together, these results would provide evidence that people can in fact “feel” the cognitions reflected on implicit measures rather than simply predicting their patterns based on cultural knowledge.

4.2.2. *Social Sensitivity: Awareness vs. Calibration*

A second distinctive feature of studying evaluative responses towards social groups is that this domain is inherently plagued with social desirability concerns: Most people would find it uncomfortable to admit to biases against social groups. On one hand, this does not seem to have hurt the accuracy with which participants predicted the patterns of their social-group biases in Hahn et al. (2014). On the other hand, however, the authors found relatively lower between-subjects correlations than within-subjects correlations between IAT score predictions and IAT scores. They explained this difference as a result of different labeling preferences when using the prediction scales. That is, social desirability concerns may lead participants to abstain from using strong labels (e.g., saying that they are “a lot more positive towards White people”). Importantly, if all participants use the same mild labels to describe their biases (e.g., predicting that their responses will be only “slightly more positive towards White”), then between-subjects variances will end up being low due to a lack of variance in the predictions, and this may explain the low between-subjects accuracy correlations Hahn et al. (2014) found. That is, for a between-subjects correlation to be high, the most biased person in the sample must know and predict that they will show more bias than other people in the sample. They would have to predict, e.g., that their IAT will show that their implicit attitude is “a lot more positive toward White people” - a potentially threatening prediction. In contrast, for a within-subjects correlation to be high, participants need to only know whether they are more biased against some groups than others (e.g. knowing that they have more positive or negative reactions toward Black people compared to Latinos, Asians, Celebrities and Children). Hence, participants in Hahn et al.’s (2014) studies showed high accuracy in the predictions of their bias *patterns*, even though they were convinced that all of their biases

were mild (and milder than the biases of others, see Study 3, Hahn et al., 2014). We refer to this difference between being able to sense one's own reactions and calibrating its social strength as the difference between awareness and calibration (Hahn & Goedderz, 2020).

We propose that in a non-social domain that is less socially sensitive, participants should be more willing to use even strong labels to describe their implicit evaluations. This should show itself in significantly more extreme predictions for similar IAT scores. As a result, participants should show similar within-subjects, but higher between-subjects correlations between IAT score predictions and IAT scores in the non-social domain compared to the social domain.

4.3. Implicit Evaluations of Food Targets

Since the invention of the IAT, researchers have used it in a variety of domains such as prejudice and stereotypes (Kurdi, Mann et al., 2019), clinical psychology (Nock et al., 2010), personality psychology (Fatfouta & Schröder-Abé, 2018), and consumer choices (Friese et al., 2006). Assessing implicit evaluations of food has become especially popular in health psychology and self-control research as a means to assess individual differences in reactions towards unhealthy food and its consumption (Richetin et al., 2007; Roefs et al., 2006; Seibt et al., 2007). For the present paper, we chose to study evaluations of baked goods. We did so for several reasons. First, baked goods are culturally important and were in fact declared UNESCO world cultural heritage in Germany (Deutsche UNESCO-Kommission e. V, 2019). Accordingly, we thought it likely that German participants would show strong average IAT preferences, similar to those in the domain of social groups. Second, and more importantly, preferences for baked goods should follow less normative patterns than social groups and there should be less cultural consensus about various baked goods such that some people prefer simple bread loafs over, e.g., sweet pastry and others prefer sweet pastry over simple bread loafs. Third, and finally, preferences for baked goods is a non-socially sensitive topic,

which should minimize participants' social desirability concerns when predicting their implicit preferences.

4.4. Overview of the Study

A sample of German university students was asked to indicate their liking for bread rolls (*Brötchen*), croissants (*Croissants*), crispbread (*Knäckebrot*), cake (*Kuchen*), sweet pastry (*Teilchen*), as well as simple bread loafs (*Brot*). Next, they were asked to predict how they would score on a computerized reaction time task measuring their reactions toward the first five categories compared with simple bread loafs. They then continued to complete five IATs towards the same categories.

To compare these results to studies where participants predicted IAT scores toward social groups, we combined samples of German participants who had completed the same social-group paradigm as participants in the studies reported by Hahn et al. (2014) but in German and in the same laboratory as participants in the present baked-goods study. We chose this sample as the most rigorous comparison group as participants from the same population completing tasks in the same language in the same laboratories seemed maximally comparable.

If it is true that participants have unique insight into the pattern of their implicit evaluations beyond knowledge about normative patterns, participants should show comparably high awareness of their reactions toward baked goods as of their reactions toward social groups, despite the fact that baked-goods preferences follow less normative patterns. This would reveal itself in similarly high within-subjects correlations between participants' person-standardized predictions for their IAT scores and their actual IAT scores. Our second hypothesis was that participants' social-group predictions may be less accurate in terms of where they rank in comparison to other participants because they tend to abstain from using strong prediction labels in this domain due to social desirability concerns. If this is true, then data patterns should look different in the less socially sensitive domain of baked goods. First,

there should be substantially more extreme predictions for similar IAT scores in the domain of baked goods compared to the domain of social groups. Second, results should show higher between-subjects correlations (but not within-subjects correlations) between IAT score predictions and actual IAT scores for baked goods than social groups.

4.5. Method

All materials, data sets, and analysis files for the analyses reported here are available on the Open Science Framework (OSF) at <https://osf.io/8ahbs/>. We report all measures and all conditions collected for the baked goods sample run for this paper. Selection of data for the social-groups comparison sample is reported below and can be read in its entirety in the respective papers from which it was drawn. The study was not preregistered.

4.5.1. Baked Goods Sample

4.5.1.1. Participants. We aimed at recruiting at least 100 participants.²⁰ One-hundred-and-five participants (85.7% female, 13.3% male, 1 “other”, ages 18-50, median=22) were approached on campus at the University of Cologne by research assistants and participated in the study for a payment of four Euros or course credit.

4.5.1.2. Materials and Procedure. After signing informed consent, participants began the study by providing explicit evaluations of the baked goods. They indicated how much they “liked” bread rolls (Brötchen), croissants (Croissants), crispbread (Knäckebrot), cake (Kuchen), sweet pastry (Teilchen), and simple bread loafs (Brot), presented in individually randomized orders, on scales ranging from 1 “not at all (überhaupt nicht gerne)” to 7 “very much (sehr gerne)”.

Next, participants read an introductory paragraph about the differences between

²⁰ The effect sizes for accuracy correlations reported by Hahn et al. (2014) are quite large, a correlation of .54 necessitates only 47 participants to show a significant relationship with a power of 99% according to the GPower program Faul et al. (2007). We opted for a larger number because we did not know which level of accuracy participants would show in this paradigm, and to have sufficient power to detect differences between the baked-goods and social-groups data.

spontaneous affective reactions as opposed to deliberately considered attitudes, similar to the explanations participants have received in the paradigm using social groups (e.g., (Hahn et al., 2014; Hahn & Gawronski, 2019)). Participants then made a trial prediction for IATs towards cats vs. dogs and flowers vs. insects (they never completed those IATs), before they were told about the IAT attitude pairs they would actually predict and complete in this study. Those were each of the first five categories above contrasted with simple bread loafs (i.e., a “CROISSANT vs. BREAD IAT”).

The predictions included the five pictures per category that would actually be used on the IATs, grouped together in categories to the left and right of the screens, with a text explaining to which categories the pictures belonged. Participants were asked to predict the spontaneous reaction they would show on an IAT contrasting these two categories on 7-point scales ranging from 1 “a lot more positive towards BREAD” to 7 “a lot more positive towards [the contrast category]” (see Figure 1 for a sample prediction slide).


After completing all predictions, participants completed the five shortened IATs in the manner developed by Hahn et al. (2014). Specifically, participants first completed a 20-trial word-sorting block (words would always be sorted to the same side in all following IATs). Next, they completed the five IATs that each consisted of 4 blocks: One 20-trial picture-sorting block where participants were trained to sort the pictures only; a 40-trial combined block in which bread was paired with positive words and the contrast category with negative words; another 40-trial picture-sorting block with reversed sorting compared to Block 1; and a final 40-trial combined block with reversed sorting compared to Block 2 (Sorting bread with negative and the contrast category with positive words). An IAT *D*-score was computed using Greenwald et al.’s (2003) scoring procedure such that higher *D*-scores reflect more positive implicit evaluations towards any of the five categories compared with bread (Cronbach’s α : bread rolls = .74, croissants = .73, crisp bread = .72, cake = .69, sweet pastry = .69). After participants had completed the 5 IATs, they completed demographic information, were

debriefed, and compensated.

Figure 1

Sample Prediction Slides for the Domain of Baked Goods (Upper Panel) and Social Groups (Lower Panel), Translated to English From German.

A: Baked Goods Sample Prediction Slide




BREAD LOAFS - CAKE

Look at these pictures. All pictures on the left belong to the category „BREAD LOAFS“. All pictures on the right belong to the category „CAKE“.

Listen to your gut reaction while you look at the pictures. What is your spontaneous reaction toward these categories? What will your IAT show?

Please click the button that best represents your implicit attitude toward BREAD LOAFS vs. CAKE.

You will complete the actual IAT after you have completed all of your predictions.



I predict that an IAT comparing my reactions to BREAD LOAFS vs. CAKE will show that my implicit attitude is...

...a lot more positive toward BREAD LOAFS

...moderately more positive toward BREAD LOAFS

...slightly more positive toward BREAD LOAFS


...the same

...slightly more positive toward CAKE

...moderately more positive toward CAKE

...a lot more positive toward CAKE

B: Social Groups Sample Prediction Slide




BLACK – WHITE

Look at these pictures. All pictures on the left belong to the category „BLACK“. All pictures on the right belong to the category „WHITE“.

Listen to your gut reaction while you look at the pictures. What is your spontaneous reaction toward these categories? What will your IAT show?

Please click the button that best represents your implicit attitude toward BLACK vs. WHITE.

You will complete the actual IAT after you have completed all of your predictions.



I predict that an IAT comparing my reactions to BLACK vs. WHITE will show that my implicit attitude is...

...a lot more positive toward BLACK

...moderately more positive toward BLACK

...slightly more positive toward BLACK

...the same

...slightly more positive toward WHITE

...moderately more positive toward WHITE

...a lot more positive toward WHITE

4.5.2. *Comparison Social Group Sample*

4.5.2.1. Participants. We pooled all participants who had completed Hahn et al.'s (2014) social group paradigm in German in the same laboratory at the University of Cologne under the same conditions. This included a hitherto unpublished pilot sample of 65 participants; 95 and 125 participants from the prediction conditions on Studies 2 and 3 from Hahn and Gawronski (2019); as well as 74 participants from the prediction-with-pictures condition in Hahn and Goedderz (2023).²¹ The three samples are analyzed and compared separately in Goedderz et al. (2023). This resulted in a sample of $N = 359$ (78.7% female, ages 17-66, median age 22)²². Racial/ethnic identification concerning the categories used in the IATs were not assessed in the pilot sample. Of the 251 participants who were asked, 85.4% identified as only White, 7.5% as Middle-Eastern or both White and Middle-Eastern, 0.7% as Black or both White and Black, 1.7% as Latino or both White and Latino, 1.0% as Asian or both White and Asian, while 10% identified with several or yet other categories. One participant did not answer the question.

4.5.2.2. Materials and Procedure. Participants who completed the social-group paradigm completed similar measures as participants in the baked-goods paradigm, with two important differences. First, the paradigm referred to the social categories ASIAN vs. WHITE, BLACK vs. WHITE, LATINO vs. WHITE, CHILD vs. ADULT, and CELEBRITY vs. REGULAR. Second, explicit ratings of these groups were done with thermometer ratings where participants were asked to assess how warm or cool their feelings were towards the groups by typing in a number between 0 and 100 (we opted to use standard 7-point "liking"

²¹The purpose of the unpublished pilot was to see if Hahn et al.'s findings replicate in Germany before running other studies, which they did. Participants in the other conditions of Studies 2 and 3 in Hahn and Gawronski (2019) did not predict their IAT scores. The participants from Hahn and Goedderz (2023) who were not included either predicted their IAT scores without pictures, or simply indicated their spontaneous reactions with no announcement of a test measuring said reactions. We only included the 74 participants who completed the exact same prediction task with pictures as all other participants included here.

²²No participants were excluded in the pilot sample. Exclusions in the samples drawn from Hahn and Gawronski (2019) and Hahn and Goedderz (2023) can be seen in the respective papers. Data were never analyzed before these exclusions were made.

scales instead of thermometer ratings in the baked-goods study for ecological validity). We used the same 10 stimuli per social group (5 male and 5 female) that Hahn et al. (2014) had used (publicly available at the Minear and Park (2004) webpage). IAT D-scores were scored such that higher scores reflected a preference for the majority group (White, regular, adult) over the target group (Asian, Black, Latino, child, or celebrity). Otherwise the procedure for the social-groups studies was the same as the baked-goods paradigm. It included (1) explicit ratings in constrained-randomized orders²³ (2) explanation of implicit attitudes as reactions contrasted from deliberate attitudes²⁴ (3) five IAT score predictions in randomized orders (see Figure 1 for a sample prediction slide), and finally, (4) five IATs completed in different randomized orders (Cronbach's α 's: Asian-White = .69, Black-White = .72, Latino-White = .67, celebrity-regular = .58, child-adult = .62). In the studies reported in Hahn and Gawronski (2019), different additional measures followed the IATs not discussed in this paper. All studies concluded with demographic information.

4.6. Results

4.6.1. Testing Assumptions About Evaluations Toward Baked Goods and Social Groups

We first examined whether our assumptions about the two domains were warranted. These were (1) that both domains should elicit comparably strong reactions, (2) that there should be more of a normative pattern for reactions toward social groups than toward baked goods, and (3) that preferences toward baked goods are a less socially sensitive topic than preferences for social groups.

4.6.1.1. Strengths of Evaluation. To test our assumption that participants had similarly strong reactions toward baked goods as toward social groups, we ran a mixed-model analysis, where participants' absolute IAT scores were predicted by domain, while domain

²³ Three blocks of groups were always rated together, but their order randomized: (1) The racial groups in random orders, (2) adults and children, and (3) celebrities and regular people.

²⁴ The explanations varied slightly between the studies. However, these differences did not produce consistent differences, such that they are combined here (see Hahn et al., 2014)

was allowed to vary both between participants and across different IATs. Average absolute IAT scores per IAT are presented in Table 1. Results showed no effect of domain, $b = -.001$, $SE = .019$, $CI_{95\%} = [-.047, .046]$, $t(5.45) = -.05$, $p = .961$, confirming that the absolute IAT scores did not differ significantly between domains (they averaged out at $D = .42$ in both domains, see Table 1). In fact, a non-significant random effect of IAT type suggested that the absolute IAT scores across the 5 different IATs in each domain did not vary significantly, IAT-specific random effect: $b = .001$, $SE = .001$, $CI_{95\%} = [.000, .006]$, $Wald Z = 1.26$, $p = .207$. As expected, however, there was significant variation in reactions between participants, as evidenced by a significant person-specific random effect, $b = .012$, $SE = .002$, $CI_{95\%} = [.009, .016]$, $Wald Z = 6.42$, $p < .001$, a point to which we turn next.

4.6.1.2. Normative Patterns. To examine whether reactions toward IATs on social groups followed a more normative pattern than IATs on baked goods we ran an independent samples t -test comparing the mean between-subject variance of the five baked-good IATs with the mean between-subject variance of the five social-group IATs (we used raw, not absolute IAT scores for this analysis). All individual variances and their mean by domain can be seen in Table 1. In line with our expectations, results showed greater variance in IAT scores for the domain of baked goods ($M = 0.20$, $SD = 0.01$) than for the domain of social groups ($M = 0.16$, $SD = 0.02$), $t(8) = 4.36$, $p = .002$.

4.6.1.3. Social Desirability. We then examined whether participants in the baked-goods study showed less of a tendency of socially desirable responding than in the social-group paradigm. First, we compared the 10 between-subjects variances in the predictions. Results confirmed that the prediction scales were used with larger variability in the baked goods as opposed to the social groups domain, $t(5.54) = 3.51$, $p = .014$ (see Table 1 for prediction variances).

Table 1

Descriptives and Degree of Social Calibration: Between-subjects Correlations between predictions and IAT scores, as well as prediction and IAT score means and (between-subjects) variances, and absolute IAT D score means and standard deviation for 5 target pairs of baked goods and social groups

Targets	Predictions (7-point scale)		IAT D scores		Absolute IAT D scores		Between- subjects Correlations Predictions and IAT Scores (Calibration)
	Mean	(Between- subjects) Variance	Mean	(Between- subjects) Variance	Mean	SD	
Baked Goods							
Bread rolls vs. Bread	5.03	1.87	0.23	0.22	0.42	0.31	.34***
Croissants vs. Bread	4.05	3.76	0.24	0.21	0.41	0.31	.43***
Crispbread vs. Bread	2.43	1.86	-0.09	0.21	0.37	0.29	.31**
Cake vs. Bread	4.78	3.12	0.32	0.21	0.47	0.31	.43***
Sweet Pastry vs. Bread	4.40	3.68	0.33	0.19	0.45	0.31	.46***
Average		2.86		0.21	0.42		.39***
Social Groups							
White vs. Asian	4.53	1.04	0.33	0.16	0.42	0.31	.19***
White vs. Black	4.29	0.98	0.36	0.18	0.46	0.31	.19***
White vs. Latino	4.35	0.88	0.31	0.18	0.32	0.23	.26***
Regular vs. celebrity	3.99	1.9	-0.02	0.16	0.47	0.29	.21***
Adult vs. child	2.38	1.43	-0.41	0.13	0.44	0.29	.19***
Average		1.25		0.16	0.42		.21***

Note. Preference scores (IAT scores and predictions) in the baked-goods domain mean preference for bread, preference scores in the social group domain mean preferences for the category, white, regular, or adult.

¹Averages are computed in a multi-level model where scores are nested under target group. The random effects of this nesting for the IAT score predictions are .0000 (*n.s.*) in both the baked goods and the social groups data, while the residuals are significant at $b = .84$, $SE = .05$, $Wald's z = 16.19$, $p < .001$ baked goods and $b = .96$, $SE = .03$, $Wald's z = 29.95$, $p < .001$ for the social groups, respectively. The -2 likelihood goodness of fit indices for the two multi-level model were 1402.17 for the baked goods and 5016.46 for the social groups data.

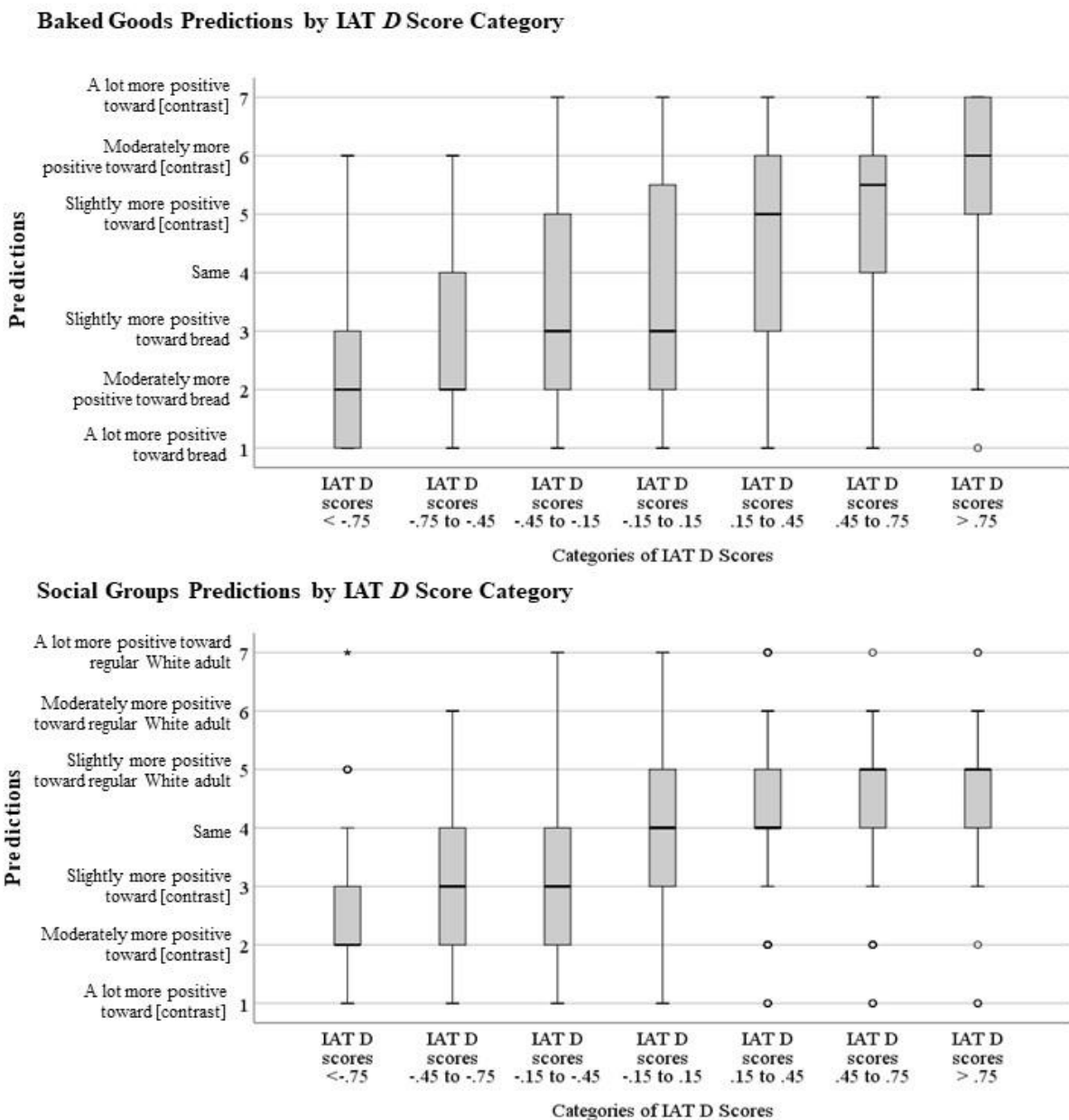
* $p < .05$, ** $p < .01$, *** $p < .001$

To further examine where this larger variability in predictions came from, we took a closer look at the prediction scale usage. To this end, we categorized all IAT scores into seven categories of equal size increments of .30, ranging from scores below -.75 to scores above .75. We then looked at the predictions participants made on the 7-point prediction scales by IAT scores category, for all 525 IAT *D* scores in the domain of baked goods, and the 1795 IAT *D* scores in the social-groups domain (N of *D* scores = amount of participants by 5 IATs).

Boxplots of the predictions by IAT score category can be seen in Figure 2. Confirming our expectations, participants labeled reactions that led to similar IAT scores quite differently in the two domains.

Figure 2

Boxplots of Predictions Made for 7 Categories of IAT Scores in the Domain of Baked Goods (Upper Panel) and Social Groups (Lower Panel).



Note. In the domain of baked goods higher scores reflect more positive evaluation of the contrast categories (bread rolls, croissants, crispbread, cake, or bread) relative to bread. In the domain of social groups higher scores mean more positive evaluations of Whites, adults, or non-celebrities relative to the contrast categories (Blacks, Asians, Latinos, children, or celebrities). Baked goods: $N = 525$ data points, social groups: $N = 1795$ data points.

Whereas participants used the full 7-point prediction scale to describe their reactions towards baked goods (upper panel), they largely abstained from any labels harsher than “slightly more positive” towards White or the other categories in the domain of social groups for IAT scores of similar size. *T*-tests accounting for unequal variances confirmed that prediction labels were less extreme for social-group as opposed to baked-good IATs in both the category of IAT scores between .45 and .75, $t(118) = 3.10, p = .002$, and the category with scores above .75, $t(87.38) = 5.12, p < .001$. These findings demonstrate that participants used different labels to describe similar reactions toward social groups as opposed to baked goods. This suggests that reporting evaluations toward baked goods may be less socially sensitive than reporting evaluations toward social groups.

4.6.2. Awareness

4.6.2.1. **Baked goods.** To assess awareness of the reactions reflected in implicit evaluations of baked goods, we regressed person-standardized IAT scores onto similarly person-standardized predictions for those IAT scores separately for each participant on Level-1 of a multi-level model. On Level-2 we looked at the fixed effect to determine the average within-subjects correlation between predictions and IAT scores. Results of this model are presented in the first column of Table 2. The fixed effect was $b = .41, SE = .05, CI_{95\%} [.32; .51], t(103) = 8.56, p < .001$. Computing correlations separately for each participant revealed a skewed distribution (Skewness = $-.89, SE = .24$) with the same mean and a median of .59. Fisher-*z*-transformed values showed a mean of $z = .62$, which back-translates to a correlation of .55. In sum, participants were able to predict their pattern of reactions on IATs toward baked goods.

Next, we tested whether the baked-goods data replicated Hahn et al.’s (2014) results that implicit-explicit relations could be entirely explained by participants’ predictions. This would suggest that part of participants’ explicit evaluations is based on their consciously

accessible gut reaction (reflected in predictions), but that they consider additional information for their final explicit report.

Results replicated Hahn et al's (2014) results. First, the zero-order relationship between explicit evaluations and IAT scores, $b = .30$, $SE = .05$, $CI_{95\%} [.21; .40]$, $t(104) = 6.20$, $p < .001$, was lower than the relationship between predictions and IAT scores (compare Columns 2 and 3 in Table 2). Second, the relationship between explicit evaluations and IAT scores went to nil when predictions were included in the model, $b = .01$, $SE = .06$, $CI_{95\%} [-.11; .13]$, $t(184.86) = .18$, $p = .861$, whereas prediction accuracy remained unchanged $b = .40$, $SE = .06$, $CI_{95\%} [.28; .53]$, $t(201.01) = 6.50$, $p < .001$.

Table 2

Awareness: IAT *D*-scores regressed on IAT score predictions and explicit thermometer ratings, simple relationships and simultaneous regressions. Relationships are calculated on standardized scores within-subjects, once per participant, and then aggregated across participants in a multi-level analysis.

Parameters (DV: IAT <i>D</i> -scores)	Baked Goods Data			Social Groups Data		
	Predictio n model estimates	Imp.-exp. model estimates	Sim. regr. model estimates	Prediction model estimates	Imp.-exp. model estimates	Sim. regr. model estimates
Fixed effects						
IAT score predictions	.41***		.40***	.52***		.48***
Explicit therm. ratings		.30***	.01		.36***	.09**
Random effect variances						
IAT score predictions	.10**		.08*	.00		.00
Explicit therm. ratings		.09*	.05		.00	.01
Residuals	.59***	.66***	.57***	.58***	.70***	.57***
Goodness of fit						
-2 log likelihood	1268.38	1318.32	1269.62	4134.23	4445.33	4117.76

Note. All variables and the dependent IAT scores are standardized for each individual participant before they are entered in the analysis. Hence, all intercepts are 0 and they are not estimated in these models

* $p < .05$, ** $p < .01$, *** $p < .001$

4.6.2.2. Social groups. Results from the same analyses conducted on the social-group data set replicated these effects and hence the effects shown in Hahn et al. (2014) in a different cultural context (see right half of Table 2). Within-subject prediction accuracy in the multi-level model was $b = .52$, $SE = .02$, $CI_{95\%} [.48; .56]$, $t(163.56) = 25.23$, $p < .001$. Computing separate correlations per participants revealed a skewed distribution (Skewness = -1.10, $SE = .13$) with the same mean, a median of $r = .65$, and a z -transformed mean of $z = .74$, which translates back into a corrected average correlation of $r = .63$. These values thus replicated the values found by Hahn et al. (2014) on a US-American sample on a German sample (.54, .67, and .65, for mean, median and corrected mean, respectively).

The relationship between explicit evaluations and IAT scores, $b = .36$, $SE = .02$, $CI_{95\%} [.32; .41]$, $t(1794,0) = 16.38$, $p < .001$ was unexpectedly higher than implicit-explicit correlations found in the literature (Hofmann, Gawronski et al., 2005) and those found by Hahn et al. (2014, both around .20-.29). The results nevertheless replicated the pattern reported above. Implicit-explicit correlations dropped when predictions were included in the model, $b = .09$, $SE = .03$, $CI_{95\%} [.04; .14]$, $t(312.33) = 3.31$, $p = .001$, whereas the relationship between predictions and IAT scores remained largely unchanged, $b = .48$, $SE = .02$, $CI_{95\%} [.43; .52]$, $t(1765.61) = 19.14$, $p < .001$.

4.6.3. Predictions Beyond Normative Patterns?

A central aim of the present study was to examine whether people would be able to predict the patterns of their IAT results in a domain where we expected less culturally normative patterns. Our analyses of between-subject variances in the IAT scores already suggested that participants showed more normative responses in the social-group domain than in the baked-goods domain.

As another test of whether participants' predictions went beyond normative patterns, we adapted a procedure employed by Hahn et al. (2014) as introduced by Rahmani Azad et al. (2022): We randomly paired each participant with another participant in the same sample and

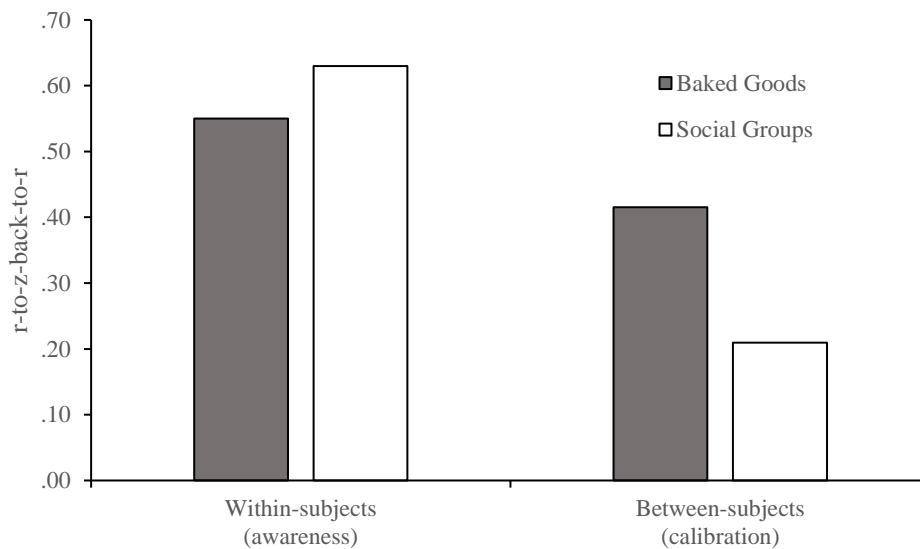
ran a model in which we predicted every participant's IAT scores from the randomly paired other participant. We iterated this process 1000 times such that we got a distribution of fixed effects that indicated how accurately the randomly paired other participant predicted another participant's pattern of IAT scores on average. If participants mostly base their predictions on their beliefs about what people would normatively show on such tests, then any participant's predictions should predict any others participant's scores as well as participants' own predictions. Results showed that in both domains, the randomly paired other participants' predictions predicted participants' IAT scores above zero in all 1000 iterations (Social-groups sample: $M_b = .415$, $\text{Range}_b [.346; .470]$; baked-goods sample: $M_b = .202$, $\text{Range}_b [.058; .321]$). This suggests that IAT scores follow normative patterns in both domains. Importantly, all 1000 random others' estimates in the social-groups domain were higher than the random others' estimates in the baked-goods domain. This is in line with our hypothesis that in the domain of baked goods there may be less of a normative pattern to base one's own prediction on than in the domain of social groups.

To further examine how much variance over and above the random others' predictions the participants' own predictions explained, we additionally simultaneously regressed IAT scores onto participants' own predictions and the randomly paired other's predictions. Results of the two fixed-effects slopes and their distributions across the 1000 iterations can be seen in Figure 3. They showed that in both domains, participants' own prediction outperformed the random others' prediction in all 1000 iterations (Social-groups sample: $M_{\text{Self } b} = .428$, $\text{Range}_{\text{Self } b} [.369; .499]$, $M_{\text{Other } b} = .223$, $\text{Range}_{\text{Other } b} [.153; .283]$; baked-goods sample: $M_{\text{Self } b} = .391$, $\text{Range}_{\text{Self } b} [.338; .447]$, $M_{\text{Other } b} = .058$, $\text{Range}_{\text{Other } b} [-.080; .200]$, see Figure 2). This replicates previous findings by Hahn et al. (2014) and Rahmani Azad et al. (2022) and suggests that participants in both domains have unique insides into their implicit evaluations beyond a culturally shared pattern. Importantly, this analysis further showed that the 99% confidence interval of the random other estimates in the baked-goods sample ($CI_{99\%} [-.053;$

.166]) did not overlap with the 99% confidence interval of the random other estimates in the social-groups sample (CI_{99%} [.171;.276]). This is in line with our hypothesis that participants based their predictions less on normative patterns in the domain of baked goods than in the domain of social goods. Additionally, the 95% confidence interval of the random other estimates still included zero (CI_{95%} = -.024 - .136). This suggests that in the baked-goods domain the random other (normative) prediction may not significantly explain variance of participants' own pattern of IAT scores above participants' own predictions in all cases. This suggests that many participants did not base their predictions on cultural knowledge in this domain, but predicted their own idiosyncratic patterns instead.

Figure 3

Correlations Between IAT Score Predictions and IAT Scores



Note. Individual correlations were Fisher- z -transformed and averaged. The averages are back-transformed to *Pearson's r* correlations for easier readability. This back-transformation precludes the usage of error bars.

4.6.4. Calibration

To determine how consistently participants labeled their preferences, we computed between-subjects correlations between participants' IAT score predictions and their IAT scores separately for each attitude pair. Results for both the baked-goods and the social-

groups data are presented in Table 1. We computed the average between-subjects correlation by standardizing IAT scores once for each target pair and regressing this score onto similarly sample-standardized predictions on Level-1 of a multi-level model, and then aggregating the results across target pairs on Level-2 (computing a simple arithmetic average of the five correlations yields equivalent results). As can be seen in Table 1, the average between-subjects accuracy in the baked-goods domain was $b = .39$, $SE = .04$, $CI_{95\%} [.31; .47]$, $t(524) = 9.74$, $p < .001$. The average correlation in the social groups domain was $b = .21$, $SE = .02$, $CI_{95\%} [.16; .25]$, $t(1794.00) = 8.94$, $p < .001$.

4.6.5. Awareness vs. Calibration in Different Domains

Our hypothesis was that people would be better calibrated when reporting their reactions toward baked goods compared to their reactions toward social groups, despite similar levels of awareness. To test this, we combined the data sets and ran a series of mixed-model analyses. In a model testing differences in calibration, we regressed sample-standardized IAT scores onto sample-standardized IAT score predictions nested under IAT type, and then looked at the interaction of the predictions with a contrast comparing the baked-goods domain (coded -1) with the social-groups domain (coded 1) on Level-2. This analysis confirmed that social calibration was significantly higher in the domain of baked goods as opposed to the domain of social groups, $b = -.09$, $SE = .02$, $CI_{95\%} [-.14; -.05]$, $t(2318.00) = -3.85$, $p < .001$. Unexpectedly, another model testing for differences in awareness, where scores were person-standardized and nested under participants, showed that within-subjects correlations between predictions and IAT scores were lower in the baked-goods as opposed to social-groups data, $b = .05$, $SE = .02$, $CI_{95\%} [.01; .10]$, $t(459.00) = 2.33$, $p = .020$.

To account for the skew in the distributions of individual correlations, we also computed each of the individual correlations that went into both of those analyses separately (1 per target = 10 in the between-subjects analyses for social calibration, 1 per participant =

464 for the within-subjects analyses for introspective awareness), then *Fisher-z*-transformed them and compared those *z*-scores in *t*-tests accounting for unequal variances. Average *Fisher z*-scores, back-transformed into correlation coefficients, are presented in Figure 2. These analyses continued to show a significant difference in calibration between the domains, $t(5.40) = 5.42, p = .002$, but no significant difference in awareness, $t(145.50) = -1.49, p = .137$. As predicted, participants were better calibrated in their reports for their attitudinal reactions in the domain of food as opposed to the domain of social groups, even though awareness tended to be similar.

4.7. Discussion

In the present study we investigated whether the findings of Hahn et al. (2014) that people are able to predict the patterns of their IAT scores toward five social groups are generalizable to a different non-social attitudinal domain that (1) may follow a less normative pattern, and (2) may be less socially sensitive. To this end, we conceptually replicated the paradigm introduced by Hahn et al. (2014) and extended it to the domain of food, specifically baked goods. We then compared the baked-goods sample from the present study to a comparable sample that completed the original social-groups prediction paradigm by Hahn et al. (2014).

In line with the hypothesis that people have unique insight into the cognitions reflected on implicit measures, results showed that participants were comparably accurate in predicting the patterns of their IAT results in the less normative domain of baked goods as in the social-group domain, where evaluations do follow normative patterns.

Confirming the hypothesis that social desirability concerns influence participants' usage of the prediction scale, participants used stronger labels to predict their implicit evaluations in the baked-goods domain than in the social-groups domain. As a result, we further hypothesized that participants would be more calibrated in communicating their biased reactions in less socially sensitive topics. Confirming this, between-subjects correlations

between predictions and IAT scores were lower in the social-groups domain than in the baked-goods domain, despite similar levels of within-subjects correlations. We discuss the evidence for each of the points and their implications for our understanding of implicit measures next.

4.7.1. Predicting IAT Scores Beyond Normative Patterns

One common explanation for how exactly people came to have accurate knowledge of the patterns they would show on their IATs in Hahn et al.'s (2014) studies is that they simply inferred what scores they would show from cultural knowledge. Counterarguing this idea, the present study showed that participants were able to predict the pattern of their IAT results on baked goods, a domain where one would not expect strong normative patterns. And indeed, in line with the assumption that evaluations toward baked goods would follow less normative patterns than evaluations toward social groups, the between-subjects variations on each IAT was larger in the baked-goods domain than in the social-groups domain. Further, the relationship between a random other participants' predictions and participants' own patterns of IAT results was smaller in the baked-goods domain than in the social-groups domain. And in a simultaneous model, the randomly paired other participants' predictions in the baked-goods study explained participants' pattern of IAT results over and above participants' own predictions in less than 95% of the iterations. Conversely, in the social-groups domain, both the random other and participants' own predictions jointly explained participants' patterns of IAT results in all iterations. Together, these findings indicate that the pattern of IAT scores participants produced were less normative in the domain of baked goods compared to the domain of social groups.

A test of whether participants were able to predict these non-normative patterns with the exact same level of accuracy as their social-groups patterns yielded mixed results. A direct comparison of raw accuracy correlations suggested that food predictions were less accurate than social-group predictions. In contrast, once the skewed distribution of raw correlations

was taken into account via a Fisher- z -transformation, this difference disappeared. Although we believe that the analysis on z -transformed values is the more accurate representation, these results remain somewhat ambiguous with respect to whether predictions of food IAT scores are similarly accurate or slightly less accurate than predictions of social-group IATs.

In addition to less normative patterns, another reason food IAT predictions might be less accurate is smaller differences in reactions between targets. It may have been harder for participants to sense the fine nuances between their preferences for, e.g., bread loafs over bread roles than to sense their reactions toward e.g. children as opposed to Black people, because the former vary less. Confirming this idea, the within-person variance for the five IATs was on average lower in the baked-goods study ($M = .17, SD = .14$) than in the social-groups sample ($M = .25, SD = .18$), $t(1041.71) = -11.51, p < .001$. Future research is needed to confirm these interpretations.

Whether or not the accuracy of predictions was exactly the same or slightly lower for food as opposed to social groups, it is important to remember that participants did predict their patterns accurately in both domains. Their predictions furthermore entirely explained relationships between IAT scores and traditional explicit measures and substantially outperformed the predictions of a random other participant. Hence, even if part of the prediction accuracy in Hahn et al. (2014) findings can be explained by replicating normative patterns, the current studies clearly show that it is possible to predict patterns of IAT scores that aren't normative and that show large between-subjects variation. Cultural knowledge might help (and the current results might be ambiguous concerning how much it helps), but it doesn't seem to be a necessary factor to observe one's own reactions.

4.7.2. Predicting IAT Score Patterns in Socially Less Sensitive Domains – Awareness vs. Calibration

One major difference and novelty in the present findings was that participants were less well calibrated in the domain of social groups as opposed to food items. This could be

seen in the fact that, even though *within*-subjects correlations between IAT score predictions and IAT scores were comparable in size in both domains, *between*-subjects correlations were substantially higher for baked goods than for social groups.

Analyses further suggested that this difference in between-subjects correlations was due to the fact that participants used only very conservative labels to describe their biases toward social groups, whereas they used the prediction scale much more liberally when predicting their IAT scores toward baked goods. The same IAT scores that were labeled “strong” in the domain of baked goods (IAT D scores $> .75$) were labeled predominantly as “mild” in the domain of social groups. Importantly, however, these mild labels sufficed to describe patterns of IAT scores in the social-groups domain, confirming that participants were generally aware of their biases, even if unwilling to name those biases anything but “mild”. This unwillingness to use harsh-sounding descriptors in socially sensitive topics may be a main reason for the low levels of calibration in the domain of social groups.

Specifically, Hahn and Goedderz (2020) have posited that awareness versus calibration of automatic cognitions depend on different psychological processes. Awareness depends on internal processes, specifically (1a) the strength of a signal a process produces and (1b) whether a person pays attention to said signal. Calibration, on the other hand, cannot depend on internal processes. Whether or not one’s reaction is stronger or weaker than the reactions of other people is information that does not reside in a person’s own cognitive system. Instead, calibration should depend on (2a) whether a person knows the social conventions of what a certain reaction is labeled by the comparison sample (e.g., what a “slight” vs. “strong” preference feels like, what other people would say), as well as (2b) willingness to apply these labels to one’s own cognitions (e.g., willingness to say “I have a strong preference for Group A”).

Concerning awareness, analyses confirmed that participants showed similar reactions towards food as towards social groups (process 1a), and all participants were asked to look at

pictures and pay attention to their own reactions, such that attention was held constant across the domains (process 1b). The data are hence compatible with the model in that we found similar awareness across domains (but see discussion on possible differences above).

Concerning calibration, we have so-far focused our discussion on process (2b); The fact that people should be much more willing to call their food preferences than their social-group preferences “strong”; and additional data analyses are compatible with the interpretation that people are unwilling to apply certain bias labels to their social-group biases. As a result, one could say that many people are “miscalibrated” about their social-group biases (Hahn & Goedderz, 2020). However, we believe process (2a) may be at play here, too. Talking about one’s food preferences and seeing associated behavior (e.g., how much a person likes and eats something) is much more common than talking about social-group preferences. As a result, people should have much more knowledgeable about what a “mild” vs. “strong” food preference feels like than how one would refer to a bias against a social group. While this interpretation remains speculative, it is compatible with the data. Future research is needed to validate other aspects of Hahn and Goedderz’s (2020) model.

4.7.3. Implications for Dual-process Models

One potential explanation for Hahn et al.’s (2014) successful theoretical dissociation of results of predictions, implicit, and explicit measures is that announcing a test score simply made participants more honest. Participants may feel the same feelings towards different attitude targets at all times, but distort those on traditional explicit measures. IAT score predictions would then only be accurate because they announce measurement, forcing participants to become honest. This explanation is reflected in the common explanation that implicit measures reflect attitudes people are “unwilling or unable to report” (<https://implicit.harvard.edu/>), but it seems unlikely in the domain of baked goods. There shouldn’t be anything threatening about “admitting” that one likes bread rolls better than bread loafs. Despite this difference, results in the baked goods domain were similar to those in

the social-group domain. Participants' explicit reports were less correlated with their IAT scores than their IAT score predictions, and the implicit-explicit relationship could be explained by their predictions. This contradicts the notion that implicit-explicit dissociations can be reduced to more or less honesty, or "willingness" to report certain attitudes.

Instead, it is compatible with dual-process models that claim that implicit measures reflect spontaneous reactions while explicit measures reflect propositional attitudes (Gawronski & Bodenhausen, 2006; Hahn & Goedderz, 2020). Spontaneous reactions can be readily observed when people are encouraged to listen to them, but they are only one piece of information that goes into explicit attitude reports. Hence, implicit-explicit relationships will tend to be low even as people can predict their implicit score patterns accurately. And this pattern holds across both more and less socially sensitive domains.

4.7.4. Limitations

The present study focused on two domains of attitudes that we felt differed maximally in terms of social-desirability concerns and normative patterns. However, these domains do of course differ on countless other dimensions, and there are countless additional attitude domains that may fall anywhere on these dimensions. As such, these studies can only be viewed as one incremental step towards understanding the different processes that factor into awareness and calibration. Additionally, we limited our implicit measure to the IAT to stay as close to the original paradigm as possible. Morris and Kurdi (2022) recently provided first evidence that the effect of awareness of implicit attitudes generalizes to other implicit measures such as the Affect Misattribution Procedure (AMP, Payne et al., 2005) and to 57 different broadly and randomly sampled attitude targets. While randomly sampled attitude targets across domains make systematic comparisons between domains more difficult, these findings make us optimistic that our assumptions may apply more broadly to the cognitions reflected in implicit evaluations independent of the measure used to capture these cognitions, and that it extends to many more attitude targets. Future studies on many more attitude targets

and domains chosen and compared on theoretical grounds are necessary to confirm our theoretical interpretations.

Another major limitation of this project is that we compared two independent samples in a quasi-experimental design rather than randomly assigning participants to conditions in the same study. This potentially invites the question of whether the differences we found (mainly on calibration) are really a result of domain or whether our baked-goods sample was simply better calibrated than our social-groups sample. While lack of random assignment makes this a theoretical possibility, we carefully selected a comparison sample across all our available data sets that was drawn from the same student population in the same lab. Additionally, the only one difference we found is theoretically consistent with our theorizing: People showed similar awareness but differential calibration. Hence, we believe our results can be attributed to true differences between the domains and not random differences between the samples, despite the quasi-experimental nature of our design. Future research is needed to confirm these points.

4.7.5. Conclusion

The purpose of the present paper was to extend Hahn et al.'s (2014) findings that people can predict the patterns of their IAT scores towards social groups to a domain that is non-social, tends to show less normative patterns, and where there are fewer concerns with social desirability. To meet these goals, we chose the domain of food items, specifically baked goods. The present study replicated findings by Hahn et al. (2014) in the domain of baked goods and showed that participants were able to accurately predict the patterns of their IAT scores toward baked goods even though the reactions toward these targets followed less normative patterns. These findings support the notion that the cognitions underlying implicit measures can be consciously perceived rather than just inferred; and that implicit measures do not capture attitudes people are "unwilling or unable to report". Instead, they are more

compatible with the notion that implicit measures capture spontaneous reactions that may sometimes evade attention, but that can be observed when a person is encouraged.

In contrast to similar levels of awareness, participants differed in their level of calibration between the domains. They freely chose more extreme labels to describe their food preferences than their social-group preferences, and these more extreme labels were better-aligned between participants. These findings suggest that people may often be aware but miscalibrated in their biases toward social groups. Most importantly, they suggest that distinguishing awareness from calibration might be important if one wants to understand what people know and don't know about their own cognitions.

Chapter 5. General Discussion

The aim of this dissertation was to provide a more nuanced perspective of when and how people are able to report on their automatic cognitions, in order to reconcile supposedly inconsistent findings regarding the conscious or unconscious nature of the cognitions reflected on implicit measures. By integrating empirical findings on awareness and implicit evaluations (Gawronski et al., 2006; Hahn et al., 2014; Hahn & Gawronski, 2019; Hahn & Goedderz, 2020) with theories of consciousness and introspection (Dehaene et al., 2006; Hofmann & Wilson, 2010), the present dissertation presented a framework of reporting on automatic cognitions and behaviors and subsequently provided evidence in line with the proposed framework. Going beyond a simple dichotomy of implicit measures as either reflecting entirely conscious or unconscious cognitions, the framework proposes (1) whether an automatic cognition should be called unconscious depends on the concept of awareness that a researcher investigates and (2) automatic cognitions are often neither entirely conscious nor unconscious but rather reside in a preconscious state until specific conditions are met. The two concepts of awareness the framework distinguishes between are *introspective awareness*, defined as the ability to sense and report on an automatic cognition, and *social calibration*, defined as the act of labeling an automatic cognition in accordance with labeling conventions in the sample. Both concepts pertain to different empirical approaches and analytical strategies and are dependent on different processes. Specifically, the framework proposes that introspective awareness is determined by (1a) the strength of the signal a cognition produces and (1b) the degree of attention paid to the signal. Social calibration, in turn, is determined by (2a) knowledge about labeling conventions and (2b) the willingness to apply these labels to one's own cognition.

In line with the framework's proposition of a conceptual difference between introspective awareness and social calibration, Chapter 2 demonstrated replications across 17 studies, in line with Hahn et al.'s (2014) findings that participants were accurate in predicting

the pattern of their own IAT results toward 5 social groups, but less accurate in placing their individual IAT results toward each social group in the sample distribution. Accordingly, while across studies the average within-subject correlation between participants' predictions and IAT scores showed a meta-analytical effect of $b = 0.44$, the average between-subject correlation only showed a meta-analytical effect of $b = 0.22$. This between-subject correlation is comparable to meta-analytical effects of between-subject correlations between implicit and explicit measures reported by other researchers ($r = .24$ reported in Hofmann, Gawronski et al., 2005). The importance of distinguishing between the concepts of introspective awareness and social calibration when studying awareness is further highlighted by Chapter 4. Here we found that people showed similar within-subject correlations between predictions and IAT scores in the domains of social groups and baked goods, while they showed considerably larger between-subject correlations in the domain of baked goods as in the domain of social groups. Only examining between-subject correlations in both domains could have led to the assumption that people are more aware of their preferences for food items than they are of their social-group biases. From the lens of the new framework, we get a much more nuanced perspective; participants seemed to be similarly introspectively aware of their own automatic cognitions in both domains, but less aware of where their own automatic cognitions rank in comparison to other people in the domain of social groups. Taken together, these findings emphasize that inferring awareness from between-subject correlations between implicit and explicit measures may lead to the assumption that people lack introspective awareness of the cognitions reflected on their implicit measure, when in fact they may just be miscalibrated in labeling them.

Chapter 4 further provided first evidence for the proposition of the framework that the different concepts of awareness may be determined by different processes. On the one hand, we predicted similar levels of introspective awareness in both domains because we expected both domains to elicit comparably strong affective reactions (process 1a), and we held

attention constant by asking participants to predict their IAT results while listening to their gut reaction (process 1b). On the other hand, we expected participants to show higher levels of social calibration in the domain of baked goods than in the domain of social groups because we hypothesized that people may have more experience in voicing their preferences toward food items than toward social groups, leading to greater knowledge of labeling conventions (process 2a). Additionally, talking about preferences toward food items should be less prone to social desirability and self-presentational concerns, making people more willing to apply those labels to their cognitions (process 2b). The latter hypothesis was supported by the fact that the variance of predictions was significantly larger in the baked goods domain than in the social groups domain, and people more often used labels toward the upper end of the predictions scale in the baked goods domain. As outlined in the discussion of Chapter 4, it is important to highlight that we did not experimentally manipulate the different processes, but that we rather chose a domain we thought would exhibit variation regarding the proposed processes involved in social calibration. However, the two domains may vary in countless other ways, such that it is possible that the pattern of results we observed stemmed from other differences between the two domains. To ultimately test the proposed mechanisms of the framework, future research would have to systematically manipulate each of the proposed processes and investigate whether these manipulations impact the different concepts of awareness in the expected way.

Chapter 3 provided an experimental approach to testing the role of attention (process 1b) in introspective awareness. The studies did not explicitly test the mechanisms of the framework by examining accurate within-subject predictions of IAT score patterns but instead inferred awareness from surprise reactions to IAT feedback. In line with the framework, we argued that people may often report surprise at IAT feedback because they rarely pay attention to their biases, but once they are encouraged to pay attention to their biases, they will report less surprise, indicating that they gained introspective awareness to their automatic

cognitions. Results supported this hypothesis and showed that participants reported less surprise in reaction to their IAT feedback when they were instructed to pay attention to their spontaneous affective reaction than when they did not pay attention to their affective reactions. Changing labels to be less threatening (Study 2) or providing a lengthy non-threatening explanation to what the IAT measures (Study 4b) were not sufficient to reduce participants' surprise. Neither was the mere prediction of IAT results without encouragement to pay attention to spontaneous affective reactions (Study 3). These findings demonstrate that participants did not simply report surprise at their feedback because they expected other labels or because of social desirability concerns.

In summary, the studies presented in Chapters 2-4 provided evidence in line with the proposed framework of reporting on automatic cognitions and behaviors in the domains of attitudes toward social groups and food items. They showed that the concepts of introspective awareness and social calibration can successfully be dissociated using different analytical approaches, and that they produce different outcomes in different domains that can be predicted by the mechanisms proposed by the framework. Further, a first experimental study provided evidence for the idea that automatic cognitions often reside in a preconscious state until certain conditions are met, and they become consciously reportable.

5.1. Introspective Awareness or Inferences from External Information?

A large body of research suggests that people may not have true introspective access to their own cognitions but instead typically infer their own cognitions from plausible lay theories about their cognitions or external information they observe about themselves (Bem, 1972; Nisbett & Wilson, 1977; Wilson & Dunn, 2004). This poses the question of whether participants in the current research were actually introspectively aware of their own cognitions reflected on their implicit evaluations, or whether they showed high prediction accuracies because they inferred their pattern of IAT results from other information (Morris & Kurdi, 2022). There are several ways participants could have inferred their own pattern of IAT

results beyond true introspective awareness. First, especially in the domain of social groups, participants could have based their predictions on knowledge about normatively expected patterns of biases in the specific context. For example, the average psychology student may have had a lecture on the concept of implicit attitudes and know that most people in western societies show biases against racial minorities and in favor of White people, and biases in favor of children over old people on the IAT (Nosek et al., 2002). In the studies by Hahn et al. (2014), the meta-analysis in Chapter 2, and the social-groups sample in Chapter 4, this knowledge may have sufficed to achieve high accuracies in predicting one's own pattern of IAT scores without participants having to have real introspective awareness of their own cognitions. Counterarguing this explanation, Hahn et al. (2014) showed that participants predicted their individual pattern of IAT results beyond their expectations of what the average student at their university would show. Additionally, Hahn et al. (2014), as well as Chapter 2, and Chapter 4, have shown that while the randomly paired other participants' prediction patterns were also related to participants' IAT results, participants' own predictions were consistently a better predictor of their own pattern of IAT results. If participants truly only based their own predictions on their knowledge about normatively expected bias reactions without having introspective awareness of their own automatic cognitions, any other participants' predictions should be just as strongly related to participants' own IAT scores as their own predictions. These findings suggest that participants prediction may be a combination of unique introspective insight and cultural knowledge, with unique insight playing a slightly larger role. Additionally, Chapter 4 showed that participants demonstrated comparable levels of introspective awareness in the domain of baked goods, even though IAT scores seemed to follow less normative patterns than in the domain of social groups. This demonstrates that introspective awareness is possible, even if cognitions are not simply inferable from cultural knowledge.

A second way participants may be able to accurately infer their IAT score patterns without introspective awareness of their own automatic cognitions is by observing their own behaviors. In this context, one could imagine that people that complete the IAT observe their own reactions during task completion and may be able to notice that it is easier for them to react when pictures of Black people and negative words are paired together than when pictures of White people and negative word are paired together (Monteith et al., 2001). In a similar vein, it could be that people are able to simulate their upcoming behavior on the IAT if they are familiar with the task, or that they are able to recollect instances of encounters with the attitudinal objects in question and infer their reactions on IATs from their past behaviors. Several findings contradict these possibilities. The fact that participants report surprise at their IAT feedback suggests that it is difficult for participants to accurately interpret their reactions on the IAT and infer their IAT results from their performance on the IAT (Goedderz & Hahn, 2022). Hahn et al. (2014) have further shown that experience with the procedure of the IAT was not a necessary precondition for accurate predictions (Study 4), a finding corroborated by Chapter 2, as almost all included studies in the meta-analysis did not include practice IATs or explanations about the IAT procedure and yet showed high prediction accuracies.

In summary, the existing evidence suggests that people may have true introspective insight into their own automatic cognitions and do not only infer them from knowledge about normatively expected bias patterns or by observing or anticipating their behavior.

5.2. What Do People Introspect Upon?

If one accepts then that people are able to gain introspective awareness to their automatic cognitions, an open question is what exactly do people introspect upon? The framework presented in this dissertation suggested that people gain introspective awareness by paying attention to the signal the cognition produces. But what kind of signal do people need to pay attention to, and which factors determine whether a signal is strong enough?

Regarding the nature of the signal, Hofmann and Wilson (2010) have postulated that the perceptible signal a cognition produces can be a variety of different experiences such as an affective reaction, a spontaneous gut reaction, or a fluency perception. For a person to accurately report on a cognition reflected on an implicit measure, the signal that the person needs to pay attention to is determined by the cognition the measure is supposed to pick up on. As such, the signal that people need to pay attention to may be different for different domains. For instance, in the domain of implicit evaluation, research suggests that implicit evaluations reflect spontaneous affective reactions (Gawronski & Bodenhausen, 2006; Gawronski & LeBel, 2008; Hofmann, Gawronski et al., 2005; Hofmann & Wilson, 2010). In the present research, instructions on predicting IAT scores often entailed a sentence telling people to pay attention to their spontaneous affective reactions, or their first gut reactions, and this may have enabled participants to pay attention to the right signal. In contrast, stereotypes are assumed to reflect more semantic associations (Amodio & Devine, 2006), such that paying attention to an affective reaction may not suffice to make accurate predictions (Rahmani Azad et al., 2022). Interestingly, research by Rahmani Azad et al. (2022) showed that people were also able to accurately predict their cognitions reflected on implicit gender stereotypes. The authors hypothesized that in this case, participants may have inferred their automatic cognitions from fluency perceptions – that is, how easily certain target-word pairs came to mind (Unkelbach & Greifeneder, 2013). Together, these findings suggest that either the cognitions reflected on different implicit measures produce different signals that people adaptively pay attention to, or the cognitions reflected on implicit measures produce several perceptible outputs which people can introspect upon. In accordance with the latter idea, Rivers and Hahn (2019) have found that participants' predictions in Hahn et al.'s (2014) original studies were best explained by a combination of activated associations and self-regulatory control processes in the quadruple process model (Conrey et al., 2005). This suggests that the signal people pick up on when predicting their automatic cognitions may be

a cumulative experience of signals related to affective reactions, fluency perceptions, and control processes. A possibility to shed light on which signals are most important for gaining introspective awareness of one's automatic cognitions would be to experimentally manipulate to which signal participants pay attention to.

A second question is, which factors may facilitate introspective awareness by enhancing the signal the cognition produces. On the one hand, the framework proposes that if an automatic cognition is weak in the first place, it may not be accessible to introspection. For instance, Nosek (2005) found that the relationship between implicit and explicit evaluations was weaker for weaker evaluations and stronger for stronger evaluations across 57 attitude domains. This could suggest that when implicit evaluations were stronger, they more likely entered conscious awareness and were more strongly considered for explicit responses. While this is indicative that evaluative strengths may facilitate introspective awareness, it is important to remember that higher correlations between implicit and explicit evaluations can have various reasons beyond increased awareness. To test the hypothesis more directly, future studies could investigate whether the strengths of the evaluations influences the within-subject correlation between participants' predictions and IAT results. If the framework holds true, people should show higher introspective awareness when evaluations are strong than when they are weak.

Another way of experimentally testing whether stronger evaluations facilitate introspective awareness is by experimentally altering the strength of the signal. The studies reported in this dissertation all adopted the prediction procedure by Hahn et al. (2014) in which participants predict their IAT results while looking at the pictures that are later used on the IAT contrasted on one slide (e.g., pictures of Black people left, pictures of White people right). Further, all studies used the IAT as the central measure of implicit evaluations. All these procedural aspects may have potentially increased the evaluative signal participants are able to feel. For instance, pictures are assumed to facilitate access to affective reactions

(Hinojosa et al., 2009; Houwer & Hermans, 1994; Kensinger & Schacter, 2006), showing pictures side by side may additionally pronounce differences between the target pairs (Gawronski et al., 2005), and the IAT is known to demonstrate stronger effects than other implicit measures (Bar-Anan & Nosek, 2012). According to the propositions of the framework, experimentally manipulating these procedural aspects to reduce the strength of the perceivable signal should decrease participants' introspective access to their automatic cognitions. Indeed, unpublished work in progress from Adam Hahn's lab in cooperation with myself supports this hypothesis. For instance, one line of research suggests that participants are more accurate in predicting their IAT score results when they see pictures during their predictions than when they do not see pictures (for a summary of the available data see Hahn & Goedderz, 2020). Another line of research suggests that participants are more accurate in predicting their results on implicit measures when two attitude objects are contrasted than when they are presented separately, and when they predict standard (contrastive) IAT results than when they predict results in separate single-category IATs (Goedderz et al., 2022). Both of these findings are in line with the idea that increasing the perceivable signal a cognition produces could increase the introspective accessibility of the cognition.

Taken together, initial evidence suggests that participants gain introspective awareness by paying attention to the signal produced by the cognitions reflected on implicit evaluations and stereotypes, which often manifest in a spontaneous affective reaction or a feeling of fluency. The strength of the signal seems to be malleable such that there may be factors that could increase introspective awareness such as showing pictures or presenting attitude objects in contrast.

5.3. Traditional Explicit Measures and Predictions

The present research has documented that participants' predictions were related to their pattern of IAT scores beyond participants' reports on traditional explicit measures. This poses the question of what differentiated traditional explicit measures from the explicit

prediction scales that made participants willing and able to report on their cognitions reflected on their implicit evaluations. Both measures differed in several important ways throughout the studies²⁵. First, the traditional explicit ratings asked participants to report their feelings toward the attitude objects abstractly, while the prediction slides featured concrete pictures of the attitude objects that would later be used on the IAT. Second, traditional explicit ratings were reported toward one attitude object at a time (e.g., “Please indicate how warmly or coolly you feel toward Black people.”), while predictions asked about both attitude objects that would later be contrasted in one IAT at once (e.g., “I predict that an IAT comparing my reactions to BLACK vs. WHITE will show that my implicit attitude is a lot more positive toward WHITE”). Third, participants were explicitly instructed to pay attention to their “gut reaction” or their “spontaneous affective reaction” when making their predictions, while there was no additional instruction before the traditional explicit ratings. These procedural differences could potentially explain why participants’ predictions were more aligned with their IAT score patterns than their traditional explicit reports. Specifically, as outlined before, the present framework suggests that predictions led to more introspective awareness because they enhanced the signal the cognitions produced (through pictures and contrast) and instructed participants to pay attention to concrete stimuli. For instance, Study 3 in Chapter 3 showed that surprise reactions were not reduced when participants made predictions in the abstract without being asked to pay attention to their spontaneous affective reactions and not seeing any pictures of the target groups. This suggests that it is not simply the act of asking participants directly to predict a specific test that differentiated the prediction task from the

²⁵ In some studies, traditional explicit measures and predictions differed in other dimensions not explicitly listed at this point because either the difference did not affect the outcome in a substantial way (e.g., lengthy explanations of the concept of implicit attitudes before IAT score prediction) or the difference was intentionally manipulated to test a specific hypothesis (e.g., only thinking about a prediction without actually completing a scale)

traditional explicit scale. Rather, as predicted by the framework, paying attention to concrete stimuli seems to be a crucial aspect to be able to introspect upon one's automatic cognitions.

An alternative explanation for why predictions were more related to participants IAT score patterns than traditional explicit measures is provided by the idea of structural fit (Payne et al., 2008). Payne et al. (2008) demonstrated that similarity in task demands led to higher correlations between implicit and explicit measures. With regard to the present research, this could suggest that showing pictures, contrasting target groups, and explicitly asking to reflect on spontaneous affective reactions made the explicit rating maximally similar in structure to the IAT instead of genuinely increasing awareness. The fact that the average between-subject correlation between predictions and IAT scores across target groups showed a comparable meta-analytical effect of $b = 0.22$ (Chapter 2) to the meta-analytical effect based on between-subject correlations between traditional explicit ratings and IAT scores of $r = .24$ (Hofmann, Gawronski et al., 2005) speaks against this idea. If the structural fit was, in fact, the main driver of higher correlations between predictions and IAT scores in the present studies, this should have manifested in between-subject correlations as well. Instead, in line with the propositions of the framework, predictions were more strongly related to IAT score patterns than traditional explicit ratings in within-subject analysis, but this difference did not seem to hold in between-subject correlations. This suggests that predictions are different from traditional explicit ratings because their procedural aspects increased introspective awareness by enhancing the signal and making people pay attention to this signal. In contrast, these procedural differences between predictions and traditional explicit ratings did not seem to influence people's ability to calibrate their responses.

5.4. Social Calibration or Socially Desirable Responding?

One may wonder whether the concept of social calibration is just another word for the problem of socially desirable responding (SDR; Crowne & Marlowe, 1960). It could be argued that the differences in labeling preferences in the domain of social groups and baked

goods in Chapter 4 may have emerged because participants were motivated to present themselves in a more positive light in the socially sensitive topic of social groups, but less so in the domain of baked goods. While this may be one reason for participants to report labels out of sync with the sample distribution, this explanation may be too narrow, and other important factors influencing participants labeling preferences may be overlooked. For example, people may more often talk to other people about their food preferences and compare their own tastes to those of others. One can easily imagine a conversation going something like “I just love croissants, they are my favorite food.” and a friend answering “Really? Well I am more of a savory person, croissants are fine, but I really prefer crackers.”. The same conversation talking about social groups, such as talking about Black and White people, feels rather awkward and would rarely take place (“I just love White people, they are my favorite people.”). As such, people may be much less experienced in reporting their own evaluations of social groups as reporting their food preferences, which may be another explanation for inconsistent labeling decisions. Of course, the fact that people rarely communicate their evaluations of social groups may be due to the sensitivity of the topic, but the ultimate reason for people to then choose different labels in the domain of social groups is not a deliberate choice to respond in a socially desirable way but rather the consequence of a lack of knowledge. Hence, the concept of social calibration goes beyond the mere problem of socially desirable responding in several ways. First, it additionally considers social knowledge and comparison processes involved in reporting on one’s automatic cognitions. And second, it includes other explanations for unwillingness to report on one’s automatic cognitions beyond deliberate dishonesty.

5.5. Implications for Theories on Implicit and Explicit Evaluations

The proposed framework along with the empirical findings has important implications for theories on implicit and explicit evaluations. As laid out before, current theories suggest that implicit evaluations either reflect fully unconscious cognitions people are unable to

introspect upon (Greenwald & Banaji, 1995), or cognitions that are fully conscious but people reject these for explicit reports and deliberately choose to reveal other information to the researcher (Fazio, 1990, 2007; Gawronski & Bodenhausen, 2006, 2011). The present findings are at odds with both conceptualizations.

First, the fact that participants in Chapters 2 and 3 were consistently able to accurately predict the pattern of their IAT results clearly contradicts the unconscious hypothesis (Hahn et al., 2014; Hahn & Goedderz, 2020). At the same time, the fact that people often report feeling surprise at IAT feedback as demonstrated by Chapter 3, is hard to reconcile with the idea of fully conscious cognitions (Gawronski, 2019; Goedderz & Hahn, 2022; Krickel, 2018). Importantly, making participants pay attention to their spontaneous affective reactions was more effective than using other labels to describe biases or introducing implicit evaluations in a non-threatening way for lowering their reported surprise. This suggests that participants did not simply pretend to be surprised at their IAT feedback, or were surprised at the labels used to describe their biases while being fully aware of their own automatic cognitions. Instead, it indicates that participants were not aware of their own cognitions reflected on their implicit evaluations until they paid attention to them. In a similar vein, Hahn and Gawronski (2019) had already documented that participants in their studies aligned their explicit reports more with implicit measures and acknowledged their biases more after IAT score predictions. This illustrates that participants seemed to have learned something new about themselves after predicting their IAT scores. In line with theories on consciousness and the proposed framework, this suggests that the cognitions reflected on implicit evaluations are consciously accessible, but often reside in a preconscious state until certain conditions are met (Dehaene et al., 2006; Goedderz & Hahn, 2022; Hahn & Goedderz, 2020; Hofmann & Wilson, 2010).

Second, the findings that people demonstrated different levels of awareness when the correlation between their predictions and their IAT scores were computed within-subject than when they were computed between-subject also contradicts theories that assume constant

awareness of implicit evaluations. That is, such models would predict that whenever people consider the same information for their explicit reports that is also reflected on their implicit measures, implicit and explicit measures should show high correlations. This prediction held true when examining within-subject correlations. However, correlations between predictions and IAT scores were less accurate when examined between-subject, even though we believe that participants considered comparable information when they made predictions as were reflected on their IAT scores. In line with the framework presented in this dissertation, this suggests that even if people consider the same information for their explicit reports and their implicit responses, reports may still differ considerably because people do not calibrate their responses consistently. This was illustrated by Chapter 4 which showed that even though participants were equally good at predicting their pattern of IAT results in the domains of social groups and baked goods, they were less good at placing their IAT results in the sample distribution in the domain of social groups due to labeling preferences. Specifically, in the social-groups domain, participants largely restricted their predictions to the middle of the scale (“little to no preference”, “mild preference”) while in the baked-goods domain, participants used the full prediction scale. This suggests that implicit and explicit evaluations are not only different because they rely on different information (automatically activated associations vs. propositional evaluations), but also because responses on explicit ratings are based on subjective labeling preferences, while implicit measures are not.

Together the present findings contradict both theories which suggest that (1) cognitions reflected in implicit evaluations are completely unconscious and (2) these cognitions are completely conscious at all times. Instead the present findings suggest that people often remain unaware of their automatic cognitions until certain conditions are met and that they are often aware of their own automatic cognitions but unaware of the social meaning of these cognitions. As such, the present framework moves the discussion on theories of implicit and explicit evaluations beyond a simple dichotomy of conscious and unconscious

cognitions and proposes a more nuanced perspective to studying the underlying cognitions reflected in implicit and explicit evaluations.

5.6. Introspective Awareness and Social Calibration of Other Automatic Cognitions

The research presented in this dissertation focused on introspective awareness and social calibration of the cognitions reflected on implicit evaluations. I deliberately chose this area for several reasons. First, the debate around awareness of the cognitions reflected on implicit measures is rooted in research on implicit evaluations and most theories revolve around the distinction between implicit and explicit evaluations. Hence, the framework seemed especially applicable and useful in this area of research. Second, the fundamental assumptions of the framework are based on Hahn et al.'s (2014) prediction studies in the domain of social groups. To examine the robustness of the initial findings and to examine individual mechanisms proposed by the framework, I aimed at sticking as closely as possible to the original paradigm while changing only small parts without introducing too many additional confounds. Nonetheless, I believe that the principles of the framework should be applicable to other automatic cognitions and behaviors. For instance, a plethora of research revolves around self-knowledge of personality dispositions and often suggests that we have only limited introspective awareness of our own personality (Vazire, 2010; Vazire & Carlson, 2010). Most of these studies infer self-knowledge from between-subject correlations between self-reports and external criteria (e.g., Back et al., 2009; Vazire, 2010; Vazire & Carlson, 2010) and may thus conflate introspective awareness and social calibration. For instance, knowing your own personality pattern (e.g., knowing that you are more extroverted than introverted, highly conscientious and agreeable, but less open to experiences and little neurotic) is something different than knowing how to label the magnitude of your personality (e.g., knowing that you are “strongly” neurotic). The framework further proposes that introspective awareness and social calibration is susceptible to different information. Whereas introspective awareness requires a focus on internal processes, social calibration requires a

focus on external processes in comparison to others. Adapting this reasoning to the area of personality suggests that personality traits that are mostly observed internally (e.g., neuroticism) should be easier to introspect upon but more difficult to calibrate, while personality traits that reflect in externally observable behavior (e.g., intellect) should be easier to calibrate but more difficult to introspect upon (see also Vazire, 2010). The example of self-knowledge in personality illustrates that the presented framework could lead to a more nuanced understanding of self-knowledge and the ability to report on one's own automatic cognitions more generally. Future research will show whether the proposed mechanisms of the framework hold when applied to other automatic cognitions and behaviors.

5.7. Generalizability to Other Implicit Measures

All studies presented in this dissertation used the IAT as the central criterion to measure implicit evaluations. I chose the IAT because it is the most widely used implicit measure (Hahn & Gawronski, 2018), it is comparably more reliable than other implicit measures (Bar-Anan & Nosek, 2012; Gawronski & De Houwer, 2014), and I wanted to stay as close as possible to the original paradigm. However, the IAT has received a lot of criticism pertaining to methodological and conceptual problems. For example, the psychometric properties of the IAT have been vigorously debated and researchers are still in disagreement about whether the IAT is suited to study individual differences in attitudes (Blanton et al., 2006; Carpenter et al., 2022; Payne et al., 2017). It has further been established that the IAT is not a process pure measure of implicit evaluations, but that it also captures task-specific variance such as control processes (Conrey et al., 2005; Payne, 2005, 2008). Further, implicit measures, including the IAT tend to show low correlation between each other (Bar-Anan & Nosek, 2012). All these aspects pose the question of whether the findings of the present dissertation are generalizable to the cognitions reflected on other implicit measures. Initial evidence that the basic prediction effect replicates for other implicit measures as well is provided by Morris and Kurdi (2022) who found that participants were also able to accurately

predict their results on the Affective Misattribution Procedure (AMP; Payne et al., 2005). While these findings support the generalizability of the present research, the different procedural aspects of different implicit measures may also show differences in how well people can report their outcomes on such measures. I believe that the propositions of the framework may also help to understand which implicit measures should be easier and which more difficult to report on. For instance, the hypothesis that stronger signals should be easier to introspect upon suggests that it will be more difficult for people to report on implicit measures that elicit less strong reactions – for example, if they measure reactions toward one attitude object at a time instead of in contrast. To gain a broad understanding of how much people know about the cognitions reflected on different implicit measures, future research applying the principles of the framework is paramount.

5.8. Practical Implications

The concept of implicit biases has received a lot of attention outside academic circles and found its way into the general public by offering a potential explanation for persisting racial discrimination (BBC News, 2017; Devlin, 2018; Grinberg, 2015). The topic has been so influential, it was even discussed in the US presidential debate of 2016, where Hillary Clinton said that all people have implicit biases (Merica, 2016). To tackle the issue of implicit biases, so-called “implicit bias trainings” are on the rise (Chamorro-Premuzic, 2020; Green & Hagiwara, 2020; Robson, 2021). Most prominently, after an incident of racial discrimination in a Starbucks coffee shop, the company would give their employees “unconscious bias training” (Basu, 2018). Lately, the effectiveness of such bias trainings has been debated, and many scholars have pointed out that it may not lead to meaningful changes (Green & Hagiwara, 2020). Among others, one problem of such implicit bias trainings in the past may have been the conceptualization of implicit biases as capturing unconscious cognitions. If one assumes that implicit biases are impossible for people to gain conscious access to, then the only way of tackling implicit biases would be to tell people about their biases or to reduce the

underlying biases altogether. Both approaches have been proven to be difficult. First, most people react defensively if they are informed about their biases, potentially making it less likely for them to acknowledge their own biases and want to change their own behaviors (Howell et al., 2013; Howell & Ratliff, 2017). Second, research by Lai et al. (2016) has documented that interventions aimed at reducing implicit racial preferences showed only short-term effects but did not lead to long-term changes in implicit preferences.

I believe that the present framework provides a new perspective on how to approach implicit bias interventions by moving beyond the narrative of unconscious biases and acknowledging that people are able to actively pay attention to their biases. The research provided in this dissertation suggests that instead of telling people about their biases or trying to change the underlying cognitions, a simple encouragement to pay attention to one's biases may be a promising road to raising awareness without making people defensive. Indeed, Hahn and Gawronski (2019) found that predicting IAT scores as means to make people pay attention to their biases was more effective to raise acknowledgment of biases than merely completing IATs or receiving feedback on IAT performance. Whether acknowledgment of biases ultimately leads to more engagement in egalitarian behavior is an open question yet to be addressed by future research, but it opens a new promising road for bias interventions.

5.9. Conclusion

The aim of the present dissertation was to provide a more nuanced understanding of when people are able to gain introspective awareness of their own automatic cognitions. To this end, based on theories of consciousness and introspection (Dehaene et al., 2006; Hofmann & Wilson, 2010), along with research on awareness and implicit evaluations (Hahn et al., 2014; Hahn & Gawronski, 2019; Hahn & Goedderz, 2020), I presented a framework on reporting on automatic cognitions proposing factors that facilitate or impede awareness of one's own automatic cognitions. Specifically, the framework introduced two concepts that are involved when people report on their automatic cognitions. It distinguishes between the

question of whether a person is aware of their own automatic cognitions toward different targets (*introspective awareness*) or aware of how their automatic cognitions should be labeled in accordance with conventions (*social calibration*). In the present thesis, it was demonstrated that both concepts of awareness can be successfully empirically distinguished and can differ substantially from one another, depending on the domain of study.

Additionally, initial evidence was presented in line with the proposed factors facilitating introspective awareness or social calibration. The findings suggested that people's automatic cognitions may often reside in a preconscious state until they are attended to, and people may often be miscalibrated in reporting their automatic cognitions because they lack knowledge about labeling conventions or are unwilling to apply strong labels in socially sensitive topics. The framework and the presented research advance the understanding of the nature of cognitions reflected on implicit measures and have important implications for theories on implicit social cognition and bias interventions. I hope that the present thesis can inspire new ways of studying awareness of automatic cognitions, and it can provide a more profound understanding of how much people know about themselves.

References

- Akram, S. (2018). Representative bureaucracy and unconscious bias: Exploring the unconscious dimension of active representation. *Public Administration*, *96*(1), 119–133. <https://doi.org/10.1111/padm.12376>
- Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & al, e. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68*(5), 804–825. <https://doi.org/10.1037/0022-3514.68.5.804>
- Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. <https://doi.org/10.1080/10463280802613866>
- Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661. <https://doi.org/10.1037/0022-3514.91.4.652>
- Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology*, *97*(3), 533–548. <https://doi.org/10.1037/a0016229>
- Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health*, *22*(4), 153–160. <https://doi.org/10.1136/ebmental-2019-300117>
- Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift Für Experimentelle Psychologie*, *48*(2), 145–160. <https://doi.org/10.1026//0949-3946.48.2.145>
- Bar-Anan, Y., & Nosek, B. A. (2012). A comparative investigation of seven implicit measures of social cognition. *SSRN Electronic Journal*, *38*, 1193. <https://doi.org/10.2139/ssrn.2074556>

- Basu, T. (2018, April 17). *Starbucks Will Give Employees Unconscious Bias Training. That May Not Help.: When it comes to implicit bias training, the science just doesn't hold up: There isn't much scientific proof that it actually works.* The Daily Beast.
<https://www.thedailybeast.com/starbucks-will-give-employees-unconscious-bias-training-that-may-not-help>
- BBC News. (2017, June 5). *Implicit bias: Is everyone racist?* BBC News.
<https://www.bbc.com/news/magazine-40124781>
- Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology*, 6, 1–62. [https://doi.org/10.1016/S0065-2601\(08\)60024-6](https://doi.org/10.1016/S0065-2601(08)60024-6)
- Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up? *Journal of Experimental Social Psychology*, 43(3), 399–409.
<https://doi.org/10.1016/j.jesp.2006.10.019>
- Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, 42(2), 192–212. <https://doi.org/10.1016/j.jesp.2005.07.003>
- Carpenter, T. P., Goedderz, A., & Lai, C. K. (2022). Individual differences in implicit bias can be measured reliably by administering the same implicit association test multiple times. *Personality & Social Psychology Bulletin*, 1461672221099372.
<https://doi.org/10.1177/01461672221099372>
- Carpenter, T. P., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S., LaBouff, J., Isenberg, N., & Chakroff, A. (2019). Survey-software implicit association tests: A methodological and empirical analysis. *Behavior Research Methods*, 51(5), 2194–2208.
<https://doi.org/10.3758/s13428-019-01293-3>
- Chamorro-Premuzic, T. (2020, January 4). *Implicit Bias Training Doesn't Work: Instead of changing how employees think, change company policies.* Bloomberg.

<https://www.bloomberg.com/opinion/articles/2020-01-04/implicit-bias-training-isn-t-improving-corporate-diversity>

Cole, K. (2018). Thinking through race: white racial identity, motivated cognition and the unconscious maintenance of white supremacy. *Politics, Groups, and Identities*, 6(2), 181–198. <https://doi.org/10.1080/21565503.2016.1198708>

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology*, 89(4), 469–487. <https://doi.org/10.1037/0022-3514.89.4.469>

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349–354. <https://doi.org/10.1037/h0047358>

Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin*, 30(10), 1332–1346. <https://doi.org/10.1177/0146167204264654>

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, 12, 163–170. <https://doi.org/10.1111/1467-9280.00328>

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, 135(3), 347–368. <https://doi.org/10.1037/a0014211>

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10(5), 204–211. <https://doi.org/10.1016/j.tics.2006.03.007>

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2), 1–37.

- Deutsche UNESCO-Kommission e. V. (2019). *Bundesweites Verzeichnis Immaterielles Kulturerbe: A bis Z = German inventory of intangible cultural heritage* (3., aktualisierte Auflage, Stand: November 2019). Deutsche UNESCO-Kommission e. V.
- Devlin, H. (2018, December 2). *Unconscious bias: what is it and can it be eliminated?* The Guardian. <https://www.theguardian.com/uk-news/2018/dec/02/unconscious-bias-what-is-it-and-can-it-be-eliminated>
- Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, 23(3), 316–326. <https://doi.org/10.1177/0146167297233009>
- Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Dryden Press.
- Fatfouta, R., & Schröder-Abé, M. (2018). A wolf in sheep's clothing? Communal narcissism and positive implicit self-views in the communal domain. *Journal of Research in Personality*, 76(7), 17–21. <https://doi.org/10.1016/j.jrp.2018.07.004>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE Model as an integrative framework. *Advances in Experimental Social Psychology*, 23, 75–109. [https://doi.org/10.1016/S0065-2601\(08\)60318-4](https://doi.org/10.1016/S0065-2601(08)60318-4)
- Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, 25, 603–637. <https://doi.org/10.1521/soco.2007.25.5.603>
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69(6), 1013–1027. <https://doi.org/10.1037/0022-3514.69.6.1013>

- Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327.
<https://doi.org/10.1146/annurev.psych.54.101601.145225>
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238.
<https://doi.org/10.1037/0022-3514.50.2.229>
- Fiske, S. T. (2017). Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion) . *Perspectives on Psychological Science*, *12*(5), 791–799. <https://doi.org/10.1177/1745691617708204>
- Friese, M., Wänke, M., & Plessner, H. (2006). Implicit consumer preferences and their influence on product choice. *Psychology & Marketing*, *23*(9), 727–740.
<https://doi.org/10.1002/mar.20126>
- Gallup. (2021). *Race Relations*. Gallup. <https://news.gallup.com/poll/1687/race-relations.aspx>
- Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, *14*(4), 574–595.
<https://doi.org/10.1177/1745691619826015>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. <https://doi.org/10.1037/0033-2909.132.5.692>
- Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology*, *44*, 59–127. <https://doi.org/10.1016/B978-0-12-385522-0.00002-0>
- Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 283–310). Cambridge University Press.

- Gawronski, B., Deutsch, R., & Seidel, O. (2005). Contextual influences on implicit evaluation: A test of additive versus contrastive effects of evaluative context stimuli in affective priming. *Personality & Social Psychology Bulletin*, *31*(9), 1226–1236. <https://doi.org/10.1177/0146167205274689>
- Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, *33*(5), 573–589. <https://doi.org/10.1002/ejsp.166>
- Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are “implicit” attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499. <https://doi.org/10.1016/j.concog.2005.11.007>
- Gawronski, B., Houwer, J. de, & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition*, *38*(Supplement), s1-s25. <https://doi.org/10.1521/soco.2020.38.suppl.s1>
- Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, *44*(5), 1355–1361. <https://doi.org/10.1016/j.jesp.2008.04.005>
- Goedderz, A., & Hahn, A. (2022). Biases left unattended: People are surprised at racial bias feedback until they pay attention to their biased reactions. *Journal of Experimental Social Psychology*, *102*, 104374. <https://doi.org/10.1016/j.jesp.2022.104374>
- Goedderz, A., & Hahn, A. (2023). Predicting implicit preferences towards social groups vs. food items: The role of normativity and social desirability in accurate IAT score predictions. [Manuscript under review].
- Goedderz, A., Rahmani-Azad, Z., & Hahn, A. (2023). Awareness of implicit attitudes revisited: Meta-analysis on replications across samples and settings. [Manuscript in preparation].

- Goedderz, A., Sperlich, L., & Hahn, A. (2022, February 18). *Bias in Contrast – How the Implicit Association Test May Exacerbate Bias* [Conference Presentation]. Society for Personality and Social Psychology's Annual Convention, San Francisco, USA.
<https://tinyurl.com/2p9x4smx>
- Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. <https://doi.org/10.1111/spc3.12267>
- Green, T. L., & Hagiwara, N. (2020, August 28). *The Problem with Implicit Bias Training*. Scientific American. <https://www.scientificamerican.com/article/the-problem-with-implicit-bias-training/>
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. <https://doi.org/10.1037/0033-295X.102.1.4>
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*(1), 3–25. <https://doi.org/10.1037/0033-295X.109.1.3>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. <https://doi.org/10.1037/0022-3514.85.2.197>
- Grinberg, E. (2015, November 25). *4 ways you might be displaying hidden bias in everyday life*. CNN. <https://edition.cnn.com/2015/11/24/living/implicit-bias-tests-feat/index.html>

- Hahn, A., & Gawronski, B. (2014). Do implicit evaluations reflect unconscious attitudes? *Behavioral and Brain Sciences*, 37(1), 28–29.
<https://doi.org/10.1017/S0140525X13000721>
- Hahn, A., & Gawronski, B. (2018). Implicit social cognition. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 4, 1–33.
<https://onlinelibrary.wiley.com/doi/full/10.1002/9781119170174.epcn412>
- Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, 116(5), 769–794.
<https://doi.org/10.1037/pspi0000155>
- Hahn, A., & Goedderz, A. (2020). Trait-unconsciousness, state-unconsciousness, preconsciousness, and social miscalibration in the context of implicit evaluation. *Social Cognition*, 38(Supplement), s115-s134.
<https://doi.org/10.1521/soco.2020.38.sup.s115>
- Hahn, A., & Goedderz, A. (2023). Beyond dishonesty and unawareness. Accuracy of IAT score predictions depends more on the concreteness of the question than on knowledge of measurement. [Manuscript in preparation].
- Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3), 1369–1392.
<https://doi.org/10.1037/a0035028>
- Haider, A. H., Schneider, E. B., Sriram, N., Dossick, D. S., Scott, V. K., Swoboda, S. M., Losonczy, L., Haut, E. R., Efron, D. T., Pronovost, P. J., Freischlag, J. A., Lipsett, P. A., Cornwell, E. E., MacKenzie, E. J., & Cooper, L. A. (2014). Unconscious race and class bias: Its association with decision making by trauma and acute care surgeons. *The Journal of Trauma and Acute Care Surgery*, 77(3), 409–416.
<https://doi.org/10.1097/TA.0000000000000392>

- Haider, A. H., Sexton, J., Sriram, N., Cooper, L. A., Efron, D. T., Swoboda, S., Villegas, C. V., Haut, E. R., Bonds, M., & Pronovost, P. J. (2011). Association of unconscious race and social class bias with vignette-based clinical assessments by medical students. *Jama*, *306*(9), 942–951.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61-83; discussion 83-135.
<https://doi.org/10.1017/S0140525X0999152X>
- Hillard, A. L., Ryan, C. S., & Gervais, S. J. (2013). Reactions to the Implicit Association Test as an educational tool: A mixed methods study. *Social Psychology of Education*, *16*(3), 495–516. <https://doi.org/10.1007/s11218-013-9219-5>
- Hinojosa, J. A., Carretié, L., Valcárcel, M. A., Méndez-Bértolo, C., & Pozo, M. A. (2009). Electrophysiological differences in the processing of affective information in words and pictures. *Cognitive, Affective & Behavioral Neuroscience*, *9*(2), 173–189.
<https://doi.org/10.3758/CABN.9.2.173>
- Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385.
<https://doi.org/10.1177/0146167205275613>
- Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On implicit-explicit consistency: the moderating role of individual differences in awareness and adjustment. *European Journal of Personality*, *19*(1), 25–49. <https://doi.org/10.1002/per.537>
- Hofmann, W., Gschwendner, T., & Schmitt, M. (2009). The road to the unconscious self not taken: Discrepancies between self- and observer-inferences about implicit dispositions from nonverbal behavioural cues. *European Journal of Personality*, *23*(4), 343–366.
<https://doi.org/10.1002/per.722>

- Hofmann, W., & Wilson, T. D. (2010). Consciousness, introspection, and the adaptive unconscious. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 197–215). Guildford Press.
- Houwer, J. de, & Hermans, D. (1994). Differences in the affective processing of words and pictures. *Cognition & Emotion*, *8*(1), 1–20.
<https://doi.org/10.1080/02699939408408925>
- Howell, J. L., Collisson, B., Crysel, L., Garrido, C. O., Newell, S. M., Cottrell, C. A., Smith, C. T., & Shepperd, J. A. (2013). Managing the threat of impending implicit attitude feedback. *Social Psychological and Personality Science*, *4*(6), 714–720.
<https://doi.org/10.1177/1948550613479803>
- Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2015). Caught in the middle: defensive responses to IAT feedback among Whites, Blacks, and biracial Black/Whites. *Social Psychological and Personality Science*, *6*(4), 373–381.
<https://doi.org/10.1177/1948550614561127>
- Howell, J. L., & Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to Implicit Association Test feedback. *British Journal of Social Psychology*, *56*(1), 125–145. <https://doi.org/10.1111/bjso.12168>
- Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive responding to IAT feedback. *Social Cognition*, *35*(5), 520–562.
<https://doi.org/10.1521/soco.2017.35.5.520>
- Jacoby, L. L., & Witherspoon, D. (1982). Remembering without awareness. *Canadian Journal of Psychology/Revue Canadienne De Psychologie*, *36*(2), 300–324.
<https://doi.org/10.1037/h0080638>
- Jost, J. T., Pelham, B. W., & Carvallo, M. R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of*

Experimental Social Psychology, 38(6), 586–602. [https://doi.org/10.1016/S0022-1031\(02\)00505-X](https://doi.org/10.1016/S0022-1031(02)00505-X)

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond (Third Edition)*. Routledge Taylor & Francis Group.

Jussim, L., Yen, H., & Aiello, J. R. (1995). Self-consistency, self-enhancement, and accuracy in reactions to feedback. *Journal of Experimental Social Psychology*, 31(4), 322–356. <https://doi.org/10.1006/jesp.1995.1015>

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, 3(3), 522–540. <https://doi.org/10.1111/j.1747-9991.2008.00138.x>

Kensinger, E. A., & Schacter, D. L. (2006). Processing emotional pictures and words: Effects of valence and arousal. *Cognitive, Affective & Behavioral Neuroscience*, 6(2), 110–126. <https://doi.org/10.3758/cabn.6.2.110>

Krickel, B. (2018). Are the states underlying implicit biases unconscious? – A Neo-Freudian answer. *Philosophical Psychology*, 31(7), 1007–1026. <https://doi.org/10.1080/09515089.2018.1470323>

Kruglanski, A. W. (1989). The psychology of being “right”: The problem of accuracy in social perception and cognition. *Psychological Bulletin*, 106(3), 395–409.

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, 116(13), 5862–5871. <https://doi.org/10.1073/pnas.1820240116>

Kurdi, B., Ratliff, K. A., & Cunningham, W. A. (2020). Can the Implicit Association Test serve as a valid measure of automatic cognition? A response to Schimmack (2020). *Perspectives on Psychological Science*, 1-13. <https://doi.org/10.1177/1745691620904080>

- Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American Psychologist*, *74*(5), 569–586. <https://doi.org/10.1037/amp0000364>
- Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, *7*(5), 315–330. <https://doi.org/10.1111/spc3.12023>
- Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology. General*, *145*(8), 1001–1016. <https://doi.org/10.1037/xge0000179>
- Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, *8*(8), 875–881. <https://doi.org/10.1177/1948550617693058>
- Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding the Implicit Association Test: IV. *Implicit Measures of Attitudes*, 59–102.
- Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016, November 30). *Research Preregistration 101*. <https://www.psychologicalscience.org/observer/research-preregistration-101>
- McConnell, A. R., Dunn, E. W., Austin, S. N., & Rawn, C. D. (2011). Blind spots in the search for happiness: Implicit attitudes and nonverbal leakage predict affective forecasting errors. *Journal of Experimental Social Psychology*, *47*(3), 628–634. <https://doi.org/10.1016/j.jesp.2010.12.018>

- Merica, D. (2016). *Hillary Clinton talks race: "We all have implicit biases"*. CNN.
<https://edition.cnn.com/2016/04/20/politics/hillary-clinton-race-implicit-biases/index.html>
- Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4), 630–633.
<https://doi.org/10.3758/BF03206543>
- Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, 19(4), 395–417. <https://doi.org/10.1521/soco.19.4.395.20759>
- Morris, A., & Kurdi, B. (2022). Awareness of implicit attitudes: Large-scale investigations of mechanism and scope. PsyArXiv. <https://doi.org/10.31234/osf.io/dmjfq>
- Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, 18(2), 107–118. <https://doi.org/10.1177/1088868313496330>
- Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations*, 8(1), 39–52.
<https://doi.org/10.1177/1368430205048615>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259. <https://doi.org/10.1037/0033-295X.84.3.231>
- Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science*, 21(4), 511–517. <https://doi.org/10.1177/0956797610364762>
- Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, 134(4), 565–584.
<https://doi.org/10.1037/0096-3445.134.4.565>

- Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69. <https://doi.org/10.1111/j.1467-8721.2007.00477.x>
- Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. <https://doi.org/10.1037//1089-2699.6.1.101>
- Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion*, *22*(4), 553–594. <https://doi.org/10.1080/02699930701438186>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, *19*(2), 172–182. <https://doi.org/10.1037/gpr0000034>
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192. <https://doi.org/10.1037/a0032734>
- Payne, B. K. (2005). Conceptualizing control in social cognition: How executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology*, *89*(4), 488–503. <https://doi.org/10.1037/0022-3514.89.4.488>
- Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass*, *2*(2), 1073–1092. <https://doi.org/10.1111/j.1751-9004.2008.00091.x>

- Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, *94*(1), 16–31. <https://doi.org/10.1037/0022-3514.94.1.16>
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. <https://doi.org/10.1037/0022-3514.89.3.277>
- Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1–15). Guildford Press.
- Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248. <https://doi.org/10.1080/1047840X.2017.1335568>
- Perry, S. P., Murphy, M. C., & Dovidio, J. F. (2015). Modern prejudice: Subtle, but unconscious? The role of Bias Awareness in Whites' perceptions of personal and others' biases. *Journal of Experimental Social Psychology*, *61*, 64–78. <https://doi.org/10.1016/j.jesp.2015.06.007>
- Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738. <https://doi.org/10.1162/089892900562552>
- Powell, J. A. (2016, September 27). *Implicit bias in the presidential debate [Blog post]*. UC Berkeley. Berkeley Blog. <https://blogs.berkeley.edu/2016/09/27/implicit-bias-in-the-presidential-debate/>
- Quillian, L. (2008). Does unconscious racism exist? *Social Psychology Quarterly*, *71*(1), 6–11. <https://doi.org/10.1177/019027250807100103>

- Rahmani Azad, Z., Goedderz, A., & Hahn, A. (2022). Self-awareness and stereotypes: Accurate prediction of implicit gender stereotyping. *Personality and Social Psychology Bulletin*, 1461672221120703. <https://doi.org/10.1177/01461672221120703>
- Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, 44(2), 386–396. <https://doi.org/10.1016/j.jesp.2006.12.008>
- Redford, L. (2018). *Mapping County-level Geographical Variation in Implicit Racial Attitudes* [November 4, 2018]. Project Implicit. <https://www.implicit.harvard.edu/implicit/blog.html>
- Richetin, J., Perugini, M., Prestwich, A., & O’Gorman, R. (2007). The IAT as a predictor of food choice: The case of fruits versus snacks. *International Journal of Psychology*, 42(3), 166–173. <https://doi.org/10.1080/00207590601067078>
- Rivers, A. M., & Hahn, A. (2019). What cognitive mechanisms do people reflect on when they predict IAT scores? *Personality & Social Psychology Bulletin*, 45(6), 878–892. <https://doi.org/10.1177/0146167218799307>
- Robson, D. (2021, April 25). *What unconscious bias training gets wrong... and how to fix it*. The Guardian. <https://www.theguardian.com/science/2021/apr/25/what-unconscious-bias-training-gets-wrong-and-how-to-fix-it>
- Roefs, A., Quaedackers, L., Werrij, M. Q., Wolters, G., Havermans, R., Nederkoorn, C., van Breukelen, G., & Jansen, A. (2006). The environment influences whether high-fat foods are associated with palatable or with unhealthy. *Behaviour Research and Therapy*, 44(5), 715–736. <https://doi.org/10.1016/j.brat.2005.05.007>
- Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit

- Association Test. *Social Cognition*, 17(4), 437–465.
<https://doi.org/10.1521/soco.1999.17.4.437>
- Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(3), 501–518.
- Schimmack, U. (2019). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*. Advance online publication.
<https://doi.org/10.1177/1745691619863798>
- Schlachter, S., & Rolf, S. (2017). Using the IAT: how do individuals respond to their results? *International Journal of Social Research Methodology*, 20(1), 77–92.
<https://doi.org/10.1080/13645579.2015.1117799>
- Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, 84(1), 60–79. <https://doi.org/10.1037/0022-3514.84.1.60>
- Seibt, B., Häfner, M., & Deutsch, R. (2007). Prepared to eat: how immediate affective and motivational responses to food cues are influenced by food deprivation. *European Journal of Social Psychology*, 37(2), 359–379. <https://doi.org/10.1002/ejsp.365>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
<https://doi.org/10.1177/0956797611417632>
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1), 76–80. <https://doi.org/10.1177/1745691613514755>
- Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology*, 42(4), 300–313.
<https://doi.org/10.1027/1864-9335/a000072>

- Stiensmeier-pelster, J., Martini, A., & Reisenzein, R. (1995). The role of surprise in the attribution process. *Cognition & Emotion*, 9(1), 5–31.
<https://doi.org/10.1080/02699939508408963>
- Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, 59(5).
<http://www.jstatsoft.org/v59/i05/>
- Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In C. Unkelbach & R. Greifeneder (Eds.), *The Experience of Thinking: How feelings from mental processes influence cognition and behaviour* (1st ed., pp. 11–32). Psychology Press.
- Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, 98(2), 281–300.
<https://doi.org/10.1037/a0017908>
- Vazire, S., & Carlson, E. N. (2010). Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass*, 4(8), 605–620.
<https://doi.org/10.1111/j.1751-9004.2010.00280.x>
- Vitriol, J. A., & Moskowitz, G. B. (2021). Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change. *Journal of Experimental Social Psychology*, 96(8), 104165.
<https://doi.org/10.1016/j.jesp.2021.104165>
- Vosgerau, J., Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). 99% impossible: A valid, or falsifiable, internal meta-analysis. *Journal of Experimental Psychology: General*, 148(9), 1628–1639. <https://doi.org/10.1037/xge0000663>
- Wen, T. (2020, August 28). *Is it possible to rid police officers of bias?* BBC.
<https://www.bbc.com/future/article/20200827-is-it-possible-to-rid-police-officers-of-bias>

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, 55, 493–518.

<https://doi.org/10.1146/annurev.psych.55.090902.141954>