Nudging, Judging, and Deciding:

Essays on Behavior and Evaluation in Organizations

Der Fakultät für Wirtschaftswissenschaften der

Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften

– Doctor rerum politicarum –

vorgelegte Dissertation von

Miro Mehic, M.Sc.

Geboren am 03.02.1996 in Brilon (Deutschland)

Erscheinungsjahr 2026

ACKNOWLEDGEMENTS

The past few years of my doctoral studies have been a meaningful chapter in my life. Throughout this journey, I had the privilege of crossing paths with many wonderful people who supported and inspired me. I would like to express my sincere gratitude to everyone who contributed to this experience and accompanied me along the way.

To begin with, I would like to express my deepest gratitude to Professor Kirsten Thommes, who has been an exceptional supervisor throughout my doctoral journey. Her great enthusiasm for research and her caring mentorship consistently motivated and inspired me. I am deeply grateful for the opportunity to have been part of her research group for such a long time, where I was able to grow both professionally and personally while gaining many valuable experiences and insights.

I would also like to sincerely thank the members of my dissertation committee, Professor Martin Schneider, Professor Philip Yang, and Professor Stefan Jungblut, for their time and insightful comments on this dissertation. I greatly appreciate their willingness to contribute to this work.

Furthermore, I am truly grateful for the opportunity to have met so many amazing colleagues throughout the years, both at the chair and beyond. Special thanks go to Conny, Elena, Glory, Jana, Jarek, Jörg, Katharina, Liv, Maurice, Olesja, Robin, and Thorsten for the inspiring discussions, shared experiences, and unforgettable moments that made this journey so special.

Beyond academia, I would like to express my heartfelt gratitude to all my friends and family. In particular, I would like to thank Toni, Bürgi, Fiete, Mo, Joe, and Gawan for always believing in me and supporting me throughout this journey. I am especially grateful to Franzi, who has had an incredibly positive impact on my life over the past year and whose constant encouragement means more to me than words can express. Lastly, I would like to thank my parents for their endless love and sacrifices. Without them, I would not have come this far.

Miro Mehic

Paderborn, May 2026

TABLE OF CONTENTS

LIST OF PAPERS

Paper 1

Mehic, M. (2026). *How effective is nudging in the long run? A meta-analysis of pro-environmental behavior* (Working Paper No. 178). Paderborn University, Faculty of Business Administration and Economics. https://ideas.repec.org/p/pdn/dispap/178.html

Paper 2

Thommes, K., & Mehic, M. (2026). The persistence of default effects: Evidence from CO_2 offsetting in cargo transportation. *Transportation Research Part A: Policy and Practice*, *204*, Article 104838. https://doi.org/10.1016/j.tra.2025.104838

Paper 3

Mehic, M., & Thommes, K. (2026). *Foreign language use, attribution error, and newcomer integration* (Working Paper No. 177). Paderborn University, Faculty of Business Administration and Economics. https://ideas.repec.org/p/pdn/dispap/177.html

Paper 4

Gutt, J. K., Thommes, K., & Mehic, M. (2025). *Can verbal performance appraisals and machine learning models improve the accuracy of performance evaluations?* (Working Paper No. 132). Paderborn University, Faculty of Business Administration and Economics. https://ideas.repec.org/p/pdn/dispap/132.html

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 | SYNOPSIS

1.1 INTRODUCTION

"We can be blind to the obvious, and we are also blind to our blindness." – Daniel Kahneman, Nobel Prize winner, 2011

Organizations are built on decisions. Employees decide which actions to take, which processes to follow, and how to allocate resources. Managers evaluate performance, assess newcomers, and determine who receives opportunities and support. These behavioral and evaluative decisions influence major organizational outcomes, such as efficiency, motivation, and long-term performance. What ties these two processes together is human decision-making.

Actual decisions often differ from the assumptions of classical economic decision models. Employees may not choose the most efficient option, managers may assess signals wrongly, and organizations may persist in ineffective routines. These deviations arise because decision-makers face cognitive constraints and rely on simplified judgment strategies. Research on bounded rationality shows that individuals use heuristics to navigate complex environments (Gigerenzer, 2020; Gigerenzer & Gaissmaier, 2011; Kahneman, 2003; Simon, 1955; Tversky & Kahneman, 1974). While these heuristics enable efficient action, they can also produce systematic biases that affect for instance the processing of information behavioral choices.

One response to these limitations is to design decision environments more deliberately. Research on choice architecture shows that arranging options in particular ways can influence behavior without limiting choice (Münscher et al., 2016; Thaler & Sunstein, 2008). These interventions, often referred to as "nudges", have been applied across various domains, including health, finance or sustainability. While numerous studies have documented their cost-effective short-term impact (Hummel & Maedche, 2019; Mertens et al., 2022), their effectiveness varies across contexts and may diminish over time (Szaszi et al., 2025).

At the same time, organizations rely on correct assessment of signals. With respect to human resources decisions for instance, performance appraisals, hiring decisions, and newcomer assessments influence resource allocation and the organization's success. However, the

effectiveness of current assessment systems is still debated (Adler et al., 2016; DeNisi & Murphy, 2017; Murphy, 2020). Critics argue that these systems often fail to accurately capture true performance and are prone to systematic bias (Angelovski et al., 2016; Brown et al., 2017; Bizzi, 2018). According to attribution research, observers tend to overemphasize dispositional explanations and underestimate situational influences, a phenomenon known as the fundamental attribution error (Ross, 1977). Additionally, appraisal formats can influence judgments and introduce distortions.

Despite their shared foundation in bounded rationality, research on behavioral interventions and research on evaluative decision-making have evolved largely in isolation from one another. Consequently, there is a lack of integrated understanding of how organizational decision environments jointly shape behavioral and evaluative outcomes and how long these effects persist. This dissertation addresses this gap by examining how organizational decision environments influence behavioral choices and performance evaluations, with particular attention to temporal dynamics. First, the dissertation analyzes how contextual features of decision environments affect behavior and whether these effects persist. Second, this work investigates how evaluation environments shape judgments of performance and how contextual cues and appraisal formats introduce bias.

To address these objectives, the dissertation draws on four empirical papers employing meta-analytic, field, and experimental methods. The studies examine (1) the long-term effectiveness of behavioral interventions, (2) the persistence of default effects among professional decision-makers, (3) the influence of contextual cues on newcomer evaluations, and (4) the role of appraisal formats in performance accuracy.

Based on these considerations, the dissertation addresses the following overarching research question: *How do organizational decision environments influence choices and the assessment of signals?*

1.2 CONCEPTUAL BACKGROUND

This dissertation draws on the concepts of bounded rationality, nudging and cognitive biases to explain how contextual structures and cognitive constraints influence decision-making processes within organizations. The next section provides an overview of these concepts.

1.2.1 Bounded rationality

The concept of bounded rationality, introduced by Simon (1955, 1972), challenges the classical economic assumption of fully rational decision-making. Traditional models based on the notion of "homo economicus" assume that individuals have complete information, unlimited cognitive capacity, and enough time to evaluate all alternatives to maximize utility. However, Simon argued that these assumptions are unrealistic. In real-world contexts, decision-making is constrained by cognitive, informational, and temporal limitations. Under these conditions, individuals do not optimize but rather engage in satisficing by selecting options that are "good enough" given these constraints. This perspective provides a central foundation for subsequent developments in behavioral economics.

Building on this foundation, research in behavioral economics has examined the cognitive mechanisms underlying bounded rationality. The heuristics-and-biases program, developed by Tversky and Kahneman (1972, 1973, 1974), demonstrates that individuals rely on heuristics, or mental shortcuts that simplify complex decisions. While these heuristics enable efficient judgments under uncertainty, they can also lead to systematic deviations from the normative standard of rationality. The availability heuristic (Tversky & Kahneman, 1973), for instance, causes individuals to assess the likelihood of events based on how easily instances come to mind, often resulting in the overestimation of recent events. Similarly, anchoring effects reflect a reliance on initial reference points, and framing effects demonstrate how the presentation of information influences decisions (Tversky & Kahneman, 1974). Importantly, such deviations are systematic and predictable rather than random.

In contrast to viewing heuristics primarily as sources of bias, a complementary perspective is offered by Gigerenzer and colleagues. The fast-and-frugal heuristics framework conceptualizes heuristics as adaptive tools well-suited to specific environments (Gigerenzer & Gaissmaier, 2011; Gigerenzer & Goldstein, 1996; Gigerenzer, 2020). From this perspective, simple decision rules can lead to accurate and efficient outcomes. Gigerenzer and Brighton (2009), for example, demonstrate "less-is-more" effects, showing that strategies ignoring part of the available information can outperform more complex models by reducing variance and avoiding overfitting.

The cognitive underpinnings of bounded rationality are further elaborated upon through dual-process theory of decision-making, which was most prominently articulated by Kahneman (2003, 2011). The framework distinguishes between two modes of thinking: System 1 is fast, automatic, and intuitive. System 2 is slow, deliberative, and cognitively demanding. Due to limited cognitive resources, individuals often rely on System 1, especially in complex or time-constrained situations. While this enables efficient decision-making, it also increases susceptibility to biases. Although System 2 is more accurate, it is often underutilized due to its higher cognitive cost. This imbalance between the two systems reinforces the notion of bounded rationality and explains why individuals often fail to engage in fully rational deliberation. Thus, decision-making is constrained not only by external factors, such as information availability or limited time, but also by internal cognitive architectures that shape how information is processed.

Bounded rationality also provides a theoretical foundation for behavioral policy interventions, particularly nudging. According to Thaler & Sunstein (2008), nudges are subtle changes in choice architecture that influence behavior without restricting freedom of choice. These interventions are effective because they leverage the heuristics and biases inherent in human decision-making processes. For example, default options exploit status quo bias (for an

overview: e.g., Lemken, 2024), framing influences how information is perceived (for an overview: e.g., Cornelissen & Werner, 2014), and social norms guide behavior by signaling what is considered typical or desirable (for an overview: e.g., Gross & Vostroknutov, 2022). Rather than assuming fully rational decision-making, nudging acknowledges cognitive limitations and seeks to design environments that facilitate better choices.

However, more recent research has critically examined the relationship between bounded rationality and nudges. For instance, Lodge and Wegrich (2016) highlight the "rationality paradox" of nudging: interventions designed to correct individuals' cognitive limitations implicitly assume that policymakers are less affected by such constraints. If bounded rationality also applies to organizations, this assumption may undermine the effectiveness and predictability of such interventions. Related debates further emphasize that the impact of nudging depends on how rationality is conceptualized. Engelen (2019) distinguishes between outcome- and process-oriented rationality, arguing that nudges can improve decision outcomes while potentially bypassing deliberative processes. At the same time, nudging does not necessarily undermine rationality, particularly when decisions would otherwise rely on similar heuristics. Building on this idea, Schmidt (2019) proposes an ecological perspective according to which rationality depends on the fit between cognitive processes and the environment. According to this view, bounded rationality is not only a limitation but also implies that well-designed environments can improve the quality of decisions.

Throughout this dissertation and its four studies, decision-makers are conceptualized as boundedly rational.

1.2.2 Nudging

Nudging refers to the deliberate design of decision environments and the presentation of options in ways that influence behavior without restricting freedom of choice or significantly altering economic incentives. Introduced by Thaler and Sunstein (2008), the concept builds on

the idea of libertarian paternalism. This approach involves providing subtle, non-intrusive guidance to steer individuals toward choices that enhance their well-being while preserving their autonomy (Sunstein & Thaler, 2003). It is grounded in the assumption that individuals often make suboptimal decisions due to bounded rationality, limited self-control, and unstable or context-dependent preferences.

Nudging emphasizes that small modifications to choice architecture, such as changes to defaults, framing, or the order and number of options, can systematically influence decisions. A substantial body of empirical research demonstrates the effectiveness of such interventions (Hummel & Maedche, 2019; Mertens et al., 2022). For instance, preselected defaults in pension plans (Madrian & Shea, 2001), organ donation programs (Johnson & Goldstein, 2003; Abadie & Gay, 2006), and green energy consumption programs (Pichert & Katsikopoulos, 2008; Ebeling & Lotz, 2015; Kaiser et al., 2020) have been shown to substantially raise participation rates, even though individuals can still opt out.

From the perspective of bounded rationality, choice architecture acknowledges that people rarely evaluate all available alternatives thoroughly. Instead, people rely on heuristics and contextual cues such as representativeness, availability, anchoring, and framing effects (Tversky & Kahneman, 1974; Kahneman, 2003, 2011). Deliberately structuring decision environments through nudges can reduce cognitive load and guide behavior toward outcomes that are beneficial from an individual or societal perspective. Consequently, choice architecture has emerged as a powerful framework for influencing decision-making in domains such as health, finance, and sustainability, where minor design changes can have significant effects.

Over the past decade, nudges have experienced rapid global diffusion. Since the creation of the first behavioral insights unit in 2010, over 300 similar teams have emerged in more than 60 countries (Hubble & Varazzani, 2023). However, parallel to this development, nudges have

become the subject of extensive academic debates (Banerjee & John, 2025; Congiu & Moscati, 2022; De Ridder et al., 2024; Lades & Delaney, 2022; Schmidt & Engelen, 2020).

On the one hand, nudges are frequently praised for their cost-effectiveness and ease of implementation. Compared to traditional policy instruments, such as regulations or financial incentives, nudges can be integrated into existing decision environments at relatively low cost and can be scaled efficiently (Benartzi et al., 2017). Moreover, nudging is closely linked to an evidence-based approach to policymaking. Interventions can be evaluated through randomized controlled trials (DellaVigna & Linos, 2022), enabling policymakers to iteratively refine and optimize their effectiveness. In addition, proponents argue that decision environments inevitably shape behavior, regardless of whether nudges are intentionally implemented or not (Sunstein, 2015).

On the other hand, nudging has attracted substantial criticism. A central concern relates to its ethical implications, particularly the potential to undermine individual autonomy. Critics argue that nudges may influence decisions in subtle and non-transparent ways, raising concerns about manipulation and the absence of informed consent (Hausman & Welch, 2010; Hansen & Jespersen, 2013; Wilkinson, 2013). Furthermore, nudging has been criticized for exploiting cognitive biases rather than engaging individuals' deliberative capacities (Bovens, 2009; Grüne-Yanoff, 2012).

Beyond ethical considerations, scholars have also questioned the effectiveness and durability of nudges (Beshears & Kosowsky, 2020). While numerous studies report positive behavioral effects, these effects are often modest, context-dependent, and may diminish over time. More recent evidence suggests that effect sizes may be overstated once publication bias is taken into account (Maier et al., 2022; Szaszi et al., 2022, Szaszi et al., 2025). Additionally, nudges may be less effective in high-stakes contexts. For instance, during the COVID-19 pandemic, nudging interventions often failed to produce meaningful behavioral change (Hume et al.,

2021). Finally, critics argue that nudging primarily targets individual behavior while neglecting broader structural determinants, thereby limiting its capacity to address systemic challenges (Chater & Loewenstein, 2023).

This dissertation examines nudges in greater detail in Studies 1 and 2, focusing on their effectiveness in Chapter 2 and their potential application in professional decision-making contexts in Chapter 3.

1.2.3 Cognitive Biases

The concept of cognitive biases stems from the heuristics-and-biases program introduced by Tversky and Kahneman (1974). Their seminal work challenged the assumption of fully rational decision-making by showing that individuals consistently differ from standard models of judgment under uncertainty. Rather than relying on complex calculations, individuals use heuristics, or simple mental shortcuts, that facilitate efficient decision-making but can also result in systematic errors.

The heuristics-and-biases program has been highly influential in behavioral economics, leading to the identification of a large number of cognitive biases. These biases are often summarized in visual overviews, such as the Cognitive Bias Codex[1] (for an overview: Benson, 2016). Depending on the underlying cognitive challenges, cognitive biases can be categorized in different ways. In a more recent review, Azzopardi (2021) distinguished four overarching categories of cognitive biases: those arising from information overload, lack of meaning, time pressure, and memory limitations.

- *Information overload:* When confronted with excessive information, individuals tend to rely on heuristics that simplify, such as availability, anchoring, and framing, which can distort judgment.

[1] The Cognitive Bias Codex by Benson (2016) illustrates over 180 types of cognitive biases. However, not all of them have been empirically validated in peer-reviewed articles.

- *Lack of meaning:* When information lacks clarity or meaning, people often use generalizations, pattern recognition, or social cues. This can lead to biases such as the bandwagon or reinforcement effects.
- *Time pressure:* When under time pressure, the need for efficiency shapes decision-making, leading to biases such as ambiguity aversion or decoy effects.
- *Memory limitations:* Limitations in memory lead individuals to selectively retain and recall information, resulting in biases such as priming or order effects.

However, the concept of cognitive biases has also been subject to criticism. Gigerenzer (2018), for example, introduced the notion of the "bias bias", which describes the tendency to interpret behavior as systematically biased even when it may in fact be rational. According to this perspective, many alleged biases arise from the use of inappropriate normative benchmarks, especially when decisions under uncertainty are evaluated against models of rational choice under risk. Gigerenzer (2018) identifies three common sources of such misinterpretations: failing to distinguish between small-sample and population statistics; mistaking individual's random error for systematic error; and treating logically equivalent frames as informationally equivalent.

Of the many cognitive biases, the *fundamental attribution error* (FAE) is particularly relevant to this dissertation. Introduced by Ross (1977), the FAE describes the tendency to overemphasize dispositional factors, such as personality traits or intentions, and underestimate situational influences when explaining the behavior of others. This bias is especially relevant in organizational and evaluative contexts, where misattributions can result in flawed judgments, such as biased performance evaluations or incorrect assessments of responsibility. In this dissertation, the FAE plays a central role in Study 3 (Chapter 4), in which attribution processes are examined in greater detail.

1.3 OVERVIEW OF STUDIES AND RESEARCH OBJECTIVES

This dissertation comprises four empirical studies examining behavioral and evaluative decision-making in organizational contexts from a complementary perspective. Together, the studies investigate how cognitive biases and contextual decision environments influence behavior and evaluations, as well as how these effects evolve over time. Table 1.1 provides an overview of the individual studies, including their respective research questions, contributions and methods.

The first study addresses the temporal dimension of behavioral interventions by synthesizing existing evidence on the long-term effectiveness of nudging. Although nudging research has demonstrated substantial short-term effects across domains, evidence on persistence remains fragmented. Therefore, Study 1 conducts a meta-analysis of pro-environmental nudging interventions to examine how and whether behavioral effects change over time. By analyzing longitudinal effect sizes across multiple studies, the meta-analysis provides systematic evidence on the persistence and decay of nudging effects and identifies factors that moderate long-term effectiveness. This study lays the temporal groundwork for the dissertation by clarifying whether behavioral interventions can produce persistent change.

Building on this foundation, the second study examines persistence in a real-world organizational context. Specifically, it investigates how default settings impact carbon offsetting decisions in cargo transportation bookings. Although defaults have been shown to be powerful nudging tools, little is known about their long-term effects in professional decision environments. Using large-scale field data and a difference-in-differences design, Study 2 examines whether an opt-out default increases pro-environmental behavior and if this effect persists over time. The findings provide field evidence on the stability of defaults in a business-to-business context and expand the scope of nudging research to include repeated professional decisions. While the first two studies focus on behavioral choices, the third and fourth studies examine evaluative decision-making. Study 3 investigates how contextual factors influence the

Table 1.1: Summary of studies included in this dissertation

Title	Research question	Contributions	Literature	Method	Sample
Study 1: How Effective Is Nudging in the Long Run? A Meta-Analysis of Pro-Environmental Behavior	Do environmental nudges produce persistent behavioral effects? How does effectiveness change over time?	First meta-analysis on long-term effects of environmental nudges; quantifies behavioral decay; identifies heterogeneity across nudge types and temporal structures	Nudging and behavioral public policy; pro-environmental behavior; long-term intervention effect	Systematic literature review; three-level random-effects meta-analysis; meta-regression and moderator analyses	42 peer-reviewed studies; 140 effect sizes; N = 613,894
Study 2: The Persistence of Default Effects: Evidence from CO_2 Offsetting in Cargo Transportation	Do opt-out defaults increase pro-environmental behavior in professional decision-making? Do these effects persist over time?	Provides field evidence on persistent default effects in a B2B context; shows partial decay and long-run stabilization; demonstrates no negative revenue effects	Default effects; carbon offsetting; persistence of behavioral interventions; B2B decision-making	Large-scale field study; difference-in-differences design; probit regressions; longitudinal analysis over 65 weeks	>200,000 bookings; 47,533 customers; 3,128 customers with repeated decisions
Study 3: Foreign Language Use, Attribution Error, and Newcomer Integration	Does foreign language use reduce attribution errors and improve fairness in performance evaluations of newcomers?	Extends attribution error research to multilingual contexts; shows reduced evaluation gaps driven by stricter evaluations of incumbents; contributes to newcomer integration literature	Fundamental attribution error; newcomer integration; organizational socialization; language in organizations	Laboratory experiment; randomized language condition; behavioral and evaluative measures; linguistic analysis (LIWC)	80 students; 26 experimental sessions; German native speakers using German or English
Study 4: Can Verbal Performance Appraisals and Machine Learning Models Improve the Accuracy of Performance Evaluations?	How accurately do different performance appraisal formats (numerical ratings vs. written and spoken comments) reflect actual performance? Which format captures performance differences most reliably within and between teams?	Compares accuracy of numerical ratings and verbal appraisal formats; shows that spoken comments track performance differences most closely within teams; introduces machine-learning approach to translate comments into comparable ratings; contributes to research on improving evaluation accuracy in organizations	Performance appraisal accuracy; rating bias; evaluation formats; subjective vs. objective performance measurement	Laboratory experiment; collection of numerical ratings, written and spoken comments; objective performance measures; between- and within-team analysis	Student participants in lab experiment; multiple teams performing collaborative task (N = 160)

performance evaluations of newcomers. Drawing on attribution theory, Study 3 examines whether using a foreign language affects attribution processes and fairness in performance evaluations. In a laboratory experiment, participants evaluate newcomers under different language conditions. The results demonstrate that contextual features, such as language use, can influence attribution processes and evaluation outcomes. This highlights how situational factors shape evaluative judgments in organizational settings.

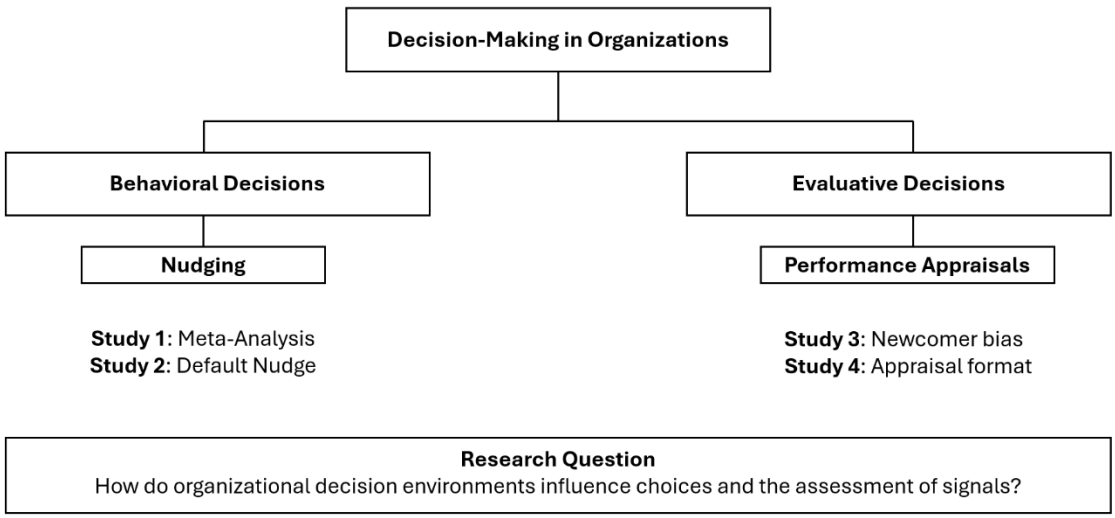
Study 4 examines how performance appraisal formats affect the accuracy of evaluations. Although performance appraisal systems are central to organizational decision-making, they are often criticized for bias and limited accuracy. This study compares numerical ratings with written and spoken evaluative comments, linking them to objective performance measures. The study uses experimental data and machine-learning methods to translate qualitative comments into comparable metrics and assess which evaluation format most accurately reflects individual performance. The findings provide insights into how evaluation systems can be designed to improve decision quality.

Together, the four studies address different aspects of behavioral decision-making in organizations. Studies 1 and 2 focus on behavioral interventions and their persistence over time. Studies 3 and 4 examine evaluative decision processes and the influence of contextual factors on performance appraisals. Integrating these perspectives, the dissertation provides a comprehensive account of how decision environments shape behavior and evaluations, and how these effects unfold over time.

Figure 1.1 illustrates the dissertation's overarching conceptual framework. The dissertation primarily examines decision-making in organizations, distinguishing between two central domains: behavioral and evaluative decisions. Behavioral decisions are action-oriented choices, such as adopting sustainable options, and are examined through the lens of choice architecture and nudging interventions (Studies 1 and 2). Evaluative decisions are judgments

about individuals, such as performance appraisals and newcomer evaluations. They are analyzed through the lens of attribution processes and evaluation formats (Studies 3 and 4).

Figure 1.1: Conceptual framework



The figure shows that both types of decisions are influenced by cognitive biases and the context of the decision environment. While behavioral interventions, such as defaults, influence choices directly, evaluation systems and communication contexts shape how performance is interpreted and assessed. Importantly, the dissertation emphasizes the temporal dimension of these effects and investigates whether and how they persist over time. Together, the four studies provide an integrated perspective on how choice architecture and cognitive biases influence behavioral and evaluative decisions in organizational settings.

The studies included in this dissertation are at various stages in the publication process (see table 1.2). Study 1, a single-author meta-analysis, is currently being prepared for submission and has been accepted for presentation at the 2026 World Meetings of the Economic Science Association (ESA). Study 2 has been published in co-authorship in Transportation Research Part A: Policy and Practice and was previously presented at the 2022 Economic Science Association Conference and the faculty research workshop, where it received the Best Paper Award. Study 3 is a co-authored experimental paper currently being prepared for journal

Table 1.2: Summary of publication stages

Chapter	Study title	Status	Conferences	Share of contribution (in %)
Chapter 2	**Study 1:** How Effective Is Nudging in the Long Run? A Meta-Analysis of Pro-Environmental Behavior	In preparation for submission to Journal of Behavioral and Experimental Economics	Economic Science Association Conference (ESA), Los Angeles, USA, 2026	Miro Mehic (100)
Chapter 3	**Study 2:** The Persistence of Default Effects: Evidence from CO_2 Offsetting in Cargo Transportation	Published in Transportation Research Part A: Policy and Practice	Economic Science Association Conference (ESA), Boston, USA, 2022; Faculty Research Workshop, Melle, Germany, 2022 (Best Paper Award Winner)	Kirsten Thommes (50) Miro Mehic (50)
Chapter 4	**Study 3:** Foreign Language Use, Attribution Error, and Newcomer Integration	In preparation for submission to Organizational Behavior and Human Decision Processes	Interdisciplinary Network for Group Research Conference (INGRoup), Seattle, USA, 2023	Miro Mehic (80) Kirsten Thommes (20)
Chapter 5	**Study 4:** Can Verbal Performance Appraisals and Machine Learning Models Improve the Accuracy of Performance Evaluations?	In preparation for submission to Human Relations	Colloquium on Personnel Economics (COPE), Amsterdam, Netherlands, 2023; Academy of Management Conference (AOM), Boston, USA, 2023 (awarded Top 10% Best HR Division Conference Papers)	Jana Kim Gutt (90) Miro Mehic (5) Kirsten Thommes (5)

submission. It was presented at the INGRoup Conference in 2023. Study 4 is a co-authored experimental study on performance appraisal formats currently being prepared for submission. The study received the Top 10% Best HR Division Conference Paper award at the 2023 Academy of Management Annual Meeting.

The author of this dissertation contributed to all the included studies. Study 1 was conducted independently. Studies 2 and 3 were developed in close collaboration with the co-author. Study 4 was primarily led by a co-author. The author contributed mainly to the study design.

1.4 RESEARCH METHODS

To address the overarching research question of this dissertation, a multi-methods research design was employed. This dissertation integrates meta-analytic, field-experimental, and laboratory-experimental approaches to examine how decision environments influence behavioral and evaluative decisions within organizational contexts, as well as the persistence of these effects over time. This combination of complementary methods enables a comprehensive analysis of decision-making processes across various levels of abstraction, contexts, and time horizons.

First, the dissertation employs a systematic meta-analytic approach to synthesize the existing evidence regarding the long-term effectiveness of behavioral interventions. Study 1 uses a systematic literature review and a three-level random-effects meta-analysis to examine how pro-environmental nudging effects persist over time. This meta-analysis allows for the aggregation of effect sizes across heterogeneous studies and the examination of temporal dynamics and moderators of intervention effectiveness, making it particularly suited to this research objective. Integrating findings from multiple longitudinal studies provides a broad, generalizable assessment of whether and how nudging effects decay or persist.

Second, the dissertation includes a large-scale field study examining behavioral decision-making in a real-world organizational setting. Study 2 uses longitudinal booking data from a

cargo transportation platform to analyze the impact of default settings on carbon offsetting decisions. This study uses a quasi-experimental difference-in-differences design combined with regression analyses to identify causal effects and examine their persistence over time. Field data allows for the observation of repeated professional decisions in natural settings, thereby improving the external validity of study results. This approach complements meta-analytic evidence by offering detailed insights into behavioral persistence in an applied business-to-business context.

Third, the dissertation uses two laboratory experiments to study evaluative decision-making processes in a controlled environment. Study 3 uses an experimental design to examine how contextual factors, particularly the use of a foreign language, influence the attribution of performance evaluations to newcomers. Participants were randomly assigned to different language conditions (native vs. foreign language) to identify effects on evaluative judgments. Laboratory experiments are useful for studying cognitive biases because they isolate specific mechanisms and control for confounding factors.

Study 4 also employs an experimental design to examine the effectiveness of various performance appraisal formats. Participants evaluated individual performance within teams using numerical ratings, written comments, and spoken feedback. These evaluations were linked to objective performance measures, which enabled an assessment of how accurately the different formats capture actual performance differences. Additionally, computational text analysis and machine learning methods translated qualitative comments into comparable metrics. This combination of experimental and computational methods enables a thorough analysis of the accuracy and bias of performance appraisal systems.

The dissertation combines complementary methods from the four studies. The meta-analysis offers a comprehensive synthesis of existing evidence on behavioral persistence. The field study offers high external validity and insights into real-world professional decision-making

processes. Laboratory experiments allow for causal inference and identification of the cognitive mechanisms underlying evaluative decisions. Together, these methods provide a comprehensive empirical basis for understanding how decision environments influence behavior and evaluations within organizations, as well as how these effects evolve over time.

1.5 RESULTS OVERVIEW

This section summarizes the central findings of the four studies included in this dissertation. Each study addresses a specific research objective related to behavioral or evaluative decision-making in organizational contexts.

Study 1 addresses whether environmental nudges produce long-lasting behavioral changes and how their effectiveness changes over time. Although prior research has demonstrated the substantial short-term effects of nudging interventions, the long-term dynamics of these effects are mostly unclear. This study examines the temporal dynamics of pro-environmental nudging effects using a systematic literature review and a three-level random-effects meta-analysis of 140 effect sizes from 42 peer-reviewed studies (N = 613,894). The results reveal that nudging interventions produce statistically significant behavioral changes that endure over time. However, effect sizes gradually decline, indicating partial behavioral decay. The persistence of these effects varies depending on the type of nudge and the temporal structure of the intervention. Defaults and structural interventions tend to produce more stable effects than informational nudges. Overall, the study demonstrates that nudges can produce long-lasting behavioral changes, but persistence depends on the characteristics of the intervention. These findings provide a quantitative basis for evaluating the long-term effectiveness of behavioral interventions.

Study 2 examines whether default settings influence professional decision-making and if these effects endure over time. Although defaults are among the most effective nudging tools, their long-term impact in business-to-business contexts is not well understood. Using a difference-

in-differences design and longitudinal booking data from over 200,000 cargo transportation transactions, the study examines the introduction of an opt-out carbon offsetting default. The analysis tracks decisions made by professional customers over a 65-week period. The findings reveal that the opt-out default significantly increases carbon offsetting rates compared to an opt-in condition. Importantly, the default effect persists over time, though a moderate decline is observed initially following implementation. After this period, offsetting rates stabilize at a higher level than under the opt-in condition. Furthermore, the intervention does not negatively impact revenues, suggesting that pro-environmental defaults can be implemented without economic disadvantages. This study provides field evidence showing that default effects remain robust even among experienced professional decision-makers and in repeated decision-making environments. Therefore, the study extends nudging research to real-world organizational contexts and complements the meta-analytic findings of Study 1.

Study 3 examines whether contextual cues influence evaluative judgments by focusing on the fundamental attribution error in evaluations of newcomers. Specifically, it investigates whether using a foreign language reduces attribution biases and improves the fairness of performance assessments. In a laboratory experiment, participants evaluated the performance of newcomers and incumbents in different language settings (native versus foreign language). Behavioral and evaluative measures were collected, and linguistic analyses were conducted. The results indicate that using a foreign language reduces the evaluation gap between newcomers and incumbents. This effect is not driven by more favorable evaluations of newcomers, but rather by stricter evaluations of incumbents. These results suggest that language context influences attribution processes and reduces dispositional bias in performance evaluations. This study demonstrates that subtle contextual factors, such as language use, can influence evaluative judgments. The study contributes to attribution theory by extending the fundamental attribution

error to multilingual organizational settings, highlighting how evaluation environments influence fairness.

Study 4 addresses how accurately different performance appraisal formats reflect actual performance. Although numerical ratings are widely used in organizations, alternative formats, such as written and spoken comments, may capture performance differences in different ways. In a laboratory experiment involving team-based tasks, participants evaluated individual performance using numerical ratings, written comments, and spoken feedback. These evaluations were linked to objective performance measures. Machine-learning methods translated the qualitative comments into comparable quantitative metrics. The findings show that appraisal format significantly influences evaluation accuracy. Specifically, spoken comments track performance differences within teams more closely than numerical ratings do. Numerical ratings tend to compress performance differences, while qualitative formats provide richer differentiation of performance. However, they also require additional interpretation and processing effort. This study demonstrates that performance appraisal systems function as decision environments that shape evaluative judgments. The study provides evidence that alternative evaluation formats can improve the accuracy of performance assessments and reduce systematic distortions.

In total, the four studies provide converging evidence that organizational decision environments influence behavioral and evaluative decisions. Studies 1 and 2 demonstrate that behavioral interventions can have long-lasting effects, especially when embedded in structured decision environments. Studies 3 and 4 demonstrate that evaluative judgments are also influenced by contextual cues and evaluation formats. Contextual design plays a central role across all studies. Defaults, language contexts, and appraisal formats influence how decisions are made and interpreted. Furthermore, the findings underscore the importance of examining not only the immediate effects, but also the persistence and stability of decision outcomes over

time. The results collectively support the dissertation's central argument that organizational decision environments systematically influence behavioral and evaluative decisions, and these effects can persist, although often in an attenuated form, over time.

1.6 DISCUSSION

This dissertation examined how decision environments influence behavioral and evaluative decision-making in organizational contexts. Based on the concepts of bounded rationality, nudging, and cognitive biases, the findings from four empirical studies show that decision-making is influenced by cognitive constraints and contextual structures.

Across all four studies, organizational decision-making is not solely determined by stable preferences or objective performance but shaped by the design of the surrounding environment. Behavioral choices, such as adopting sustainable options, and evaluative judgments, such as performance appraisals, are influenced by contextual features that guide attention, interpretation, and action. Furthermore, the dissertation emphasizes that these effects persist over time, although they may attenuate or stabilize at different levels.

1.6.1 Discussion of Empirical Findings

Studies 1 and 2 contribute to the debate on the longevity of nudging effects by explicitly examining their temporal dynamics and contextual stability. Although the short-term effectiveness of nudging is well-established, concerns remain about attenuation (Allcott & Rogers, 2014; Beshears & Kosowsky, 2020).

Study 1 is the first meta-analytic synthesis of the long-term effects of pro-environmental nudges. Nudges produce small and significant effects that persist beyond the intervention period. However, meta-regression analyses reveal gradual attenuation over time, indicating behavioral decay rather than abrupt disappearance. These findings refine earlier conclusions that pro-environmental nudges have limited durability (Nisa et al., 2019) and qualify claims that specific nudge types are superior based primarily on short-term evidence (Mertens et al.,

2022). The findings also align with field evidence showing smaller real-world effects compared to those observed in academic trials (DellaVigna & Linos, 2022) and with recent concerns about the robustness of aggregate results (Hu et al., 2025). Therefore, persistence of nudging interventions emerges as a context-dependent outcome rather than a simple binary one.

Study 2 complements this long-term evidence with longitudinal field data from a professional, business-to-business context. In line with previous studies (Araña & León, 2013; Ebeling & Lotz, 2015; He et al., 2023; Kaiser et al., 2020), switching from an opt-in to an opt-out system significantly increased participation. Contrary to the claim that defaults weaken with expertise (Löfgren et al., 2012), this effect persisted for at least eleven months and remained substantial, even among frequent, price-sensitive users. These findings contribute to the body of research on repeated decision contexts (Egebark & Ekström, 2016; Kesternich et al., 2019; Liebe et al., 2021) and are consistent with research on the long-lasting effects of default nudges (Liebe et al., 2021). Consistent with the concept of bounded rationality (Simon, 1955), even experienced, price-sensitive decision-makers rely on choice architectures to simplify complex decisions.

Studies 3 and 4 examine how evaluation environments influence performance evaluations and demonstrate how contextual features can bias evaluative outcomes.

Study 3 shows that working in a second language does not decrease team productivity. The study further replicates the coordination results of Weber and Camerer (2003) and contradicts previous evidence of communication breakdowns in multilingual settings (Neeley, 2013; Tenzer et al., 2014). However, language context significantly alters evaluations. The fundamental attribution error emerges in the first-language condition (Weber & Camerer, 2003) but disappears in the second-language condition due to the compression of ratings downward. Building on research about cognitive load in foreign language use (Piekkari et al., 2005; Harzing & Feely, 2008; Wilmot et al., 2024), these results imply a change in evaluative reference points rather than enhanced precision. Although attribution bias is reduced,

compressed evaluations may lead to dissatisfaction and the perception of unfairness (Santos et al., 2019; Sharma & Kulshreshtha, 2023).

Study 4 shows that the appraisal format further shapes evaluative accuracy. While numerical ratings do not predict objective performance differences between teams, algorithmic ratings derived from spoken comments can capture such differences within teams where evaluators have a reference point. These findings are consistent with research on comparative frames in performance evaluations (Blume et al., 2009; Roch et al., 2007) and positivity bias in isolated rating contexts (Schoenmueller et al., 2020).

1.6.2 Discussion of Theoretical Contributions

This dissertation's findings contribute to and expand upon the theoretical perspectives outlined in Section 1.2 by incorporating insights from research on bounded rationality, nudging, and cognitive biases.

First, the results reinforce and expand upon the concept of bounded rationality, which was introduced by Simon (1955, 1972). Across all studies, decision-makers rely on simplified strategies and respond to environmental cues rather than engaging in fully rational optimization. Importantly, the findings demonstrate that bounded rationality applies not only to individual choice behavior, but also to evaluative judgments within organizations. Performance evaluations are influenced by contextual factors, such as language use and appraisal format. This indicates that cognitive constraints shape both judgments about others and decisions about actions.

Second, the dissertation contributes to the literature on nudging, as developed by Thaler and Sunstein (2008), by emphasizing the temporal dimension of behavioral interventions. While previous studies have primarily examined short-term effects, this research reveals that nudging interventions exhibit dynamic trajectories over time. Study 1 shows that behavioral effects persist but gradually decrease, and Study 2 demonstrates that default effects can remain stable

in real-world organizational settings. Together, these studies suggest that the effectiveness of nudging is a temporal and context-dependent process rather than a static outcome.

Additionally, the results address ongoing debates about the limits of nudging, particularly the "rationality paradox" identified by Lodge and Wegrich (2016). The persistence and context-dependence of intervention effects imply that decision environments are embedded within organizational structures subject to bounded rationality. Consequently, the effectiveness of interventions cannot be fully separated from the broader context in which they are implemented.

Third, the dissertation contributes to the literature on cognitive biases by expanding its scope to include evaluative decision-making in organizations. Based on the heuristics-and-biases framework developed by Tversky and Kahneman (1974), the findings demonstrate that biases, such as the fundamental attribution error (Ross, 1977), influence performance evaluations and are sensitive to contextual conditions. Study 3 demonstrates that language context can alter attribution processes, and Study 4 shows that evaluation formats influence the accuracy of evaluations. Together, these studies suggest that biases are not only cognitive phenomena, but also structurally embedded in organizational decision environments.

The results also align with the ecological perspective on rationality proposed by Gigerenzer (2020) and Schmidt (2019). Rather than viewing all deviations from normative standards as irrational, the findings imply that the quality of decisions depends on the alignment between cognitive processes and the environment. Well-designed decision environments, such as appropriate appraisal formats, can improve decision quality, whereas poorly aligned environments may distort outcomes.

1.6.3 Practical Implications

The findings of this dissertation have important implications for organizations and policymakers. First, the results suggest that behavioral interventions, such as defaults, can

effectively influence decision-making, even in professional and repeated contexts. Organizations seeking to promote sustainable behavior can use choice architecture to guide decisions without restricting autonomy. However, the observed attenuation of effects over time indicates that interventions should be monitored and adjusted as needed to maintain their effectiveness.

Second, the findings emphasize that evaluation systems should not be viewed as neutral measurement tools. The language used and the format of the appraisal significantly influence how performance is evaluated. Organizations can improve the fairness and accuracy of performance evaluations by carefully designing communication and feedback processes. For example, they can complement numerical ratings with qualitative assessments or consider the role of language in multinational contexts.

1.6.4 Limitations and Future Research

Despite its contributions, this dissertation has several limitations. First, the four studies address related but distinct aspects of decision-making, so they do not constitute a single, unified empirical design. While this multi-study approach broadens the dissertation's scope, it also limits direct comparability across studies.

Second, the studies focus on specific contexts and intervention types. Study 1 is limited to pro-environmental nudges. Studies 3 and 4 rely on laboratory experiments, which may not fully capture the complexity of real organizational environments. Future research could build on these findings by examining similar mechanisms in different domains and real-world contexts.

Third, while the dissertation emphasizes the role of decision environments, they are operationalized differently in different studies. This includes differences in nudges, language use, and appraisal formats. While this diversity highlights the concept's broad applicability, it also complicates the isolation of underlying mechanisms.

Future research could expand on these findings by investigating the temporal dynamics of behavioral interventions, examining evaluative decision-making in field settings, and exploring the interaction between different types of decision environments.

1.7 CONCLUSION

This dissertation examines how organizational decision environments influence behavioral and evaluative decisions and whether these effects persist over time. Across four studies, the findings show that decision environments systematically shape both what individuals do and how they are evaluated.

Studies 1 and 2 demonstrate that nudging interventions can produce persistent behavioral changes. However, persistence is not indefinite. The effects gradually attenuate and vary across contexts yet remain influential, even in professional and repeated decision-making settings. These findings indicate that bounded rationality continues to apply to experienced decision-makers and that choice architecture can influence behavior beyond the immediate decision moment.

Studies 3 and 4 reveal that evaluation outcomes are also dependent on context. Language environments and appraisal formats influence performance evaluations by shifting reference points and affecting the accuracy of differentiation. Thus, performance evaluations reflect not only individual performance, but also the structural characteristics of the evaluation system.

Overall, this dissertation emphasizes that organizational outcomes are co-determined by the architecture of decision environments. Improving organizational decision-making requires capable individuals and carefully designed contexts that account for bounded rationality and its dynamic effects over time.

REFERENCES

Abadie, A., & Gay, S. (2006). The impact of presumed consent legislation on cadaveric organ donation: a cross-country study. *Journal of Health Economics*, *25*(4), 599–620. https://doi.org/10.1016/j.jhealeco.2006.01.003

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting rid of performance ratings: Genius or folly? A debate. *Industrial and Organizational Psychology*, *9*(2), 219–252. https://doi.org/10.1017/iop.2015.106

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037. https://doi.org/10.3386/w18492

Angelovski, A., Brandts, J., & Sola, C. (2016). Hiring and escalation bias in subjective performance evaluations: A laboratory experiment. *Journal of Economic Behavior & Organization*, *121*, 114–129. https://doi.org/10.1016/j.jebo.2015.10.012

Araña, J. E., & León, C. J. (2013). Can defaults save the climate? Evidence from a field experiment on carbon offsetting programs. *Environmental and Resource Economics*, *54*(4), 613–626. https://doi.org/10.1007/s10640-012-9615-x

Azzopardi, L. (2021, March). Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (pp. 27–37). Association for Computing Machinery. https://doi.org/10.1145/3406522.3446023

Banerjee, S., & John, P. (2025). Behavioral public policy: past, present, & future. *Policy and Society*, Article puaf012. https://doi.org/10.1093/polsoc/puaf012

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., ... & Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, *28*(8), 1041–1055. https://doi.org/10.1177/0956797617702501

Benson, B. (2016, September 1). *Cognitive bias cheat sheet: An organized list of cognitive biases because thinking is hard*. Medium. https://buster.medium.com/cognitive-bias-cheat-sheet-55a472476b18

Beshears, J., & Kosowsky, H. (2020). Nudging: Progress to date and future directions. *Organizational Behavior and Human Decision Processes*, *161*, 3–19. https://doi.org/10.1016/j.obhdp.2020.09.001

Bizzi, L. (2018). The problem of employees' network centrality and supervisors' error in performance appraisal: A multilevel theory. *Human Resource Management*, *57*(2), 515–528. https://doi.org/10.1002/hrm.21880

Blume, B. D., Baldwin, T. T., & Rubin, R. S. (2009). Reactions to different types of forced distribution performance evaluation systems. *Journal of Business and Psychology*, *24*(1), 77–91. https://doi.org/10.1007/s10869-009-9093-5

Bovens, L. (2009). The ethics of nudge. In T. Grüne-Yanoff & S. Ove Hansson (Eds.), *Preference Change: Approaches from philosophy, economics and psychology* (pp. 207–219). Springer.

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods*, *20*(1), 121–148. https://doi.org/10.1177/1094428116668036

Chater, N., & Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, *46*, Article e147. https://doi.org/10.1017/S0140525X22002023

Congiu, L., & Moscati, I. (2022). A review of nudges: Definitions, justifications, effectiveness. *Journal of Economic Surveys*, *36*(1), 188–213. https://doi.org/10.1111/joes.12453

Cornelissen, J. P., & Werner, M. D. (2014). Putting framing in perspective: A review of framing and frame analysis across the management and organizational literature. *Academy of Management Annals*, *8*(1), 181–235. https://doi.org/10.5465/19416520.2014.875669

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ECTA18709

DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology*, *102*(3), 421–433. https://doi.org/10.1037/apl0000085

De Ridder, D., Feitsma, J., Van Den Hoven, M., Kroese, F., Schillemans, T., Verweij, M., ... & De Vet, E. (2024). Simple nudges that are not so easy. *Behavioural Public Policy*, *8*(1), 154–172. https://doi.org/10.1017/bpp.2020.36

Ebeling, F., & Lotz, S. (2015). Domestic uptake of green energy promoted by opt-out tariffs. *Nature Climate Change*, *5*(9), 868–871. https://doi.org/10.1038/nclimate2681

Egebark, J., & Ekström, M. (2016). Can indifference make the world greener? *Journal of Environmental Economics and Management*, *76*, 1–13. https://doi.org/10.1016/j.jeem.2015.11.004

Engelen, B. (2019). Nudging and rationality: What is there to worry? *Rationality and Society*, *31*(2), 204–232. https://doi.org/10.1177/1043463119846743

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3-4), 303–336. https://doi.org/10.1561/105.00000092

Gigerenzer, G. (2020). What is bounded rationality? In R. Viale (Ed.), *Routledge handbook of bounded rationality* (pp. 55–69). Routledge.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.
https://doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650.

Gross, J., & Vostroknutov, A. (2022). Why do people follow social norms? *Current Opinion in Psychology*, *44*, 1–6. https://doi.org/10.1016/j.copsyc.2021.08.016

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare*, *38*(4), 635–645. https://doi.org/10.1007/s00355-011-0636-0

Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, *4*(1), 3–28.

Hausman, D. M., & Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, *18*(1), 123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

He, G., Pan, Y., Park, A., Sawada, Y., & Tan, E. S. (2023). Reducing single-use cutlery with green nudges: Evidence from China's food-delivery industry. *Science*, *381*(6662), Article eadd9884. https://doi.org/10.1126/science.add9884

Hu, B., Xia, Z., Guo, Q., Lu, C., Constantino, S. M., & Ju, X. (2025). Assessing Nudge Impact: A Comprehensive Second-Order Meta-Analysis. *Journal of Behavioral Decision Making*, *38*(5), Article e70053. https://doi.org/10.1002/bdm.70053

Hubble, C., & Varazzani, C. (2023, May 10). *Mapping the global behavioural insights community*. OECD. https://oecd-opsi.org/blog/mapping-behavioural-insights/

Hume, S., John, P., Sanders, M., & Stockdale, E. (2021). Nudge in the time of coronavirus: The compliance to behavioural messages during crisis [Manuscript]. https://dx.doi.org/10.2139/ssrn.3644165

Hummel, D., & Maedche, A. (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, *80(1)*, 47–58. https://doi.org/10.1016/j.socec.2019.03.005

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339. https://doi.org/10.1126/science.1091721

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, *93*(5), 1449–1475.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kaiser, M., Bernauer, M., Sunstein, C. R., & Reisch, L. A. (2020). The power of green defaults: The impact of regional variation of opt-out tariffs on green energy demand in Germany. *Ecological Economics*, *174*, Article 106685. https://doi.org/10.1016/j.ecolecon.2020.106685

Kesternich, M., Römer, D., & Flues, F. (2019). The power of active choice: field experimental evidence on repeated contribution decisions to a carbon offsetting program. *European Economic Review*, *114*, 76–91. https://doi.org/10.1016/j.euroecorev.2019.02.001

Lades, L. K., & Delaney, L. (2022). Nudge forgood. *Behavioural Public Policy*, *6*(1), 75–94. https://doi.org/10.1017/bpp.2019.53

Lemken, D. (2024). Options to design more ethical and still successful default nudges: a review and recommendations. *Behavioural Public Policy*, *8*(2), 349–381.

Liebe, U., Gewinner, J., & Diekmann, A. (2021). Large and persistent effects of green energy defaults in the household and business sectors. *Nature Human Behaviour*, *5*(5), 576–585. https://doi.org/10.1038/s41562-021-01070-3

Löfgren, Å., Martinsson, P., Hennlock, M., & Sterner, T. (2012). Are experienced people affected by a pre-set default option—Results from a field experiment. *Journal of Environmental Economics and Management*, *63*(1), 66–72. https://doi.org/10.1016/j.jeem.2011.06.002

Lodge, M., & Wegrich, K. (2016). The rationality paradox of nudge: Rational tools of government in a world of bounded rationality. *Law & Policy*, *38*(3), 250–267. https://doi.org/10.1111/lapo.12056

Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly Journal of Economics*, *116*(4), 1149–1187. https://doi.org/10.1162/003355301753265543

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J., & Wagenmakers, E. J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), Article e2200300119. https://doi.org/10.1073/pnas.2200300119

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences, 119*(1), Article e2107346118. https://doi.org/10.1073/pnas.2107346118

Münscher, R., Vetter, M., & Scheuerle, T. (2016). A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making*, *29*(5), 511–524. https://doi.org/10.1002/bdm.1897

Murphy, K. R. (2020). Performance evaluation will not die, but it should. *Human Resource Management Journal*, *30*(1), 13–31. https://doi.org/10.1111/1748-8583.12259

Neeley, T. B. (2013). Language matters: Status loss and achieved status distinctions in global organizations. *Organization Science*, *24*(2), 476–497. https://doi.org/10.1287/orsc.1120.0739

Nisa, C. F., Bélanger, J. J., Schumpe, B. M., & Faller, D. G. (2019). Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. *Nature Communications*, *10*(1), Article 4545. https://doi.org/10.1038/s41467-019-12457-2

Pichert, D., & Katsikopoulos, K. V. (2008). Green defaults: Information presentation and pro-environmental behaviour. *Journal of Environmental Psychology*, *28*(1), 63–73. https://doi.org/10.1016/j.jenvp.2007.09.004

Piekkari, R., Vaara, E., Tienari, J., & Säntti, R. (2005). Integration or disintegration? Human resource implications of a common corporate language decision in a cross-border merger. *The International Journal of Human Resource Management*, *16*(3), 330–344. https://doi.org/10.1080/0958519042000339534

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. (2007). Absolute vs relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment*, *15*(3), 302–316. https://doi.org/10.1111/j.1468-2389.2007.00390.x

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 10, pp. 173–220). Academic Press. https://doi.org/10.1016/S0065-2601(08)60357-3

Santos, J. B., Hernandez, J. M., & Leão, W. (2019). Do recovery processes need empowered frontline employees? *International Journal of Operations & Production Management*, *39*(11), 1260-1279. https://doi.org/10.1108/IJOPM-12-2018-0745

Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research*, *57*(5), 853-877 https://doi.org/10.1177/00222437209418

Schmidt, A. T. (2019). Getting real on rationality-behavioral science, nudging, and public policy. *Ethics*, *129*(4), 511–543. https://doi.org/10.1086/702970

Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Philosophy Compass*, *15*(4), Article e12658. https://doi.org/10.1111/phc3.12658

Sharma, G., & Kulshreshtha, K. (2023). Why some leaders qualify for hate: an empirical examination through the lens of followers' perspective. *International Journal of Organizational Analysis*, *31*(2), 430-461. https://doi.org/10.1108/IJOA-08-2020-2369

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*, 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1972). Theories of bounded rationality. In C. B. McGuire & R. Radner (Eds.), *Decision and Organization* (pp. 161–176). Elsevier.

Sunstein, C. R. (2015). Nudges, agency, and abstraction: A reply to critics. *Review of Philosophy and Psychology*, *6*(3), 511–529. https://doi.org/10.1007/s13164-015-0266-z

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review, 70*(4), 1159–1202.

Szaszi, B., Goldstein, D. G., Soman, D., & Michie, S. (2025). Generalizability of choice architecture interventions. *Nature Reviews Psychology*, *4*(8), 518–529. https://doi.org/10.1038/s44159-025-00471-9

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., ... & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, *119*(31), Article e2200732119. https://doi.org/10.1073/pnas.2200732119

Tenzer, H., Pudelko, M., & Harzing, A. W. (2014). The impact of language barriers on trust formation in multinational teams. *Journal of International Business Studies*, *45*(5), 508–535. https://doi.org/10.1057/jibs.2013.64

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, *49*(4), 400–415.

Wilkinson, T. M. (2013). Nudging and manipulation. *Political Studies*, *61*(2), 341–355. https://doi.org/10.1111/j.1467-9248.2012.00974.x

Wilmot, N. V., Vigier, M., & Humonen, K. (2024). Language as a source of otherness. *International Journal of Cross Cultural Management*, *24*(1), 59–80. https://doi.org/10.1177/14705958231216936

CHAPTER 2 | HOW EFFECTIVE IS NUDGING IN THE LONG RUN? A META-ANALYSIS OF PRO-ENVIRONMENTAL BEHAVIOR

Chapter 2:
How Effective is Nudging in the Long Run? A Meta-Analysis of Pro-Environmental Behavior

This paper was written in single authorship. It is currently being prepared for submission. Additionally, the paper has been accepted for presentation at the World Meetings of the Economic Science Association (ESA) in Los Angeles, USA, from July 14 to 17, 2026.

Abstract

Nudging has become a widely used policy instrument for promoting pro-environmental behavior. Although extensive evidence demonstrates that nudges are effective in the short term, far less is known about whether these effects persist over time. This study presents the first systematic literature review and meta-analysis focused explicitly on the long-term effects of environmental nudges. A total of 42 publications reporting 140 effect sizes (N = 613,894) were synthesized using a three-level random-effects model. Across all studies, nudging interventions produced a small overall effect (d = 0.30, 95% CI [0.19, 0.41]). An analysis of the temporal dynamics based on 61 effect sizes showed that long-term effects remained positive and statistically significant, although smaller in magnitude than short-term effects. A continuous meta-regression revealed a significant decline in effectiveness over time (β = –0.0113), indicating gradual behavioral decay. Quartile analyses confirmed this pattern, with significant effects up to 60 days after intervention and increasing heterogeneity in effectiveness thereafter. Moderator analyses revealed substantial variation across nudge types. The findings provide systematic evidence that environmental nudges do lead to persistent behavior change, though with clear attenuation over time. These results offer important implications for policymakers and organizations designing long-term behavioral interventions and highlight the need for future research on mechanisms that enhance temporal stability.

2.1 INTRODUCTION

Nudges have become one of the most widely discussed instruments in the quest for changing behavior in public policy and organizational practice. More than 200 governmental behavioral insights units operate across over 50 countries (Hubble & Varazzani, 2023). Yet, despite this substantial expansion, a central question remains unresolved: Do nudges produce lasting changes in behavior, or are their effects merely short-lived? Scholars have repeatedly emphasized this gap. For example, Beshears and Kosowsky (2020, p. 5) explicitly call for more research on long-term outcomes, noting that "researchers should place greater emphasis on studying the extent to which nudges lead to cumulative long-run effects on outcomes."

While the short-term effectiveness of nudges is well established in a variety of fields, their long-term impact is far less understood (Hu et al., 2025). The majority of experimental and field studies focus on immediate behavioral responses—such as a single choice, a short-term shift in consumption, or a one-time compliance outcome—leaving the durability of these interventions unexplored. Among the studies that do include follow-up periods, findings are mixed (Beshears & Kosowsky, 2020; Congiu & Moscati, 2022). Some studies report persistence: energy conservation nudges continue to influence households after the intervention has ended (Allcott & Rogers, 2014), reminder postcards maintain effects on appointment adherence for several months (Altmann & Traxler, 2014), and default-based interventions can shape workplace behavior beyond their removal (Venema et al., 2018). However, other longitudinal evidence raises doubts about such durability. Polman and Maglio (2024) show that although default, decoy, and compromise nudges influence initial choices, individuals subsequently engage less with the nudged option over multiple months. Likewise, Van Rookhuijzen et al. (2021) find that nudge effects in prosocial contexts may spill over, but do not persist in domains such as healthy eating. Together, these findings suggest that while

nudges reliably shape short-term behavior, their ability to generate persistent change remains inconsistent and insufficiently understood.

Existing meta-analyses provide valuable insights into the overall effectiveness of nudges, but they do not resolve the question of persistence. Comprehensive syntheses have examined the average short-term impact of nudges across domains (e.g., Beshears & Kosowsky, 2020; Hummel & Maedche, 2019; Mertens et al., 2022), while several reviews focus on specific behavioral areas such as health (Arno & Thomas, 2016; Broers et al., 2017; Cadario & Chandon, 2020), environmental behavior (Nisa et al., 2019), or financial decision-making (Antinyan & Asatryan, 2024). Others analyze the relative performance of particular nudge types—for example, defaults (Jachimowicz et al., 2019) or social norms (Melnyk et al., 2022)—or compare effects across laboratory and field settings (DellaVigna & Linos, 2022). Yet none of these syntheses explicitly evaluate whether nudge effects persist beyond the immediate intervention period. As a result, the literature offers robust evidence on what nudges can achieve in the short term, but no systematic assessment of whether and under which conditions these effects endure.

To address this gap, this meta-analysis concentrates on nudging interventions within the environmental domain. This focus is motivated by both theoretical and empirical considerations. Environmental behavior represents a domain in which the long-term impact of interventions is particularly decisive: actions such as conserving energy, reducing waste, or adopting sustainable consumption patterns only generate environmental benefits if behavioral changes endure beyond the immediate intervention period. Moreover, compared with other domains, environmental nudging research offers a relatively rich body of studies that include follow-up measurements over weeks, months, or even years—providing the empirical basis to meaningfully assess the persistence of nudge effects. Finally, environmental nudges are widely implemented in public policy and organizational practice, making this domain especially

relevant for policymakers seeking to design interventions that achieve lasting behavioral change.

This study contributes to the literature in four important ways. First, it provides the first systematic review and meta-analysis that focuses explicitly on the long-term effectiveness of nudges in the environmental domain, thereby addressing a critical gap identified by policymakers and researchers alike. Second, it offers a quantitative synthesis of short-term and long-term effect sizes, enabling a direct comparison of how nudge impacts evolve over time. Third, the analysis examines moderators that may influence persistence, including nudge type, intervention duration, and study design, offering theoretical insights into the mechanisms that support or undermine durable behavior change. Fourth, the findings generate practical recommendations for policymakers and organizations seeking to design nudging interventions that produce persistent environmental outcomes, as well as identify promising avenues for future research on how behavioral interventions can support long-term change.

2.2 METHODS

For this systematic literature review and meta-analysis, I developed a search strategy to identify relevant studies on nudging, following the guidelines suggested by vom Brocke et al. (2008), Pickering and Byrne (2014) as well as Borenstein et al. (2021). The search was conducted in two major databases: Scopus and Web of Science. The search terms “nudge” OR “nudging” were applied to the title, abstract, and keywords of the studies. Full search strings for Scopus and Web of Science are provided in Appendix. The search covered all records available from 2008 to 2025. The starting point of 2008 was chosen because it marks the publication of the foundational book Nudge: Improving Decisions About Health, Wealth, and Happiness by Thaler and Sunstein (2008). To ensure quality and relevance, only peer-reviewed journal articles were included, while other types of publications (e.g., conference proceedings, grey literature) were excluded. Additionally, articles were limited to those published in English.

Literature reviews were excluded, as the focus was on original research employing quantitative designs (e.g., randomized controlled experiments), aligning with the key objective of this review: estimating the long-term effectiveness of nudging interventions.

The screening process was structured according to the PRISMA Statement (Page et al., 2021) and followed a multi-phase approach. During the initial screening, broad inclusion criteria were applied. These criteria were based on relevance as indicated by the title, abstract, and research domain. Only studies examining environmental behavior (e.g., energy use, waste reduction, sustainable consumption) were considered. In the second screening, studies were further assessed for methodological rigor, ensuring they met specific criteria such as reporting effect sizes or providing sufficient data for their calculation. Only studies that employed interventions consistent with the definition of nudging[2] were included. Additionally, studies had to measure actual behavior change rather than self-reported behavior. In the final screening phase, contextual factors of each study were evaluated, such as the sample information (e.g., students or households). The search, conducted in December 2024[3], identified a total of 4,294 unique publications. Ultimately, 42 articles containing 140 effect sizes were included in the meta-analysis, covering a pooled sample size of 613,894 participants (ranging from 21 to 246,940 participants). Figure 2.1 illustrates the PRISMA framework detailing the study identification and screening process.

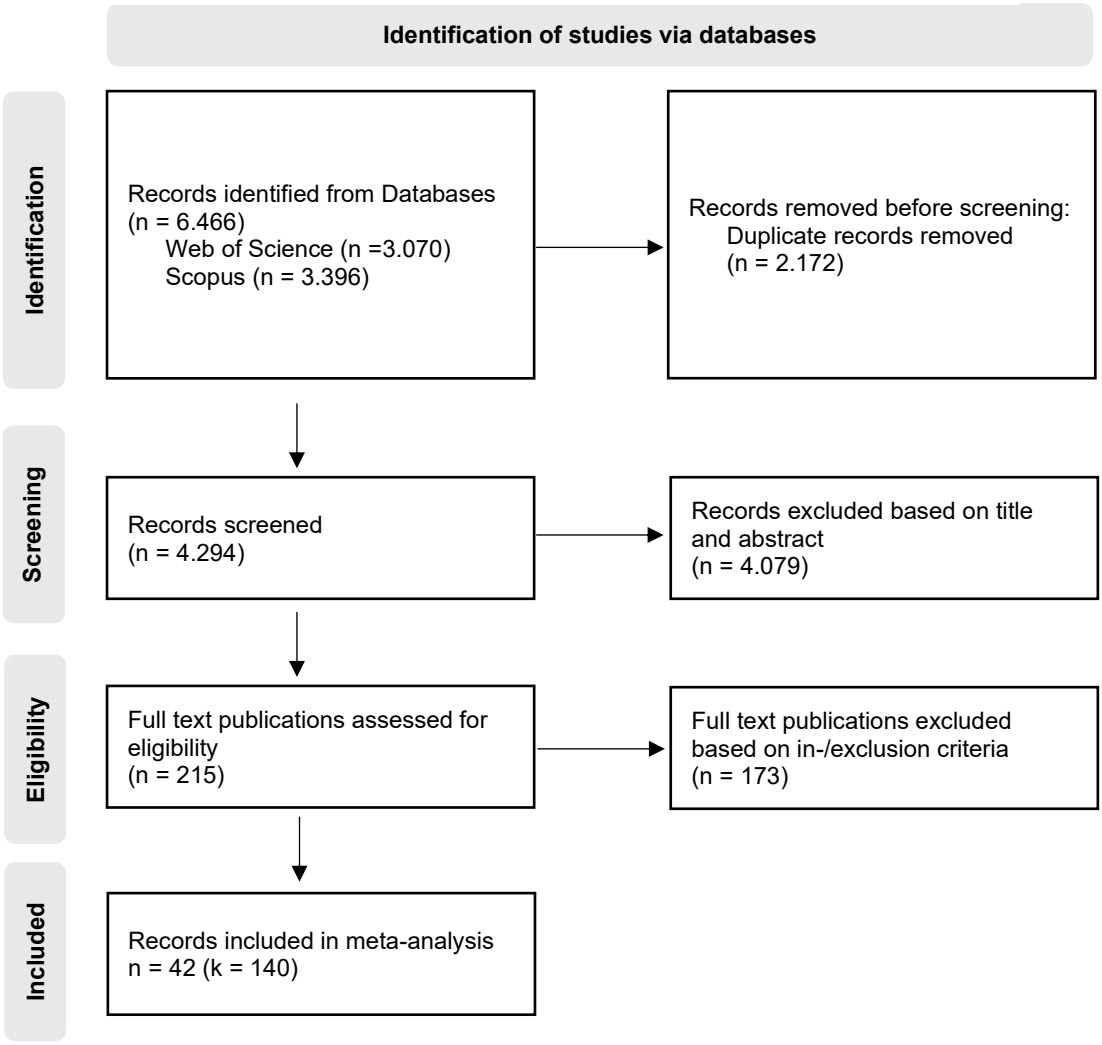
All included publications were coded using a structured coding manual (see Appendix), which I developed based on the codebook from the most recent and highly cited meta-analysis by Mertens et al. (2022). That codebook is itself grounded in the taxonomy of Münscher et al. (2016) for classifying nudging interventions and the experimental framework of Harrison and

[2] Following Thaler and Sunstein (2008), nudges were defined as interventions that alter choice architecture without restricting options or significantly changing economic incentives.

[3] Several studies with 2025 publication dates were already available at the time of the December 2024 search because Scopus and Web of Science index forthcoming articles ahead of print. These studies were accordingly included.

List (2004) for categorizing study designs. Building on this foundation, the coding manual for the present study was extended to capture additional features relevant for assessing long-term nudge effectiveness.

Figure 2.1: PRISMA framework



2.2.1 Codebook

For each publication, bibliographic and contextual characteristics were extracted, including author(s), title, year of publication, journal name, keywords, country of origin, and sample information (overall sample size; sample sizes for intervention and control groups; and population type such as general population, students, working population, households, or consumers).

Behavioral interventions were categorized according to Münscher et al.'s (2016) taxonomy, which provides a comprehensive classification of nudging techniques. This taxonomy distinguishes between three categories—decision information, decision structure, and decision assistance—each capturing different mechanisms through which nudges influence behavior.

- **Decision information** interventions target the way information is presented, without altering the underlying choice set. These techniques include translating information (e.g., simplifying or reframing messages to enhance comprehensibility), making information visible (e.g., providing feedback, labels, or real-time consumption displays), and providing social reference points (e.g., descriptive norms, peer comparisons). Such interventions aim to make relevant information more salient or easier to interpret, thereby facilitating more informed decision-making.
- **Decision structure** interventions modify the architecture of the choice environment. These include changing choice defaults, modifying option-related effort (e.g., making undesirable options harder to select), altering the range or composition of options (e.g., adding or removing alternatives), and changing option consequences (e.g., adjusting how costs or benefits are framed). These techniques influence behavior by altering the ease, attractiveness, or perceived outcomes associated with different options—often without changing the underlying incentives.
- **Decision assistance** interventions support individuals in following through on intended behaviors. This category includes providing reminders (e.g., letters, emails, or SMS prompts) and facilitating commitment (e.g., allowing pre-commitment or public pledges). Such interventions help to overcome forgetfulness, inertia, or self-regulation barriers.

Study types were classified following the experimental taxonomy of Harrison and List (2004), which provides a framework for differentiating levels of experimental control. This classification distinguishes among four types of experiments:

- **Conventional laboratory experiments**, in which participants (often students) make decisions in abstract, highly controlled settings. These experiments offer strong internal validity but may lack generalizability to real-world contexts.
- **Artefactual field experiments**, which involve non-student populations but still rely on controlled, laboratory-like decision tasks. They improve external validity while retaining a structured experimental setup.
- **Framed field experiments**, in which real-world participants make decisions in realistic contexts involving actual goods, tasks, or environments. These studies introduce naturalistic elements (e.g., real products, actual monetary outcomes) and thereby enhance validity.
- **Natural field experiments**, in which participants are unaware they are part of a study and behave in fully natural environments (e.g., grocery stores, public services). These studies are generally considered to offer the strongest external validity.

To address the research question concerning the long-term effectiveness of nudging interventions, additional variables capturing the temporal structure of each study were coded. Specifically, I recorded whether the study employed a cross-sectional design, which measures outcomes at a single point in time, or a longitudinal design, which tracks behavioral outcomes across multiple time points to assess persistence. When the study design was longitudinal, I recorded the duration of the study in days. Moreover, to account for temporal variation in the intervention itself, each nudge was coded as either permanent—continuously present throughout the intervention period (e.g., a default embedded in the choice architecture)—or temporary, meaning it appeared only once or for a limited time (e.g., a one-time feedback letter). This distinction is important for interpreting persistence, as permanent and temporary nudges differ substantially in their potential to produce lasting effects.

To further assess publication quality, I expanded the dataset to include the journal ranking (h-index) and number of citations for each study.

2.2.2 Analysis procedure

To synthesize the results of the included studies and compare the effectiveness of different nudge interventions, I calculated standardized effect sizes. For continuous outcome variables (e.g., energy consumption or amount of money saved), Cohen's d was computed by dividing the difference between treatment and control group means by the pooled standard deviation. For dichotomous outcomes (e.g., opting into a default plan: yes/no), effect sizes were calculated using the arcsine transformation, following procedures commonly applied in previous meta-analyses (Beshears & Kosowsky, 2020; Mertens et al., 2022) and consistent with the recommendations of Borenstein et al. (2021). For studies assessing changes over time within the same individuals, I additionally computed Morris's d, which accounts for the dependence between repeated measurements. Morris's d was calculated as the mean pre-post change in the treatment group minus the mean pre-post change in the control group, divided by the pooled pretest standard deviation (Morris, 2008).

Positive effect sizes indicate behavior change in the intended direction, meaning that the nudge successfully shifted behavior toward the target outcome defined by the intervention (e.g., reducing energy consumption, increasing recycling, or choosing a more sustainable option). Negative effect sizes, in contrast, reflect unintended or counterproductive changes in behavior. When studies reported multiple effect sizes, each was coded as a separate entry in the dataset. For studies that did not report separate sample sizes for treatment and control groups, sample sizes were estimated by dividing the total number of participants equally across experimental conditions. In cases where the available information was insufficient to calculate effect sizes, I reached out to the study authors to request the missing data. To ensure coding accuracy, papers were coded independently by two raters, including a trained student assistant. After the

independent coding, all discrepancies were reviewed and resolved through discussion until full consensus was reached.

Statistical analyses were performed in R using the metafor package (Viechtbauer, 2010). A three-level random-effects meta-analysis was performed to account for the hierarchical structure of the data, as many publications reported multiple studies and multiple effect sizes, which would otherwise violate the assumption of independent effect sizes (Borenstein et al., 2021). In this model, level 1 captured variance between effect sizes within the same sample, level 2 captured variance between different samples within a study, and level 3 captured variance between studies. To further address potential dependencies among effect sizes originating from the same study, cluster-robust standard errors were applied. All confidence intervals are based on robust variance estimates to ensure more reliable inference. Variance components were estimated using Restricted Maximum Likelihood (REML). Overall heterogeneity was quantified using τ^2 and I^2 estimates. Potential moderators were further examined using mixed-effects meta-regression models. To assess the robustness of the findings, several sensitivity and publication bias analyses were conducted. These include visual inspection of funnel plot asymmetry, leave-one-out analyses, Cook's distance, and statistical tests for publication bias.

2.3 RESULTS

A total of 42 publications reporting 140 effect sizes were included in the systematic review, covering a combined sample of 613,894 participants. Across these 140 effect sizes, a three-level random-effects meta-analysis indicates a statistically significant overall effect of nudging interventions on pro-environmental behavior (Cohen's $d = 0.30$, 95% CI [0.19, 0.41], $p < .001$). This finding indicates that, on average, nudges produce small improvements in environmentally relevant decisions, consistent with previous meta-analyses focusing on short-term effects (e.g., Mertens et al., 2022).

However, since the primary focus of this meta-analysis is the persistence of nudge effects over time, the analyses directly addressing temporal dynamics are presented first (see table 2.1). Three aspects of persistence were examined: (1) differences between short-term and long-term study designs, (2) the role of intervention duration, and (3) differences between temporary and permanent nudges.

Table 2.1: Analysis of temporary dynamics

Category	k	n	d	95% CI
Full sample:				
Overall effect size	140	613,894	0.2982	0.1881; 0.4082
Study type				
• Cross-sectional	79	79,195	0.3189	0.1818; 0.4560
• Longitudinal	61	534,699	0.2883	0.1092; 0.4674
Longitudinal studies only:				
Duration days	61	534,699	-0.0113	-0.0141; -0.0085
Duration days in quartiles				
• 2-30 days	16	2.317	0.2405	0.0494; 0.4317
• 31-60 days	15	117.994	0.2209	0.0184; 0.4233
• 61-90 days	15	302.424	0.4925	-0.1948; 1.1798
• > 90 days	15	111.964	0.1822	-0.1933; 0.5577
Permanent nudge				
• Temporary	45	168,413	0.4505	-0.0418; 0.9429
• Permanent	16	366,286	0.0556	-0.2715; 0.3827

2.3.1 Differences between short-term and long-term study designs

The initial evidence of temporal persistence comes from comparing cross-sectional and longitudinal study designs. Cross-sectional designs measure behavior immediately following the intervention and produce an estimated effect size: d = 0.3189 (95% CI [0.1818, 0.4560]). In contrast, longitudinal designs, which measure outcomes after a delay ranging from several days to several months, produced a slightly smaller but still significant effect size: d = 0.2883 (95% CI [0.1092, 0.4674]). The modest difference between these two estimates suggests that,

although nudges tend to have the strongest impact immediately after implementation, environmental nudges maintain a significant degree of effectiveness over time. Notably, the long-term estimate remains positive and statistically significant, suggesting that nudges can lead to lasting behavioral changes, not just ephemeral ones.

2.3.2 Role of intervention duration (longitudinal studies only)

The assessment of the role of intervention duration involved the combination of meta-regression, quartile analyses, and visual analyses.

2.3.2.1 Meta-regression analysis

To more precisely model temporal dynamics, a continuous meta-regression was conducted, using follow-up duration (in days) as a predictor of effect size. This analysis was restricted to longitudinal studies, as only these studies provide the repeated measurements necessary to estimate decay trajectories over time. The results revealed a significant negative relationship ($\beta = –0.0113$, 95% CI [–0.0141, –0.0085]). This indicates that, on average, the expected effect size declines slightly with each additional day between the intervention and outcome measurement. Although the daily decrease is small, the cumulative decay pattern is significant. Nudges gradually lose strength over time yet continue to exert a detectable influence, even after extended periods.

2.3.2.2 Quartile analyses

To examine whether temporal dynamics follow a nonlinear pattern, follow-up durations in longitudinal studies were grouped into quartiles. Across the four intervals, the effect size remained positive but varied substantially in magnitude.

In the first quartile (2–30 days), nudges produced an effect size of $d = 0.2405$ (95% CI [0.0494, 0.4317]). This indicates that nudges continue to have a significant impact within the first month

after implementation. The confidence interval excludes zero, suggesting robust persistence of the intervention effect.

In the second quartile (days 31–60), the effect size remained positive and comparable in magnitude (d = 0.2209, 95% CI [0.0184, 0.4233]). Although the point estimate is slightly lower than in the first quartile, the results suggest that nudges maintain their influence for up to two months. The confidence interval is wider than in the first quartile, but it still excludes zero, indicating a statistically significant effect.

The third quartile (61–90 days) showed the largest point estimate (d = 0.4925, 95% CI [-0.1948, 1.1798]). However, the confidence interval is extremely wide, and it includes zero. This indicates that the effect in this interval is not statistically significant. This lack of statistical significance suggests that the apparent increase in the point estimate is likely driven by a small subset of studies with unusually strong effects combined with substantial heterogeneity and limited statistical power within this timeframe. Therefore, this quartile should be interpreted with caution.

Finally, in the fourth quartile (>90 days), the effect size was smaller (d = 0.1822, 95% CI [–0.1933, 0.5577]) and characterized by substantial uncertainty. The confidence interval includes zero, indicating that long-term persistence beyond three months is more variable and less detectable across the available evidence. The positive direction of the estimate suggests that nudges may continue to influence behavior even after extended periods, though with attenuated magnitude. Taken together, the quartile analysis aligns with the broader temporal pattern observed in the continuous meta-regression: nudge effects tend to weaken gradually as follow-up duration increases. The first two quartiles demonstrate statistically significant persistence up to 60 days. Beyond this period, however, the effects become more heterogeneous and imprecise and are no longer statistically significant, suggesting that long-term persistence is possible, but more dependent on context.

2.3.2.3 Visual analyses of intervention period

Visual analyses were performed to complement the quantitative meta-analytic results and explore the relationship between intervention duration and effect magnitude in longitudinal nudging studies. Figure 2.2 shows the distribution of effect sizes according to intervention duration in all longitudinal studies, using a logarithmic scale. Logarithmic transformation was applied because intervention durations are highly right-skewed, ranging from a few days to several years. The log scale allows both short- and long-term studies to be visualized within a coherent framework without distortion. The smoothed trend indicates a nonlinear relationship between duration and effect size. Effectiveness increases during the early intervention period, peaks between days 60 and 120, and then declines. Very long-lasting interventions show comparatively small average effects, suggesting that the behavioral impact of nudges diminishes over extended time periods.

Figure 2.2: Distribution of effect sizes across all longitudinal studies

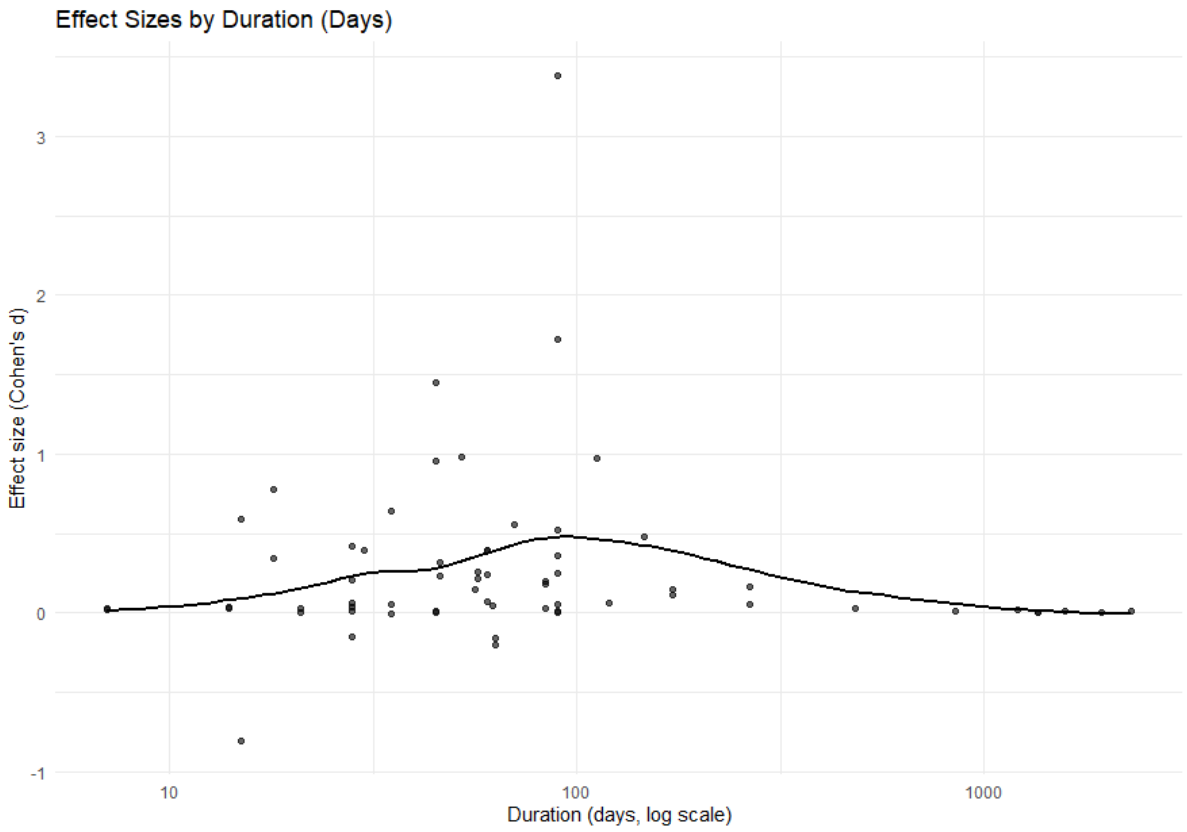
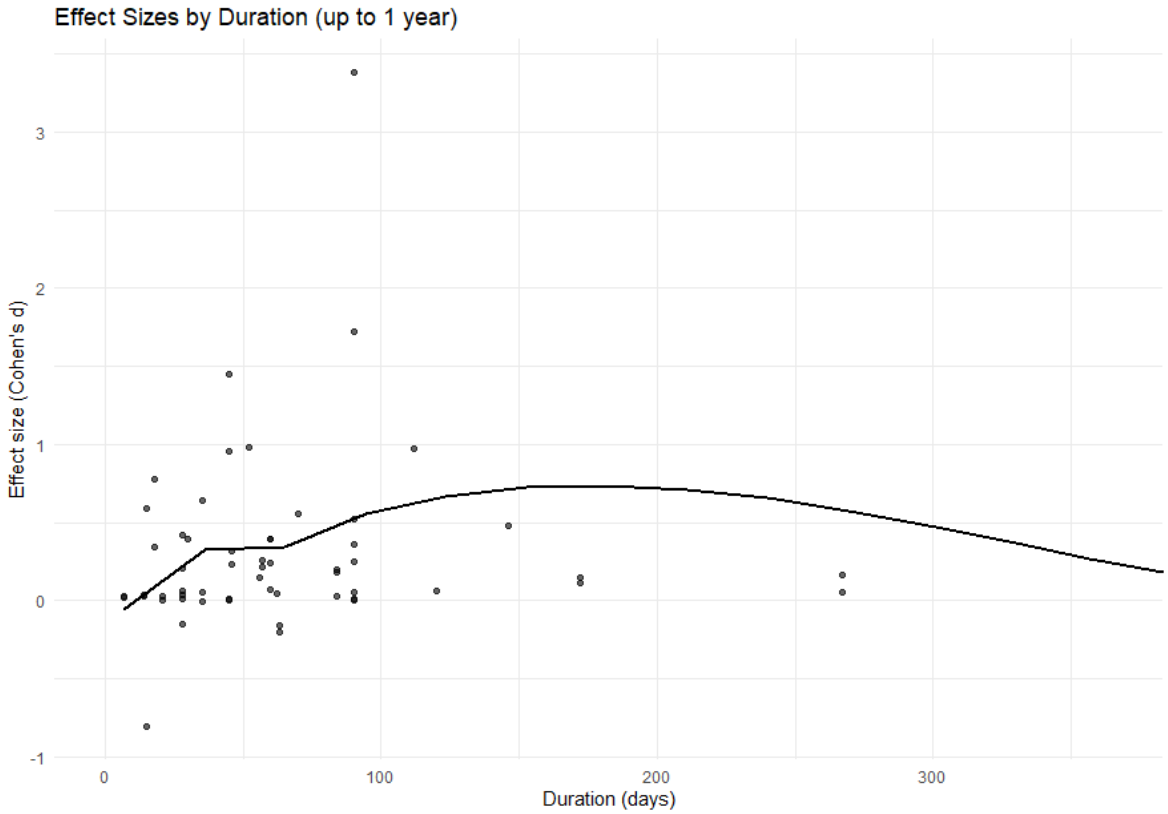


Figure 2.3 shows the same relationship for interventions lasting up to one year. The upper limit of one year was chosen because the majority of longitudinal studies fall within this duration range, ensuring that the visualization is not disproportionately influenced by a small number of very long-term interventions. Within this range, the trend is consistent with the full analysis, showing growing effect sizes initially that reach their maximum within the medium-duration range (roughly 60–120 days) before declining. This pattern suggests that nudges are most effective during short to medium-length implementation periods and that their influence weakens with too short or too prolonged exposure.

Figure 2.3: Distribution of effect sizes across longitudinal studies lasting up to 1 year



Visual analyses are not intended as inferential tests but rather as exploratory diagnostics that characterize temporal patterns not easily captured by point estimates or linear meta-regression coefficients. The visual findings, however, are consistent with the results of the meta-regression and quartile analyses. Together, these analyses provide convergent evidence that the effects of

nudging gradually attenuate over time and are strongest during the early to medium stages of implementation.

2.3.3 Differences between permanent and temporary nudges (longitudinal studies only)

In addition to the duration of temporal follow-up, the persistence of nudge effects may depend on whether the intervention is implemented as a temporary or permanent change to the choice architecture. To examine this, effect sizes were compared across nudges that were applied at a single point in time (e.g., reminders or feedback) versus those embedded continuously into the choice architecture (e.g., default modifications).

Temporary nudges yielded a larger effect size (d = 0.4505, 95% CI [–0.0418, 0.9429]). However, the wide confidence interval, which spans from slightly negative to nearly one standard deviation, indicates substantial heterogeneity and a lack of statistical significance. The inclusion of zero in the confidence interval suggests that, despite often being attention-grabbing, temporary nudges may vary considerably in their effectiveness depending on context, design, and exposure frequency.

Permanent nudges, in contrast, produced a smaller effect (d = 0.0556, 95% CI [–0.2715, 0.3827]). This estimate is close to zero, and the confidence interval encompasses a broad spectrum of potential effects, ranging from negative to positive. Accordingly, no statistically reliable effect is detectable for permanent nudges in this dataset. The attenuation of effects for permanent interventions may reflect habituation processes. Once individuals become accustomed to a stable feature of their decision environment, its psychological salience—and thus its behavioral influence—may diminish over time. Alternatively, the choice architecture in which such nudges are implemented may impose structural constraints that limit the magnitude of observable behavioral change.

2.3.4 Additional moderator analysis

Several study- and intervention-level moderators were analyzed to examine under which conditions nudges are more or less effective (see table 2.2). Due to the substantial heterogeneity observed in overall effect sizes, the goal of these analyses was to identify systematic patterns in effectiveness rather than to establish definitive causal moderators. Moderator analyses were conducted separately for longitudinal and cross-sectional studies to account for their fundamental differences in design and temporal structure. Due to the limited number of effect sizes in some subgroups, the results should be interpreted as descriptive patterns rather than conclusive evidence of differential effectiveness.

A) Moderator analysis of longitudinal studies

Intervention category

First, moderator analyses were conducted for longitudinal studies by intervention category, as these studies are central to assessing the persistence of nudge effects over time. Following the taxonomy of Münscher et al. (2016), nudges were classified into decision assistance, decision information, and decision structure interventions.

Decision assistance nudges produced a small but statistically significant average effect size (d = 0.2807, 95% CI [0.0609, 0.5004]). Similarly, decision information nudges (k = 47) produced a comparable, statistically significant effect size (d = 0.2755, 95% CI [0.0534, 0.4976]). These findings suggest that assistance- and information-based nudges can generate behavioral changes that persist beyond the intervention period. Decision structure nudges showed a larger point estimate (d = 0.4460, 95% CI [0.2956, 0.5963]). However, this result is based on only two effect sizes. Consequently, despite the relatively large estimate, this finding should be interpreted with considerable caution as it may reflect idiosyncratic study characteristics rather than a generalizable pattern.

Table 2.2: Additional moderator analysis

Moderator Analysis	k	n	d	95% CI
Longitudinal studies only:	61	534,699	0.2883	0.1092; 0.4674
Intervention category				
• Assistance	12	123264	0.2807	0.0609; 0.5004
• Information	47	410830	0.2755	0.0534; 0.4976
• Structure	2	605	0.4460	0.2956; 0.5963
Experiment Type				
• Conventional lab	2	42	-0.1897	-1.3477; 0.9683
• Artefactual field	-	-	-	-
• Framed field	2	49	0.8245	0.5892; 1.0598
• Natural field	57	534608	0.0170	0.0550; 0.4919
Population				
• General population	5	55384	0.3727	0.0479; 0.6976
• Students	12	616	0.0129	-0.3825; 0.4083
• Working population	23	1671	0.5479	0.0834; 1.0125
• Households	14	106879	-0.0844	-0.6919; 0.5230
• Consumers	7	370149	0.2793	0.0525; 0.5061
Cross-sectional studies only:	79	79,195	0.3189	0.1818; 0.4560
Intervention category				
• Assistance	4	6092	0.4157	0.2666; 0.5649
• Information	45	13032	0.2843	0.1436; 0.4251
• Structure	30	60071	0.3607	0.2023; 0.5191
Experiment Type				
• Conventional lab	32	2951	0.3591	0.0494; 0.6687
• Artefactual field	28	25885	0.2596	-0.3798; 0.8990
• Framed field	4	684	0.0192	-0.1364; 0.1748
• Natural field	15	49675	0.4214	-0.5717; 1.4145
Population				
• General population	36	24965	0.3594	0.0269; 0.6918
• Students	24	2603	0.2307	-0.1038; 0.5651
• Working population	9	2717	0.1894	0.0658; 0.3130
• Households	4	42552	0.5583	-0.9402; 2.0569
• Consumers	6	6358	0.2892	0.0144; 0.5639

Experiment type

Next, moderator analyses examined differences across experimental designs classified according to the framework of Harrison and List (2004). Most of the longitudinal evidence came from natural field experiments (k = 57), which produced a small, yet statistically significant, effect size (d = 0.0170, 95% CI [0.0550, 0.4919]).

Other experimental types were represented by a small number of studies. Conventional laboratory experiments (k = 2) produced a negative but highly uncertain estimate (d = −0.1897, 95% CI [−1.3477, 0.9683]), and framed field experiments (k = 2) yielded a large point estimate (d = 0.8245, 95% CI [0.5892, 1.0598]). Due to the extremely limited number of studies in these categories, these estimates are not suitable for meaningful comparison.

Population

Moderator analyses by population type reveal additional heterogeneity in long-term nudge effectiveness. Studies targeting working populations produced a relatively large and statistically significant effect size (d = 0.5479, 95% CI [0.0834, 1.0125]). Similarly, studies conducted in the general population (k = 5) produced a significant average effect size (d = 0.3727, 95% CI [0.0479, 0.6976]). In contrast, effects among students (k = 12) were small and not statistically significant (d = 0.0129, 95% CI [-0.3825, 0.4083]). Likewise, studies focusing on households (k = 14) showed a negative but highly uncertain estimate (d = −0.0844, 95% CI [−0.6919, 0.5230]). Studies focusing on consumers (k = 7) produced a small, yet statistically significant, effect size (d = 0.2793, 95% CI [0.0525, 0.5061]).

B) Moderator analysis of cross-sectional studies

The primary focus of this study is the persistence of nudge effects over time. However, analyzing moderators in cross-sectional studies provides an important point of reference for interpreting longitudinal results. Specifically, this analysis allows for an evaluation of whether the characteristics of the intervention, the population, and the study context that drive short-term effectiveness also explain long-term persistence or if different patterns emerge over time.

Intervention category

For cross-sectional studies, which capture immediate behavioral responses, moderator analyses by intervention category revealed statistically significant effects across all nudge types. Decision assistance nudges (k = 4) produced an average effect size of d = 0.4157 (95% CI [0.2666, 0.5649]), decision information nudges (k = 45) yielded d = 0.2843 (95% CI [0.1436, 0.4251]), and decision structure nudges (k = 30) produced d = 0.3607 (95% CI [0.2023, 0.5191]).

Experiment type

Cross-sectional moderator analyses by experimental design revealed positive average effects across most study types. Conventional laboratory experiments (k = 32) produced a statistically significant effect (d = 0.3591, 95% CI [0.0494, 0.6687]). Artefactual field experiments (k = 28) showed a positive but non-significant estimate (d = 0.2596, 95% CI [–0.3798, 0.8990]). Framed field experiments (k = 4) yielded a small and non-significant effect (d = 0.0192, 95% CI [–0.1364, 0.1748]), while natural field experiments (k = 15) produced a positive but highly uncertain estimate (d = 0.4214, 95% CI [–0.5717, 1.4145]).

Population

Finally, moderator analyses by population type for cross-sectional studies showed positive average effects across all populations examined. Nudges targeting the general population (k = 36) produced a statistically significant effect (d = 0.3594, 95% CI [0.0269, 0.6918]). Consumer-focused studies (k = 6) similarly yielded a significant effect (d = 0.2892, 95% CI [0.0144, 0.5639]). Effects among students (k = 24), working populations (k = 9), and households (k = 4) were positive but not statistically significant.

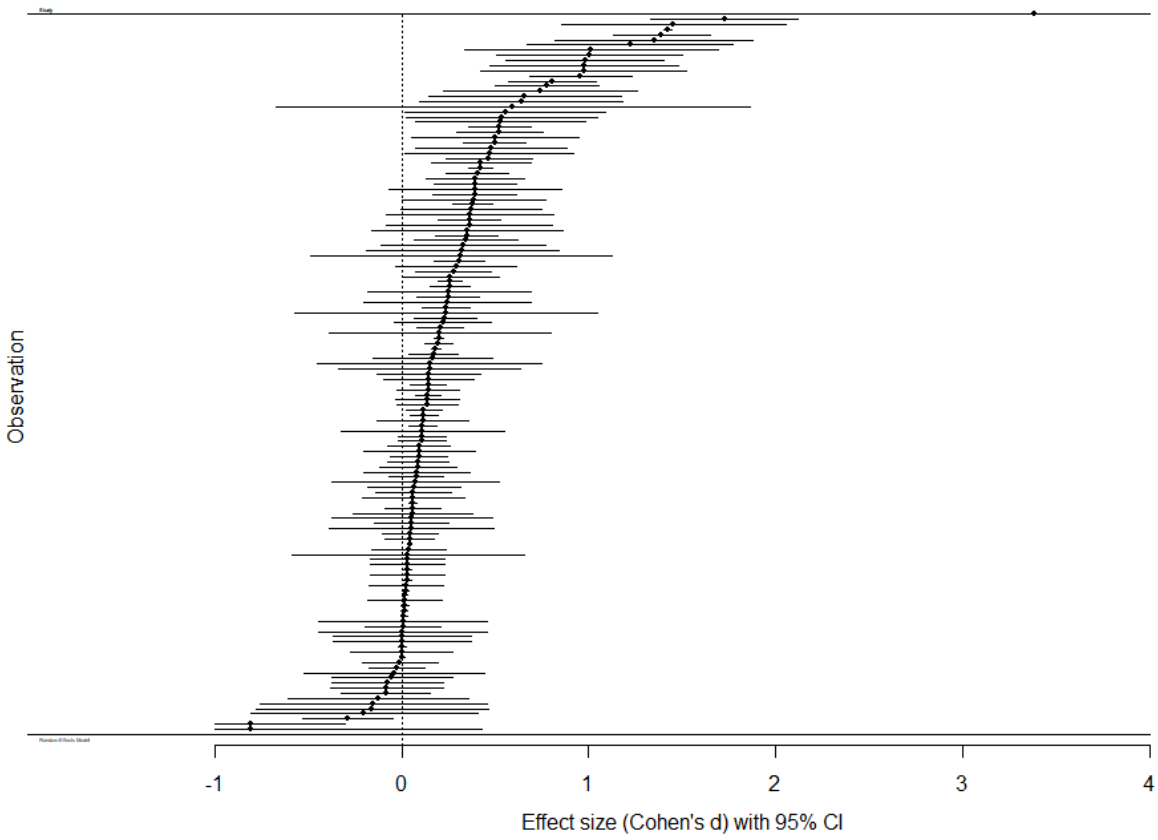
Across both longitudinal and cross-sectional studies, moderator analyses reveal substantial heterogeneity in nudge effectiveness across intervention types, populations, and study contexts. While short-term effects appear robust across most categories, long-term persistence is more context-dependent and supported by uneven evidence across subgroups. No single intervention

category or population consistently dominates in producing persistent effects, underscoring that the long-term effectiveness of environmental nudges depends on how, where, and for whom they are implemented.

2.3.5 Robustness check and sensitivity analysis (full data set)

A corresponding forest plot illustrates the wide dispersion of effect sizes across all studies (Figure 2.4). Heterogeneity across all studies is substantial ($Q(139) = 16{,}437.72$, $p < .001$), with considerable between-publication variance ($\tau^2 = 0.122$) and only minimal within-publication variance ($\tau^2 = 0.002$). To account for dependent effect sizes, cluster-robust standard errors based on 49 study clusters were applied. The 95% prediction interval ranges from −0.42 to 1.02, indicating that the true effect of a choice architecture intervention in a new study may vary widely, ranging from negative to very strong positive effects.

Figure 2.4: Forest plot



To evaluate the necessity of the variance components, the full three-level model is compared with reduced models in which one of the random-effects variances is fixed to zero. Removing the publication-level variance component (Level 3) significantly worsens model fit (LRT = 110.4885, $p < .0001$). Likewise, removing the within-study effect-size variance component (Level 2) also results in a substantially poorer fit (LRT = 145.2395, $p < .0001$). These tests indicate that both variance components are statistically significant and that a three-level random-effects structure is required to appropriately account for the dependence among effect sizes.

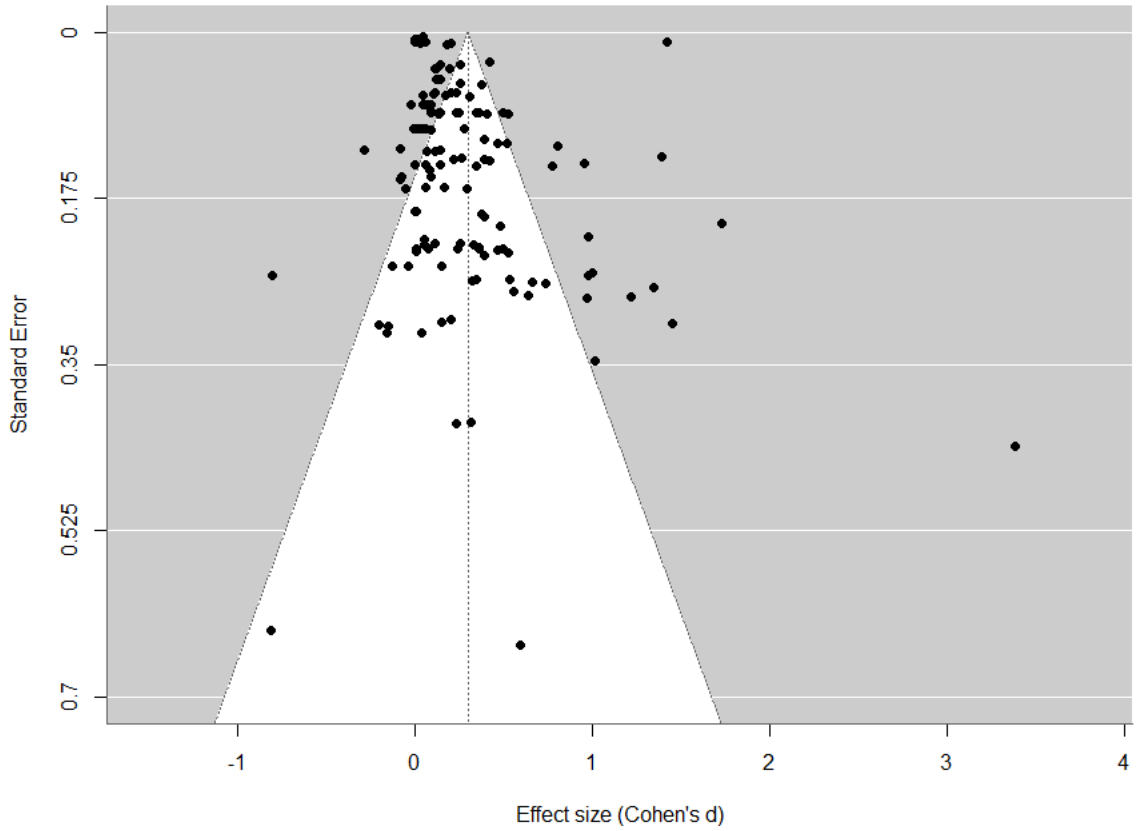
The total amount of true heterogeneity was $\tau^2 = 0.124$. The overall I^2 statistic indicates that 99.36% of the observed variance reflects true heterogeneity rather than sampling error. Of this total heterogeneity, 97.43% is attributable to between-study differences (Level 3), whereas only 1.93% originates from variation between effect sizes within the same study (Level 2). This pattern demonstrates that almost all heterogeneity arises at the study level, with minimal dependence among effect sizes within studies.

As part of the sensitivity analysis, standardized residuals were examined to detect potential outliers. Two effect sizes exceeded the conventional $|z| > 3$ threshold (Chakravarty & Mishra, 2019; Visser & Schoormans, 2023). Influence diagnostics based on Cook's distance further indicated that one effect size from Chakravarty and Mishra (2019) slightly exceeded the recommended cutoff value ($4 / (k - p - 1)$), suggesting that this estimate may exert disproportionate influence on the model. Therefore, one outlier was excluded from the analysis. Overall, only a very small subset of effect sizes showed indications of outlying or influential behavior.

The funnel plot, Figure 2.5, illustrates the expected funnel-shaped pattern but also suggests considerable variability among studies with larger standard errors. To formally test for small-study effects, an Egger-type regression was fitted by regressing effect sizes on the square root

of their sampling variance in a three-level meta-analytic model with cluster-robust standard errors. The moderator was not statistically significant, $F(1, 47) = 0.28$, $p = .60$, indicating no evidence of small-study effects or publication bias according to this test.

Figure 2.5: Funnel plot



2.4 DISCUSSION

To the best of my knowledge, this meta-analysis is the first to systematically analyze the long-term effectiveness of environmental nudges. Across 42 publications and 140 effect sizes, nudging interventions produce a small average effect on pro-environmental behavior ($d = 0.30$, 95% CI [0.19, 0.41]). Notably, this effect was not limited to immediate post-intervention outcomes. Longitudinal analyses demonstrate that nudge effects persist over time, though their effectiveness gradually declines as the temporal distance between the intervention and the outcome measurement increases.

Meta-regression results reveal a negative association ($\beta = -0.0113$) between follow-up duration and effect size. This indicates a gradual behavioral decay rather than abrupt extinction. Quartile analyses confirm this pattern, showing that nudges produce statistically significant effects for up to 60 days. However, beyond this period, the effects become increasingly heterogeneous. Visual diagnostics reveal a nonlinear trajectory, indicating that effectiveness peaks during short- to medium-term implementation periods and then decreases over longer periods. Together, these findings suggest that environmental nudges can generate persistent behavioral changes, though these changes are neither uniform nor indefinite.

The present study extends prior meta-analytic evidence in several ways. The findings directly address the call by Beshears and Kosowsky (2020) for systematic evidence on long-run outcomes, as the literature has convincingly established short-term effectiveness. In line with this perspective, the present analysis shows that nudges can influence behavior beyond the intervention period. It also confirms that long-term effectiveness cannot be inferred from short-term success.

In the environmental domain, Nisa et al. (2019) report small average effects of behavioral interventions on household in climate action. They conclude that such interventions are unlikely to produce lasting change once exposure ends. The present results do not contradict this conclusion but refine it by showing that nudge effects tend to decay gradually rather than disappear abruptly. Modeling persistence as a continuous function of time reveals that attenuation and increasing heterogeneity characterize longer follow-up periods.

In contrast to the comprehensive meta-analysis by Mertens et al. (2022), which found that decision-structure nudges outperformed other intervention types based largely on short-term outcomes, the present study calls into question claims of persistent superiority. When considering long-term outcomes, the differences between nudge categories become less pronounced, and the evidence supporting sustained advantages of specific mechanisms remains

limited. However, the small number of longitudinal effect sizes in several moderator categories substantially limits statistical power and warrants cautious interpretation of these findings.

Evidence from large-scale policy implementations further supports this interpretation. DellaVigna and Linos (2022) demonstrate that the effects of nudges implemented by government nudge units are smaller than those reported in academic trials. The present results suggest that temporal attenuation may contribute to this discrepancy, as even initially effective nudges lose their effectiveness over time in complex real-world environments.

Recent second-order meta-analytic research has raised concerns about the robustness of average nudge effects when corrected for publication bias (Hu et al., 2025). While these concerns are valid, the current findings suggest that persistence cannot be determined based solely on short-term aggregate effect sizes. This meta-analysis explicitly models time as a continuous dimension and demonstrates that nudges do not suddenly become ineffective after implementation. Rather, they exhibit gradual decay with substantial heterogeneity across contexts.

From a policy perspective, the findings suggest that nudges are a cost-effective tool for promoting pro-environmental behavior. However, they also caution against expecting permanent effects. The observed decay indicates that nudges should be incorporated into adaptive intervention strategies that may include periodic reinforcement or integration with other policies. Notably, the substantial heterogeneity observed at longer time horizons underscores the necessity of continuous evaluation rather than one-time implementation.

Several limitations warrant consideration. Despite the focus on long-term outcomes, the number of studies with extended follow-up periods is limited. This contributes to wide confidence intervals. Additionally, the small number of effect sizes in some moderator categories limits the ability to detect reliable differences. Future research should prioritize

longitudinal experiments with repeated measurements and explicitly test mechanisms such as habituation or learning effects.

2.5 CONCLUSION

This meta-analysis provides evidence that environmental nudges have small yet significant effects that persist beyond the intervention period. Although nudges can lead to long-lasting behavior changes, their effectiveness gradually declines over time, suggesting a process of attenuation rather than abrupt disappearance.

The findings also show that the long-term effectiveness of nudges varies depending on the type of intervention, the population, and the study context. This suggests that the effectiveness of nudges is not inherent, but rather contingent on the specific design and implementation conditions. Therefore, nudges should not be viewed as self-sustaining interventions but rather as tools that may require reinforcement to maintain their impact.

Future research should focus on identifying the mechanisms that support long-lasting behavior change and designing interventions that enhance the temporal stability of nudge effects.

2.6 APPENDIX

Inlcuded studies in Meta-Analysis

1. Antinyan, A., & Corazzini, L. (2025). Breaking the bag habit: Testing interventions to reduce plastic bag demand. *Ecological Economics*, *228*, Article 108454. https://doi.org/10.1016/j.ecolecon.2024.108454
2. Banerjee, S., & Picard, J. (2023). Thinking through norms can make them more effective. Experimental evidence on reflective climate policies in the UK. *Journal of Behavioral and Experimental Economics*, *106*, Article 102024. https://doi.org/10.1016/j.socec.2023.102024
3. Bastini, K., Kerschreiter, R., Lachmann, M., Ziegler, M., & Sawert, T. (2024). Encouraging individual contributions to net-zero organizations: Effects of behavioral policy interventions and social norms. *Journal of Business Ethics*, *192*(3), 543–560. https://doi.org/10.1007/s10551-023-05516-8
4. Bauer, J. M., Aarestrup, S. C., Hansen, P. G., & Reisch, L. A. (2022). Nudging more sustainable grocery purchases: Behavioural innovations in a supermarket setting. *Technological Forecasting and Social Change*, *179*, Article 121605. https://doi.org/10.1016/j.techfore.2022.121605
5. Belaïd, F., & Flambard, V. (2023). Boosting buildings energy efficiency: the impact of social norms and motivational feedback. *Journal of Economic Behavior & Organization*, *215*, 26–39. https://doi.org/10.1016/j.jebo.2023.09.003
6. Bernedo, M., Ferraro, P. J., & Price, M. (2014). The persistent impacts of norm-based messaging and their implications for water conservation. *Journal of Consumer Policy*, *37*(3), 437–452. https://doi.org/10.1007/s10603-014-9266-0
7. Bruns, H., Kantorowicz-Reznichenko, E., Klement, K., Jonsson, M. L., & Rahali, B. (2018). Can nudges be transparent and yet effective? *Journal of Economic Psychology*, *65*, 41–59. https://doi.org/10.2139/ssrn.2816227
8. Buckley, P., & Llerena, D. (2022). Nudges and peak pricing: A common pool resource energy conservation experiment. *Journal of Behavioral and Experimental Economics*, *101*, Article 101928. https://doi.org/10.1016/j.socec.2022.101928
9. Caballero, N., & Ploner, M. (2022). Boosting or nudging energy consumption? The importance of cognitive aspects when adopting non-monetary interventions. *Energy Research & Social Science*, *91*, Article 102734. https://doi.org/10.1016/j.erss.2022.102734

10. Cappa, F., Rosso, F., Giustiniano, L., & Porfiri, M. (2020). Nudging and citizen science: The effectiveness of feedback in energy-demand management. *Journal of Environmental Management*, *269*, Article 110759. https://doi.org/10.1016/j.jenvman.2020.110759
11. Carattini, S., & Blasch, J. (2024). Nudging when the descriptive norm is low: Evidence from a carbon offsetting field experiment. *Journal of Behavioral and Experimental Economics*, *110*, Article 102194. https://doi.org/10.1016/j.socec.2024.102194
12. Chakravarty, S., & Mishra, R. (2019). Using social norms to reduce paper waste: Results from a field experiment in the Indian Information Technology sector. *Ecological Economics*, *164*, Article 106356. https://doi.org/10.1016/j.ecolecon.2019.106356
13. Chandra, G. (2023). Non-monetary intervention to discourage consumption of single-use plastic bags. *Behavioural Public Policy*, *7*(1), 143–156. https://doi.org/10.1017/bpp.2020.9
14. Choudhary, V., Shunko, M., Netessine, S., & Koo, S. (2022). Nudging drivers to safety: Evidence from a field experiment. *Management Science*, *68*(6), 4196–4214. https://doi.org/10.1287/mnsc.2021.4063
15. Crago, C. L., Spraggon, J. M., & Hunter, E. (2020). Motivating non-ratepaying households with feedback and social nudges: A cautionary tale. *Energy Policy*, *145*, Article 111764. https://doi.org/10.1016/j.enpol.2020.111764
16. Dannenberg, A., & Weingärtner, E. (2023). The effects of observability and an information nudge on food choice. *Journal of Environmental Economics and Management*, *120*, Article 102829. https://doi.org/10.1016/j.jeem.2023.102829
17. Decrinis, L., Freibichler, W., Kaiser, M., Sunstein, C. R., & Reisch, L. A. (2023). Sustainable behaviour at work: How message framing encourages employees to choose electric vehicles. *Business Strategy and the Environment*, *32*(8), 5650–5668. https://doi.org/10.1002/bse.3441
18. Demarque, C., Charalambides, L., Hilton, D. J., & Waroquier, L. (2015). Nudging sustainable consumption: The use of descriptive norms to promote a minority behavior in a realistic online shopping environment. *Journal of Environmental Psychology*, *43*(1), 166–174. https://doi.org/10.1016/j.jenvp.2015.06.008
19. Dorigoni, A., & Bonini, N. (2023). Water bottles or tap water? A descriptive-social-norm based intervention to increase a pro-environmental behavior in a restaurant. *Journal of Environmental Psychology*, *86*, Article 101971. https://doi.org/10.1016/j.jenvp.2023.101971

20. Ebeling, F., & Lotz, S. (2015). Domestic uptake of green energy promoted by opt-out tariffs. *Nature Climate Change*, *5*(9), 868–871. https://doi.org/10.1038/nclimate2681
21. Filippini, M., Kumar, N., & Srinivasan, S. (2021). Nudging adoption of electric vehicles: Evidence from an information-based intervention in Nepal. *Transportation Research Part D: Transport and Environment*, *97*, Article 102951. https://doi.org/10.1016/j.trd.2021.102951
22. Gessner, J., Habla, W., & Wagner, U. J. (2024). Can social comparisons and moral appeals encourage low-emission transport use? *Transportation Research Part D: Transport and Environment*, *133*, Article 104289. https://doi.org/10.1016/j.trd.2024.104289
23. Gustavsson, P., & Ljung Aust, M. (2024). In-vehicle nudging for increased Adaptive Cruise Control use: a field study. *Journal on Multimodal User Interfaces*, *18*(2), 257–271. https://doi.org/10.1007/s12193-024-00434-z
24. Hauslbauer, A. L., Schade, J., Drexler, C. E., & Petzoldt, T. (2022). Extending the theory of planned behavior to predict and nudge toward the subscription to a public transport ticket. *European Transport Research Review*, *14*(1), Article 5. https://doi.org/10.1186/s12544-022-00528-3
25. Hilton, D., Charalambides, L., Demarque, C., Waroquier, L., & Raux, C. (2014). A tax can nudge: The impact of an environmentally motivated bonus/malus fiscal system on transport preferences. *Journal of Economic Psychology*, *42*, 17–27. https://doi.org/10.1016/j.joep.2014.02.007
26. Howley, P., & Ocean, N. (2022). Can nudging only get you so far? Testing for nudge combination effects. *European Review of Agricultural Economics*, *49*(5), 1086–1112. https://doi.org/10.1093/erae/jbab041
27. Kallbekken, S., & Sælen, H. (2013). 'Nudging'hotel guests to reduce food waste as a win–win environmental measure. *Economics Letters, 119*(3), 325–327. https://doi.org/10.1016/j.econlet.2013.03.019
28. Lindahl, T., & Linder, N. (2024). What factors influence choosing fish over meat among grocery shoppers? Insights from an unsuccessful nudge intervention. *Ecological Economics*, *224*, Article 108297. https://doi.org/10.1016/j.ecolecon.2024.108297
29. Lopes, A. A., Tasneem, D., & Viriyavipart, A. (2023). Nudges and compensation: Evaluating experimental evidence on controlling rice straw burning. *Ecological Economics*, *204*, Article 107677. https://doi.org/1010.1016/j.ecolecon.2022.107677

30. Luo, Y., & Zhao, J. (2023). Using behavioral interventions to reduce single-use produce bags. *Resources, Conservation and Recycling*, *193*(19), Article 106942. https://doi.org/10.1016/j.resconrec.2023.106942
31. My, K. B., & Ouvrard, B. (2019). Nudge and tax in an environmental public goods experiment: Does environmental sensitivity matter? *Resource and Energy Economics*, *55*, 24–48. https://doi.org/10.1016/j.reseneeco.2018.10.003
32. Namazu, M., Zhao, J., & Dowlatabadi, H. (2018). Nudging for responsible carsharing: using behavioral economics to change transportation behavior. *Transportation*, *45*(4), 105–119. https://doi.org/10.1007/s11116-016-9727-1
33. Neumann, O., Gonin, A., Pfalzgraf, M., & Patt, A. (2023). Governments can nudge household solar energy adoption: Evidence from a field experiment in Switzerland. *Energy Research & Social Science*, *105*, Article 103293. https://doi.org/10.1016/j.erss.2023.103293
34. Neumayr, L., & Moosauer, C. (2021). How to induce sales of sustainable and organic food: The case of a traffic light eco-label in online grocery shopping. *Journal of Cleaner Production*, *328*(10), Article 129584. https://doi.org/10.1016/j.jclepro.2021.129584
35. Qin, B., & Chen, H. (2022). Does the nudge effect persist? Evidence from a field experiment using social comparison message in China. *Bulletin of Economic Research*, *74*(3), 689–703. https://doi.org/10.1111/boer.12313
36. Raineau, Y., Giraud-Héraud, É., & Lecocq, S. (2025). Social comparison nudges: What actually happens when we are told what others do *Ecological Economics*, *228*, Article 108436. https://doi.org/10.1016/j.ecolecon.2024.108436
37. Randers, L., & Thøgersen, J. (2023). From attitude to identity? A field experiment on attitude activation, identity formation, and meat reduction. *Journal of Environmental Psychology*, *87*, Article 101996. https://doi.org/10.1016/j.jenvp.2023.101996
38. Shearer, L., Gatersleben, B., Morse, S., Smyth, M., & Hunt, S. (2017). A problem unstuck? Evaluating the effectiveness of sticker prompts for encouraging household food waste recycling behaviour. *Waste Management*, *60*(6), 164–172. https://doi.org/10.1016/j.wasman.2016.09.036
39. Silvi, M., & Rosa, E. P. (2021). Reversing impatience: Framing mechanisms to increase the purchase of energy-saving appliances. *Energy Economics*, *103*, Article 105563. https://doi.org/10.1016/j.eneco.2021.105563

40. Sloboda, M., Pavlovský, P., & Sičáková-Beblavá, E. (2024). Simplify and Deter: Nudging waste collection fee debtors. *Journal of Behavioral and Experimental Economics*, *111*, Article 102225. https://doi.org/10.1016/j.socec.2024.102225
41. von Zahn, M., Bauer, K., Mihale-Wilson, C., Jagow, J., Speicher, M., & Hinz, O. (2025). Smart green nudging: Reducing product returns through digital footprints and causal machine learning. *Marketing Science*, *44*(4), 954–969. https://doi.org/10.1287/mksc.2022.0393
42. Wong-Parodi, G., Krishnamurti, T., Gluck, J., & Agarwal, Y. (2019). Encouraging energy conservation at work: A field study testing social norm feedback and awareness of monitoring. *Energy Policy*, *130*, 197–205. https://doi.org/10.1016/j.enpol.2019.03.028

Search String for Web of Science:
"Nudge" OR "Nudging" (Topic) and 2025 or 2024 or 2023 or 2022 or 2021 or 2020 or 2019 or 2018 or 2017 or 2016 or 2015 or 2014 or 2013 or 2012 or 2011 or 2010 or 2009 or 2008 (Publication Years) and Article (Document Types) and English (Languages) and Business Economics or Environmental Sciences Ecology or Psychology or Government Law or Public Environmental Occupational Health or Social Sciences Other Topics or Public Administration or Nutrition Dietetics or Computer Science or Health Care Sciences Services or Education Educational Research or Food Science Technology or Behavioral Sciences or Sociology or Philosophy or Communication or Biomedical Social Sciences or Social Issues or Transportation (Research Areas)

Search String for Scopus:
TITLE-ABS-KEY ("nudge" OR "nudging") AND PUBYEAR > 2007 AND PUBYEAR < 2026 AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (EXCLUDE (SUBJAREA , "EART") OR EXCLUDE (SUBJAREA , "ARTS") OR EXCLUDE (SUBJAREA , "ENGI") OR EXCLUDE (SUBJAREA , "AGRI") OR EXCLUDE (SUBJAREA , "MATH") OR EXCLUDE (SUBJAREA , "PHYS") OR EXCLUDE (SUBJAREA , "BIOC") OR EXCLUDE (SUBJAREA , "MATE") OR EXCLUDE (SUBJAREA , "CHEM") OR EXCLUDE (SUBJAREA , "IMMU") OR EXCLUDE (SUBJAREA , "CENG") OR EXCLUDE (SUBJAREA , "VETE") OR EXCLUDE (SUBJAREA , "PHAR") OR EXCLUDE (SUBJAREA , "DENT"))

Table 2.3: Codebook Meta-Analysis

Variable	Description	Coding
Authors	Authors of paper	
Title	Title of paper	
Year	Year of publication	
Country	Country in which study was conducted	
Intervention category	Intervention category based on taxonomy by Münscher et al. (2016)	- Information = decision information - Assistance = decision assistance - Structure = decision structure
Intervention technique	Intervention technique based on taxonomy by Münscher et al. (2016)	- Translation = translate information - Visibility = make information visible - Social reference = provide social reference point - Default = change choice default - Effort = change option-related effort - Composition = change range or composition of options - Consequence = change option consequences - Reminder = provide reminders - Commitment = facilitate commitment
Permanent intervention	Nudge is continuously present throughout the intervention period (e.g., a default change embedded in the choice architecture) or temporary, appearing only once or for a limited time (e.g., feedback provided via letter)	- Permanent - Temporary
Type experiment	Type of experiment as defined by Harrison and List (2004)	- Conventional lab = Conventional lab experiment

		- Artefactual field = artefactual field experiment - Framed field = framed field experiment - Natural field = natural field experiment
Population	Target population of intervention	- General population - Students - Working population - Households - Consumers
N study	Overall sample size of study	
N comparison	Sample size of control + intervention condition	
N control	Sample size of control group	
N intervention	Sample size of intervention condition	
Binary outcome	Scale of outcome variable	- continuous - binary
Mean control	Mean of outcome variable in control condition	
SD control	Standard deviation of outcome variable in control condition	
Mean intervention	Mean of outcome variable in intervention condition	
SD intervention	Standard deviation of outcome variable in intervention condition	
Cohens d	Extracted effect size of intervention (Cohen's d)	
Variance d	Variance around extracted effect size of intervention	
Longitudinal	Design of the experiment	- Cross-sectional data - Longitudinal data
Duration	Duration of the experiment in days	
Ranking	H Index Journal ranking	
Citations	Number of citations in database	

Table 2.4: Taxonomy of Münscher et al. (2016)

Category	Techniques
Decision information	- Translate information (e.g., reframe, simplify) - Make information visible (e.g., feedback, labeling) - Provide social reference point (e.g., norms, peer comparison)
Decision structure	- Change choice defaults - Change option-related effort - Change range/composition of options - Change option consequences (e.g., highlight risks or rewards)
Decision assistance	- Provide reminders - Facilitate commitment (e.g., allow public pledges or pre-commitment)

Table 2.5: Taxonomy of Harrison & List (2004)

Type	Definition	Example
Conventional lab	Standard student pool, abstract tasks, controlled setting	University lab with economic games
Artefactual field	Field (non-student) subjects, but still in lab-like conditions	Farmers doing simplified tasks
Framed field	Real-world subjects + realistic context (task, good, or environment)	Real consumers choosing real products
Natural field	Subjects are unaware they are in an experiment, setting and behavior are completely natural	Testing nudges in a grocery store

REFERENCES

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037. https://doi.org/10.3386/w18492

Altmann, S., & Traxler, C. (2014). Nudges at the dentist. *European Economic Review*, *72*, 19–38. https://doi.org/10.1016/j.euroecorev.2014.07.007

Antinyan, A., & Asatryan, Z. (2024). Nudging for tax compliance: A meta-analysis. *The Economic Journal*, 135(668), 1033–1068. https://doi.org/10.1093/ej/ueae088

Arno, A., & Thomas, S. (2016). The efficacy of nudge theory strategies in influencing adult dietary behaviour: A systematic review and meta-analysis. *BMC Public Health, 16*, 1–11. https://doi.org/10.1186/s12889-016-3272-x

Beshears, J., & Kosowsky, H. (2020). Nudging: Progress to date and future directions. *Organizational Behavior and Human Decision Processes*, *161*, 3–19. https://doi.org/10.1016/j.obhdp.2020.09.001

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley & Sons.

Broers, V. J., De Breucker, C., Van den Broucke, S., & Luminet, O. (2017). A systematic review and meta-analysis of the effectiveness of nudging to increase fruit and vegetable choice. *The European Journal of Public Health, 27*(5), 912–920. https://doi.org/10.1093/eurpub/ckx085

vom Brocke, J. V., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. In Newell, S., Whitley, E. A., Pouloudi, N., Wareham, J., & Mathiassen, L. (Eds.), *Proceedings of the 17th European Conference on Information Systems* (pp. 2603–2614). Association for Information Systems.

Cadario, R., & Chandon, P. (2020). Which healthy eating nudges work best? A meta-analysis of field experiments. *Marketing Science, 39*(3), 465–486. https://doi.org/10.1287/mksc.2018.1128

Chakravarty, S., & Mishra, R. (2019). Using social norms to reduce paper waste: Results from a field experiment in the Indian Information Technology sector. *Ecological Economics*, *164*, 106356. https://doi.org/10.1016/j.ecolecon.2019.106356

Congiu, L., & Moscati, I. (2022). A review of nudges: Definitions, justifications, effectiveness. *Journal of Economic Surveys*, *36*(1), 188–213. https://doi.org/10.1111/joes.12453

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ECTA18709

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009–1055.

Hu, B., Xia, Z., Guo, Q., Lu, C., Constantino, S. M., & Ju, X. (2025). Assessing Nudge Impact: A Comprehensive Second-Order Meta-Analysis. *Journal of Behavioral Decision Making*, *38*(5), Article e70053. https://doi.org/10.1002/bdm.70053

Hubble, C., & Varazzani, C. (2023, May 10). *Mapping the global behavioural insights community*. OECD. https://oecd-opsi.org/blog/mapping-behavioural-insights/

Hummel, D., & Maedche, A. (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, *80(1)*, 47–58. https://doi.org/10.1016/j.socec.2019.03.005

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy, 3*(2), 159–186. https://doi.org/10.1017/bpp.2018.43

Melnyk, V., Carrillat, F. A., & Melnyk, V. (2022). The influence of social norms on consumer behavior: A meta-analysis. *Journal of Marketing, 86*(3), 98–120. https://doi.org/10.1177/00222429211029199

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences, 119*(1), Article e2107346118. https://doi.org/10.1073/pnas.2107346118

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Münscher, R., Vetter, M., & Scheuerle, T. (2016). A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making*, *29*(5), 511–524. https://doi.org/10.1002/bdm.1897

Nisa, C. F., Bélanger, J. J., Schumpe, B. M., & Faller, D. G. (2019). Meta-analysis of randomised controlled trials testing behavioural interventions to promote household

action on climate change. *Nature Communications*, *10*(1), Article 4545. https://doi.org/10.1038/s41467-019-12457-2

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*(71). https://doi.org/10.1136/bmj.n71

Polman, E., & Maglio, S. J. (2024). Nudges increase choosing but decrease consuming: Longitudinal studies of the decoy, default, and compromise effects. *Journal of Consumer Research, 51*(3), 542–551. https://doi.org/10.1093/jcr/ucad081

Pickering, C., & Byrne, J. (2014). The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. H*igher Education Research & Development, 33*(3), 534–548. https://doi.org/10.1080/07294360.2013.841651

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Van Rookhuijzen, M., De Vet, E., & Adriaanse, M. A. (2021). The effects of nudges: One-shot only? Exploring the temporal spillover effects of a default nudge. Frontiers in Psychology, 12, Article 683262. https://psycnet.apa.org/doi/10.3389/fpsyg.2021.683262

Venema, T. A., Kroese, F. M., & De Ridder, D. T. (2018). I'm still standing: A longitudinal study on the effect of a default nudge. *Psychology & Health, 33*(5), 669-681. https://doi.org/10.1080/08870446.2017.1385786

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Visser, M., & Schoormans, J. (2023). Get rid of the eco-button! Design interventions to steer sustainable use of washing machines. *Cleaner and Responsible Consumption*, *8*, Article 100096. https://doi.org/10.1016/j.clrc.2022.100096

CHAPTER 3 | THE PERSISTENCE OF DEFAULT EFFECTS: EVIDENCE FROM CO_2 OFFSETTING IN CARGO TRANSPORTATION

Chapter 3:
The Persistence of Default Effects: Evidence from CO_2 Offsetting in Cargo Transportation

This paper was co-authored with Kirsten Thommes. It was presented at the 2022 World Economic Science Association Conference, held at the MIT Sloan School of Management in Boston, USA, from June 13–16, 2022. The paper was also awarded the Best Paper Award at the faculty research workshop of Paderborn University, which took place from September 26–28, 2022, in Melle, Germany. The article has been published in *Transportation Research Part A: Policy and Practice* and is available online at: https://doi.org/10.1016/j.tra.2025.104838

Abstract

Research has shown that default settings can substantially influence pro-environmental behavior. While prior studies have largely focused on individual, one-time decisions in private households, little is known about the persistence of such effects in professional contexts. This paper examines how default settings affect business procurement decisions in cargo transportation and how these effects evolve over time. We analyze field data from more than 200,000 transportation bookings over a 15-month period on an online platform where customers could choose to offset their CO_2 emissions. The platform changed its default option from opt-in to opt-out for CO_2 offsetting. Following this change, offsetting rates increased sharply from about 12% to about 85% of all bookings. After roughly 12 weeks, the effect stabilized at around 50% and remained steady thereafter. Additional analyses show that CO_2 offsetting was not related to the sender's regional environmental consciousness, local emission levels, or inattentive purchasing behavior (e.g., nighttime bookings), but was instead associated with the perceived pro-environmental orientation of recipients. Moreover, the default change did not reduce the cargo firm's revenues. These findings indicate that implementing opt-out defaults can effectively foster lasting behavioral change even among professional decision-makers.

3.1 INTRODUCTION

Nudging—defined as an intervention that influences behavior without restricting options or substantially altering economic incentives (Thaler & Sunstein, 2008)—has been proposed as a promising tool for promoting pro-environmental behavior (Carlsson et al., 2021). It preserves freedom of choice while guiding individuals and organizations toward decisions aligned with policymakers' intentions (Thaler & Sunstein, 2003). However, the effectiveness and persistence of nudging remain subjects of debate. While a meta-analysis by Mertens et al. (2022) finds small to medium positive effects of nudges across domains such as health, food, and the environment, subsequent re-analyses suggest that these effects may be overstated (Maier et al., 2022; Szaszi et al., 2022). Moreover, DellaVigna and Linos (2022) caution that experimental estimates may not fully translate to real-world settings, where long-term and professional decision-making dynamics differ from those of student or lay samples (Beshears & Kosowsky, 2020; Löfgren et al., 2012).

This paper investigates whether nudges can produce persistent behavioral effects in a professional context—specifically, in the transportation sector. Firms in this industry operate under tight cost constraints, leaving limited room for sustainability considerations, even though the transport sector accounts for roughly 23% of global emissions (Jaramillo et al., 2022). Carbon offsetting provides one of the few short-term mechanisms to mitigate emissions, allowing organizations to compensate for their carbon footprint by investing in sustainable projects (Blasch & Farsi, 2014). Yet, recent studies have questioned the effectiveness of common offsetting mechanisms, highlighting issues such as additionality and behavioral rebound effects (Calel et al., 2025; Günther et al., 2020; Probst et al. 2023). Despite these concerns, offsetting remains a widely used instrument in sectors where direct emission reductions are difficult to achieve.

We examine whether professional customers choose to offset emissions and how default settings affect this decision. Using field data from a large European freight forwarding company operating one of the continent's major logistics platforms, we analyze over 200,000 cargo transportation bookings made between February 2021 and April 2022. During this period, the platform changed its CO_2 offsetting default from opt-in to opt-out. Customers—mainly small and medium-sized enterprises—could decide whether to offset emissions associated with their shipments, with the sender bearing the cost. Our data allow us to observe both pre- and post-intervention behavior, as well as repeated decisions by 3,128 customers who placed orders under both conditions.

Our results reveal a strong and persistent effect of default settings on offsetting behavior. When the default changed from opt-in to opt-out, the share of offset bookings increased from about 12% to about 85%. After approximately 12 weeks, the effect stabilized at around 50% and remained constant thereafter. We find similar behavioral patterns among one-time and frequent customers, suggesting that professional experience does not eliminate default effects. The timing of orders—such as night versus daytime—does not influence offsetting decisions, indicating that inattentiveness is not driving the effect. However, regional heterogeneity exists: Senders appear to consider the presumed environmental attitudes of recipients, with shipments to regions characterized by stronger Green party support exhibiting higher offset rates. Likewise, sending to more polluting regions results in a higher likelihood of offsetting. In contrast, the sender's own regional environmental consciousness does not significantly predict offsetting behavior. Importantly, the firm did not experience any loss in revenue following the default change.

Our findings contribute to the literature by demonstrating that opt-out defaults can induce substantial and persistent pro-environmental behavior even among professional decision-makers in cost-sensitive markets. This challenges the assumption that business contexts are

immune to behavioral interventions and suggests that carefully designed defaults can serve as an effective policy tool for promoting sustainable practices in the transport sector and beyond.

3.2 RELATED LITERATURE

3.2.1 Defaults as nudging interventions

One of the most widely used nudging techniques is the implementation of default settings, in which individuals are presented with a pre-selected option and must actively opt for an alternative if desired (Sunstein, 2014). Changing defaults preserves individual freedom of choice while influencing behavior without relying on regulation or monetary incentives (Sunstein & Reisch, 2014). Defaults are typically easy to implement, low-cost, and pose minimal ethical concerns.

The choice architect—the entity designing the decision context—determines how options are presented (Thaler et al. 2013). In many cases, one option is preselected, and the decision-maker can either remain with it or actively switch. Such dichotomous designs often produce striking effects: opt-out systems tend to yield very high participation rates, whereas opt-in systems result in substantially lower uptake (e.g., Johnson & Goldstein, 2003). Defaults have proven powerful in diverse domains, including retirement savings (Madrian & Shea, 2001), organ donation (Abadie & Gay, 2006; Johnson & Goldstein, 2003), and green energy consumption (Ebeling & Lotz, 2015; Kaiser et al., 2020; Pichert & Katsikopoulos, 2008).

Several mechanisms explain the default effect, defined as the behavioral difference between opt-in and opt-out conditions (Jachimowicz et al., 2019). First, defaults may be perceived as implicit recommendations from a knowledgeable authority or expert (McKenzie et al., 2006), leading individuals to accept them with minimal cognitive effort—especially when decisions are complex or uncertain (Serra-Garcia & Szech, 2023). Second, defaults act as reference points, and individuals may stay with them to avoid potential losses associated with switching (Blumenstock et al., 2018; Dinner et al., 2011). This reflects loss aversion, whereby the pain

of losses outweighs the pleasure of equivalent gains (Fryer et al., 2012; Kahneman et al., 1991; Tversky & Kahnemann, 1991). Third, inertia and convenience play a role: changing the default requires effort, so individuals often stick with it (Handel, 2013; Madrian & Shae, 2001). Finally, moral and social considerations may reinforce defaults. Opting out of a prosocial default—such as donating organs or offsetting carbon—can evoke guilt, as doing so conflicts with social or personal norms (Farrow et al., 2017; Hedlin & Sunstein, 2016).

In the context of carbon offsetting, Rodemeier (2023) finds that individuals are generally willing to offset emissions, yet they are highly price-elastic. Consequently, increasing costs through an opt-in design may suppress participation. By contrast, inertia and moral costs associated with opting out of a default offset option could increase offsetting rates even when prices are higher.

3.2.2 Defaults in carbon offsetting

Defaults have been extensively examined in carbon offsetting research, but most studies focus on short-term effects in single decision contexts.

For example, Araña and Léon (2013) conducted a field experiment with 1,680 conference attendees in Gran Canaria (2009–2011) and found that participants were significantly more likely to accept a carbon offset when it was included by default (opt-out) compared to when they were asked to opt-in. Similarly, Löfgren et al. (2012) investigated 240 environmental economists at the 2008 EAERE conference and found no significant default effect, suggesting that expert decision-makers are less susceptible to nudges when they possess strong domain knowledge.

Other studies have tested active choice designs, where no option is pre-selected. In a large-scale field experiment on long-distance bus ticket purchases, Kesternich et al. (2019) show that travelers who had to actively choose whether to offset CO_2 emissions were almost 50% more likely to participate than those who could ignore the offer. However, among repeat customers,

offsetting rates declined slightly across multiple bookings (from 26.5% to 21%), indicating diminishing effects over time even when choices are repeated within short intervals.

A more recent study by Berger et al. (2022) extends the concept of defaults beyond binary choices. Using a movable slider to select between fast (costly) and slow (cheap) carbon offsetting options, the authors analyze over 30,000 flight bookings and find that 43% of customers adhered exactly to the default setting. Price differences influenced behavior: when the default was only moderately more expensive (less than €50 difference), about half of the participants stuck with it; when the price gap widened, adherence decreased.

Taken together, these findings confirm that defaults can significantly increase participation in carbon offsetting programs. However, the magnitude and persistence of these effects appear sensitive to cost, decision context, and user expertise. Moreover, most evidence concerns private consumers and single-shot decisions, leaving the long-term effects in professional or repeated-choice contexts largely unexplored.

3.2.3 Long-term effects of defaults

Although a vast body of evidence supports the short-term effectiveness of nudges, including defaults (Beshears & Kosowsky, 2020; Hummel & Maedche, 2019; Jachimowicz et al., 2019; Mertens et al., 2022), the durability of these effects remains uncertain. Beshears and Kosowsky (2020) highlight that relatively few studies have examined whether nudges maintain their impact over time.

Evidence from other behavioral interventions suggests that nudge effectiveness can fade. For instance, Allcott and Rogers (2014) studied the Opower program, which provided households with social comparison energy reports. When the intervention ended after two years, its effect declined by 10–20% per year, indicating partial but not complete persistence.

In the context of default settings, three notable studies address long-term effects. First, Liebe et al. (2021) examined default changes in green energy procurement among over 200,000 Swiss households and 8,000 businesses. When the default was switched from conventional to renewable energy, adoption rates jumped from 3% to over 80%, and remained remarkably stable over six years (80% among households, 71% among businesses). This suggests strong and lasting default effects even in repeated annual decisions, though the low decision frequency limits external validity for high-frequency contexts. Second, Egebark and Ekström (2016) investigated a default change in printer settings (simplex vs. duplex printing) at a Swedish university. Paper consumption fell by 15%, and the effect persisted for at least six months. However, since users faced no personal cost and often made unconscious decisions, the persistence may largely reflect inertia rather than deliberate behavioral change. Third, He et al. (2023) studied behavioral interventions to reduce single-use cutlery on a Chinese food delivery platform. Setting "no cutlery" as the default and offering green reward points increased no-cutlery orders by 20 percentage points (a 648% rise) without harming platform revenue. Effects persisted across a one-year period and were strongest among frequent users, women, and higher-income customers. Still, as the study involved B2C consumers making cost-free decisions, generalizing to B2B contexts where monetary trade-offs exist remains challenging.

In summary, the literature demonstrates that defaults can effectively influence behavior—including in pro-environmental domains such as carbon offsetting and green energy. Yet, major research gaps remain:

- The long-term persistence of default effects in repeated decision-making contexts is not well understood.
- It is unclear whether experienced or professional decision-makers respond differently than consumers.

- The role of costs and conscious versus unconscious decision-making mechanisms remains ambiguous.

Addressing these gaps is essential for understanding when and why defaults lead to persistent behavioral change and how they can inform effective environmental policy design. Accordingly, we propose the following hypotheses:

- *H1: When the default option for carbon-neutral bookings is set to opt-out, customers are more likely to select carbon-neutral options than when the default is set to opt-in.*
- *H2: The effectiveness of the opt-out default in promoting carbon-neutral bookings decreases over time.*

We further explore whether these temporal changes differ across customer types—such as frequent versus infrequent users—given that decision frequency may influence the strength and persistence of nudges. Finally, we examine whether carbon offsetting represents a conscious behavioral adaptation or rather reflects inertia-driven default compliance.

3.3 RESEARCH DESIGN

For our analysis, we partnered with a freight forwarding company in Europe that offers transportation services. The company operates a digital platform that sells transportation capacity in heavy goods vehicles across Europe. Customers are mainly small- to medium-sized companies that do not manage their own logistics fleets and ship smaller volumes of goods to their customers. Customers can choose whether to offset the generated CO_2-emissions when ordering space in a truck via www.myclimate.org. Myclimate.org uses a calculator to estimate the CO_2-emissions associated with each shipment. They recalculate how many trees need to be planted to offset the CO_2-emissions of a shipment and determine a corresponding price. The price is a direct function of the cargo's distance, volume, and weight, and paid by the seller (sender), even though the costs will eventually be shifted to the buyer (receiver).

We observe a total of 201,343 shipments from 47,533 customers. Of these customers, 44,405 (93.42%) placed orders only once or singularly before or after the intervention. These customers account for 81,313 of the bookings (40.39%). On the other hand, 3,128 customers (6.58%) placed orders before and after the intervention. They account for 120,030 bookings (59.61%). For these customers who placed orders both before and after the intervention, the mean number of bookings is 38, and the median is 11. Our data set covers a period of 65 weeks, from February 1st, 2021, to April 29th, 2022. On June 4th, 2021, the company changed its default CO_2 offsetting policy from opt-in to opt-out. Thus, we observed 18 weeks with opt-in mode and 47 weeks with opt-out mode. For each shipment, we observe the exact geographic locations of both sender and receiver, as well as the precise timestamp of each booking.

Figure 3.1: Screen design

Figure 3.1 illustrates the booking interface used in this study. In the post-intervention phase, the "climate neutral"-button ("Klimaneutral") for climate-compensated bookings is pre-selected for customers. A corresponding price for the offsetting fee is also calculated, displayed on the right of the button, and highlighted in green. When a booking is compensated, a green leaf symbol appears next to the product type. Conversely, in the pre-intervention phase, users had to manually click the green button to offset manually, as no climate compensation was pre-selected. Upon clicking, the corresponding price would appear on the screen, and a green leaf

would appear next to the product type. Apart from the change in default, the screens looked identical before and after the intervention.

Concerning product types (right side of the screen), customers' options differed only in terms of other options, such as delivery time. In addition to regular cargo transportation, fast delivery ("Express"), transportation on a preferred date ("Wunschtermin"), or a direct route between sender and receiver ("Direktfahrt") could be selected. Prices for these options differed but did not impact the climate compensation fee. CO_2-emission and their calculation were independent of the delivery date or speed.

Our primary research objective is to investigate the impact of changing the default CO_2 offsetting policy from opt-in to opt-out on customers' likelihood of choosing climate-neutral bookings. Our partner company ensured that all data in our analysis was anonymized. The freight forwarding company provided the data with the understanding that the analysis would be used solely for research purposes, and we signed a non-disclosure agreement. The university's ethical committee approved our research and our data protection strategy. Furthermore, there were no conflicts of interest between the researchers and the company.

Table 3.1: Descriptive statistics

Variable	Mean	Std. Dev.	Min	Max
Nudge (0/1)	0.730	0.443	0.000	1
Price (€)	92.70	74.65	0.00	2,354.36
Compensation CO_2 (€)	0.46	1.235	0.00	74.45
Climate neutral (0/1)	0.435	0.495	0.000	1
Distance (km)	392.996	192.400	0.000	1,155.887

Summary statistics for N = 201,343 observations

Table 3.1 provides descriptive data on the most relevant variables. The transportation price (in €) depends on the type of cargo in terms of volume, weight, and distance, and varies between €0.00 and €2,354.36. The fees charged for CO_2 compensation (in €) range from €0.00 to €74.45. Before the change in the default setting, the average offsetting fee per booking was €0.04. After the intervention, this average rose to €0.61, not because individual offset prices increased, but because a substantially larger share of customers chose the offsetting option.

We calculated the distance between sender and receiver postal codes using the Google Maps API, based on the fastest highway route in Germany. When both postal codes were identical, the distance was set to zero, which occurred in 132 bookings. Across all observations, the average shipping distance was 393 km, with a maximum distance of 1,156 km.

For the main analysis examining the intervention effect of opt-in versus opt-out defaults (hypothesis 1, see section 3.4.2) and the persistence of this effect over time (hypothesis 2, see section 3.4.2), we focus on the regular customer sample. To evaluate the impact of the intervention on customer behavior, we apply a difference-in-differences (DiD) approach, comparing the share of climate-neutral bookings before and after the policy change of customers who have ordered before and after the default change. We estimate the likelihood of selecting a climate-neutral booking using probit regression models, controlling for transportation price, distance, and cargo characteristics.

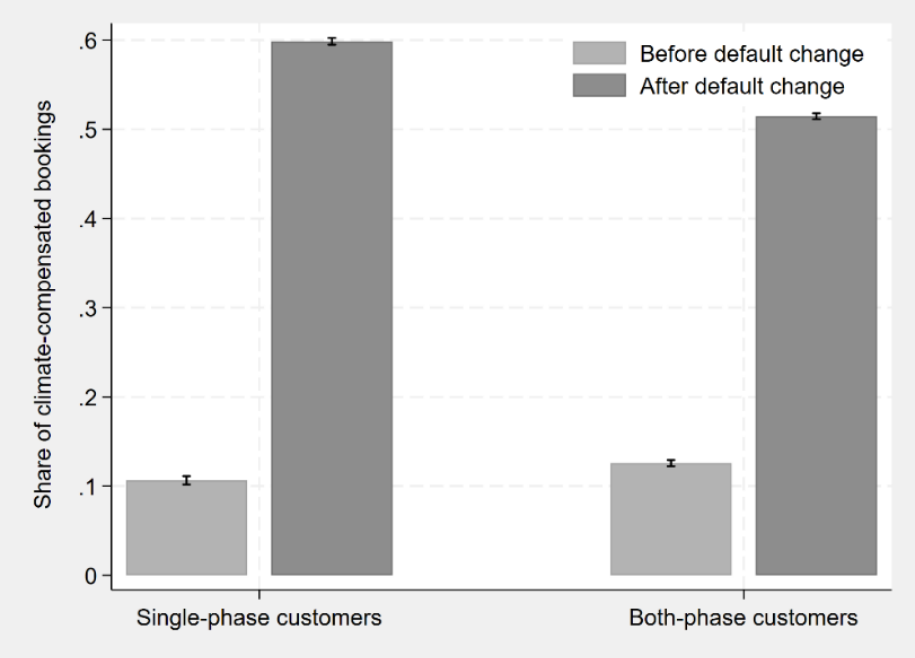
In section 3.4.3, we explore heterogeneity in treatment effects by comparing regular customers with one-time customers, as well as by differentiating regular customers based on their purchasing frequency. To better understand the drivers of behavioral change, we assess the predictive power of several contextual variables: regional traffic-related emissions (as a proxy for regional guilt), regional vote shares for the Green Party (as a proxy for environmental consciousness), and the time of booking (daytime vs. nighttime) as an indicator of decision deliberateness. The results of these analyses are presented in section 3.4.4.

3.4 RESULTS

3.4.1 Descriptive results

After analyzing all 201,343 bookings on a booking platform for cargo transport, we observe a clear trend in user behavior toward carbon offsetting. After the change from an opt-in design, where users have to actively choose to offset their carbon emissions, to an opt-out design, where carbon offsetting is the default option, we observe a significant increase in the proportion of bookings that are offset (from 11.99% to 55.17%). This highlights the impact of default settings on user behavior and the potential for nudging users toward more sustainable choices. We find a similar pattern when categorizing customers into groups based on their bookings, as illustrated in Figure 3.2. Our analysis reveals a significant offset increase after changing to an opt-out design. Single-phase users, defined as customers who placed orders before or after the intervention, experienced an increase in the proportion of offsetting from 10.65% to 59.86% ($p < 0.001$, chi-square test). For both-phase users, defined as customers who placed orders before and after the intervention, we observe that the proportion of offsetting increases from 12.57% to 51.47% ($p < 0.001$, chi-square test).

Figure 3.2: Share of climate-compensated bookings before and after default change



3.4.2 Difference-in-Difference analysis

To test Hypothesis 1, which states that customers are more likely to choose carbon-neutral options when the default option is set to opt-out rather than opt-in, we conduct a difference-in-differences analysis to examine the impact of the default setting change on carbon offsetting (see Table 3.2). This analysis enables us to control for potentially confounding variables, including price, distance, package type, sender region, and receiver region. We provide the full results in the Appendix (table 3.8). Regardless of the use of control variables, we notice that the effectiveness of changing default settings from opt-in to opt-out is significant ($p < 0.001$), indicating the robustness of the default effect across different contexts. This highlights the substantial impact of changing default settings on user behavior, even among both-phase users, and underscores the potential of leveraging such interventions to promote persistent decision-making.

Table 3.2: Difference-in-differences results

Control	Treatment	DID	p-value	Controls
0.142	0.583	0.441	0.0000	No
0.099	0.543	0.444	0.0000	Yes

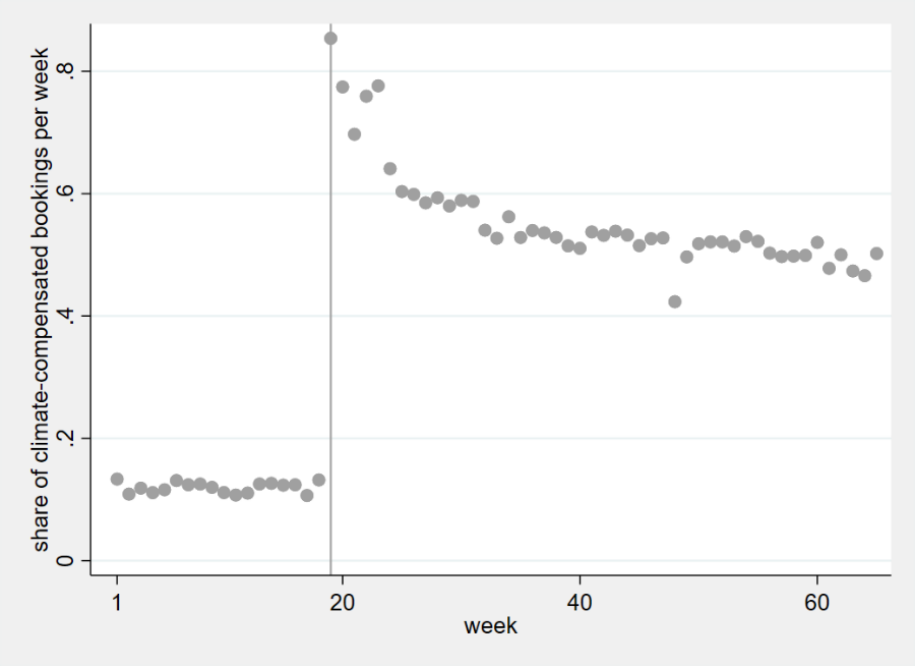
N = 120,030 orders and 3,128 customers. Control variables include price, distance, package type, sender region, and receiver region.

To test hypothesis 2, the declining effect over time, we rely on a visual interpretation and the estimate of the combined probability of climate offsetting after intervention. First, we display the probability per week. Before the intervention, the percentage of climate offsetting among platform users remained relatively stable at around 12% over time combined for both customer types. After the intervention, the combined share of offsetting orders jumps to 85% and plateaus at about 50%. This trend is depicted in figure 3.3.

It is worth noting that week 48 of the observation period represents an outlier. This week coincided with the final week of 2021 and the holiday season between Christmas and New Year's Eve. During this period, the platform experienced a marked decline in orders compared

to the surrounding weeks, which may have temporarily affected the share of climate-compensated shipments. Nevertheless, this single deviation does not materially alter the overall upward trend in climate offsetting observed in the data.

Figure 3.3: Share of climate-compensated bookings by week



To complement the descriptive findings, we estimate two probit models to formally test whether the likelihood of selecting a climate-compensated shipment declines over time following the default change. Table 3.3 reports the estimation results. Across both customer groups, the coefficients on the week variable and its squared term are highly significant and opposite in sign, indicating a non-linear time effect. The negative linear and positive quadratic components suggest an initial decline in offsetting behavior that later stabilizes and partially rebounds.

Price exhibits a negative and statistically significant coefficient, implying that higher prices slightly reduce the probability of climate offsetting. Distance has a positive and significant effect, meaning that shipments over longer distances are more likely to be compensated. The package type variable is significant only for both-phase customers, suggesting some

heterogeneity in preferences across segments. Other controls, such as sender and receiver regions, show no statistically significant effect on offsetting decisions after the default change.

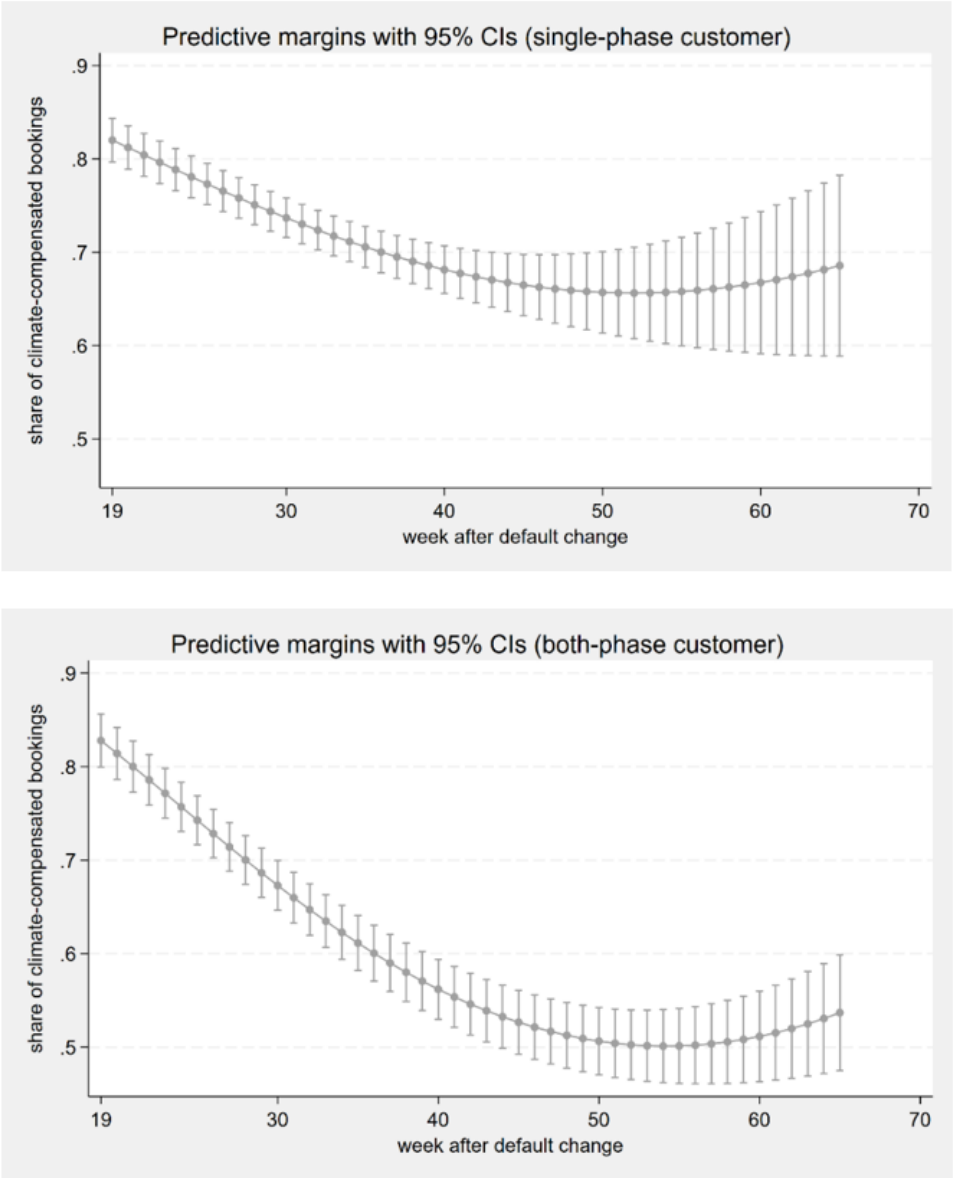
Table 3.3: Probit regressions of climate-compensated bookings after default change by customer

	Single-phase		Both-phase	
Variable	**Coef.**	**(SE)**	**Coef.**	**(SE)**
Price	-0.004***	(0.001)	-0.001***	(0.000)
Distance	0.000**	(0.000)	0.000*	(0.000)
Type	0.009	(0.006)	0.004**	(0.002)
Region sender	0.004	(0.008)	0.001	(0.002)
Region receiver	0.002	(0.003)	0.001	(0.001)
Week	-0.118**	(0.047)	-0.083***	(0.009)
$Week^2$	0.001**	(0.001)	0.001***	(0.000)
Constant	4,129**	(1.381)	2.264***	(0.183)
Observations	64,849		82,166	
Groups	30,968		3,128	
Wald X^2	325.88		263,81	

Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Figure 3.4 visualizes these dynamics by depicting the predicted margins by week after the default change. The predicted probabilities decrease from roughly 0.82 to 0.66 among single-phase customers and from 0.83 to 0.50 among both-phase customers, before slightly increasing again after week 50. This U-shaped trajectory supports Hypothesis 2, demonstrating that the impact of the default intervention diminishes over time but remains substantially above the pre-intervention baseline. Detailed weekly margins are reported in table 3.9.

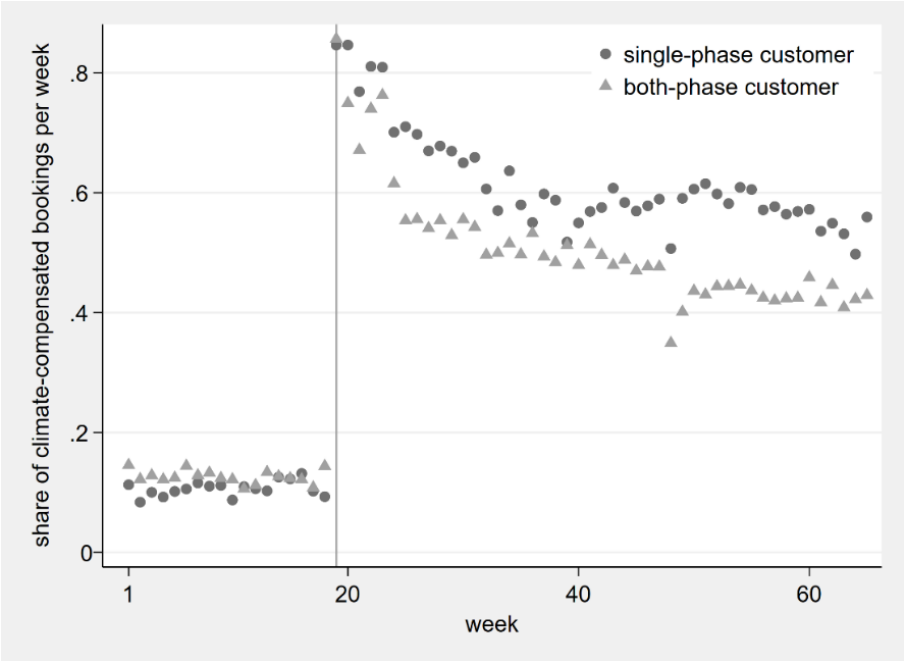
Figure 3.4: Predictive margins after default change by customer



3.4.3 Group comparison

We find that customers with greater platform experience—those active in both phases—are less likely to offset their emissions than single-phase customers (see figure 3.5).

Figure 3.5: Share of climate-compensated bookings by week and customer



To assess whether experience moderates the impact of the default change, we estimate a series of panel probit regression models examining how the opt-out intervention influenced the likelihood of choosing carbon offsetting among experienced users.

Our sample comprises 3,128 customers who completed a total of 120,030 bookings before and after the intervention. The dependent variable is the probability of choosing offsetting, while price, distance (in 1,000 km), and the nudge indicator (default change) serve as independent variables. In Model (II), we include a proxy for experience—the number of bookings made before the intervention—to capture customers' familiarity and habitual use of the platform. Model (III) adds an interaction term between the nudge and experience variables. The regression results are summarized in table 3.4.

The panel probit models offer detailed insights into how default settings shape climate-neutral booking behavior, particularly among more experienced users.

Table 3.4: Probit models

	(I)	(II)	(III)
Price	-.0031***	-.0019***	-.0019***
	(.0001)	(.0001)	(.0001)
	[-.0005]	[-.0002]	[-.0003]
Distance in 1,000km	.0052***	0.0042***	0.0039***
	(.0001)	(.0001)	(.0001)
	[.0005]	[.0005]	[.0005]
Nudge	1.0801***	2.4042***	1.0801***
	(.0560)	(.0013)	(.0560)
	[.4289]	[.2897]	[.4289]
Experience	-	-.0041***	-.0039***
		(.0005)	(.0004)
		[-.0005]	[-.0006]
Nudge*Experience	-	-	0.0048***
			(.0011)
Constant	-.6142***	-1.6023***	-1.5734***
	(.0740)	(.0012)	(.1095)
Type	Yes	Yes	Yes
Region sender	Yes	Yes	Yes
Region receiver	Yes	Yes	Yes
(N)	120,030	120,030	120,030
(N Groups)	3,128	3,128	3,128
(Wald Chi)	588.14	506.22	488.12

Robust standard error for customer-level in parentheses, marginal effects in square brackets, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In Model (I), both price and distance significantly influence offsetting behavior. The negative price coefficient indicates that higher prices reduce the likelihood of selecting a climate-neutral booking, whereas the positive distance coefficient suggests that longer transport distances

increase the probability of offsetting. As expected, the nudge variable is positive and highly significant, confirming that the default change substantially increased offsetting rates.

In Model (II), the coefficient for experience is negative and statistically significant, implying that customers who place more orders are less likely to choose offsetting. This effect may arise from greater price sensitivity or increased awareness of the opt-out option among experienced users.

Model (III) introduces the interaction between the nudge and experience variables. The positive and significant interaction term indicates that the opt-out default becomes more effective among experienced customers, counteracting the negative main effect of experience. This suggests that familiar users may better recognize the benefits or legitimacy of the offsetting option, making them more responsive to the default intervention.

Overall, the results show that the default change significantly increased the likelihood of climate-neutral bookings, even among experienced customers. These findings underscore the robustness of default effects and highlight their potential to promote sustainable decision-making in professional contexts.

To further investigate whether the default change affects customers differently depending on their experience with the platform, we analyze offsetting behavior by booking frequency. Specifically, we categorize customers who are active in both phases into four percentile groups based on their total number of bookings during the 65-week observation period:

- 1st percentile: 2-44 orders
- 2nd percentile: 45-138 orders
- 3rd percentile: 139-467 orders
- 4th percentile: more than 467 orders

Figure 3.6 illustrates the evolution of the default intervention's effect over time across these groups. The results reveal that the effectiveness of the default intervention differs by booking

frequency. While all groups exhibit a similar temporal pattern—characterized by an initial increase in offsetting rates followed by stabilization—the magnitude of the effect decreases as booking frequency increases. In other words, customers with fewer orders show a stronger and more immediate response to the default change, whereas among high-frequency users, the effect is weaker and levels off more quickly. This pattern suggests that the default intervention exerts the greatest influence during customers' early interactions with the platform, but its impact diminishes as behavior becomes more routine and habitual.

Figure 3.6: Share of climate-compensated bookings for each of the percentiles

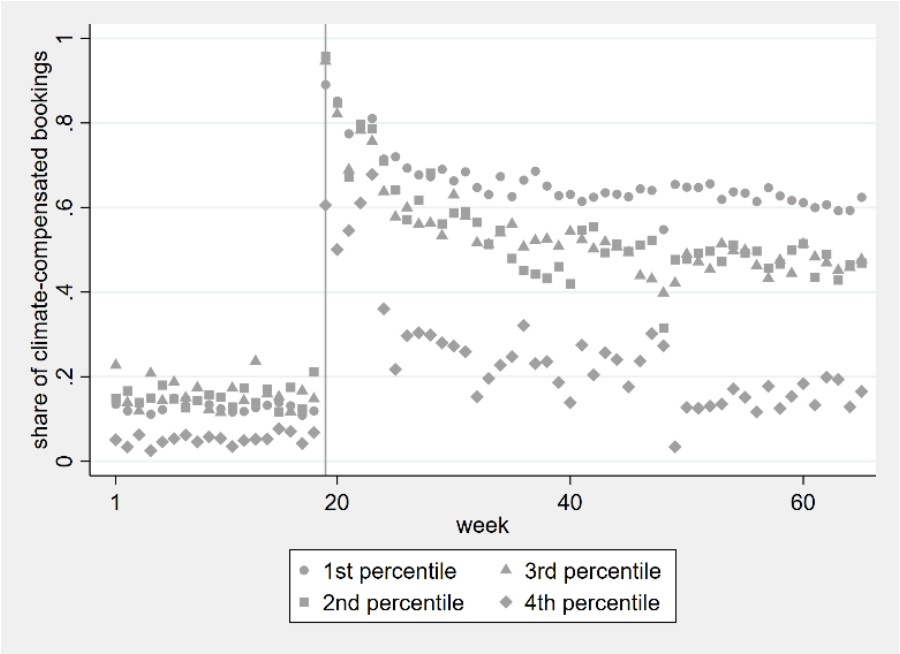


Table 3.5 displays the results of the difference-in-difference analysis, which helps us determine the disparities between the subgroups. Upon comparing the opt-in mode with the opt-out mode and controlling for variables such as price, distance, type of package, sender region, and receiver region, we observe a significant increase in offsetting across all subgroups ($p = .000$). Before the intervention, only 17.6% of bookings the first percentile were offset. In contrast, after the intervention, 74.4% of bookings were offset. In the second percentile, 12.8% of all bookings were offset before, but after the intervention, 64.2% of the bookings were offset. In

the third percentile, 12.1% of bookings were offset before, while after the intervention, 61.9% of bookings were offset. In the fourth percentile, 0% of bookings were offset before, but after the intervention, 20.0% of bookings were offset. These changes in default settings remain significant among all subgroups (p = .000). We observe similar effectiveness in changing default settings from opt-in to opt-out mode even without controlling for variables, suggesting that the default effect is robust and holds across different contexts.

Table 3.5: Diff-in-Diff for experience percentiles

	Control	Treat.	DID	p-value	Controls
25% percentile	0.144	0.706	0.562	0.0000	no
(2-44 orders per customer)	0.176	0.744	0.569	0.0000	yes
50% percentile	0.144	0.653	0.509	0.0000	no
(45-138 orders per customer)	0.128	0.642	0.514	0.0000	yes
75% percentile	0.133	0.627	0.494	0.0000	no
(139-467 orders per customer)	0.121	0.619	0.498	0.0000	yes
100% percentile	0.091	0.417	0.327	0.0000	no
(> 467 orders per customer)	-.134	0.200	0.334	0.0000	yes

Control variables include price, distance, package type, sender region, and receiver region.

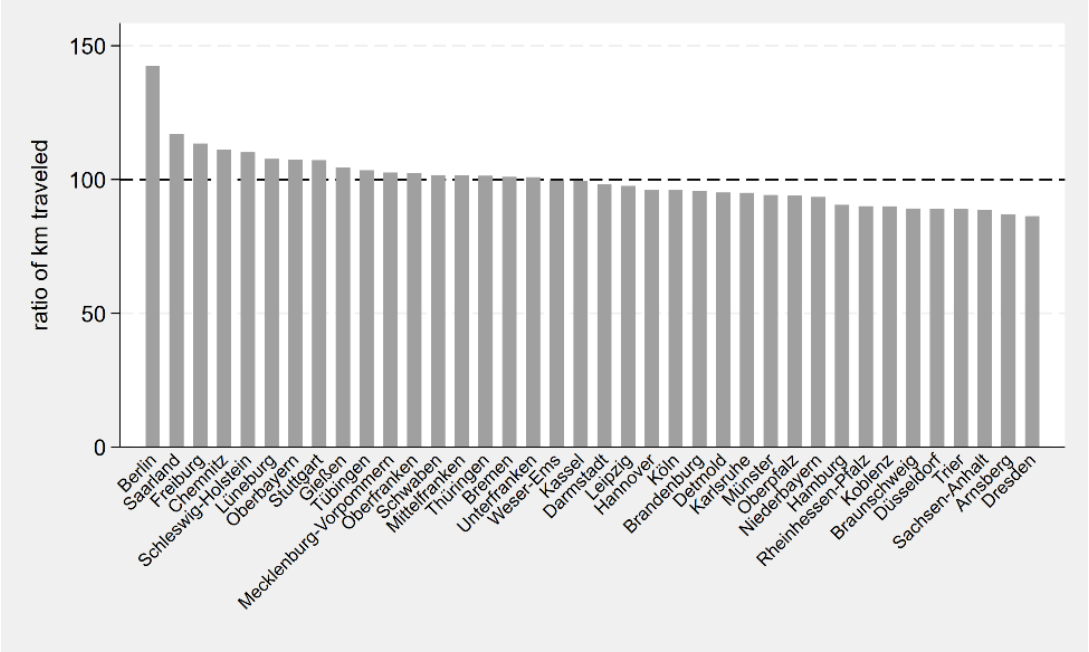
3.4.4 Potential reasons for offsetting

3.4.4.1 Offsetting by region

The likelihood of CO_2-offsetting may also depend on the sender's and the receiver's region. This is due to a significant imbalance in transportation in and out of specific areas. One of the most extreme cases in Germany is the region of Berlin: Here, the amount of goods transported to Berlin exceeds the number of goods transported out of the city by 1.5. Roughly speaking, out of three trucks fully loaded traveling to Berlin, one must leave empty as Berlin receives a lot of goods from other German places but does not produce much in return. This results in a remarkably negative ratio and unproductive CO_2-emissions as trucks have to travel back empty.

Most Germans are aware of this ratio qualitatively and the specific environmental burden that Berlin causes in Germany. One may assume that senders shipping to Berlin might be more inclined to opt for climate compensation to compensate for the higher levels of perceived guilt when shipping into this area. Similarly, those senders from Berlin, aware of the city's transportation imbalance, might also be more likely to compensate. Figure 3.7 gives an overview of the In-/Out-ratios for the different German regions.

Figure 3.7: Distribution of In/Out transportation ratio across German regions



Distribution of the In/Out transportation ratio across German regions, measured in kilometers driven. Data sourced from the Federal Ministry of Transport, 2022.

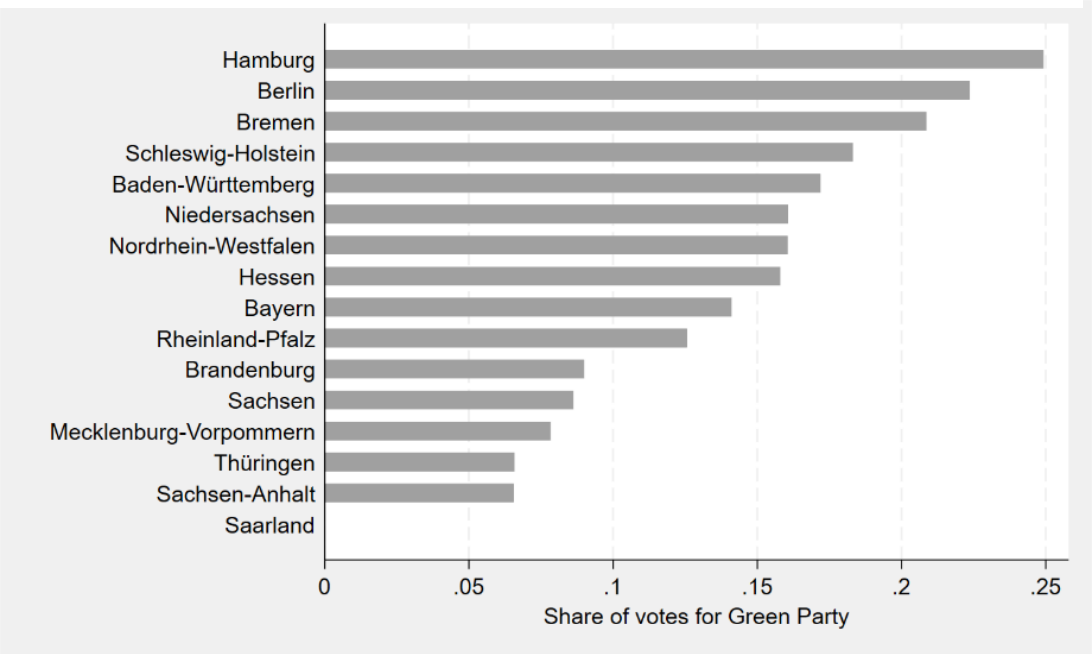
We first investigate the so-called "Berlin effect" in transportation by including binary indicators identifying whether the sender or the receiver is located in Berlin[4].

In addition to potential transportation imbalances, political preferences may also shape the likelihood of carbon offsetting. Given the Green Party's strong pro-environmental orientation, senders from regions with higher Green Party vote shares may be more inclined toward offsetting their shipments. Likewise, senders dispatching goods to regions with higher Green

[4] Table 3.7 also reports an alternative specification using the ratio of inbound-to-outbound freight flows for each German region as a continuous measure; this specification does not materially alter the results.

Party support may anticipate that their customers place greater value on environmentally responsible behavior, thereby increasing the propensity to offset. To examine these relationships, we include the Green Party's vote share in both the sender's and receiver's regions as explanatory variables. Figure 3.8 illustrates the distribution of Green Party vote shares across the 16 German federal states in the 2021 federal election. Notably, the Green Party did not appear on the ballot in Saarland, which is therefore excluded from our analysis[5].

Figure 3.8: Shares of votes for Green Party in last federal election 2021



Share of votes for Green Party in last federal election 2021. Data sourced from the federal returning officer.

Germany is divided into 16 NUTS1 regions, corresponding to 16 federal states. These 16 federal states are, in turn, subdivided into 38 regions at the NUTS2 level. Therefore, we use NUTS1 regions for the upper hierarchy level and NUTS2 regions for the lower hierarchy level to run multilevel logistic regression models. Table 3.6 presents the results for both regional indicators separately (Model I-II) and jointly (Model III).

[5] Table 3.7 additionally presents a specification using the presence of Green Party members in the respective state governments as an alternative indicator of political influence.

Table 3.6: Offsetting by region

	(I)	(II)	(III)
Berlin sender	-.1696		0.0222
	(.3082)		(.7147)
Berlin receiver	.2566***		0.1924***
	(.0305)		(.0326)
Green party votes (sender region)		-1.8214***	-2.0322
		(.8829)	(2.7868)
Green party votes (receiver region)		1.1574***	0.7901***
		(.1357)	(.1455)
Nudge	2.2683***	2.2707***	2.2627***
	(.0145)	(.0145)	(.0147)
Price	-.0032***	-.0030***	-.0032***
	(.0001)	(.0001)	(.0001)
Distance in 1,000km	0.5564***	0.5481***	0.5625***
	(.0270)	(.0266)	(.0272)
Type	0.0098***	0.0076***	0.0090***
	(.0013)	(.0013)	(.0013)
Constant	-2.0838***	-2.0106***	-1.9174***
	(.0559)	(.1406)	(.4169)
Random intercept (sender NUTS1 region)		0.0357	
		(.0136)	
Random intercept (sender NUTS2 region)	0.0919		
	(.0214)		
Random intercept (sender NUTS2 region| sender NUTS1 region)			0.4504
			(.1109)
(N)	201,343	197,639	197,639
(N Groups)	38	15	37 | 46
(Wald Chi)	25102.60	25120.25	24659.18

Standard errors in parentheses, * $p < 0.100$, ** $p < 0.050$, *** $p < 0.010$. Note that the number of observations drops compared to the previous models because we had to exclude one German region as there was some inconsistency in elections (The Green party failed to register for the election in the region of Saarland).

The first multilevel logit model in table 3.6 examines the "Berlin effect," referring to the pronounced imbalance between incoming and outgoing cargo volumes in Berlin. Since the city receives substantially more cargo than it sends, transportation to Berlin is comparatively inefficient in terms of emissions. Model I reveals a significant effect: shipments to Berlin are considerably more likely to be accompanied by carbon offsetting. In contrast, firms based in

Berlin appear less inclined to acknowledge or compensate for their own emissions through offsetting.

Model II evaluates the influence of political orientation by incorporating the share of Green Party votes in both the sending and receiving regions. The results indicate that a higher proportion of Green Party votes in the sending region is associated with a lower likelihood of CO_2 offsetting. Conversely, a higher share of Green Party votes in the receiving region corresponds to an increased tendency to offset emissions. As Green Party vote shares and the Berlin effect are uncorrelated, Model III jointly includes both variables to assess their combined impact.

We interpret these findings as suggesting that companies may employ carbon offsetting as a marketing instrument, particularly when they believe that recipients—such as customers in regions with stronger Green Party support—are more likely to value environmental responsibility. However, these associations may also reflect selection effects: eco-friendly products from firms with explicit sustainability missions could be more frequently ordered by customers in regions with higher Green Party support. Moreover, since wealthier areas tend to both vote Green and tolerate higher shipping costs, omitted variable bias cannot be ruled out.

Although we cannot disentangle these mechanisms with the available data, two robust patterns emerge: (1) senders to Berlin are more likely to offset emissions, possibly due to awareness of the city's transportation imbalance and its associated emission burden; and (2) shipments to regions with a higher share of Green Party voters are more likely to be offset, reflecting either normative expectations or strategic corporate behavior.

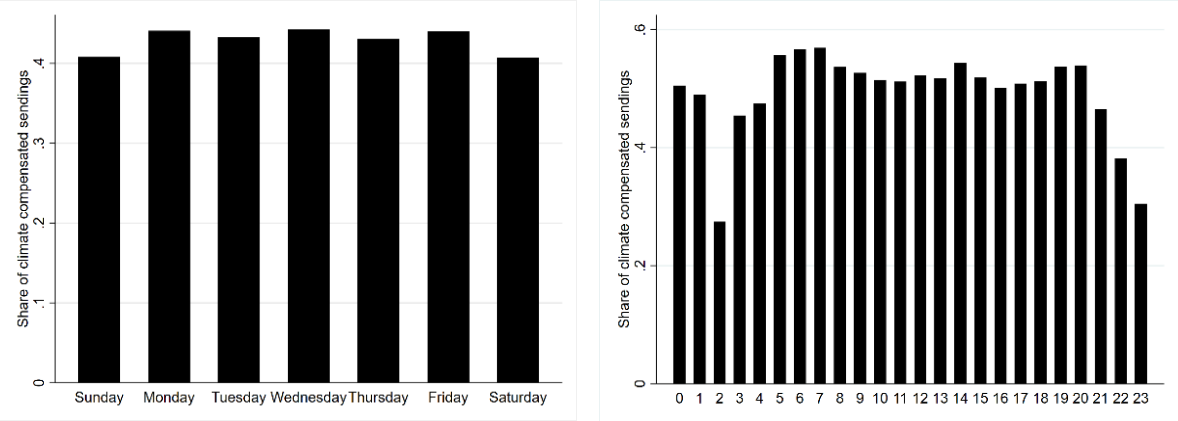
3.4.4.2 Offsetting by date and time

One might argue that customers' potential inattention during the booking process influences their likelihood of opting out of carbon offsetting. Unfortunately, we cannot fully address this question, as detailed booking timestamps are not available for the entire observation period.

The platform owner began recording booking times only after the intervention, primarily to support dynamic pricing simulations. Consequently, our analysis covers bookings made between September 2021 and April 2022, comprising a total of 113,281 observations. We present these results with caution, acknowledging that they can provide only indicative evidence and may not fully capture temporal effects.

To explore potential temporal patterns, we divided booking times into three-hour intervals throughout the day and examined whether the likelihood of choosing carbon offsetting varied by time of day. As illustrated in figure 3.9, we find no systematic variation in offsetting behavior across different time periods or days. Platform users were generally equally likely to select the offsetting option regardless of when they booked, with the exception of slightly lower activity during nighttime off-peak hours. Overall, these findings suggest that the effectiveness of the default intervention is stable across booking times, indicating that the default mechanism functions intuitively well throughout the day and night.

Figure 3.9: Share of carbon offsetting per day and hour



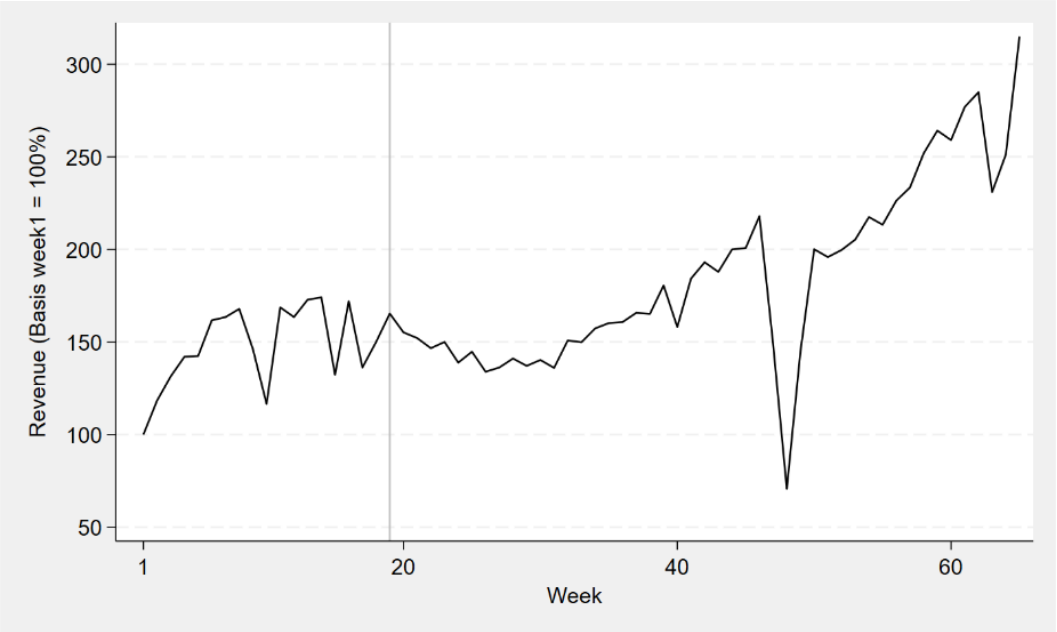
3.4.4.3 Revenue effect of offsetting

Although He et al. (2023) find that nudging does not negatively affect firm revenue, in our case—where customers are nudged to pay an additional fee for carbon offsetting—it could theoretically reduce revenue. Moreover, Rodemeier (2023) shows that private customers in

one-shot purchasing decisions are price-elastic but impact-inelastic, suggesting that while they might support offsetting in principle, their voluntary contributions would likely be zero.

To assess whether the company's carbon offsetting initiative affected revenues, we aggregated weekly revenues before and after the intervention. As shown in figure 3.10, revenues continued to increase despite the introduction of the offsetting nudge, indicating no apparent negative impact.

Figure 3.10: Revenue over time



Note that we have omitted the absolute values due to our non-disclosure agreement. We show a relative comparison instead of the first values relative to the first week of the observational period. The only week below the first week is week 48 (Christmas break).

However, this finding should be interpreted with caution. Germany experienced a COVID-19 lockdown from approximately week 1 to week 20, which strongly affected the transportation sector and small B2B businesses. Additionally, the partner firm was a relatively young startup, making revenue forecasts inherently uncertain. Therefore, we conservatively conclude that the default change did not harm business revenues.

3.5 DISCUSSION AND CONCLUSION

Our study demonstrates that changing default settings to nudge individuals toward carbon offsetting—despite higher per-unit costs—can substantially increase participation. Switching the default option from opt-in to opt-out raises the share of climate-compensated bookings from 12% to 55% across all observations. Importantly, the intervention did not harm business performance. As long as offsetting remains the most viable mitigation tool in transportation, implementing default-based policy interventions in cargo logistics appears both effective and practical for reducing transport-related emissions. Despite ongoing debates about CO_2 offsetting programs in terms of their effectiveness, the default change encouraged customers to compensate for their shipment emissions, increasing total contributions from €2,312.94 to €89,781.65 and average compensation per booking from €0.042 to €0.610. The average payment per booking rose from €0.125 before the intervention to €0.514 after. At a rate of €30 per ton of CO_2, this corresponds to an increase in compensated emissions from 77 tons to 2,992 tons of CO_2.

Our findings reinforce prior evidence that default settings are a powerful lever for promoting sustainable behavior (e.g., Ebeling & Lotz, 2015; He et al., 2023; Kaiser et al., 2020), including in the domain of CO_2 offsetting (Araña & Léon, 2013). Yet, our results also challenge existing assumptions about the limits of this effect. While earlier field experiments suggested that the influence of opt-out defaults weakens as individuals gain experience with repeated decisions (Löfgren et al., 2012), our data reveal a persistent — and even robust —impact among experienced logistics managers.

This persistence suggests that defaults may operate through mechanisms beyond mere inattention or decision fatigue. In professional contexts, where decisions are made frequently and systematically, defaults may come to represent a social or organizational norm rather than a cognitive shortcut. Over time, the default option might signal what is considered the

"appropriate" or "responsible" choice, embedding pro-environmental behavior into routine decision-making.

The contrast with earlier findings highlights the importance of context in understanding how defaults influence sustainable behavior. Factors such as professional norms, organizational accountability, and the framing of environmental responsibility may moderate the persistence of default effects. Future research should therefore explore how experience interacts with contextual and motivational factors—such as perceived legitimacy, moral framing, and institutional incentives—to determine when and why defaults maintain their influence.

Ultimately, our study suggests that in organizational settings, defaults can serve not only as behavioral nudges but also as instruments for shaping enduring sustainability norms and practices.

Our findings reveal that most platform users adhered to the default option, and this behavior persisted for at least eleven months. Although the initial proportion of compensated bookings exceeded 85% and later declined, the offsetting rate remained consistently above 50% throughout the observation period. The default effect was evident across all customer groups, including frequent users categorized by booking frequency. This indicates that the observed behavior cannot be attributed solely to occasional customers; rather, the switch from an opt-in to an opt-out default was the key driver. These results extend prior research (Egebark & Ekström, 2016; He et al., 2023; Kesternich et al., 2019) by demonstrating that default effects can persist across repeated decisions and over extended periods. They highlight the enduring influence of defaults on sustainable behavior, suggesting that such interventions can serve as powerful tools for promoting long-term environmental engagement.

However, the initial spike in compensated bookings gradually decreased, stabilizing at around 50% after twelve weeks. Several mechanisms may account for this decline. First, experienced users may have become more familiar with the platform and actively deselected the pre-set

option once they recognized the associated costs. Second, some firms may operate under procurement rules or cost-control procedures that discourage recurring optional payments. Third, part of the early increase likely stemmed from user inattention immediately after the policy change, which diminished as users became more aware of the default design.

In contrast to Liebe et al. (2021), who observed persistent effects for green electricity contracts, our results suggest that decision frequency plays a critical moderating role. Electricity contracts are revisited infrequently, whereas our participants made booking decisions repeatedly, offering more opportunities for deliberation and cost-conscious adjustment. This repeated exposure may explain the gradual decline and subsequent stabilization of the default effect over time.

To sustain higher offsetting rates, default interventions could be complemented with reinforcing behavioral strategies. Prior research shows that personalized feedback can maintain pro-environmental actions in energy conservation (Allcott & Rogers, 2014), while social-norm messages enhance commitment to sustainable choices (Farrow et al., 2017). Similarly, logistics platforms could inform users about their cumulative CO_2 savings or compare their offsetting performance with that of comparable firms. Such complementary measures may help counteract behavioral fatigue and sustain long-term engagement with carbon offsetting.

The persistence of the default effect is particularly striking given the professional context of our study. Our data come from small and medium-sized enterprises (SMEs), where decision-makers are typically more price-sensitive and might be expected to scrutinize default options more carefully. Yet, our analysis reveals that the default effect exerts a stronger influence than price considerations. While price sensitivity does vary among customers—with higher booking costs reducing the likelihood of offsetting—the overall impact of the default remains dominant. These findings align with Liebe et al. (2021), who observed a robust default effect on green energy adoption among private households and businesses in Switzerland. Together, the

evidence underscores the considerable potential of default settings to promote sustainable behavior even in cost-conscious, professional decision-making environments.

We find no evidence that negligence drives the higher likelihood of offsetting observed after the default change. Within our data, neither the time nor the date of booking correlates with the default effect, suggesting that the increase in offsetting behavior is not merely a byproduct of users' inattention during the booking process. Instead, we observe a modest experience effect: regular customers are slightly more likely to override the default compared to less experienced users, although this effect remains limited in magnitude.

These findings suggest that the persistence of the default effect reflects a genuine behavioral adjustment rather than automatic or inattentive compliance. In line with theories of bounded rationality (Simon, 1957), decision-makers may rely on defaults as heuristics that simplify complex choices under time and information constraints. Over time, repeated exposure to the booking environment likely transforms the default into a habitual anchor, reinforcing sustainable behavior even among cost-sensitive professional users. This interpretation aligns with prior research on habit formation and the durability of nudges in repeated decision settings (e.g., Allcott & Rogers, 2014; Loewenstein et al., 2016), emphasizing that defaults can remain effective even when users become more experienced and deliberate in their choices.

We also identify distinct regional patterns in our data. Shipments to areas with higher traffic volumes are more likely to be accompanied by CO_2 compensation. Moreover, decision-makers appear to consider the perceived environmental or political attitudes of the recipients. For example, deliveries to regions with a higher share of Green Party voters are more likely to include CO_2 offsetting. In contrast, the senders' own regional political leanings seem to play a less significant role in their decision-making. Finally, we find that changing the default settings did not negatively affect the company's revenues—consistent with the findings of He et al. (2023).

The use of nudges also raises important ethical considerations. According to the principles outlined by Sunstein (2015), the effectiveness of nudge interventions is inherently limited, and such measures should be complemented by traditional, incentive-based approaches within a transparent decision architecture to foster lasting behavioral change. Moreover, nudges can sometimes yield unintended negative consequences. For instance, by offering seemingly effortless solutions, they may inadvertently reduce public support for more comprehensive and effective policies, such as carbon taxation (Hagmann et al., 2019). Consequently, the ethical application of nudges requires careful reflection on their broader societal implications, a strong commitment to transparency and respect for individual autonomy, and continuous evaluation to ensure that they genuinely contribute to promoting sustainable behavior.

While our findings underscore the powerful role of default settings in promoting sustainable behavior, it is important to recognize that they are not a one-size-fits-all solution. Defaults should be viewed as one component within a broader policy toolkit to foster sustainability. To achieve more comprehensive and lasting change, default interventions should be complemented by other measures such as communication (Hoffmann & Thommes, 2024), financial incentives (Hoffmann & Thommes, 2022), and regulatory frameworks. Defaults can deliver quick and effective short-term results ("quick wins") by guiding behavior with minimal resistance. However, addressing long-term sustainability challenges requires combining these interventions with educational initiatives and structural policies that cultivate a deeper understanding of environmental issues and support persistent behavioral change. A multifaceted approach will thus be essential for driving sustained progress toward sustainability goals.

Our results also carry important policy implications for organizations and regulators in the cargo transportation and logistics sectors. The substantial rise in climate-compensated bookings following the default change demonstrates the scalability of this approach for

promoting sustainable practices across the broader transportation industry. Implementing default-based interventions across similar platforms could lead to significant reductions in CO_2-emissions and contribute meaningfully to climate targets. Moreover, our findings suggest that default settings can effectively complement other policy tools—such as carbon pricing, subsidies for green transport solutions, and stricter emission standards—by subtly reinforcing pro-environmental choices. Policymakers should therefore consider defaults as part of an integrated strategy to advance environmentally responsible behavior and accelerate the transition to a more sustainable transport system.

Nonetheless, our study has several limitations. The data stem from a single freight-forwarding company operating in Europe, which may constrain the generalizability of our results. Moreover, the analysis focuses on a specific time period and customer base, with limited information about individual decision-makers or firm-level characteristics. These constraints may limit our ability to assess longer-term effects of the default intervention or to account for broader contextual factors such as market conditions and customer preferences. Future research should examine the effectiveness of default settings across different industries and cultural contexts, as well as explore potential interactions between defaults and complementary policy measures. Such investigations would provide valuable insights into when and how defaults can most effectively contribute to persistent behavior change.

3.6 APPENDIX

Table 3.7 is an alternative specification to table 3.6 in the paper: Here, we use green party governmental participation in a federal state instead of Green votes and the metric variable ratio in- to outflow instead of the "Berlin-Effect" on NUTS2-level. The results remain the same compared to the specifications in the paper.

Table 3.7: Offsetting by region models IV-VI.

	(IV)	(V)	(VI)
Ratio km traveled sender	-.0004		0.0059
	(.0048)		(.0104)
Ration km traveled receiver	.0031***		0.0033***
	(.0005)		(.0005)
Green party government participation (sender region)		-.1456	-.2481
		(.1121)	(.2049)
Green party government participation (receiver region)		-.0341***	-.0293***
		(.0099)	(.0100)
Nudge	2.2672***	2.2711***	2.2624***
	(.0145)	(.0145)	(.0147)
Price	-.0031***	-.0030***	-.0032***
	(.0001)	(.0001)	(.0001)
Distance in 1,000km	0.5430***	0.5456***	0.5418***
	(.0273)	(.0266)	(.0275)
Type	0.0098***	0.0077***	0.0091***
	(.0013)	(.0013)	(.0013)
Constant	-2.3505***	-1.9887***	-2.8600***
	(.4841)	(.0949)	(1.0536)
Random intercept (sender NUTS1 region)		0.0409	
		(.0156)	
Random intercept (sender NUTS2 region)	0.0924		
	(.0216)		
Random intercept (sender NUTS2 region| sender NUTS1 region)			0.4329
			(.1073)
(N)	201,343	197,639	197,639
(N Groups)	38	15	37 | 46
(Wald Chi)	25091.62	25077.01	24636.13

Standard errors in parentheses, * $p < 0.100$, ** $p < 0.050$, *** $p < 0.010$.

Table 3.8: Coefficients of control variables in the DID model

Variable	Coeff.	Std. Err.	t	P>t
Price	-0.000	0.000	-25.248	0.000***
Distance	0.000	0.000	10.787	0.000***
Package type	0.005	0.000	14.520	0.000***
Region sender	-0.001	0.000	-4.566	0.000***
Region receiver	0.000	0.000	1.747	0.081*
R^2		0.20		
Observations		120,030		

Notes: Dependent variable is opting for carbon offsetting. Sample restricted to repeated users. The model includes covariates for price, distance, package type, sender region, and receiver region. Significance levels: *** $p<.01$, ** $p<.05$, * $p<.10$.

Table 3.9: Comparison of predictive margins for customer groups by week after default change

Week	Single-phase Margin (95% CI)	Both-phase Margin (95% CI)
19	0.820 (0.797, 0.844)	0.828 (0.800, 0.856)
20	0.812 (0.789, 0.835)	0.814 (0.786, 0.842)
21	0.804 (0.781, 0.827)	0.800 (0.773, 0.827)
22	0.796 (0.774, 0.819)	0.786 (0.759, 0.813)
23	0.788 (0.766, 0.811)	0.771 (0.745, 0.798)
24	0.781 (0.758, 0.803)	0.757 (0.731, 0.783)
25	0.773 (0.751, 0.795)	0.743 (0.717, 0.769)
26	0.766 (0.744, 0.787)	0.728 (0.702, 0.754)
27	0.758 (0.736, 0.780)	0.714 (0.688, 0.740)
28	0.751 (0.729, 0.772)	0.700 (0.674, 0.726)
29	0.744 (0.722, 0.765)	0.686 (0.660, 0.713)
30	0.737 (0.716, 0.758)	0.673 (0.646, 0.700)
31	0.730 (0.709, 0.751)	0.660 (0.633, 0.687)
32	0.724 (0.702, 0.745)	0.647 (0.620, 0.675)
33	0.717 (0.696, 0.739)	0.635 (0.607, 0.663)
34	0.711 (0.690, 0.733)	0.623 (0.594, 0.652)
35	0.706 (0.684, 0.728)	0.611 (0.582, 0.641)
36	0.700 (0.678, 0.723)	0.600 (0.570, 0.630)
37	0.695 (0.672, 0.718)	0.590 (0.559, 0.620)
38	0.690 (0.666, 0.714)	0.580 (0.549, 0.611)
39	0.686 (0.661, 0.710)	0.571 (0.539, 0.602)
40	0.681 (0.656, 0.707)	0.562 (0.530, 0.594)

41	0.677 (0.651, 0.704)	0.554 (0.521, 0.586)
42	0.674 (0.646, 0.702)	0.546 (0.513, 0.579)
43	0.670 (0.641, 0.699)	0.539 (0.506, 0.572)
44	0.667 (0.636, 0.699)	0.532 (0.499, 0.566)
45	0.665 (0.632, 0.698)	0.527 (0.493, 0.561)
46	0.663 (0.628, 0.697)	0.521 (0.487, 0.556)
47	0.661 (0.624, 0.697)	0.517 (0.482, 0.551)
48	0.659 (0.620, 0.698)	0.513 (0.478, 0.548)
49	0.658 (0.617, 0.699)	0.509 (0.474, 0.545)
50	0.657 (0.613, 0.701)	0.506 (0.470, 0.542)
51	0.656 (0.610, 0.703)	0.504 (0.468, 0.541)
52	0.656 (0.607, 0.705)	0.503 (0.465, 0.540)
53	0.656 (0.605, 0.708)	0.501 (0.463, 0.540)
54	0.657 (0.602, 0.712)	0.501 (0.462, 0.540)
55	0.658 (0.600, 0.716)	0.501 (0.461, 0.541)
56	0.659 (0.598, 0.721)	0.501 (0.461, 0.543)
57	0.661 (0.596, 0.726)	0.504 (0.461, 0.546)
58	0.663 (0.594, 0.731)	0.506 (0.461, 0.550)
59	0.665 (0.593, 0.737)	0.508 (0.462, 0.555)
60	0.667 (0.591, 0.744)	0.511 (0.463, 0.560)
61	0.670 (0.590, 0.751)	0.515 (0.465, 0.566)
62	0.674 (0.590, 0.758)	0.520 (0.467, 0.573)
63	0.677 (0.589, 0.766)	0.525 (0.469, 0.581)
64	0.681 (0.589, 0.774)	0.531 (0.472, 0.589)
65	0.686 (0.589, 0.782)	0.537 (0.475, 0.599)

REFERENCES

Abadie, A., & Gay, S. (2006). The impact of presumed consent legislation on cadaveric organ donation: A cross-country study. *Journal of Health Economics, 25*(4), 599–620. https://doi.org/10.1016/j.jhealeco.2006.01.003

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037. https://doi.org/10.3386/w18492

Araña, J. E., & León, C. J. (2013). Can defaults save the climate? Evidence from a field experiment on carbon offsetting programs. *Environmental and Resource Economics*, *54*(4), 613–626. https://doi.org/10.1007/s10640-012-9615-x

Berger, S., Kilchenmann, A., Lenz, O., Ockenfels, A., Schlöder, F., & Wyss, A. M. (2022). Large but diminishing effects of climate action nudges under rising costs. *Nature Human Behaviour, 6*(10), 1381–1385. https://doi.org/10.1038/s41562-022-01379-7

Beshears, J., & Kosowsky, H. (2020). Nudging: Progress to date and future directions. *Organizational Behavior and Human Decision Processes, 161*, 3–19. https://doi.org/10.1016/j.obhdp.2020.09.001

Blasch, J., & Farsi, M. (2014). Context effects and heterogeneity in voluntary carbon offsetting: A choice experiment in Switzerland. *Journal of Environmental Economics and Policy, 3*(1), 1–24. https://doi.org/10.1080/21606544.2013.842938

Blumenstock, J., Callen, M., & Ghani, T. (2018). Why do defaults affect behavior? Experimental evidence from Afghanistan. *The American Economic Review, 108*(10), 2868–2901.

Calel, R.; Colmer, J.; Dechezleprêtre, A.; Glachant, M. (2025). Do carbon offsets offset carbon? *American Economic Journal: Applied Economics*, *17*(1), 1–40. https://doi.org/10.1257/app.20230052

Carlsson, F., Gravert, C., Johansson-Stenman, O., & Kurz, V. (2021). The use of green nudges as an environmental policy instrument. *Review of Environmental Economics and Policy*, *15*(2), 216–237. https://doi.org/10.1086/715524

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ECTA18709

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology: Applied*, *17*(4), 332–341. https://doi.org/10.1037/a0024354

Ebeling, F., & Lotz, S. (2015). Domestic uptake of green energy promoted by opt-out tariffs. *Nature Climate Change*, *5*(9), 868–871. https://doi.org/10.1038/nclimate2681

Egebark, J., & Ekström, M. (2016). Can indifference make the world greener? *Journal of Environmental Economics and Management*, *76*, 1–13. https://doi.org/10.1016/j.jeem.2015.11.004

Farrow, K., Grolleau, G., & Ibanez, L. (2017). Social norms and pro-environmental behavior: A review of the evidence. *Ecological Economics*, 140, 1–13.

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2022). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. *American Economic Journal: Economic Policy, 14*(4), 269–299. https://doi.org/10.1257/pol.20190287

Günther, S. A., Staake, T., Schöb, S., & Tiefenbeck, V. (2020). The behavioral response to a corporate carbon offset program: A field experiment on adverse effects and mitigation strategies. *Global Environmental Change, 64*, Article 102123. https://doi.org/10.1016/j.gloenvcha.2020.102123

Hagmann, D., Ho, E. H., & Loewenstein, G. (2019). Nudging out support for a carbon tax. *Nature Climate Change*, 9(6), 484–489. https://doi.org/10.1038/s41558-019-0474-0

Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, *103*(7), 2643–2682. http://dx.doi.org/10.1257/aer.103.7.2643

He, G., Pan, Y., Park, A., Sawada, Y., & Tan, E. S. (2023). Reducing single-use cutlery with green nudges: Evidence from China's food-delivery industry. *Science*, *381*(6662), Article eadd9884. https://doi.org/10.1126/science.add9884

Hedlin, S., & Sunstein, C. R. (2016). Does active choosing promote green energy use? Experimental evidence. *Ecology Law Quarterly, 43*(1), 107–142.

Hoffmann, C., & Thommes, K. (2022). Combining egalitarian and proportional sharing rules in team tournaments to incentivize energy-efficient behavior in a principal-agent context. *Organization & Environment*, *35*(2), 307–331. https://doi.org/10.1177/1086026620945343

Hoffmann, C., & Thommes, K. (2024). Can leaders motivate employees' energy-efficient behavior with thoughtful communication? *Journal of Environmental Economics and Management, 125*, Article 102990. https://doi.org/10.1016/j.jeem.2024.102990

Hummel, D., & Maedche, A. (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, *80(1)*, 47–58. https://doi.org/10.1016/j.socec.2019.03.005

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy, 3*(2), 159–186. https://doi.org/10.1017/bpp.2018.43

Jaramillo, P., Kahn Ribeiro, S., Newman, P., Dhar, S., Diemuodeke, O. E., Kajino, T., Lee, D. S., Nugroho, S. B., Ou, X., Hammer Strømman, A., & Whitehead, J. (2022). Transport. In P. R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, & J. Malley (Eds.), *Climate Change 2022: Mitigation of climate change*. Cambridge University Press. https://doi.org/10.1017/9781009157926

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339. https://doi.org/10.1126/science.1091721

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193–206. https://doi.org/10.1257/jep.5.1.193

Kaiser, M., Bernauer, M., Sunstein, C. R., & Reisch, L. A. (2020). The power of green defaults: The impact of regional variation of opt-out tariffs on green energy demand in Germany. *Ecological Economics*, *174*, Article 106685. https://doi.org/10.1016/j.ecolecon.2020.106685

Kesternich, M., Römer, D., & Flues, F. (2019). The power of active choice: field experimental evidence on repeated contribution decisions to a carbon offsetting program. *European Economic Review*, *114*, 76–91. https://doi.org/10.1016/j.euroecorev.2019.02.001

Liebe, U., Gewinner, J., & Diekmann, A. (2021). Large and persistent effects of green energy defaults in the household and business sectors. *Nature Human Behaviour*, *5*(5), 576–585. https://doi.org/10.1038/s41562-021-01070-3

Loewenstein, G., Price, J., & Volpp, K. (2016). Habit formation in children: Evidence from incentives for healthy eating. *Journal of Health Economics*, *45*, 47–54. https://doi.org/10.1016/j.jhealeco.2015.11.004

Löfgren, Å., Martinsson, P., Hennlock, M., & Sterner, T. (2012). Are experienced people affected by a pre-set default option—Results from a field experiment. *Journal of*

Environmental Economics and Management, *63*(1), 66–72. https://doi.org/10.1016/j.jeem.2011.06.002

Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, *116*(4), 1149–1187. https://dx.doi.org/10.2139/ssrn.223635

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), Article e2200300119. https://doi.org/10.1073/pnas.2200300119

McKenzie, C. R. M., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, *17*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2006.01721.x

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences, 119*(1), Article e2107346118. https://doi.org/10.1073/pnas.2107346118

Pichert, D., & Katsikopoulos, K. V. (2008). Green defaults: Information presentation and pro-environmental behaviour. *Journal of Environmental Psychology*, *28*(1), 63–73. https://doi.org/10.1016/j.jenvp.2007.09.004

Probst, B., Toetzke, M., Kontoleon, A., Diaz Anadon, L., Minx, J. C., Haya, B. K., Schneider, L., Trotter, P. A., West, T., Gill-Wiehl, A., & Hoffmann, V. H. (2023). Systematic assessment of the achieved emission reductions of carbon crediting projects. *Nature Communications, 15*(1), 9562. https://doi.org/10.1038/s41467-024-53645-z

Rodemeier, M. (2022). Willingness to pay for carbon mitigation: Field evidence from the market for carbon offsets. *The Review of Financial Studies, 39*(1), 1–29. https://doi.org/10.1093/rfs/hhaf054

Serra-Garcia, M., & Szech, N. (2023). Incentives and defaults can increase COVID-19 vaccine intentions and test demand. *Management Science*, *69*(2), 1037–1049. https://doi.org/10.1287/mnsc.2022.4405

Simon, H. A. (1957). *Models of man: Social and rational. Mathematical essays on rational human behavior in a social setting* (pp. 241–260). Wiley & Sons.

Sunstein, C. R. (2014). Nudging: A very short guide. *Journal of Consumer Policy, 37*(4), 583–588. https://doi.org/10.1007/s10603-014-9273-1

Sunstein, C. R. (2015). The ethics of nudging. *Yale Journal on Regulation*, *32*(2), 413–450.

Sunstein, C. R., & Reisch, L. A. (2014). Automatically green: Behavioral economics and environmental protection. *Harvard Environmental Law Review, 38*(1), 127–158.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., ... & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, *119*(31), Article e2200732119. https://doi.org/10.1073/pnas.2200732119

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review*, *93*(2), 175–179. https://doi.org/10.1257/000282803321947001

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2013). *Choice architecture*. In E. Shafir (Ed.), *The behavioral Foundations of Public Policy* (pp. 428–439). Princeton University Press.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, *106*(4), 1039–1061. https://doi.org/10.2307/2937956

CHAPTER 4 | FOREIGN LANGUAGE USE, ATTRIBUTION ERROR, AND NEWCOMER INTEGRATION

Chapter 4:

Foreign Language Use, Attribution Error, and Newcomer Integration

This paper, co-authored with Kirsten Thommes, is currently being prepared for submission to a journal. An earlier version was presented at the INGRoup Conference in Seattle, USA, from July 20 to 22, 2023.

Abstract

When newcomers fail to perform well at first, incumbents often misattribute this underperformance to a general lack of ability rather than to limited organization-specific knowledge. As a result, performance rating gaps between new and experienced employees tend to persist. Using the example of learning organization-specific language, we examine this problem and a potential mitigation strategy: If language codes are difficult for everyone to master, will managers be more lenient with newcomers and less likely to misattribute poor performance to low ability? We test this through an experimental study in which some groups communicated in their native language while others were required to use a second language. Under the second-language condition, the performance assessment gap between newcomers and incumbents narrows. However, this convergence does not stem from newcomers receiving better evaluations. Instead, incumbents are rated more negatively. Importantly, these lower ratings are not tied to actual performance but rather to managers' evaluations of their own job.

4.1 INTRODUCTION

The integration of newcomers into established organizations involves a complex interplay of cultural, linguistic, and social dynamics. A central challenge in this process is the acquisition of organization-specific skills—not only shared knowledge, beliefs, and assumptions, but also specialized vocabulary and codes, such as abbreviations, shortcuts, or technical jargon. As Markmann (2022) notes, "people who are fluent in their office jargon can spit out sentences that are completely incomprehensible to the uninitiated," for example: *"I had to get EVPP and VPR to approve a PAR before sending it to OSP."*

Research has shown that organization-specific codes play a crucial role in facilitating efficient communication within groups (Crémer et al., 2007). Accordingly, the rapid acquisition of organization-specific language during the initial six months of employment is a strong predictor of newcomers' retention and reduces the likelihood of involuntary exits (Srivastava et al., 2018; Liu et al., 2024). However, learning and adapting to this specialized language is often challenging and not always straightforward (Galantucci et al., 2012; Hong et al., 2017; Koçak & Puranam, 2023; Selten & Warglien, 2007).

Managers often underestimate these difficulties, as illustrated by Weber and Camerer (2003). Their study found evidence of a fundamental attribution error (FAE): although newcomers learned the group's vocabulary relatively quickly, managers attributed newcomers' slightly weaker performance compared to incumbents to lower cognitive ability rather than to the challenges of learning new terminology. Strikingly, this negative assessment was not linked to actual performance, as newcomers did not diminish group outcomes after only a short adjustment period.

The influence of language on social perception extends beyond technical codes. Research shows that linguistic strategies shape how newcomers are received. For instance, newcomers who employ integrative language strategies—such as using plural pronouns ("we," "our")—are evaluated more positively by incumbents and their knowledge is more readily incorporated

into group work. In contrast, newcomers who rely on differentiating strategies (e.g., "I," "my") tend to be received less favorably (Kane & Rink, 2015). While these distinctions are salient in one's native language, they may be less clear in a foreign language. Given the rise of English as a global lingua franca, many organizations now operate in multilingual environments, further complicating newcomer integration. Since it is widely recognized that communicating precisely in a foreign language is more difficult than in one's native tongue, one might expect managers to be more lenient when evaluating communication in a second language. Indeed, native English speakers perceive grammatical errors as less bothersome when made by non-natives (Wolfe et al., 2016), leading to more positive social evaluations (Geipel et al., 2016; Hadjichristidis et al., 2017).

Whether communication in a second language results in more lenient—and potentially fairer—evaluations of newcomers is an important question. If evaluations differ systematically across language contexts, multinational companies may face challenges in ensuring comparable assessments across their global branches. Moreover, newcomer integration could be facilitated in second-language environments, particularly if the informational costs of foreign-language use are small but the motivational benefits of more positive feedback are high. Despite the frequency of second-language communication in organizational settings, research on its implications for newcomer integration remains limited.

To address this gap, we employ an experimental design comparing first- and second-language settings. Our study examines how language conditions shape incumbents' perceptions of newcomers and whether increased language difficulty reduces evaluation bias. Understanding these dynamics is critical for improving newcomer integration and promoting fairness in performance evaluations.

Our results show that group performance remains equally strong after newcomer integration, regardless of whether communication occurs in a first or second language. When

communication takes place in the native language, we replicate Weber and Camerer's (2003) findings: newcomers receive disproportionately negative evaluations that are not justified by actual performance differences. By contrast, in the second-language condition, evaluations of newcomers and incumbents are more balanced. However, this fairness effect is not driven by greater leniency toward newcomers but rather by more negative evaluations of incumbents when communication does not occur in their native language. Notably, this effect is independent of actual performance. We discuss potential explanations, such as managers' self-assessment in relation to employees and reduced communicative fluency among all group members.

4.2 RELATED LITERATURE

The integration of newcomers into established organizations has long been a central concern in organizational research, particularly with respect to processes of socialization and the acquisition of organization-specific knowledge. Scholars have examined how both surface-level and deep-level similarities shape reactions between newcomers and incumbents. For example, relational dissimilarities, such as trait likability, tend to trigger stronger negative responses, while task-related dissimilarities, such as educational background, can elicit positive evaluations (Liu et al., 2023). Incumbents' perceptions are also shaped by attributes such as newcomers' attractiveness and gender, which influence behaviors like mimicry, ingratiation, and challenge (Min et al., 2021).

Besides personal attributes, newcomers' behavioral styles and prior experiences further shape their integration trajectories. Assertive newcomers, for instance, are more likely to successfully introduce new ideas, particularly when groups are moderately optimistic about their future performance (Hansen & Levine, 2009). Prior work and transition experience also enhance performance, both at entry and over time (Beus et al., 2014). Yet, while prior knowledge and experience are valuable, successful integration ultimately depends on newcomers' ability to

rapidly acquire organization-specific knowledge. This process is critical for performance, satisfaction, and retention (Bauer et al., 2007). Supervisors play a pivotal role in this adjustment process by providing information, feedback, and training, clarifying roles, and supporting newcomers' sensemaking (Lee, 2024). Proactive newcomer behaviors, such as seeking feedback, signal commitment to adapt and increase supervisor support, which in turn facilitates task mastery and social adjustment while reducing turnover (Benson & Eys, 2017; Ellis et al., 2017).

A particularly demanding aspect of organizational socialization is mastering organization-specific language and codes. This includes jargon, abbreviations, and technical terminology that incumbents may take for granted but which newcomers must learn to fully integrate (Weber & Camerer, 2003). Over time, such codes become embedded in organizational culture, facilitating efficient communication through shared shortcuts and meanings (Crémer et al., 2007; Selten & Warglien, 2007). However, code clashes, where different groups label the same stimulus differently, present challenges, as do the difficulties of unlearning established codes to achieve alignment (Koçak & Puranam, 2022). Social dynamics further complicate this process: the adoption and diffusion of codes are shaped by relational histories, power dynamics, and role structures within groups (Koçak & Warglien, 2020; Galantucci et al., 2012).

Importantly, the time-consuming process of learning new organizational codes is frequently underestimated by those who already master the codes. When newcomers in groups lack knowledge of organization-specific routines, terminology, and tacit practices, the fundamental attribution error (FAE) may occur. This error describes the systematic tendency to overemphasize dispositional explanations of others' behavior while underestimating situational constraints (Jones & Harris, 1967; Ross, 1977). In organizational contexts, this bias is of particular concern because evaluative judgments of competence, reliability, or professionalism often shape access to resources, career opportunities, and social inclusion. When evaluators

attribute observable behavior to stable personal characteristics rather than contextual contingencies, assessments risk becoming distorted and unjust.

As Van Maanen and Schein (1979) argue, organizational cultures rely heavily on localized conventions that are not immediately transparent to outsiders. Consequently, when newcomers fail to immediately master these conventions, their hesitations or errors may be misinterpreted as evidence of limited competence or motivation. The situational explanation—namely, that their performance reflects incomplete exposure to the organization's symbolic and procedural order—is easily overlooked. Instead, observers may infer dispositional deficiencies as the root cause of lower performance in the beginning (Lyubykh et al., 2022).

This tendency is intensified during evaluative encounters, in which new members are subjected to heightened scrutiny. According to expectancy-violation theory (Burgoon, 1993), even minor deviations from expected communicative or procedural norms draw a disproportionate degree of attention. When evaluators are well-versed in organizational routines, they may suffer from the "curse of knowledge" (Camerer, Loewenstein, & Weber, 1989). This phenomenon occurs when these evaluators underestimate the degree to which these routines are opaque to external observers. The result is a systematic bias: newcomers are judged as lacking ability or diligence when, in fact, their performance primarily reflects their stage in the socialization process. Together, these mechanisms demonstrate how the FAE can create disadvantages for new organizational members, as already found by Camerer and Weber (2003). By conflating situationally induced challenges with dispositional shortcomings, evaluators risk reinforcing inaccurate assessments and, in turn, undermining the integration of newcomers into the organization. Consequently, such distorted evaluations may also provoke strong negative feelings in employees toward their supervisors, resulting in ongoing organizational conflict at worst (Santos et al., 2019; Sharma & Kulshreshtha, 2023).

The FAE often arises when group members fail to recognize a newcomer's lack of organization-specific knowledge. Although the negative consequences of such misjudgments are well-documented, strategies to mitigate them remain less understood. One possible avenue is to emphasize the inherent challenges of communication. For example, when employees must interact in a second language, communication tends to suffer even if all parties are proficient (Harzing & Feely, 2008; Piekkari et al., 2005; Wilmot et al., 2024). In such contexts, heightened awareness of language barriers may reduce the tendency to make attribution errors. Supervisors and incumbents, drawing on their own experiences with non-native communication, may show greater leniency toward newcomers and recognize the difficulties they face.

Research demonstrates the complex role of language proficiency in organizational life. Non-native speakers often face status loss and heightened anxiety when evaluated on their fluency (Neeley, 2013), while language barriers can foster misunderstandings and conflict within teams (Tenzer et al., 2014). Even when employees share a common language, cultural differences can persist, creating subtle challenges for trust and interpretation (Arnulf et al., 2021; Henderson, 2005). The use of a foreign language has been linked to reduced participation in decision-making, higher stress, and lower satisfaction (Liu & Sekiguchi, 2019). Moreover, differences in proficiency can generate power imbalances, thereby enabling fluent speakers to dominate discussions while marginalizing less fluent members (Marschan-Piekkari et al., 1999; Tenzer & Pudelko, 2017). Language difficulties also shape perceptions of competence and trustworthiness, reinforcing both cognitive and emotional barriers to integration (Tenzer et al., 2014).

Taken together, prior research suggests that language profoundly influences newcomer integration, both through the challenge of mastering organization-specific terminology and through the broader dynamics of multilingual communication. While existing studies have

largely highlighted the negative consequences of foreign language use for communication, performance, and trust, much less is known about its potential benefits. In particular, foreign language use may foster leniency toward newcomers and moderate the FAE. This leads us to the central question of our study: Does communication in a foreign language improve newcomer integration within teams? To address this question, we employ the experimental design by Weber and Camerer (2003) and test the potential mitigation strategy of communication in a second language.

4.3 METHOD

The main task of the study by Weber and Camerer (2003) is a picture description task, participants had to explain 16 unique black and white office scene images to each other (see appendix for instructions and pictures). After consultation with Weber and Camerer via email, we modified the pictures to display more up-to-date technology (and fashion). In total, 80 students participated in our experiment over 26 sessions.

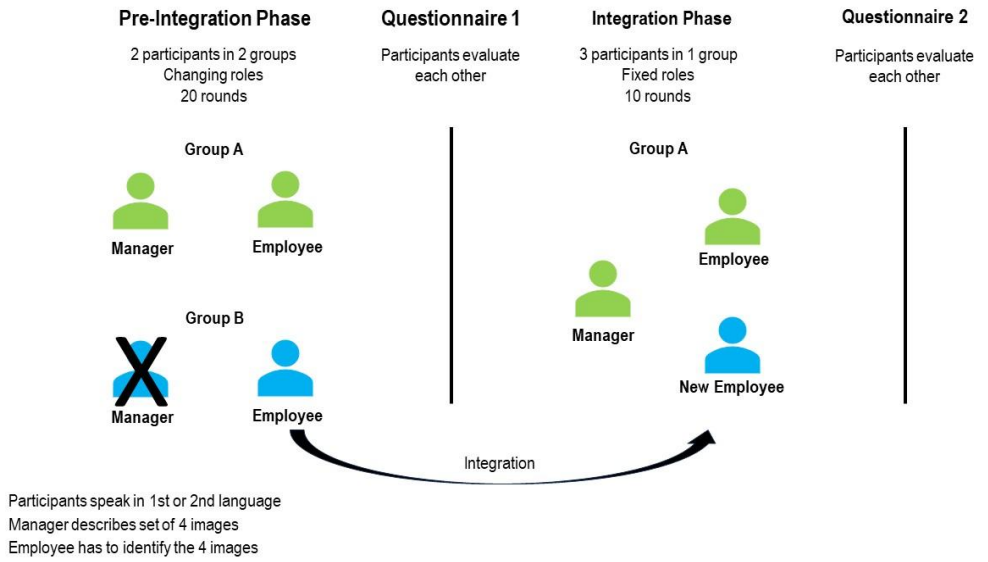
Students could apply to participate in the experiment via an online tool at the university. The pre-requisite for participating was that German was their mother tongue, all participants confirmed that. For being enrolled at the University, C1-level proficiency is required, all participants also confirmed during their application procedure in our system that their English language skills were at least on C1-level. 47,5% of the students identified as female, the average age was 22.23 years. The experiments took place in person between May and November 2022. Upon arrival, participants were randomly allocated to form a duo. Each duo was placed in a separate room, and one of the experimenters was present in the room at all times. Per round, one of the two participants, named the “manager”, described four specific images from the set to the other participant, the “employee”. The employee’s task was to correctly identify these four images in the order the manager described. Manager and employee were placed behind a

partition wall so that they could neither see each other nor each other's desk (and pictures on the desk).

We conducted 20 rounds in total per duo. After each round, the experimenter provided feedback on time and accuracy and the roles of manager and employee changed. Upon the completion of these 20 rounds, some of the duos were randomly dissolved. The two respective members joined one of the existing duos so that every duo became a trio. In each trio, again, one participant was randomly chosen to become the new manager for the remaining time, and the other group members became employees. In this phase, the manager and the employees stayed in their role over the course of the rounds. In the subsequent ten rounds, the manager had to describe the randomly chosen pictures, and the two employees had to identify the picture the manager had described. Again, all participants were separated by partition walls so that they could not see each other or each other's desks. As performance indicators, we use time used in a round for a correct identification of pictures and errors per group.

In the second phase, twelve sessions were conducted in the participants' first language (German) and fourteen sessions in the participants second language (English). We used an incentive scheme pre- and post-merger so that participants had the incentive to work as quickly as possible. If the employee would solve the description correctly, both the employee and the manager would receive a bonus of 0.93€ if the task was correctly solved within 10 seconds. The bonus decreased by €0.07 for every additional 10 seconds, reaching €0 at 150 seconds. Time was rounded up to the next 10-second interval. If the task was not completed within 150 seconds, no bonus was paid for that round. On top, participants earned a show-up fee of 3.00€, resulting in an average pay of 20.08€, the total average participation duration was 59 minutes. Figure 4.1 provides an overview of the experimental design of the study.

Figure 4.1: Experimental design



In the questionnaires after the first 20 and the second 10 rounds, participants had to rate the performance of their group members on a scale from 1 (very bad) to 9 (very good) and the difficulty of their own task and the task of every group member separately from 1 (very difficult) to 9 (very easy).

During the whole experiment, verbal interactions were recorded and later transcribed. We use this information to further analyze the performance, cooperation, and ratings in our experiment. We specifically apply Linguistic Inquiry and Word Count (LIWC) software to the recorded verbal interactions between participants (Tausczik & Pennebaker, 2010). LIWC allowed us to quantify and categorize participants' linguistic patterns along several psychological dimensions as attentional focus, emotionality, social relationships, thinking styles. We specifically analyze the following language patterns:

- Efficiency of communication: We use two indicators here, namely word count (number of words spent by manager, incumbent and newcomer) and analytical language use, which provides insight into the extent to which participants use words that suggest formal and logical patterns of thinking. The analytical thinking variable in the LIWC is a measure based on several categories of function words. Function words include

articles (e.g., a, an, the), prepositions (e.g., in, on, at), auxiliary verbs (e.g., is, has, be), or quantifiers (e.g., few, many, much).

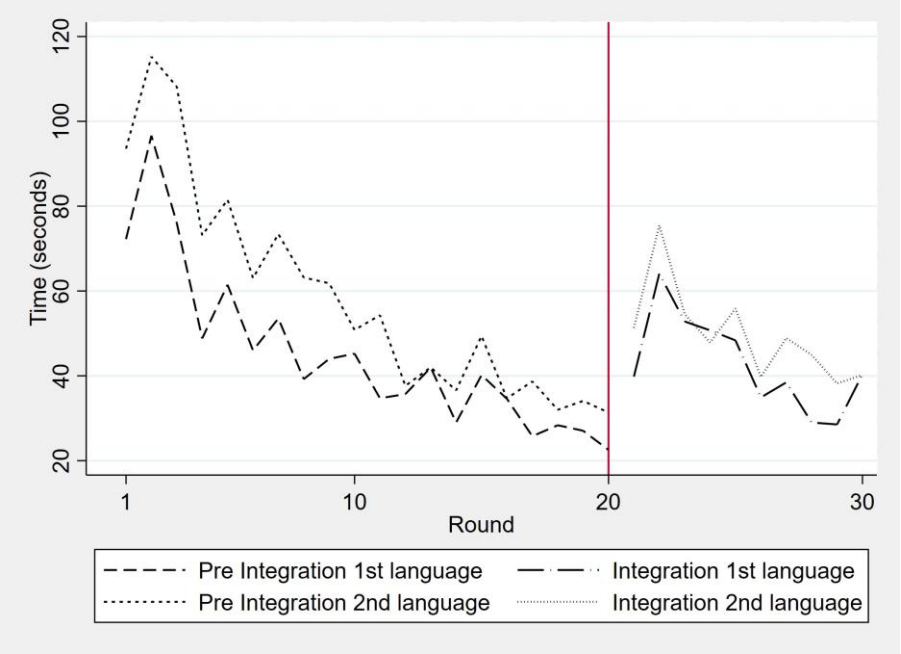
- Communication helping the manager as feedback, specifically assent variable, which measures the percentage of agreement words in each participant's transcript, e.g. yes, yeah, okay, and ok and interrogatives (e.g., how, when, why) which may help the manager to add clarifications in their descriptions.

4.4 RESULTS I – Replication of Weber/Camerer 2003

4.4.1 Group efficiency

First, we were generally able to replicate the results by Weber and Camerer (2003) concerning increasing efficiency in duos (round 1 – 20) and also trios (round 21 – 30). Figure 4.2 displays the average completion time needed by round.

Figure 4.2: Average completion time by round and language



The reduction in completion time is mainly achieved by the establishment of short cuts in groups. For example, in the early rounds, one group's description of an image showing people around a table was: "Second picture. It looks like in Grey's Anatomy. The picture is very white and a man is sitting in the middle, like a chief physician, with women around him." – ID

33386368. However, after several rounds of running the experiment, the group's code for this image became simple: "Grey's Anatomy." – ID 33386368

As can be already drawn from figure 4.2, the merger in groups by adding a newcomer to an old group increased completion time in the first rounds, but already from about round 25 on, the groups reached roughly the former speed from round 15. We observe that groups in the second language were slower throughout the experiment compared to groups who were able to communicate in their mother tongue, especially during the pre-integration phase (rounds 1-5, 6-10, 11-15, 16-20). After the integration of a newcomer, however, completion times between first and second language groups no longer differed significantly (rounds 21-25, 26-30).

Table 4.1: Completion time medians

Rounds	Group size	Time in sec. 1st Lang.	Time in sec. 2nd Lang.	p-value
1-20	2	Median = 44.5 CI = 40.92 - 51.37	Median = 55.1 CI = 50.89 - 62.75	0.0002
21-30	3	Median = 37.2 CI = 30.54-61.63	Median = 52.9 CI = 36.22-62.26	0.2812
p-value	Round 1-20 versus 21-30:	0.3203	0.0256	
1-5	2	Median = 71.4 CI = 59.86 - 79.78	Median = 87.6 CI = 80.57 - 108.86	0.0005
6-10	2	Median = 43 CI = 34.04 - 50.37	Median = 62.4 CI = 54.40 - 71.02	0.0030
11-15	2	Median = 35.8 CI = 28.23 - 40.78	Median = 40.4 CI = 37.22 - 51.23	0.0269
16-20	2	Median = 25.2 CI = 20.41 - 33.40	Median = 32.2 CI = 27.25 - 35.40	0.0385
21-25	3	Median = 47.8 CI = 35.97 - 68.73	Median = 58.8 CI = 42.16 - 66.90	0.4430
26-30	3	Median = 32.4 CI = 23.54 - 45.56	Median = 41.80 CI = 28.80 - 52.92	0.1679

During the first 20 rounds, duos were significantly faster (p=0.0002, Mann-Whitney U-test) in solving the task in their first language than in their second language. For the single team however, the improvement must have felt more significant when groups were using their second language. In the second language, groups improved from 55.1 seconds to 52.9 seconds (p=0.0256, Wilcoxon-Signed Rank).

4.4.2 Participant ratings

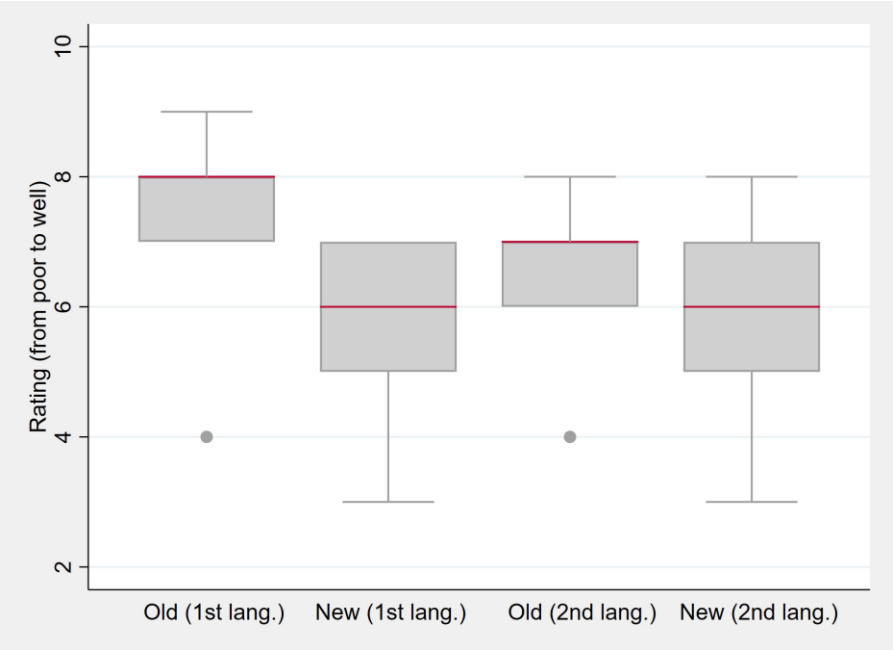
Even though the improvement in speed was much more significant in the second language, this is not reflected by the performance appraisals of managers after the second round: Regardless of the language being used, managers consistently rated the established employee (their prior collaborator) better than the newcomer (figure 4.3). In sessions conducted in the first language, managers assigned a median rating of eight points to established employees, compared to a median rating of six points for newcomers (Signrank test, p = 0.0137).

A similar pattern was observed in the second language sessions, where established employees received a median rating of seven points, while newcomers were rated at a median of six (Signrank test, p = 0.0352). Thus, newcomers received similar median ratings of six in both language settings (Mann-Whitney test, p = 0.4812).

However, the spread between established and new group member was less pronounced in the second-language setting. Notably, using a second language reduces the disparity in ratings between established employees and newcomers. The difference in median ratings is significantly smaller in the second language environment (a difference of one point) compared to the first language environment (a difference of two) (Mann-Whitney test, p = 0.0554). This could be associated with a leniency bias in ratings, where managers may be more forgiving or understanding when evaluating employees. When communicating in a non-native language, managers might view the established employee's performance slightly less favorably than when the communication takes place in the participants' mother tongue. Importantly, this less

pronounced gap is not a result of better grades in general, even though they would have been justified by the improvement in time used for solving the task, but root back to other factors. In the following, we shall try to disentangle potential reasons.

Figure 4.3: Managers' ratings of established & new employees



4.5 RESULTS II – Potential reasons for less disparity

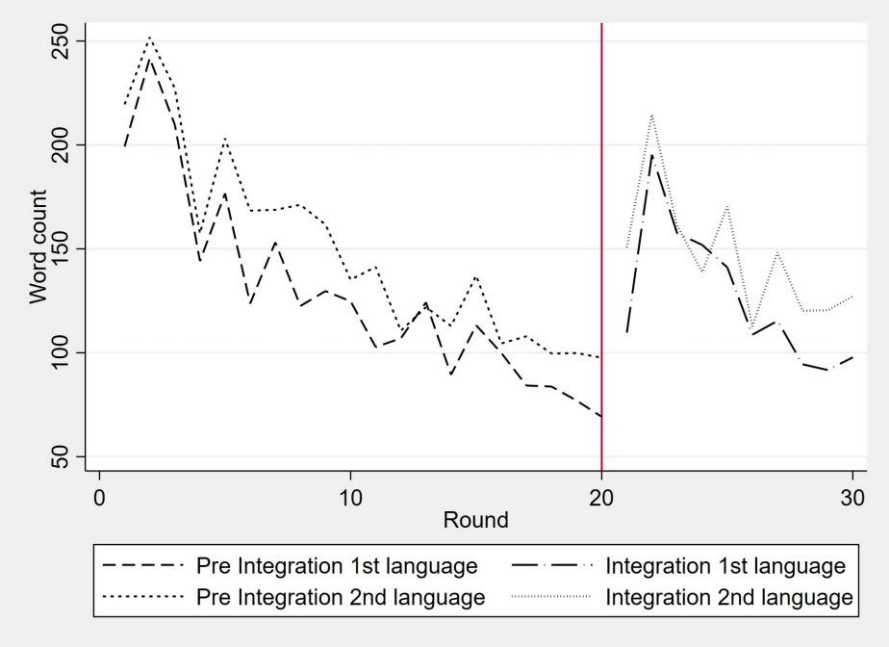
Unlike our initial idea that communication in a foreign language may prevent newcomer discrimination in appraisals, we observe that communication in a second language instead results in less discrimination between incumbents and newcomers, but only because the incumbent is judged harsher in the second language. In the following, we try to disentangle two of the potential mechanisms: (1) Incumbents may behave differently in the second language when another person is added to the team, or (2) managers may judge differently because of the focal point they use.

4.5.1 Do incumbents behave differently in a second language when a new team member is added?

One particular change in behavior might stem from the participants communication. First, the data suggests that groups consistently reduced word count which is in line with the reduced

time groups needed for completing a round (figure 4.2, see also table 4.1). In general, the advantage of the duo diminishes during the integration phase and the reduction is less pronounced in the second language compared to the first language. Figure 4.4 provides an overview of average word count over time, which displays the communicative advantage of native duos and trios.

Figure 4.4: Average word count



In general, groups speak less in the first than in the second language, but the difference between groups remains constant (table 4.2).

Table 4.2: Word count medians by rounds

Rounds	Group size	Word count 1st Lang.	Word count 2nd Lang.	Word count p-value
1-20	2	Median = 129.9 CI = 115.09 - 140.12	Median = 146.6 CI = 130.89 - 157.54	0.0420
21-30	3	Median = 119.5 CI = 102.97 - 157.77	Median = 158 CI = 112.45 - 180.57	0.1484
p-value	2 vs 3	0.5771	0.5614	

When comparing the old and the new employee in the second part of the experiment, we observe a strong difference in round 21-30 (table 4.3).

Table 4.3: Word count medians by role

Role	Post-integration 1st Language	Post-integration 2nd Language	p-value
Incumbent	Median = 75 CI = 52.265 – 82.873	Median = 70 CI = 51.247 – 86.931	0.970
Newcomer	Median = 171 CI = 100.266 – 197.735	Median = 133 CI = 72.178 – 166.011	0.194
p-value	0.000	0.004	

In both languages, new employees speak much more than old employees, resulting in a possible perception of them being more active.

One option in changed behavior may stem from less or more usage of analytical thinking words, that is detected by aggregating small function words such as articles and prepositions (table 4.4). Analytical language use was significantly higher in the first language setting for both managers and incumbents. In the post integration phase, managers continue to use more analytical language in their first language (median: 99) than in their second language (median: 96.09, p=0.0018). A similar, though marginally significant, pattern was observed for incumbents (p = 0.0558), while newcomers showed no significant difference between language settings (p = 0.5357).

Table 4.4: Analytical language use

Role	Post-integration 1st Language	Post-integration 2nd Language	p-value
Incumbent	Median = 96.03 CI = 48.50 - 99	Median = 86.56 CI = 37.50 – 90.45	0.0558
Newcomer	Median = 91.56 CI = 26.61 - 95.12	Median = 84.12 CI = 62.71 - 96.59	0.5357
Manager	Median = 99 CI = 97.88 - 99	Median = 96.09 CI = 90.52 - 98.35	0.0018
Incumbent vs. Newcomer	p = 0.3728	p = 0.2057	
Manager vs. Incumbent	p = 0.0077	p = 0.0008	
Manager vs. Newcomer	p = 0.0003	p = 0.0304	

Comparing roles, no significant differences in analytical language use were found between incumbents and newcomers in either language condition. However, managers consistently used

significantly more analytical language than both incumbents and newcomers across both first and second language settings.

A second language marker concerns the use of assent language, which reflects agreement or approval and includes words such as “yes” or “okay” (table 4.5). Across all roles, the use of assent language tended to be higher in the first language setting compared to the second, although these within-group differences were not statistically significant ($p = 0.0954$ for incumbents, $p = 0.4668$ for newcomers, and $p = 0.1492$ for managers). Incumbents showed the largest numerical difference (median = 52.22 vs. 32.27, $p = 0.0954$), but this trend did not reach conventional significance.

In contrast, between-role comparisons revealed significant differences. Managers used less assent language than both incumbents and newcomers in both language conditions ($p = 0.0000$ in all comparisons). Moreover, incumbents also used significantly more assent language than newcomers in the first language condition ($p = 0.0032$), while no significant difference was observed between them in the second language setting ($p = 0.4544$).

Table 4.5: Assent language use

Role	Post-integration 1st Language	Post-integration 2nd Language	p-value
Incumbent	Median = 52.22 CI = 40.08 – 59.30	Median = 32.27 CI = 20.68 – 54.30	0.0954
Newcomer	Median = 34.38 CI = 21.42 - 46.72	Median = 26.14 CI = 17.40 - 37.40	0.4668
Manager	Median = 2.4 CI = 1.09 – 3.98	Median = 1.6 CI = 0.93 – 2.48	0.1492
Incumbent vs. Newcomer	p = 0.0032	p = 0.4544	
Manager vs. Incumbent	p = 0.0000	p = 0.0000	
Manager vs. Newcomer	p = 0.0000	p = 0.0000	

A third linguistic marker examined was the use of interrogative language, which reflects the tendency to ask questions or seek information (table 4.6). Across all roles, the frequency of interrogative words was generally low. Incumbents showed no usage of interrogative language in either language setting (median = 0 in both), with no significant difference between first and

second language (p = 0.6265). Newcomers and managers showed slightly higher median values, but again, the differences between language settings were not statistically significant (p = 0.1313 and p = 0.1040, respectively).

Between-role comparisons showed no consistent pattern. Managers used significantly more interrogative language than incumbents in the second language condition (p = 0.0232), but not in the first (p = 0.2071). No significant differences were observed between managers and newcomers or between incumbents and newcomers in either language condition.

Table 4.6: Interrogatives

Role	Post-integration 1st Language	Post-integration 2nd Language	p-value
Incumbent	Median = 0 CI = 0 – 1.290	Median = 0 CI = 0 – 1.188	0.6265
Newcomer	Median = 0.97 CI = 0 – 1.4	Median = 0.48 CI = 0 – 0.77	0.1313
Manager	Median = 0.58 CI = 0.24 – 1.35	Median = 0.305 CI = 0.11 – 1.06	0.1040
Incumbent vs. Newcomer	p = 0.4609	p = 0.3799	
Manager vs. Incumbent	p = 0.2071	p = 0.0232	
Manager vs. Newcomer	p = 0.9872	p = 0.4739	

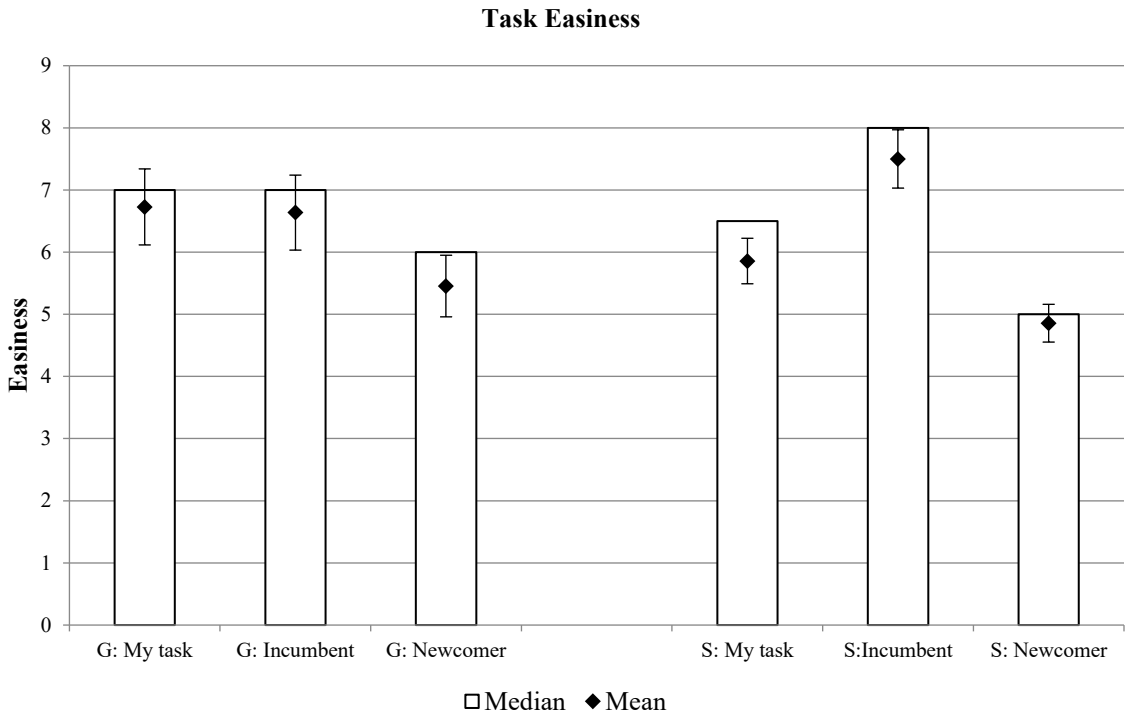
4.5.2 Do managers use a new focal point in a second language environment?

The manager may be harsher in the second language because (s)he feels that they had to work even harder in the second language compared to the first language setting and may now instead underestimate the cognitive effort of the incumbent. The task easiness assessment supports this idea (Figure 4.5).

In the German setting (first language, left hand side): managers assess their task and the task of the old group member as similarly easy (all Wilcoxon Rank Test, p = 0.8594), while they acknowledge that new members have a slightly harder task then the old group member (p = 0.0078) and themselves (p = 0.0156). In the second language setting, managers assess especially the old group members task as very easy (second bar from the right; for both comparisons against their own task (p = 0.0059) and the newcomers task Wilcoxon Rank Test

tests, p = 0.0002). Managers also assess their own task as easier than those of newcomers (Wilcoxon Rank Test, p = 0.0256).

Figure 4.5: Managers' assessment of task easiness



4.6 DISCUSSION

This study investigated whether communication in a second language can reduce the fundamental attribution error (FAE) in newcomer evaluation and whether group productivity is affected by operating in a foreign language. Our findings contribute to both streams of research.

With respect to productivity, we find no evidence that second-language communication hinders team performance. Teams in both the first and second language conditions rapidly attained high levels of coordination, replicating the results of Weber and Camerer (2003) for dyads and triads. Contrary to prior work documenting productivity losses, communication breakdowns, and voluntary withdrawal from conversations in multilingual settings (Liu & Sekiguchi, 2019; Neeley, 2013; Tenzer et al., 2014), our experimental design required continuous participation, thereby eliminating dropout effects. Moreover, the organizational code employed in our task—

shortcuts for describing images—was relatively straightforward to acquire, which may explain why no meaningful barriers to efficient communication emerged. This contrasts with studies showing significant coordination challenges when shared codes are more complex (Galantucci et al., 2012; Hong et al., 2017; Koçak & Puranam, 2023; Selten & Warglien, 2007).

Turning to evaluation, our findings reveal a more complex picture. Consistent with Weber and Camerer (2003), we replicate the presence of the fundamental attribution error in the first language condition: newcomers were systematically underrated compared to incumbents. In the second language condition, however, this evaluation gap disappeared. At first glance, this might suggest a positive effect of second-language communication on newcomer integration. Yet closer examination shows that the equalization of evaluations is not driven by more favorable assessments of newcomers, but rather by systematically lower ratings for incumbents.

We considered whether behavioral changes in communication could explain these lower ratings. However, an analysis of linguistic performance does not support this interpretation. Instead, we propose that second-language use shifted supervisors' evaluative reference points. Whereas in the first language condition evaluations appear to be anchored in comparisons between employees, in the second language condition supervisors may have drawn comparisons with their own more demanding task experience. This shift, which was likely triggered by the added psychological burden of operating in a non-native language, led to systematically harsher evaluations of all employees. Importantly, this aligns with prior evidence that communication in a foreign language imposes additional mental load even among proficient speakers (Harzing & Feely, 2008; Piekkari et al., 2005; Wilmot et al., 2024).

These findings carry important implications. On the one hand, second-language communication appears to attenuate attribution error in newcomer evaluations. On the other hand, it does so by compressing performance ratings downward, which may generate

unintended consequences in organizational contexts. Over time, such compressed evaluations could foster dissatisfaction with supervisors, trigger perceptions of unfairness, and ultimately increase conflict and turnover intentions among both newcomers and incumbents (Benson & Eys, 2017; Ellis et al., 2017; Santos et al., 2019; Sharma & Kulshreshtha, 2023).

4.7 CONCLUSION

In sum, our results suggest that communication in a second language has a dual effect on group dynamics. On the positive side, groups perform effectively regardless of language, and second-language use reduces attribution bias that disadvantages newcomers. On the negative side, however, second-language communication induces a shift in evaluative reference points, leading supervisors to downgrade both incumbents and newcomers. This effect likely stems from the psychological burden associated with non-native communication rather than from observable performance differences.

These findings extend research on multilingual communication by highlighting not only the absence of productivity costs in structured tasks, but also the subtle evaluative distortions introduced by second-language use. For organizations, this implies that while operating in a foreign language may level the playing field for newcomers, it may also risk creating dissatisfaction and conflict if performance evaluations become systematically compressed. Future research should investigate how these effects unfold in more complex organizational settings, whether training can alleviate the evaluative burden of second-language communication, and how firms might balance the potential integration benefits with the risks of evaluation bias.

4.8 APPENDIX

Appendix I: Instructions Pre-Integration Phase

This is an experiment on decision-making. After the experiment is completed, you will receive the amount you earned via bank transfer. The exact amount you will receive will be determined during the experiment and will depend on your decisions as well as the decisions of the other participants.
In this experiment, you will be an employee of one of two companies: Company A or Company B. There are six participants in this experiment. Four participants are employees of Company A, and two participants are employees of Company B. The instructions and procedures of the experiment are identical for both companies.
During the experiment, you and the other employee will work as a team for several rounds. Your role in each round of the task will be either "Manager" or "Employee." The difference between these two roles is described below.
In each round, you will be required to solve a similar task. The task is as follows:

1. Each of you has the same set of 16 images in front of you. The images are identical for both Company A and Company B.

2. On the protocol sheet of the respective Manager, 4 of the 16 images are listed in a specific order for each round. These are the target images for that round.

3. In each round, the Manager's task is to guide the Employee to select the 4 correct images from the 16 images in front of them, in the correct order. Before the Manager starts describing the images, they say "go" and the timer starts. The Manager and Employee may communicate verbally but are not allowed to see each other. Referring to the number on the back of the image is forbidden.

4. After the Employee has selected 4 images, they should write the numbers in the correct order on their protocol sheet and say "done." Only then does the experimenter stop the timer. After the time has been announced, the Employee reads the numbers from their protocol sheet out loud. The experimenter then informs them whether the set of 4 images is correct or not and points out any incorrect images.

Since the company you work for is interested in the task being completed as quickly as possible, your bonus for each round depends on how quickly you and the other employee complete the task. Specifically, your bonus in each round is calculated as follows:
Bonus = €1 - (Time taken / 150)
The time taken corresponds to the number of seconds required to complete the task. We round the time up to the next 10-second interval. So, for example, if you finish in 53 seconds, your recorded time will be 60 seconds, and your bonus for that round will be €0.60. You and the other employee will receive the same bonus in each round. The table on the next page shows the bonus amount for each completion interval. If you do not complete the task within 150 seconds (2.5 minutes), neither of you will receive a bonus for that round.
Your conversation during the task will be recorded with a voice recorder.
We are now ready to begin. From this point onward, if you have any questions, please raise your hand and wait for one of the experimenters to come to you. You should not speak out loud

or attempt to communicate directly with any of the other participants, except during the execution of the task. If you violate these rules, you will be asked to leave the experiment and will not receive any payment.
We will conduct this task over 20 rounds. At the end of the experiment, you will receive the total amount of your bonuses via bank transfer.

Time to complete the task (in seconds)	**Bonus (€)**
10	0.93
20	0.87
30	0.80
40	0.73
50	0.67
60	0.60
70	0.53
80	0.47
90	0.40
100	0.33
110	0.27
120	0.20
130	0.13
140	0.07
150	0.00

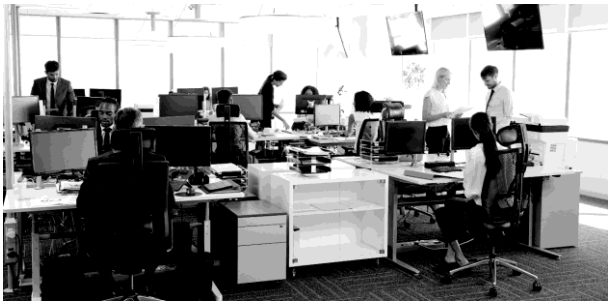
Appendix II: Instructions Integration Phase

For the remainder of the experiment, there will be 10 additional rounds in which we will carry out the same activity as before, but with one key difference. One of the companies will now take over the other company. This means that there will now be only one company, consisting of three participants: one Manager and two Employees. The roles in this part of the experiment are fixed. The Manager and one Employee come from the acquiring company, while the other Employee comes from the acquired company. In each round, the Manager will attempt to communicate the target images in the correct order to both Employees. Each of the two Employees will receive a bonus in each round, calculated in the same way as before. The bonus amounts that the Employees receive do not have to be the same. The Manager will now receive the sum of the two Employees' earnings divided by 1.5. Roles for the final 10 rounds will be assigned randomly.
You will now each write down a prediction of the average number of seconds the Employees will need to complete the task in the next 10 rounds.
We will now assign you a role for the remainder of the experiment. Please choose an envelope from the experimenter. Inside the envelope you selected, you will find a card indicating your role: Manager or Employee.
We are now ready to begin. From this point on, if you have any questions, please raise your hand and wait for one of the experimenters to come to you. You should not speak out loud or attempt to communicate directly with any of the other participants, except during the execution of the tasks. If you violate these rules, you will be asked to leave the experiment and will not receive any payment. We will conduct this task for 10 rounds. The amount you have earned in both parts of the experiment will be transferred to you via bank transfer.

Appendix III: 16 Pictures





Appendix IV: Analytical Language pre- and post-integration:

Role	**Dimension**	**Pre-integration:** 1st Language median	**Pre-integration:** 2nd Language median	**Post-integration:** 1st Language median	**Post-integration:** 2nd Language median
Managers	Analytical Language	98.72	86.93	99	96.09
Old employee	Analytical Language	99	91.71	96.03	86.56
Newcomer	Analytical Language	/	/	91.56	84.12

REFERENCES

Arnulf, J. K., Dai, W., Lu, H., & Niu, Z. (2021). Limits of a second language: Native and second languages in management team communication. *Frontiers in Psychology*, *12*, Article 580946. https://doi.org/10.3389/fpsyg.2021.580946

Bauer, T. N., Bodner, T., Erdogan, B., Truxillo, D. M., & Tucker, J. S. (2007). Newcomer adjustment during organizational socialization: A meta-analytic review of antecedents, outcomes, and methods. *Journal of Applied Psychology*, *92*(3), 707–721. https://doi.org/10.1037/0021-9010.92.3.707

Benson, A. J., & Eys, M. (2017). Understanding the consequences of newcomer integration processes: The sport team socialization tactics questionnaire. *Journal of Sport and Exercise Psychology*, *39*(1), 13–28. https://doi.org/10.1123/jsep.2016-0182

Beus, J. M., Jarrett, S. M., Taylor, A. B., & Wiese, C. W. (2014). Adjusting to new work teams: Testing work experience as a multidimensional resource for newcomers. *Journal of Organizational Behavior*, *35*(4), 489–506. https://doi.org/10.1002/job.1903

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology*, *12*(1-2), 30–48. https://doi.org/10.1177/0261927X93121003

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*(5), 1232–1254.

Crémer, J., Garicano, L., & Prat, A. (2007). Language and the theory of the firm. *The Quarterly Journal of Economics*, *122*(1), 373–407. https://doi.org/10.1162/qjec.122.1.373

Ellis, A. M., Nifadkar, S. S., Bauer, T. N., & Erdogan, B. (2017). Newcomer adjustment: Examining the role of managers' perception of newcomer proactive behavior during organizational socialization. *Journal of Applied Psychology*, *102*(6), 993–1001. https://doi.org/10.1037/apl0000201

Galantucci, B., Theisen, C., Gutierrez, E. D., Kroos, C., & Rhodes, T. (2012). The diffusion of novel signs beyond the dyad. *Language Sciences*, *34*(5), 583–590. https://doi.org/10.1016/j.langsci.2012.03.013

Geipel, J., Hadjichristidis, C., & Surian, L. (2016). Foreign language affects the contribution of intentions and outcomes to moral judgment. *Cognition*, *154*, 34–39. https://doi.org/10.1016/j.cognition.2016.05.010

Hadjichristidis, C., Geipel, J., & Surian, L. (2017). How foreign language affects decisions: Rethinking the brain-drain model. *Journal of International Business Studies*, *48*(5), 645–651. https://doi.org/10.1057/s41267-016-0040-1

Hansen, T., & Levine, J. M. (2009). Newcomers as change agents: Effects of newcomers' behavioral style and teams' performance optimism. *Social Influence*, *4*(1), 46–61. https://doi.org/10.1080/15534510802280827

Harzing, A. W., & Feely, A. J. (2008). The language barrier and its implications for HQ-subsidiary relationships. *Cross Cultural Management: An International Journal*, *15*(1), 49–61. https://doi.org/10.1108/13527600810848827

Henderson, J. K. (2005). Language diversity in international management teams. *International Studies of Management & Organization*, *35*(1), 66–82. https://doi.org/10.1080/00208825.2005.11043722

Hong, F., Lim, W., & Zhao, X. (2017). The emergence of compositional grammars in artificial codes. *Games and Economic Behavior*, *102*, 255–268. https://doi.org/10.1016/j.geb.2016.12.009

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24. https://doi.org/10.1016/0022-1031(67)90034-0

Kane, A., & Rink, F. (2015). How newcomers influence group utilization of their knowledge: Integrating versus differentiating Strategies. *Group Dynamics: Theory, Research, and Practice*, *19*, 91–105. https://doi.org/10.1037/gdn0000024

Koçak, Ö., & Puranam, P. (2022). Separated by a common language: How the nature of code differences shapes communication success and code convergence. *Management Science*, *68*(7), 5287–5310. https://doi.org/10.1287/mnsc.2021.4157

Koçak, Ö., & Puranam, P. (2023). Decoding culture: Tools for behavioral strategists. *Strategy Science*, *9*(1), 1–20. https://doi.org/10.1287/stsc.2022.0008

Koçak, Ö., & Warglien, M. (2020). When three's a crowd: How relational structure and social history shape organizational codes in triads. *Journal of Organization Design, 9*(1), 1–27. https://doi.org/10.1186/s41469-020-00078-9

Lee, A. S. (2024). Supervisors' roles for newcomer adjustment: Review of supervisors' impact on newcomer organizational socialization outcomes. *European Journal of Training and Development*, *48*(5/6), 521–539. https://doi.org/10.1108/EJTD-10-2022-0107

Liu, T., & Sekiguchi, T. (2019). How does the use of a foreign language affect team processes and member stress and satisfaction? *Japanese Journal of Administrative Science, 31*(1-2), 33–55. https://doi.org/10.5651/jaas.31.33

Liu, Y., Song, Y., Trainer, H., Carter, D., Zhou, L., Wang, Z., & Chiang, J. T. J. (2023). Feeling negative or positive about fresh blood? Understanding veterans' affective reactions toward newcomer entry in teams from an affective events perspective. *Journal of Applied Psychology*, *108*(5), 728–749. https://doi.org/10.1037/apl0001044

Liu, S., Watts, D., Feng, J., Wu, Y., & Yin, J. (2024). Unpacking the effects of socialization programs on newcomer retention: A meta-analytic review of field experiments. *Psychological Bulletin*, *150*(1), 1–26. https://doi.org/10.1037/bul0000422

Lyubykh, Z., Bozeman, J., Hershcovis, M. S., Turner, N., & Shan, J. V. (2022). Employee performance and abusive supervision: The role of supervisor over-attributions. *Journal of Organizational Behavior*, *43*(1), 125–145. https://doi.org/10.1002/job.2560

Markmann, A. (2022, February). Why starting a new job feels so awkward. Harvard Business Review. https://hbr.org/2022/02/why-starting-a-new-job-feels-so-awkward

Marschan-Piekkari, R., Welch, D., & Welch, L. (1999). In the shadow: The impact of language on structure, power and communication in the multinational. *International Business Review*, *8*(4), 421–440. https://doi.org/10.1016/S0969-5931(99)00015-3

Min, S., Humphrey, S., Aime, F., Petrenko, O., Quade, M., & Fu, S. (2021). Dealing with new members: Team members' reactions to newcomer's attractiveness and sex. *The Journal of Applied Psychology, 107*(7), 1115–1129. https://doi.org/10.1037/apl0000872

Neeley, T. B. (2013). Language matters: Status loss and achieved status distinctions in global organizations. *Organization Science*, *24*(2), 476–497. https://doi.org/10.1287/orsc.1120.0739

Piekkari, R., Vaara, E., Tienari, J., & Säntti, R. (2005). Integration or disintegration? Human resource implications of a common corporate language decision in a cross-border merger. *The International Journal of Human Resource Management*, *16*(3), 330–344. https://doi.org/10.1080/0958519042000339534

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 10, pp. 173–220). Academic Press. https://doi.org/10.1016/S0065-2601(08)60357-3

Santos, J. B., Hernandez, J. M., & Leão, W. (2019). Do recovery processes need empowered frontline employees? *International Journal of Operations & Production Management*, *39*(11), 1260–1279. https://doi.org/10.1108/IJOPM-12-2018-0745

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of National Academy Sciences*, *104*(18), 7361–7366. https://doi.org/10.1073/pnas.0702077104

Sharma, G., & Kulshreshtha, K. (2023). Why some leaders qualify for hate: an empirical examination through the lens of followers' perspective. *International Journal of Organizational Analysis*, *31*(2), 430–461. https://doi.org/10.1108/IJOA-08-2020-2369

Srivastava, S. B., Goldberg, A., Manian, V. G., & Potts, C. (2018). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science*, *64*(3), 1348–1364. https://doi.org/10.1287/mnsc.2016.2671

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tenzer, H., & Pudelko, M. (2017). The influence of language differences on power dynamics in multinational teams. *Journal of World Business*, *52*(1), 45–61. https://doi.org/10.1016/j.jwb.2016.11.002

Tenzer, H., Pudelko, M., & Harzing, A. W. (2014). The impact of language barriers on trust formation in multinational teams. *Journal of International Business Studies*, *45*(5), 508–535. https://doi.org/10.1057/jibs.2013.64

Van Maanen, J., & Schein, E. H. (1979). Toward a theory of organizational socialization. *Research in Organizational Behavior, 1*, 209–264.

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, *49*(4), 400–415. https://doi.org/10.1287/mnsc.49.4.400.14430

Wilmot, N. V., Vigier, M., & Humonen, K. (2024). Language as a source of otherness. *International Journal of Cross Cultural Management*, *24*(1), 59–80. https://doi.org/10.1177/14705958231216936

Wolfe, J., Shanmugaraj, N., & Sipe, J. (2016). Grammatical versus pragmatic error: Employer perceptions of nonnative and native English speakers. *Business and Professional Communication Quarterly*, *79*(4), 397–415. https://doi.org/10.1177/2329490616671133

CHAPTER 5 | CAN VERBAL PERFORMANCE APPRAISALS AND MACHINE LEARNING MODELS IMPROVE THE ACCURACY OF PERFORMANCE EVALUATIONS?

Chapter 5:

Can Verbal Performance Appraisals and Machine Learning Models Improve the Accuracy of Performance Evaluations?

This paper, co-authored with Jana Kim Gutt and Kirsten Thommes, is currently being prepared for submission to a journal. An earlier version was presented at the Colloquium on Personnel Economics (COPE) in Amsterdam, the Netherlands, from March 23 to 24, 2023. The paper was also presented at the Academy of Management Conference (AOM) in Boston, USA, from August 4 to 8, 2023, where it ranked among the top 10% of HR Division papers.

Abstract

Performance appraisals are subject to recent debates with one common denominator: most discussions point to their lack of accuracy. In theory, performance appraisals aim to reflect an employee's performance over a certain period of time. However, recent research shows that appraisals fall short in reaching this goal. Although many studies acknowledge the benefits of performance comments over ratings on a scale, research has paid little attention to the potential of performance comments to achieve higher accuracy in performance evaluations. To approach this issue, we conducted a laboratory experiment and collected objective performance data as well as numerical and verbal performance appraisals. In particular, we compile numerical ratings, written comments, and spoken comments on performance from independent evaluators. To make the numbers (assigned ratings) and the comments comparable, we applied a Random Forest algorithm to transfer the comments into numerical ratings (algorithmic ratings). By analyzing each rating (assigned and algorithmic) in relation to the performance, we find evidence that spoken comments reflect performance differences most accurately within a team. Our research offer important insights into how performance appraisals may be approached to reflect objective performance more accurately.

5.1 INTRODUCTION

In recent debates, performance appraisal systems and their shortcomings have been widely discussed (Adler et al., 2016; Brown et al., 2017; Murphy, 2020). In general, performance appraisals aim to reflect an employee's performance during a certain time period. However, low reliability of performance ratings and underlying biases complicate the process of reflecting true performance (Angelovski et al., 2016; Biernat et al., 2012; Bizzi, 2018; Catano et al., 2007; Kamphorst & Swank, 2018; Speer, 2018). In general, one can distinguish between issues concerning evaluations on the general level and fairness considerations. The former make it challenging to identify the strongest and the weakest performing employees, e.g., because of too much compression in ratings or too much leniency. Contrarily, the latter deal with the blurry lines on the group level as team outcomes cannot be attributed to one particular group member, e.g., the fair acknowledgement of achievements or the misattribution of mistakes. Taken together, these factors lead to a considerable decrease in the utility of performance appraisals for individuals and organizations (Murphy, 2020).

Although performance appraisals typically consist of both quantitative rating scales and qualitative comment sections, appraisals have become synonymous with numerical ratings (Brutus, 2010). In many practical approaches, comment sections are intended to provide reasoning for a given rating. However, because comments are time-consuming to process, they often remain blank or go unanalyzed. In an effort to address the shortcomings of performance appraisals, improving numerical rating scales has been the most frequently proposed solution (Adler et al., 2016). Yet, research has shown that improving rating scales does not necessarily reduce bias or enhance accuracy, which has led to increased attention on rater training (Sánchez et al., 2019). Despite this, studies also indicate that training alone is not sufficient to achieve greater accuracy (e.g., Bernardin & Pence, 1980; Sánchez et al., 2019). Recognizing this

limitation early on, Henemann et al. (1987) emphasized the need to give greater consideration to the appraisal format when investigating rating accuracy.

However, few studies in the field of performance appraisals have sought to examine the role of performance appraisal comments when examining rating accuracy. One reason for the neglect of performance comments certainly lies in manageability. Numerical ratings are much easier to produce and process than comments (Brutus, 2010; Speer, 2018). Yet, with recent advances in natural language processing (NLP), texts can be collected, analyzed, and operationalized with low time effort. The benefits of comment over rating scales have been widely acknowledged. Previous studies point out that employees pay more attention to performance comments than to ratings (Smither & Walker, 2004) and that comments provide more reasoning and give context (Speer, 2018). Accordingly, employees have more information on why a certain assessment was made and get a clearer idea of the areas they need to improve.

Nonetheless, the advantages of one over the other do not allow for judgment on the accuracy of either appraisal format. If rating scales and narrative comments are both unrelated to the true performance, they neither hold value for the evaluated employee nor for individuals working with the appraisal data (Decotiis & Petit Andre, 1978). Thus, it is imperative to shed light on the relation between the appraisal format (i.e., ratings and comments) and the true performance. In performance appraisal literature, the term "format" predominantly indicates variations in the design of graphic rating scales (Brutus, 2010). We use a broader definition of the term in this paper, including numerical but also verbal[6] elements. Numerical formats are limited in the topics they address and in their response options (e.g., rating the communication skills of a co-worker on a five-point scale). Contrarily, verbal formats also address predefined topics but

[6] In this paper, we adhere to the Merriam-Webster definition of "verbal" as "relating to or consisting of words" (Merriam-Webster, n.d.). Accordingly, "verbal" includes both spoken and written comments.

respondents are free to choose which aspects they want to mention and to what degree (e.g., evaluating the communication skills of a co-worker by writing a comment).

To judge the accuracy of numerical and verbal evaluation formats, it is crucial to know the extent to which each format is connected to the ground truth, i.e., the real performance that it aims to reflect. Yet, measures of individual performance are usually not available for employees, as they often work in teams, jobs require multi-tasking or the work output cannot directly be measured (e.g., in creative tasks).

With the aim of investigating the relationship between the appraisal format and the performance, we conducted a laboratory experiment. For this purpose, we collected spoken and written comments as well as numerical ratings to test different evaluation formats. Through the use of a trained Random Forest algorithm[7], we predict numerical ratings for the written and spoken comments (algorithmic ratings) that allow for a comparison with the assigned ratings on the one hand and the performance measure on the other. We measure the participants' true performance by documenting their individual number of errors during the task. One important differentiation that we make in our study is analyzing between and within teams. This allows us to examine the evaluation behavior on a broader level but also on a more granular scale.

Our results fall into two broad categories, regarding (1) the level of analysis and (2) the appraisal format. The findings indicate that (1) neither assigned nor algorithmic ratings correlate with performance between teams. Yet, our within-teams-analysis reveals that evaluators differentiate the most (2) when speaking about performance, i.e., the differences in algorithmic ratings derived from spoken comments correlate with performance differences.

The results from the between- and within-team analysis are largely supported by the robustness check conducted using a General AI approach (ChatGPT-4). The robustness check indicates

[7] Within the scope of a previous paper, we have already trained a Random Forest algorithm to translate verbal assessments into numbers (Source blinded to allow a double-blind review process here, source will be added later).

that spoken comments most closely reflect the number of errors at both the between- and within-team levels. In sum, our analysis suggests that the degree of differentiation in evaluations strongly depends on the chosen reference points (between or within teams) and the evaluation format (numerical ratings, written comments, or spoken comments).

Based on our findings, we suggest that performance comments, when combined with machine learning models, can significantly improve the accuracy of performance appraisals. By applying both a widely used supervised learning model (Random Forest) and a Large Language Model (ChatGPT), we account for the interpretability-accuracy trade-off in machine learning models (Feuerriegel et al., 2025). Our findings show that both models yield the same results, indicating that spoken comments relate most closely to performance. Our paper is a first step toward increasing the accuracy of performance appraisals through verbal comments and machine learning.

This paper is organized into the following sections: First, we review the existing literature on rating accuracy in performance appraisals. In the next step, we describe the data and our method. Specifically, we present our experimental design, empirical strategy, and the Random Forest algorithm which we applied to translate the appraisal comments into numerical ratings. Also, we depict our data collection and processing. Next, we describe our data analysis and show our results. In the last section of our paper, we discuss our results and conclude our work.

5.2 RATING ACCURACY

Rating accuracy is understood as a measure that reflects the degree to which behavior is accurately credited (Martell & Leavitt, 2002), or in other words, it describes the extent to which the rating reflects the true score (Funder, 1995; Speer, 2020). The accuracy of performance ratings has previously been described as “an issue of major concern” (Henemann et al., 1987, p. 431) in personnel decision-making, especially since accurate assessments of employee performance seem to pose an “impossible task” (Decotiis & Petit Andre, 1978, p. 635). One

factor influencing rating accuracy is the approach to the true value, i.e. the measures that are applied. In the context of performance assessments, the measures are predominantly distinguished in terms of subjectivity or objectivity (Bayo-Moriones et al., 2020; Hoffman et al., 1991; Rojon et al., 2015). In the following sections, we will review the literature with regard to the dimensions of both approaches.

5.2.1 Subjective measures

Over the last years, numerous studies have tried to estimate rating accuracy. For example, Borman (1979) investigated the relationship between rater training and rating biases and let participants watch videos that showed individuals performing a task. One group of participants underwent rater training in advance, while the other did not. Results showed that prior training did not improve accuracy in performance appraisals. To estimate which group rated the observed performance more accurately, expert ratings served as a true value (Borman, 1979). Similarly, Mero et al. (2003) explored the impact of accountability on rating accuracy and used expert ratings as the true score. In their study, participants also watched videotapes of employees performing a work task. If they were assigned to the treatment group, participants had to justify their performance rating. Being accountable improved the accuracy and the effect was moderated by behavior during the observational period. In line with this approach, Yun et al. (2005) drew on expert ratings when examining the effects of personality, social context, and the rating format on ratings. The performance that the participants were asked to assess was a video-taped teamwork task. They found that a rater's personality (i.e., agreeableness) affects rating accuracy.

In all of these examples, the accuracy of participants' ratings is measured through inter-rater reliability. Expert ratings serve as a reference point and both ratings (participant – expert) focus on one point in time, i.e., work performance shown in a video. In general, rating accuracy has predominantly been determined with different types of inter-rater agreement (Vanhove et al.,

2016). In this respect, inter-rater reliability is not limited to experts' and participants' ratings. Self-ratings as well as supervisor-, subordinate-, or peer-ratings can also be considered when examining inter-rater agreement (Conway & Huffcutt, 1997).

Yet, self-ratings have been found to be more lenient than peer- and supervisor-ratings (e.g., Heidemeier & Moser, 2009; Hoffman et al., 1991; Ohland et al., 2012). This potentially leads to discrepancies between self-ratings and third-party-ratings and ultimately affects the reliability. Funder (1995) calls the inter-rater approach "constructivist" (p. 666) as it only requires inter-rater consensus. It ignores that raters may suffer from the same biases and therefore also overlooks that the inter-rater overlap is not equal to objective accuracy. Similarly, Funder and West (1993) argue that: "two judges can agree with each other perfectly yet both be perfectly wrong" (p. 458). Still, inter-rater agreement influences the confidence in accuracy (Funder, 2012). When different raters come to very different conclusions, the disagreement indicates that one rater must be wrong and vice versa (Funder, 2012). This means that inter-rater agreement serves as an approach for estimating accuracy but does not objectively reflect the rated criteria.

Beyond inter-rater comparisons, intra-rater reliability is another way to estimate rating accuracy (Viswesvaran et al., 1996). For ratings over a period of time, rate-rerate reliability serves as a measure of accuracy, or more specifically for consistency (Dierdorff & Wilson, 2003). Here, the ratings of one rater are compared at different points in time (Viswesvaran et al., 1996). However, performance may change with time and variance in the (re-) rating process may be ascribed to differences in performance (Sturman et al., 2005). Additionally, Cronbach's alpha represents a way to compute the intra-rater reliability for multi-item scales (Viswesvaran et al., 1996).

Yet, all of the presented methods would fail if there were persistent and systematic measurement biases among raters or across time. In other words, when the true value is not really true, this complicates the process of measuring rating accuracy.

5.2.2 Objective measures

Apart from subjective approaches, performance can also be estimated by applying objective measures. Sundvik and Lindeman (1998) investigated rating accuracy and its predictors in a field setting. In particular, they collected supervisory performance ratings and used sales productivity as the true performance value. Since the company gave each salesperson points for every sold service, these points served as a measure of sales productivity. The authors found that rating accuracy varied significantly and that ratings were more accurate the longer the supervisor–subordinate relationship. More recently, Kusterer and Sliwka (2022) studied biases in performance evaluations by comparing subjective appraisals to objective data from a text-entering task. Their results indicate that when accuracy is rewarded, evaluations are less lenient. Also, their study shows that ratings are less accurate and more lenient when bonus payments depend on the evaluation.

Other studies have not explored rating accuracy but applied objective performance measures for different purposes. For example, Sturman and Trevor (2001) examined the relationship between turnover and performance. They used data from a financial services organization and measured performance in terms of fees that are generated from sold loans. Another example of applying objective measures of performance is the well-known study of Lazear (2000) who investigated the influence of monetary incentives on performance. In his work, performance was measured in terms of the number of pieces that employees installed, or more specifically the number of glass units. Mas and Moretti (2009) studied peer effects and assessed performance by counting the number of scanned products per second in a supermarket chain. In their study, Bandiera et al. (2005) investigated the role of relative incentives. By analyzing

data from a fruit farm, they studied the effects of relative daily pay on performance. Here, the daily pay depended on the own performance relative to the average co-worker's performance. The authors determined individual performance by evaluating the kilograms of fruit that workers picked per hour.

Although these studies show that there are certainly ways to objectively measure individual performance, literature also demonstrates that these measures predominantly rely on quantity in a given time. Frederiksen et al. (2017) refer to these settings as "simple settings where simple measures can readily quantify the overall performance of employees reasonably well" (pp. 408). Yet, in most work settings, supervisors can only observe a total work result and cannot observe the individual contribution that an employee made to this result (Mas & Moretti, 2009). This explains why most studies in the performance appraisal literature focused mainly on subjective approaches to estimate the true value.

5.2.3 Relation between subjective and objective measures

Subjective appraisals and objective productivity may be very different measures and there is some evidence that supervisors inflate ratings for low-performers (Bol, 2011; Golman & Bhatia, 2012; Varma et al., 2005; Woods, 2012), especially when signals are noisy (Bolton et al., 2019). While objective measures may only represent one single task, subjective measures are more holistic, yet prone to biases and therefore not necessarily more accurate.

We, therefore, test a different approach to capture performance, namely comments. Since objective key figures remain unavailable in most organizations (Kusterer & Sliwka, 2022), we explore the extent to which some subjective evaluations (i.e., comments) might be more accurate than others (i.e., ratings). Notably, the comments are made by individuals and underlie subjectivity. Yet, they might relate more strongly to performance than numerical ratings because they contain more information on the context and on factors influencing the evaluation. To the best of our knowledge, this method has not fully been explored in the past, typically

because comments cannot be as easily compared to each other as numbers. In comments, evaluation criteria might vary and accordingly, employees' performance cannot be ranked intuitively. However, recent advancements in natural language processing may change the picture (Haller et al., 2022; Hirschberg & Manning, 2015).

The benefits of comments have been widely acknowledged. Smither and Walker (2004) report that employees pay more attention to written texts than to numerical ratings (Smither & Walker, 2004), potentially because texts put the evaluation into context and provide reasons for certain decisions (Speer, 2018). Yet, text processing is demanding as the text needs to be written and coded. This concerns the evaluator writing texts but also managers in human resources and employees reading and processing these reports to take adequate action. For this reason, natural language processing has been on the rise (Speer, 2018). However, the content of narrative comments and the identification of analysis methods have received little research attention (Doldor et al., 2019; Silva & Ribeiro, 2021).

Related research from online reviews suggests that textual assessments might be a richer source of information (e.g., Archak et al., 2011). Brutus (2010), for instance, shows exemplarily how a written comment provides more reasoning and reveals more improvement potential than just a low rating: "Bob is rarely willing to accept our input as to the best strategy to take. In the weekly meetings, he never seriously considers our suggestions about budget allocation. Most of us do not even prepare our numbers anymore" (p. 145).

Appraisal comments may also be advantageous in the sense that linguistic cues may convey the evaluator's level of uncertainty and reveal more information about potential biases and arguments taken into consideration. Accordingly, speaking or writing performance evaluations may influence accuracy. In this paper, we explore two ways of dealing with verbal formats, including written and spoken comments. Written comments may be more feasible in many situations; however, they are time-consuming and the wording may be more reflected. Spoken

comments may be less time-consuming and bear more linguistic cues than thought-through written comments. However, they may be also more ad hoc and less prepared.

For both approaches, we use machine learning techniques to translate the comments into numerical ratings. To test numerical and verbal appraisal formats in terms of rating accuracy, we apply objective measures to approach the true value. For this purpose, we conducted a laboratory experiment in which we documented participants' individual number of errors in a picture-matching task. We use the experimental design introduced by Weber and Camerer (2003). Their experiment is divided into a pre- and post-merger phase. For our study, we use the experimental design of the pre-merger stage.

5.3 DATA AND METHOD

5.3.1 Experimental Design

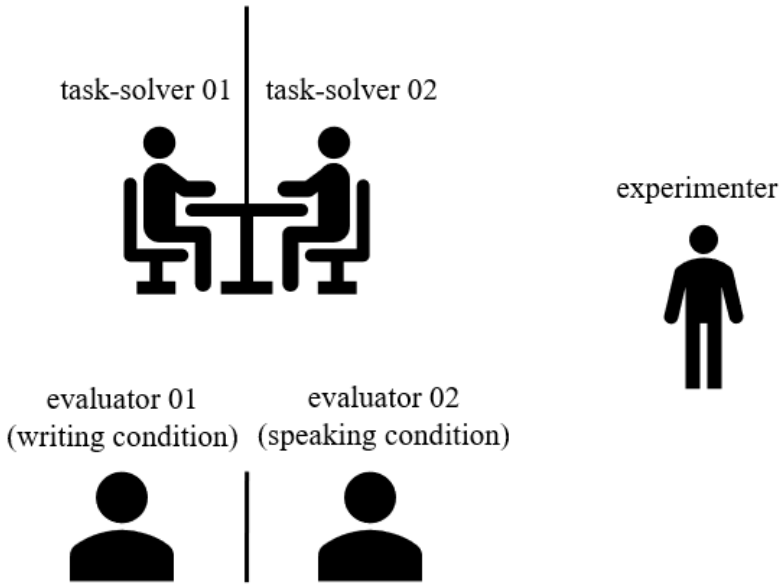
We conducted a laboratory experiment, in which participants were required to solve a picture-matching task. Originally, Weber and Camerer (2003)[8] used the task to investigate the influence of organizational cultures on the success of organizational mergers. In each session, we had four subjects: two serving as task-solvers and two as evaluators. The task-solvers were asked to perform the picture-matching task for 20 rounds in teams of two with a time limit of 150 seconds per round. As they were solving the task, they were observed by the two evaluators. The experimental design is depicted in Figure 5.1.

When arriving at the experiment, participants randomly drew a number, which determined their role (task-solver, evaluator in the writing condition, or evaluator in the speaking condition). Task-solving participants received a performance-based compensation $((1\,Euro - \frac{required\ time}{150}) * accuracy)$, while the evaluators were compensated with 12 Euros per hour.

[8] We use the original task and procedures from Weber and Camerer (2003). However, the authors recommended using updated pictures which we did.

The performance-based compensation was calculated for each of the 20 rounds. Accuracy was coded as a dummy variable that took the value 0 if errors occurred and 1 if no errors occurred in a round.

Figure 5.1: Experimental design



The two task solvers had the same selection of 16 pictures laid out on the table in front of them. All pictures showed different office environments in black and white. The subjects were seated opposite each other but could neither see one another nor one another's pictures. Each subject was assigned one role during each round: manager or employee. The roles switched after every single round, i.e., the participant who acted as a manager in round 1, performed as the employee in round 2 and vice versa. It was the manager's task to describe four predefined pictures to the employee. By listening actively and checking back, the employee tried to identify the four described pictures in the correct order within the time limit of 150 seconds. When the employee indicated that the four pictures were identified, the experimenter gave feedback on the required time and accuracy of the selected pictures. Since the roles changed after each round, the participants performed as the manager and the employee in equal parts. During the task, the

interaction between both participants was audio-recorded. The goal was for the manager to describe and for the employee to identify the four correct pictures in a short amount of time.

Two evaluators were observing the task-solvers during the picture-matching task. Both evaluators were placed so that they could see both picture-matching partners, but could not have made eye contact with one another. The evaluators were not allowed to talk until we asked them for their assessment after the task was finished. During the working phase of the picture-matching partners, the evaluators could observe the whole working performance. After the working phase, both evaluators were brought to separate rooms. Here, they were asked for their assessment of the individual performance of both partners solving the picture-matching task.

To cover both numerical and verbal assessment formats, one evaluator was asked for a written assessment, while the other gave a spoken assessment of the observed performance. The two evaluators were asked about the overall performance of each individual task-solver and their communication skills during the task. We decided to ask for an assessment of the overall performance and the communication skills in order to animate the evaluators to give their appraisal some thought and reflect on both task-solvers in a detailed way. Following performance appraisals in organizations, the written assessments consisted of a numerical rating scale and a comment section. The rating scale ranged from 1 to 6, with 1 indicating a very good and 6 implying a very poor performance. This rating scheme is based on the German school grading system which we expected the participants to be familiar with.

Evaluators who were assigned to the spoken evaluation format answered the same questions by speaking about the performance. The comments were recorded and transcribed. Each picture-matching team was assessed by a new pair of performance evaluators, meaning that each evaluator only evaluated one team instead of one after the other.

The experimenter documented the completion time and the errors during the 20 rounds. The timer was started as soon as the manager said “go” and stopped as soon as the employee

indicated that they identified all four pictures in a round. After each round, the experimenter announced the completion time to the task-solvers and the evaluators simultaneously. Then the task-solver who acted in the role of the employee in that particular round was asked to read the pictures that they identified out loud, e.g., "16, 4, 9, and 3". The experimenter gave immediate feedback by either stating that the pictures were correct or by reporting the order that would have been correct, e.g., "That is incorrect. The correct order would have been 7, 4, 9, and 3".[9] This way, every participant in the room was aware of the performance in each round. Additionally, the experimenter documented the number of errors and who they thought was responsible (see chapter "Data processing"). The number of errors varied between 0 and 4 in each round as the task-solvers had to identify four pictures. We documented three types of errors: incorrect pictures, invalid wording, and time-outs. The experimenter assigned the responsibility for the respective errors on their protocol sheet and did not declare their decision audibly.

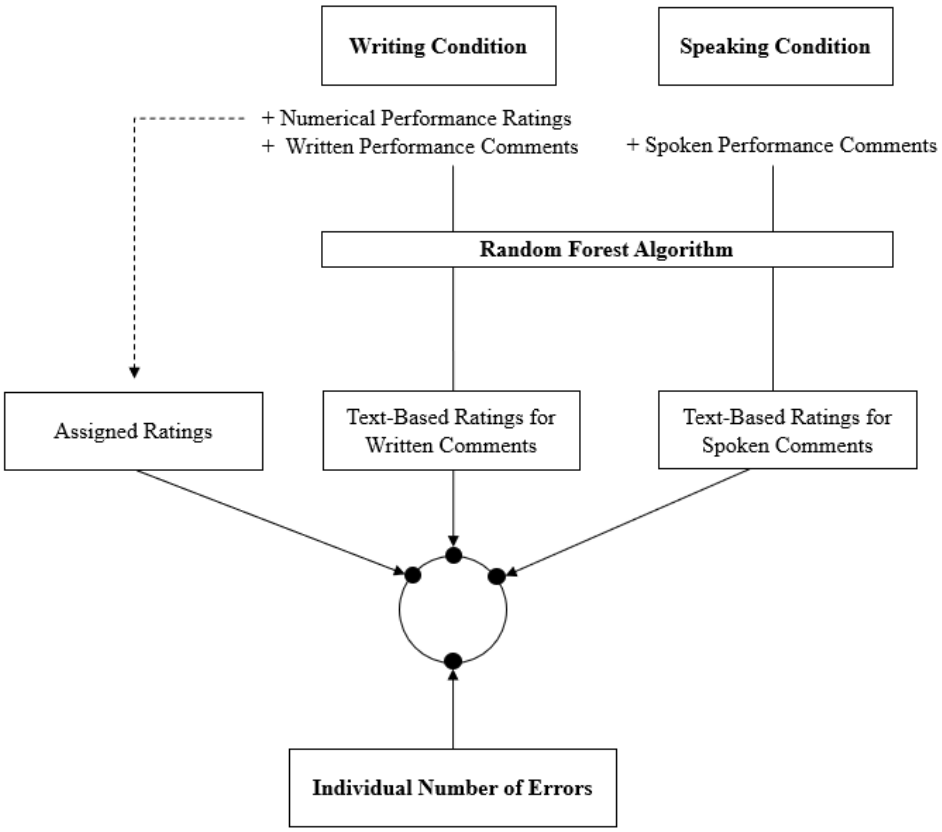
5.3.2 Empirical Strategy

As a true value, we use the individual number of errors during the picture-matching task. Figure 5.2 illustrates our approach.

Through the picture-matching task, we collect two types of data: performance data from the participants who perform the task ("task-solvers") and appraisal data from the participants who evaluate the performance of the task-solvers ("evaluators"). The appraisal data consists of written and spoken evaluations. In the writing condition, participants rated the performance of the task-solvers on a rating scale and were asked to write a short comment on the individual performance. Contrarily, in the speaking condition, participants were only asked for a comment, without giving a numerical rating for the respective performance.

[9] By giving the performance feedback after each round, we made sure that participants did not give up on the round after one error.

Figure 5.2: Empirical approach



To make the three appraisal formats comparable, it is necessary to translate the written and spoken comments into a number. For this purpose, we use a Random Forest algorithm (see chapter "Algorithmic solution") that predicts a numerical rating for each comment (hereafter, algorithmic rating). An alternative strategy that has frequently been used in the literature would be to manually code all comments (e.g., David, 2013; Silva & Ribeiro, 2021; Wilson, 2010). Yet, this approach can be prone to subjectivity and does not scale well for large datasets. Our strategy circumvents these issues as the algorithm does not rely on subjective researcher choices.

With the numerical ratings that were directly assigned by the evaluators and the algorithmic ratings that were predicted by the Random Forest (based on the written and spoken comments), we compare how the ratings relate to true performance (see chapter "Data analysis"). In this

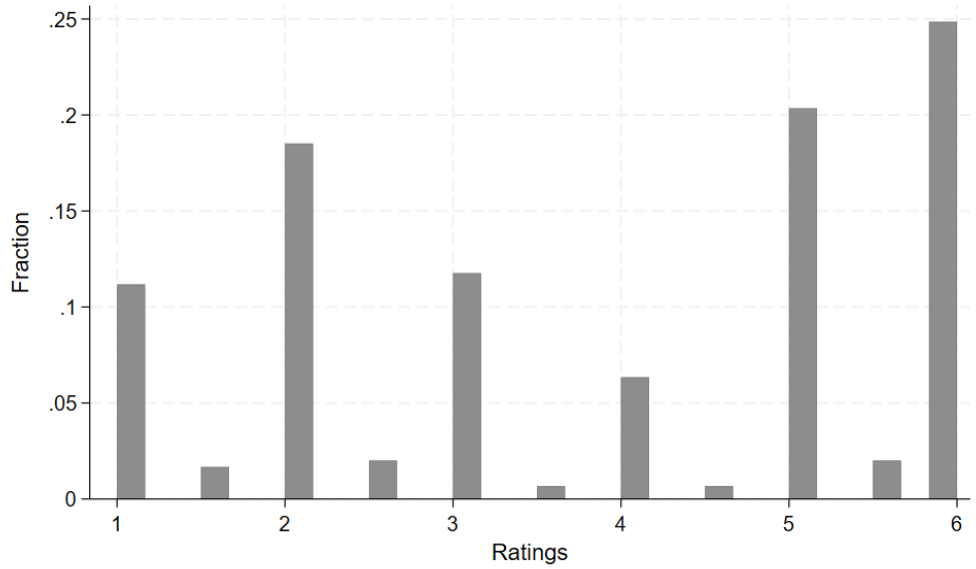
study, we estimate true performance by documenting the individual number of errors caused by the task-solvers during the picture-matching task.

5.3.3 Algorithmic solution for ratings

To connect the richness of qualitative comments with the high usability of numbers, we trained and tested a Random Forest algorithm in a previous work (blind for review). The algorithm was trained on 1,193 audio responses and respective numerical ratings. Specifically, respondents were asked to describe their experiences with the strongest and weakest team members they have worked with in the past in terms of eight predefined subject areas. When speaking about the subject area, e.g., communication skills, participants were also asked to indicate which numerical rating they would give for the described experience. The rating system was based on German school grades. Accordingly, the scale covered the ratings from 1 to 6, with 1 indicating very good and 6 implying very poor performance. To ensure easier interpretation of the results, the ratings were inverted for the analyses (1 = very poor, 6 = very good).

Participants were free to decide who they wanted to describe. For example, they could choose one person for all eight subject areas or different people. Also, they could speak about someone they worked with in the past or somebody they currently work with. By directing the questions to whomever the respondents could think of, we received relatively equally distributed data (Figure 5.3). This data distribution offered a good basis for training an algorithm that can differentiate between good and bad performance. Had we trained it on a dataset that is skewed towards high ratings, the algorithm would arguably be worse at differentiating between good and bad performances.

Figure 5.3: Distribution of numerical ratings in the training data collection



The audio comments were recorded, transcribed, and translated into English. To process the transcripts and develop our model, we used the scikit-learn library in Python. For the prediction of the ratings, we operationalized a feature vector by using Term Frequency Inverse Document Frequency (TF-IDF). Drawing on previous studies, we employed 80 percent of the data for training and 20 percent for testing purposes (e.g., Lei et al., 2016). Concerning the prediction performance of the Random Forest, we received a root mean square error (RMSE) of 0.94. Lower RMSEs stand for higher accuracy. When comparing our prediction performance to the reported performance in other scientific studies (e.g., Ganu et al., 2013), our RMSE falls in the same range. As the Random Forest showed promising performance on a separate test dataset, we are confident in applying the algorithm to the evaluation comments of the picture-matching task.

5.3.4 Data collection

We recruited participants in lectures at a medium-sized university in Europe. For participation, students received monetary compensation. The first round of data collection started in May 2022 and the last round finished in November 2022. From 184 participants who took part in

the experiment, the data of 160 subjects were considered for further analysis. We matched the performance data of 80 task-solvers to the appraisals of 80 evaluators. 40 evaluators were in the writing and 40 in the speaking condition. Consequently, each of the 80 task-solvers was assessed through one written and one spoken evaluation because the evaluators observed both participants in the working team. The descriptive statistics are summarized in table 5.1.

Table 5.1: Descriptive statistics

Variable	N	Mean	Std. Dev.	Min.	Max.
Task-solver gender (female)	80	.50	.50	0	1
Evaluator speaking condition (female)	80	.53	.50	0	1
Evaluator writing condition (female)	80	.58	.50	0	1
No. errors	80	3.05	3.89	0	22
Completion time	80	1014.48	251.27	586	1798
Assigned ratings (scale)	80	5.2	.70	3	6
Algorithmic ratings (written comments)	80	4.12	.32	3.49	4.78
Algorithmic ratings (spoken comments)	80	4.28	.35	3.34	4.87

On average, task-solvers made 3.05 mistakes (SD=3.89) in 20 rounds. This shows that most participants performed quite well. The average completion time for the task was 1,014 seconds (SD=251). With an overall time limit of 3,000 seconds (20 rounds*150 seconds), the fastest team completed the task in 586 seconds, while the slowest team required 1,798 seconds.

5.3.5 Data processing

To explore the relationship between the appraisal format and the true performance, we first analyzed the numerical ratings that were given by the evaluators in the writing condition, the written evaluation comments, and the spoken evaluations. In their evaluations, participants answered two main questions concerning the task-solvers general performance and their communication skills. For both questions, evaluators in the writing condition gave a numerical rating and wrote a comment. Evaluators in the speaking condition answered both questions in a comment without giving a numerical rating.

The spoken comments were recorded, transcribed, and translated into English. We used the machine translator DeepL since it has been found to achieve outstanding translation results (Reber, 2019). Although a word-by-word translation cannot be achieved without changing the meaning of the transcript, these inaccuracies have been reported to be of less concern when words are analyzed individually like in our approach (Reber, 2019). Just like the spoken evaluations, the written comments were also translated using DeepL. Student assistants checked the translations for errors and made sure that the original meaning of the transcripts was not changed. Additionally, the author team performed a final check. With the revised transcripts, we applied the Random Forest algorithm that predicted a numerical rating for each comment.

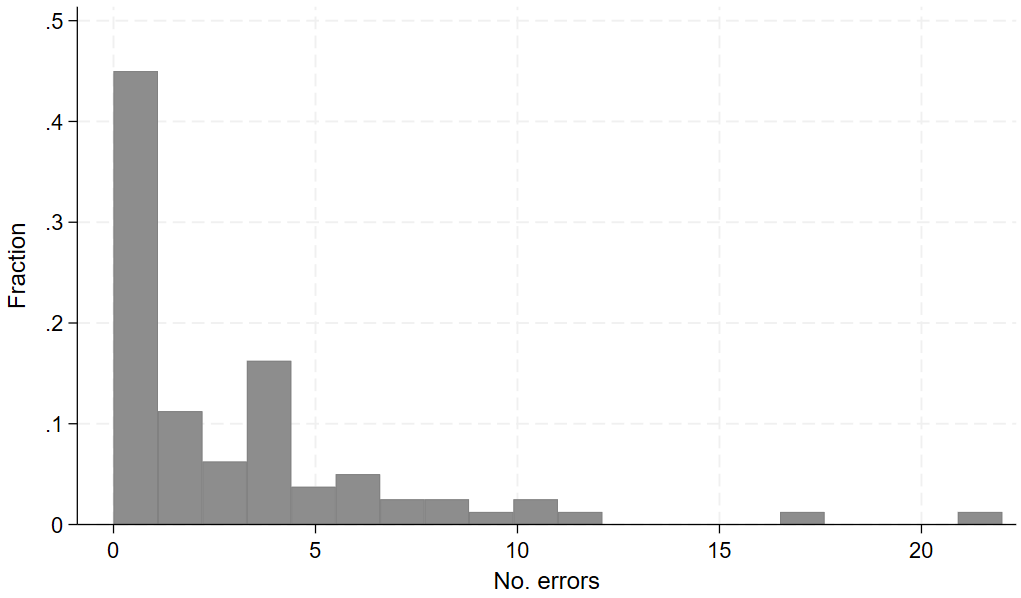
To estimate the true performance, we used the individual number of errors. For each team, the number of errors was documented by the experimenter who also recorded the error responsibility. In some cases, the responsibility was clear (e.g., “Sorry, I described the wrong pictures”), in others, it was more difficult to assign the error to one task-solver. In these cases, the experimenter assigned the error to the whole team.

5.4 DATA ANALYSIS

5.4.1 Between-team analysis

As the data from the descriptive statistics already suggested, the participants performed the task quite well with an average of 3.05 errors in 20 rounds. Figure 5.4 shows that 45 percent of the task-solvers completed the 20 rounds with one error or less. Overall, the majority of participants (56.25 percent) completed the task with 2 errors or less.

Figure 5.4: Distribution of number of errors in 20 Rounds



In comparison, the numerical ratings directly assigned by the evaluators in the writing condition (Figure 5.5), the algorithmic ratings for the written comments (Figure 5.6), and the algorithmic ratings for the spoken comments (Figure 5.7) are distributed as follows:

Figure 5.5: Distribution of the ratings assigned by the evaluators in the writing condition

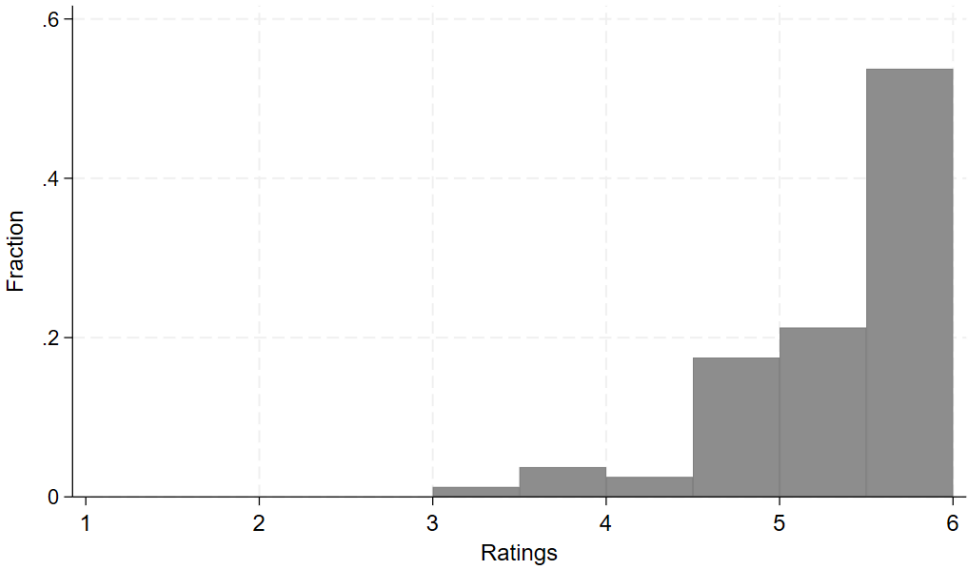


Figure 5.6: Distribution of the text-based ratings for the written comments

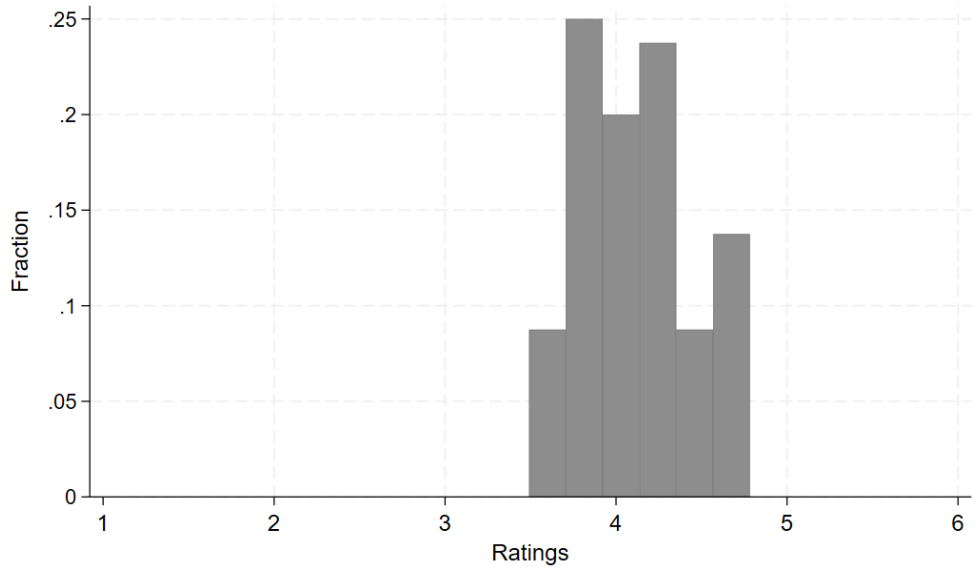
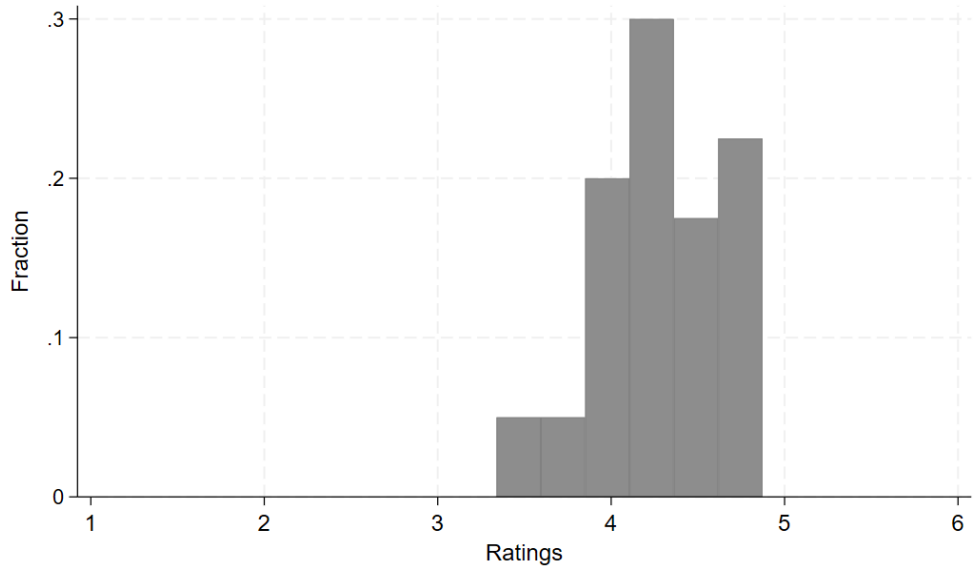


Figure 5.7: Distribution of the text-based ratings for the spoken comments



It is important to note that the distribution of the assigned ratings (Figure 5.5) largely resembles the rating distribution commonly found in the literature: it is highly skewed towards positive ratings (Hu et al., 2017). Despite the fact that the errors are distributed in a similar way (many teams with very few errors), this does not permit us to conclude that numerical ratings assigned by evaluators are an appropriate measure to evaluate performance. For example, even though restaurant ratings are positively skewed, it would be misleading to conclude that there are only

outstanding restaurants in a market. Schoenmueller et al. (2020) present evidence of such skewness in a wide variety of cases, e.g., BlablaCar, Goodreads, and Netflix.

Although the distribution of the performance data and the assigned ratings look roughly similar, the ratings do not correlate with the number of errors (Table 5.2).

Table 5.2: Between-team analysis: correlation matrix for no. errors and ratings

Variable	1.	2.	3.	4.
1. No. errors	1			
2. Assigned ratings (writing condition)	-.14	1		
3. Algorithmic ratings (writing condition)	-.06	.17	1	
4. Algorithmic ratings (speaking condition)	.03	.03	-.00	1

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The general picture emerging from the between-team analysis is that evaluators do not systematically differentiate between high and low performance between teams; regardless of using numerical ratings, written comments, or spoken comments.[10]

5.4.2 Within-team analysis

Since evaluators only observed a single team performing the task and did not get an overview of the other teams' performance, evaluators only had one reference point (the other task-solver). By watching two participants solve the task, evaluators were able to compare the performance of one task-solver relative to the other. This way, evaluators possibly determine high or low performance in relation to the other task-solver and differentiate on a much smaller scale, namely the team level instead of the whole sample. Consequently, we take a closer look inside the teams and investigate to what extent evaluators differentiate between the two task-solvers that they observed. For this purpose, we calculate the performance gap in each team

[10] The results, presented as scatter plots, are included in the Appendix (see figure 5.11 – figure 5.13).

(errors task-solver A – errors task-solver B) and compare this difference to the difference in the two ratings.

The analysis reveals that the difference in the number of errors between two task-solvers positively correlates with the difference in the algorithmic ratings in the speaking condition (Table 5.3).

Table 5.3: Within-team analysis: Correlation matrix for differences in no. errors and ratings

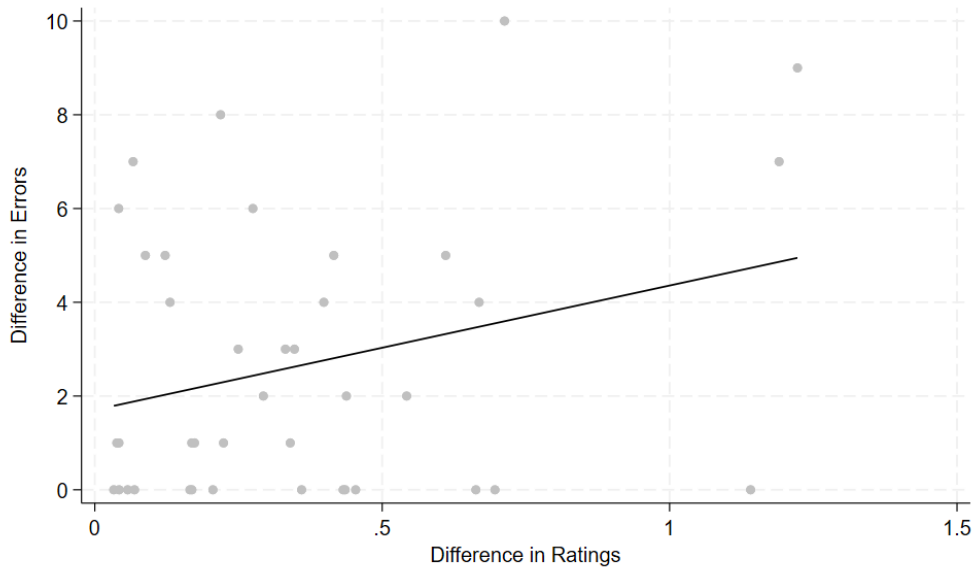
Variable	1.	2.	3.	4.
1. Differences in no. errors	1			
2. Differences in assigned ratings (writing condition)	.14	1		
3. Differences in algorithmic ratings (writing condition)	-.01	.01	1	
4. Differences in algorithmic ratings (speaking condition)	.29*	.07	-.25	1

*p < 0.1, **p<0.05, ***p<0.01

This correlation indicates that the evaluators in the speaking condition differentiate more strongly between two task-solvers the bigger their difference in the number of errors. Large differences in the number of errors imply that one task-solver performed much weaker than the other. Figure 5.8 illustrates the results for the algorithmic ratings in the speaking condition in a scatter plot.[11]

[11] The scatter plots for the assigned ratings and algorithmic ratings (writing condition) are attached in the Appendix (see figures 5.14 and 5.15).

Figure 5.8: Within-team analysis: The relationship between



5.4.3 Robustness check

In the next step, we conduct a robustness check to address concerns that our results may be merely an artifact of our prediction algorithm rather than a general pattern. To this end, we analyze the written and spoken comments using ChatGPT-4. Specifically, we prompt ChatGPT to assess the described performance on a scale from 1 (very good) to 6 (very poor). For the analyses, we inverted the ratings so that 1 corresponds to very poor and 6 to very good. The results of the between-team analysis are presented in table 5.4.

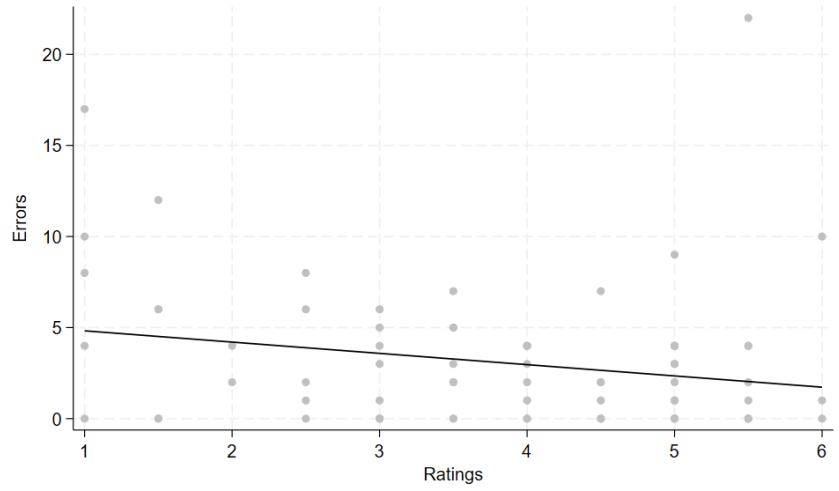
Table 5.4: Between-team analysis: Correlation matrix for no. errors and ratings (using ChatGPT)

Variable	1.	2.	3.	4.
1. No. errors	1			
2. Assigned ratings (writing condition)	-.14	1		
3. Algorithmic ratings (writing condition)	-.10	.05	1	
4. Algorithmic ratings (speaking condition)	-0.22**	.18	-.04	1

*p < 0.1, **p<0.05, ***p<0.01

While most results remain consistent, we find a significant negative correlation between the number of errors and the algorithmic ratings for spoken comments, indicating that evaluators assess task-solvers more negatively when their number of errors is high. Figure 5.9 illustrates the results for the spoken comments using a scatter plot.[12]

Figure 5.9: Between-team analysis: The relationship between errors and ChatGPT ratings (spoken comments)



For the within-team analysis, the results remain largely consistent as well. However, the positive correlation between the difference in the number of errors and the difference in ratings for spoken comments becomes even more pronounced (table 5.5).

Table 5.5: Within-team analysis: Correlation matrix for differences in no. errors and ratings (using ChatGPT)

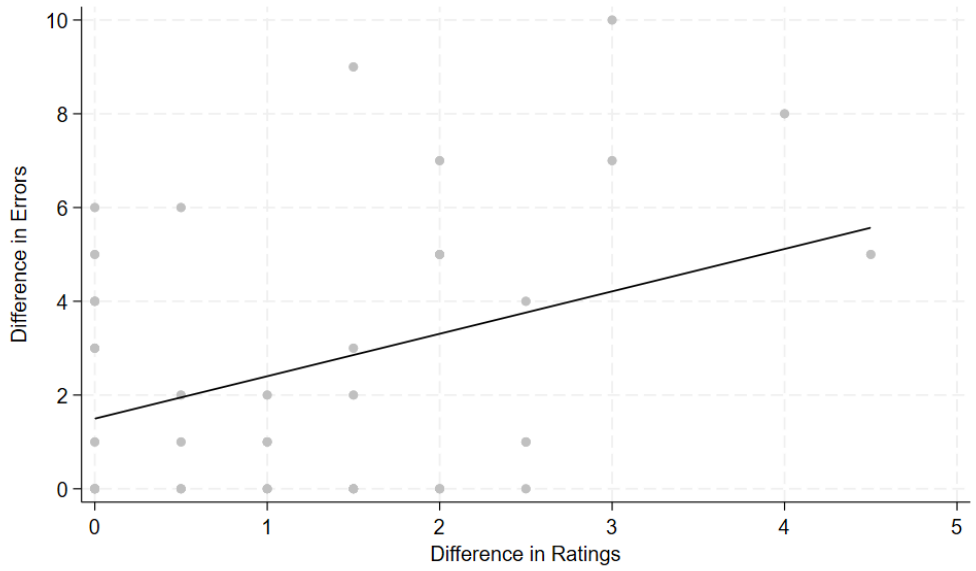
Variable	1.	2.	3.	4.
1. Differences in no. errors	1			
2. Differences in assigned ratings (writing condition)	.14	1		
3. Differences in algorithmic ratings (writing condition)	.15	.23	1	
4. Differences in algorithmic ratings (speaking condition)	.37**	.05	-.15	1

*p < 0.1, **p<0.05, ***p<0.01

[12] The scatter plot for the ChatGPT ratings in the writing condition is attached in the Appendix (see figure 5.16).

The results indicate that evaluators in the speaking condition incorporated differences in the number of errors within a team into their assessments, whereas those in the writing condition did not. Figure 5.10 presents the results for spoken comments as a scatter plot.[13]

Figure 5.10: Within-team analysis: Relationship between difference in errors & difference in ChatGPT ratings (spoken comments)



The robustness check provides strong support for the stability of our findings. Specifically, the results remain consistent when using an alternative algorithm for rating prediction, reinforcing the credibility of the approach. In our specific setting, spoken comments have shown to reflect performance most accurately.

5.4.4 Summary of results

In summary, the between- and within-team analyses provide evidence that spoken appraisal comments are a richer source of performance information than numerical ratings assigned by evaluators. Although the distributions of performance data and assigned ratings look roughly similar, we do not find support for the assumption that assigned numerical ratings are indeed an accurate measure for assessing performance. In particular, our data reveal no correlations

[13] The scatter plot for the ChatGPT ratings in the writing condition is attached in the Appendix (see figure 5.17).

between performance and the three appraisal formats on the between-team level (assigned ratings, algorithmic ratings for written comments, and algorithmic ratings for spoken comments).

To gain a deeper understanding of the underlying dynamics of the performance evaluations, we conducted a more detailed analysis within teams. Between teams, evaluators had no reference point, meaning it was challenging for them to determine good and poor performance as they only observed one team. However, they had a reference point inside their team: the other task-solver. Accordingly, we tested the extent to which evaluators differentiated between both task-solvers in their evaluations with regard to the difference in their performance. First, we analyzed the appraisal formats in terms of the performance difference within teams. We find that the difference in the algorithmic ratings for spoken comments significantly relates to the difference in the number of errors within teams.

Table 5.6: Overview of results

	Between teams	Within teams
Ratings	no significant correlations between task-solvers' no. errors and ratings	significant positive correlation between the differences in no. errors and differences in algorithmic ratings for spoken comments
ChatGPT	significant negative correlation between the differences in no. errors and differences in ratings for spoken comments	significant positive correlation between the differences in no. errors and differences in ratings for spoken comments

To assess the robustness of these results, we conducted an additional analysis using ChatGPT. This analysis reveals a significant negative correlation between the number of individual errors and ratings for spoken comments in the between-team analysis, indicating that evaluators take the number of errors into account when speaking of performance. Additionally, we observe a significant positive correlation between error differences and corresponding rating differences

for spoken comments. Our findings provide evidence that evaluators differentiate more strongly when they have a reference point and that these differentiations predominantly become obvious in spoken comments. Our results are summarized in table 5.6.

5.5 DISCUSSION

Our study aimed to explore the accuracy of appraisal formats with regard to capturing objective performance indicators. While rating accuracy is an issue of high relevance (e.g., Rojon et al., 2015; Speer, 2020; Spence & Keeping, 2011), to the best of our knowledge, most research has investigated the accuracy of performance appraisals from a subjective standpoint. In this context, comparing participants' ratings to expert ratings has been the most common approach (e.g., Borman, 1979; Mero et al., 2003; Yun et al., 2005). As objective performance measures are more challenging to obtain, only a few studies have investigated rating accuracy in performance appraisals by applying objective performance indicators (e.g., Kusterer & Sliwka, 2022; Sundvik & Lindeman, 1998). However, we are not aware of studies that explored the accuracy of numerical and verbal appraisal formats as compared to objective performance.

To explore the relationship between numerical and verbal appraisal formats and objective performance, we let participants solve a picture-matching task under the observation of two evaluators. One evaluator was asked to assess the performance of both task-solvers in writing and the other evaluated the performance by speaking. The individual number of errors served as an objective indicator of performance. We decided to let one pair of evaluators observe one pair of task-solvers only, as opposed to multiple task-solvers in a row, to circumvent sequential order bias (e.g., Chernev, 2011; Criscuolo et al., 2021; Highhouse & Gallo, 1997; Page & Page, 2010). By only watching one team solve the picture-matching task, evaluators solely have one reference point to classify high and low performance, namely the other task-solver in the team. We find evidence that participants tend to give positive ratings when they do not have an overview of the entire participant field but only one team, especially when participants make

relatively few mistakes. Accordingly, the differences that we discover, are most pronounced on the within-team level as this is the only way for evaluators to make comparisons.

In general, this raises the widely discussed question if performance evaluations should rather be made in absolute (compared to a predefined standard) or relative terms (comparing to other employees) (e.g., Blume et al., 2009; Chattopadhayay & Ghosh, 2012; Roch et al., 2007). Ideally, performance ratings would stand for themselves and give individuals the chance to make judgments on qualifications based on the rating at hand. When reading that an employee received a high rating for communication skills, this rating should be a reliable indicator that this particular employee is well suited to work in settings where high communication skills are required. However, individuals have a strong tendency towards positive ratings (e.g., Schoenmueller et al., 2020). Thus, the school grade B might express good communication skills in itself but when the rest of the team receives straight As, the grade B gets a new context and actually indicates that the ratee performed weakest in the team. Although performance appraisals should certainly stand for themselves, we find evidence that it is challenging for individuals to define high and low levels of performance when they do not have a reference point. Only when analyzing the evaluations on the team level, we see that evaluators differentiate between different levels of performance.

Additionally, our paper demonstrates that a Large Language Model (LLM) such as ChatGPT and a supervised learning model (i.e., our algorithm) arrive at the same results when predicting performance ratings. In conducting the robustness check, we follow the recommendation to assess model reliability by comparing complex models (ChatGPT) with simpler, more interpretable approaches (Random Forest) (Feuerriegel et al., 2025). Our finding strengthens confidence in the robustness of the results, as it suggests that both generalized language understanding (LLM) and task-specific learning (supervised model) detect the same underlying patterns. By showing this alignment, the study provides evidence that LLMs may potentially

serve as viable alternatives to supervised models in performance assessment tasks, particularly when the availability of labeled training data is limited. However, it is important to consider that, unlike supervised models, LLMs do not rely on explicitly defined features but rather on large-scale data, making it difficult to understand how they arrive at their predictions (e.g., Budhwar et al., 2023). Additionally, data security is not guaranteed, and firms may ultimately prefer to train their own models based on their specific evaluation criteria. Nonetheless, our findings demonstrate that machine learning approaches may generally contribute to enhancing accuracy and that self-trained models can perform effectively even with small training datasets. Our first approach to investigating the relationship between appraisal formats with regard to objective performance has some limitations. In our study, participants solved the task quite well and only made a few errors. Correspondingly, studies with a bigger heterogeneity in performance data might come to different results. Additionally, our results may depend on the operationalization of objective performance. In this study, we used the individual number of errors and assigned equal weight to the errors. However, organizational settings have a much higher complexity than laboratory experiments. Accordingly, some errors have more serious consequences than others and are weighted differently. Also, errors are more difficult to measure, and cannot easily be assigned to one person. As a consequence, our findings may have limited generalizability to settings where errors are harder to assess and need to be tested in a field experiment.

5.6 CONCLUSION

In this paper, we sought to investigate the accuracy of numerical and verbal appraisal formats in the reflection of objective performance. Our results show that between teams, neither evaluation format accurately predicts objective performance. Within teams, however, algorithmic ratings for spoken comments are superior in predicting objective performance compared to the other evaluation formats. This underscores the importance of harnessing

qualitative information in the evaluation process. Although texts are more time-consuming regarding production and processing, they have a higher informative value when it comes to differentiating performance. Accordingly, this also underlines the importance of testing methods to quantify texts in organizational settings: performance appraisals aim to get accurate information on individual performance while still obtaining data that are easy to manage. Without any insight into who performed better than the other, organizations are missing an essential basis for decision-making. This not only concerns decisions regarding who is the best fit for a certain task but also about who requires training and lacks important skills.

Especially when applying algorithms to transfer textual information into numerical ratings, it is crucial to build trust in machine learning and to highlight the connection between the human evaluation and the numerical prediction of the algorithm. The evaluation still stems from a human evaluator while the algorithm is used to assign a number to the described performance. Also, the prediction by the algorithm does not categorically exclude the option to conduct final manual checks on the ratings.

While there is wide research interest in rating accuracy, the literature on performance appraisals shows a gap in estimating true performance with objective measures. With this study, we take a first step in closing this gap and compare the accuracy of three different appraisal formats (assigned ratings, algorithmic ratings for written comments, and algorithmic ratings for spoken comments) with the individual number of errors in a picture-matching task. By using an algorithm to transfer text into numerical ratings, we point out the usefulness of machine learning and show an exemplary use case of machine learning in an organizational setting.

Acknowledgements

This research was funded by the Ministry of Economic Affairs, Innovation, Digitalization, and Energy of the State of North Rhine-Westphalia [005-2001-0017].

5.7 APPENDIX

Figure 5.11: Between-team analysis: The relationship between errors and assigned ratings

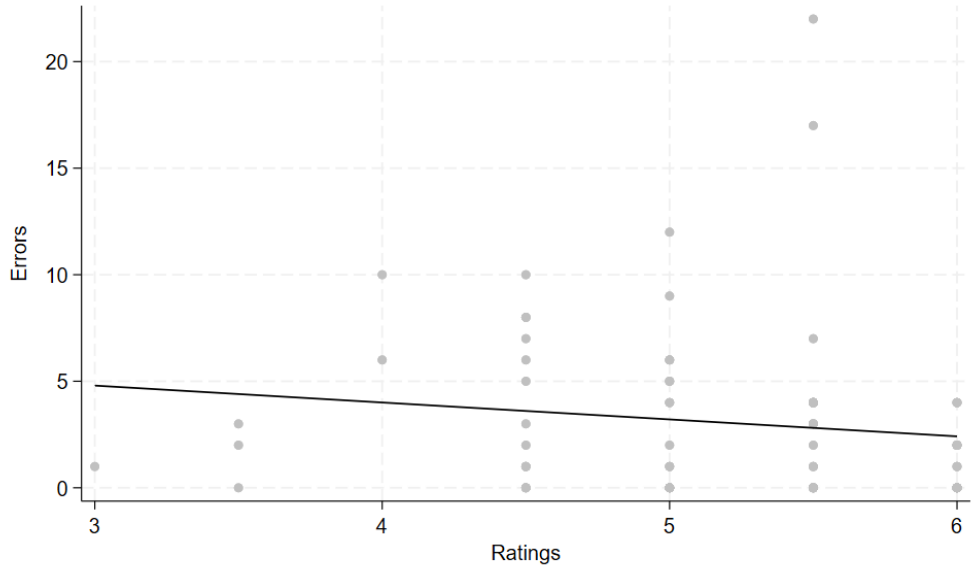


Figure 5.12: Between-team analysis: The relationship between errors and algorithmic ratings (written comments)

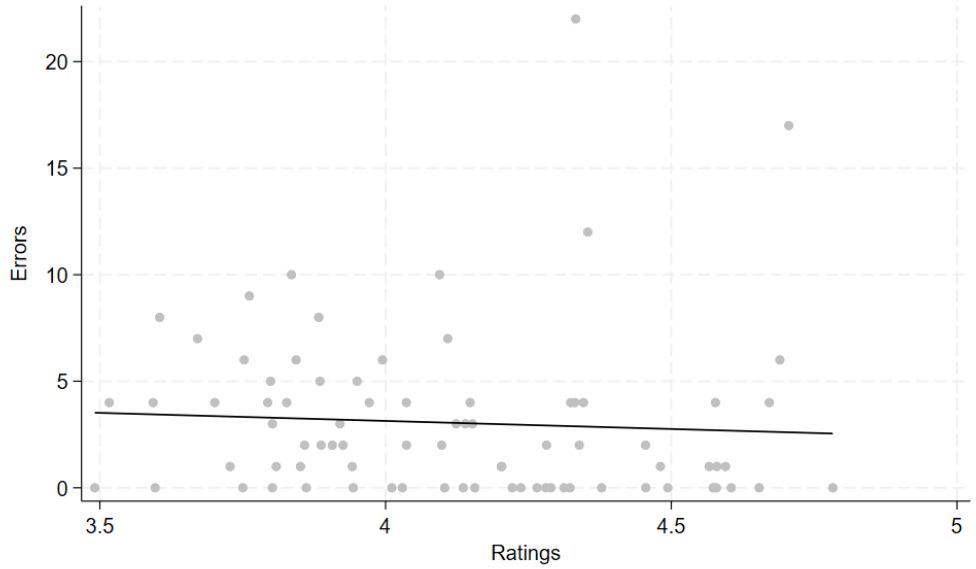


Figure 5.13: Between-team analysis: The relationship between errors and algorithmic ratings (spoken comments)

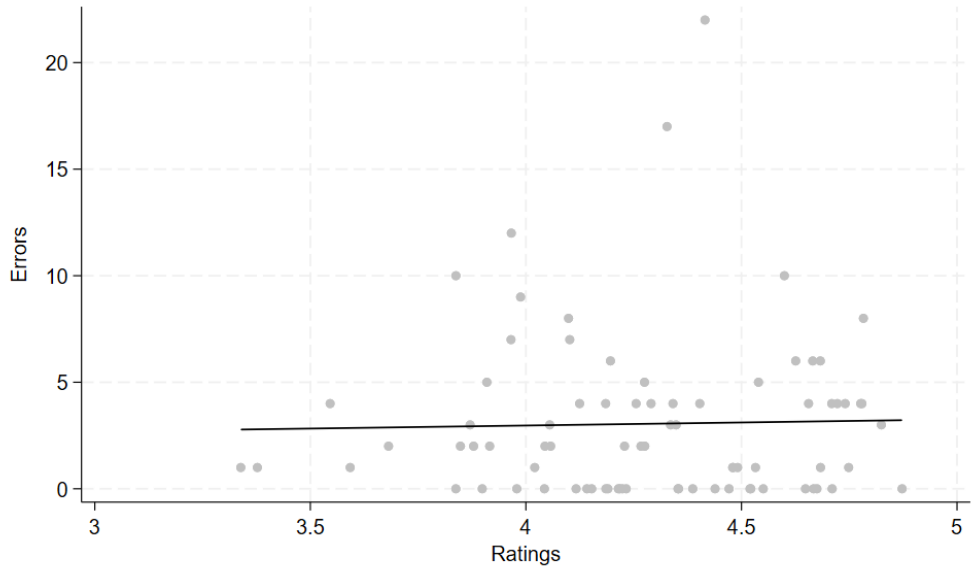


Figure 5.14: Within-team analysis: The relationship between the difference in errors and the difference in assigned ratings

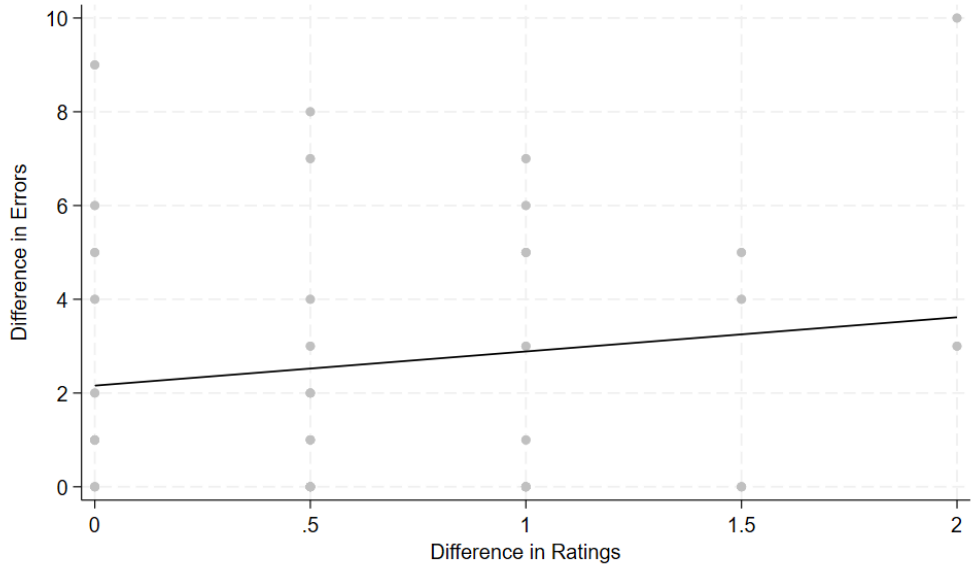


Figure 5.15: Within-team analysis: The relationship between the difference in errors and the difference in algorithmic ratings (written comments)

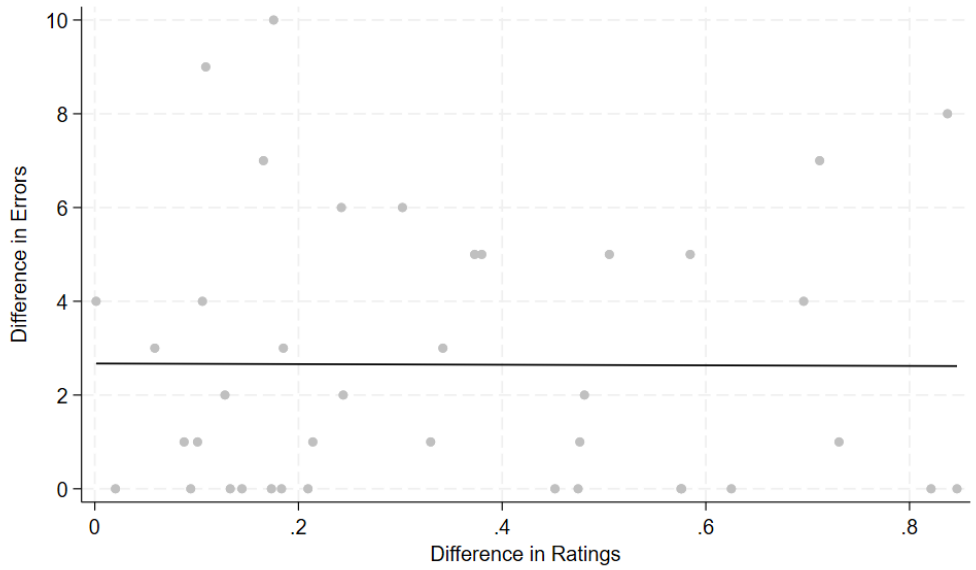


Figure 5.16: Between-team analysis: The relationship between errors and ChatGPT ratings (written comments)

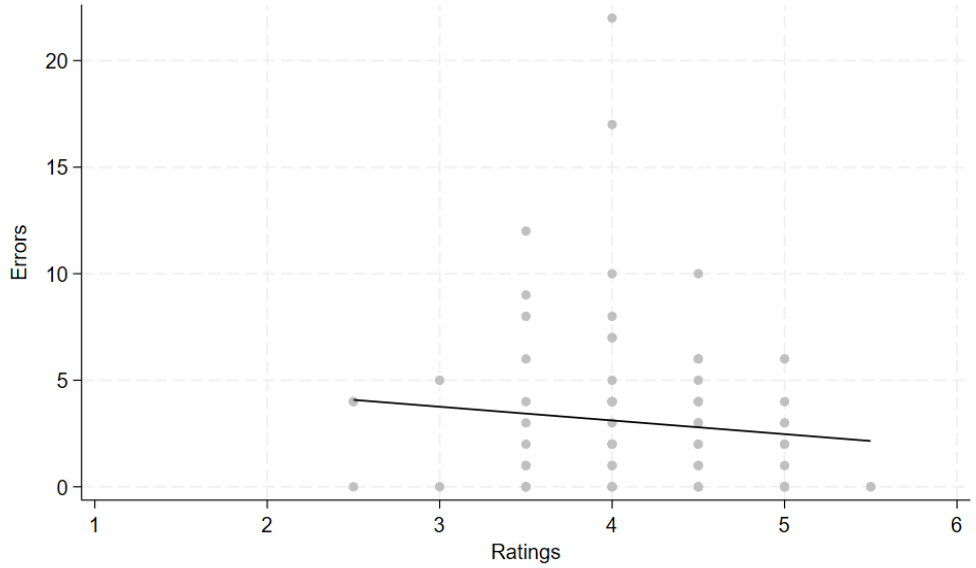
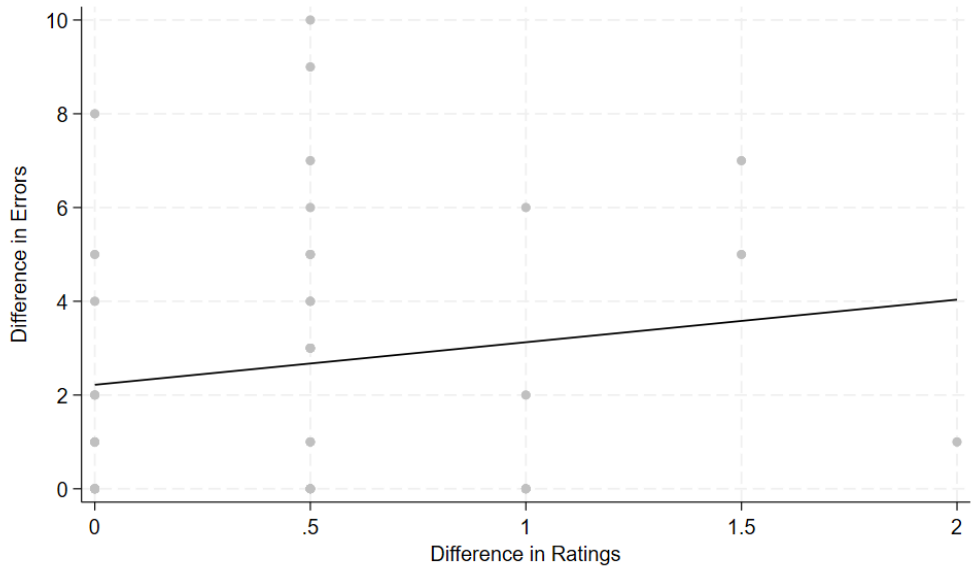


Figure 5.17: Within-team analysis: The relationship between the difference in errors and the difference in ChatGPT ratings (written comments)



REFERENCES

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting rid of performance ratings: Genius or folly? A debate. *Industrial and Organizational Psychology*, *9*(2), 219–252. https://doi.org/10.1017/iop.2015.106

Angelovski, A., Brandts, J., & Sola, C. (2016). Hiring and Escalation Bias in Subjective Performance Evaluations: A Laboratory Experiment. *Journal of Economic Behavior & Organization*, *121*, 114–129. https://doi.org/10.1016/j.jebo.2015.10.012

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485–1509. https://doi.org/10.1287/mnsc.1110.1370

Bandiera, O., Barankay, I., & Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics, 120*(3), 917–962. https://doi.org/10.1162/003355305774268192

Bayo-Moriones, A., Galdon-Sanchez, J. E., & Martinez-de-Morentin, S. (2020). Performance appraisal: Dimensions and determinants. *The International Journal of Human Resource Management, 31*(15), 1984–2015. https://doi.org/10.1080/09585192.2018.1500387

Bernardin, H. J., & Pence, E. C. 1980. Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*(1), 60–66. https://doi.org/10.1037/0021-9010.65.1.60

Biernat, M., Tocci, M. J., & Williams, J. C. (2012). The language of performance evaluations: Gender-based shifts in content and consistency of judgement. *Social Psychological and Personality Science*, *3*(2), 186–192. https:///doi.org/10.1177/1948550611415693

Bizzi, L. (2018). The problem of employees' network centrality and supervisors' error in performance appraisal: A multilevel theory. *Human Resource Management*, *57*(2), 515–528. https://doi.org/10.1002/hrm.21880

Blume, B. D., Baldwin, T. T., & Rubin, R. S. (2009). Reactions to different types of forced distribution performance evaluation systems. *Journal of Business and Psychology*, *24*(1), 77–91. https://doi.org/10.1007/s10869-009-9093-5

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review, 86*(5), 1549–1575. https://doi.org/10.2308/accr-10099

Bolton, G. E., Kusterer, D. J., & Mans, J. (2019). Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science, 65*(11), 5371–5391. https://doi.org/10.1287/mnsc.2018.3191

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*(4), 410–421. https://doi.org/10.1037/0021-9010.64.4.410

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods, 20*(1), 121–148. https://doi.org/10.1177/1094428116668036

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, *20*(2), 144–157. https://doi.org/10.1016/j.hrmr.2009.06.003

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Lee Cooke, F., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., Rogelberg, S., Saunders, M. N. K., Tung, R. L., & Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal, 33*(3), 606–659. https://doi.org/10.1111/1748-8583.12524

Catano, V. M., Darr, W., & Campbell, C. A. (2007). Performance appraisal of behavior-based competencies: A reliable and valid procedure. *Personnel Psychology, 60*(1), 201–230. https://doi.org/10.1111/j.1744-6570.2007.00070.x

Chattopadhayay, R., & Ghosh, A. K. (2012). Performance appraisal based on a forced distribution system: Its drawbacks and remedies. *International Journal of Productivity and Performance Management, 61*(8), 881–896. https://doi.org/10.1108/17410401211277138

Chernev, A. (2011). Semantic anchoring in sequential evaluations of vices and virtues. *Journal of Consumer Research*, *37*(5), 761–774. https://doi.org/10.1086/656731

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, *10*(4), 331–360. https://doi.org/10.1207/s15327043hup1004_2

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2021). The sequence effect in panel decisions: Evidence from the evaluation of research and development projects. *Organization Science, 32*(4), 987–1008. https://doi.org/10.1287/orsc.2020.1413

David, E. M. (2013). Examining the role of narrative performance appraisal comments on performance. *Human Performance, 26*(5), 430–450. https://doi.org/10.1080/08959285.2013.836197

Decotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *The Academy of Management Review, 3*(3), 635–646. https://doi.org/10.2307/257552

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology*, *88*(4), 635–646. https://doi.org/10.1037/0021-9010.88.4.635

Doldor, E., Wyatt, M., & Silvester, J. (2019). Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders. *The Leadership Quarterly*, *30*(5), Article 101308. https://doi.org/10.1016/j.leaqua.2019.101308

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 96–111. https://doi.org/10.1038/s44159-024-00392-z

Frederiksen, A., Lange, F., & Kriechel, B. (2017). Subjective performance evaluations and employee careers. *Journal of Economic Behavior & Organization*, 134, 408–429. https://doi.org/10.1016/j.jebo.2016.12.016

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177–182. https://doi.org/10.1177/0963721412445309

Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality, 61*(4), 457–476. https://doi.org/10.1111/j.1467-6494.1993.tb00778.x

Ganu, G., Kakodkar, Y., & Marian, A. (2013). Improving the quality of predictions using textual information in online user reviews. *Information Systems, 38*(1), 1–15. https://doi.org/10.1016/j.is.2012.03.001

Golman, R., & Bhatia, S. (2012). Performance evaluation inflation and compression. *Accounting, Organizations and Society, 37*(8), 534–543. https://doi.org/10.1016/j.aos.2012.09.001

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). Survey on automated short answer grading with deep learning: From word embeddings to transformers (Working Paper). https://doi.org/10.48550/arXiv.2204.03503

Heidemeier, H., & Moser, K. (2009). Self-other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology, 94*(2), 353–370. https://doi.org/10.1037/0021-9010.94.2.353

Henemann, R. L., Moore, M. L., & Wexley, K. N. (1987). Performance rating accuracy: A critical review. *Journal of Business Research, 15*(5), 431–448.

Highhouse, S., & Gallo, A. (1997). Order effects in personnel decision making. *Human Performance, 10*(1), 31–46. https://doi.org/10.1207/s15327043hup1001_2

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hoffman, C. C., Nathan, B. R., & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. *Personnel Psychology, 44*(3), 601–618. https://doi.org/10.1111/j.1744-6570.1991.tb02405.x

Hu, N., Pavlou, P. A., & Zhang, J. (2017). On self-selection biases in online product reviews. *MIS Quarterly, 41*(2), 449–471. https://doi.org/10.25300/MISQ/2017/41.2.06

Kamphorst, J. J.A., & Swank, O. H. (2018). The role of performance appraisals in motivating employees. *Journal of Economics & Management Strategy, 27*(2), 251–269. https://doi.org/10.1111/jems.12241

Kusterer, D., & Sliwka, D. (2022*). Social preferences and rating biases in subjective performance evaluations* (IZA Discussion Paper No. 15496). Institute of Labor Economics.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review, 90*(5). 1346–1361

Lei, X., Qian, X., & Zhao, G. (2016). Rating prediction based on social sentiment from textual reviews. *IEEE Transactions on Multimedia, 18*(9), 1910–1921. https://doi.org/10.1109/TMM.2016.2575738

Martell, R. F., & Leavitt, K. N. (2002). Reducing the performance-cue bias in work behavior ratings: Can groups help? *Journal of Applied Psychology, 87*(6), 1032–1041. https://doi.org /10.1037/0021-9010.87.6.1032

Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review, 99*(1), 112–145.

Mero, N. P., Motowidlo, S. J., & Anna, A. L. (2003). Effects of accountability on rating behavior and rater accuracy. *Journal of Applied Social Psychology, 33*(12), 2493–2514. https://doi.org/10.1111/j.1559-1816.2003.tb02777.x

Merriam-Webster. (n.d.). *Verbal*. *Merriam-Webster dictionary*. Retrieved February 9, 2025, from https://www.merriam-webster.com/dictionary/verbal

Murphy, K. R. (2020). Performance evaluation will not die, but it should. *Human Resource Management Journal, 30*(1), 13–31. https://doi.org/10.1111/1748-8583.12259

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., Layton, R. A., Pomeranz, H. R., & Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation. *Academy of Management Learning & Education, 11*(4), 609–630. https://doi.org/10.5465/amle.2010.0177

Page, L., & Page K. (2010). Last shall be first: A field study of biases in sequential performance evaluation on the idol series. *Journal of Economic Behavior & Organization, 73*(2), 186–198. https://doi.org/10.1016/j.jebo.2009.08.012

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text

corpora. *Communication Methods and Measures, 13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. (2007). Absolute vs relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15*(3), 302–316. https://doi.org/10.1111/j.1468-2389.2007.00390.x

Rojon, C., McDowall, A., & Saunders, M. N. K. (2015). The relationships between traditional selection assessments and workplace performance criteria specificity: A comparative meta-analysis. *Human Performance, 28*(1), 1–25. https://doi.org/10.1080/08959285.2014.974757

Sánchez, C. R., Díaz-Cabrera, D., & Hernández-Fernaud, E. (2019). Does effectiveness in performance appraisal improve with rater training? *PLoS One, 14*(9), Article e0222694. https://doi.org/10.1371/journal.pone.0222694

Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research, 57*(5), 853–877. https://doi.org/10.1177/0022243720941832

Silva, V. V. M., & Ribeiro, J. L. D. (2021). A discussion on using quantitative or qualitative data for assessment of individual competencies. *Personnel Review, 50*(6), 1460–1478. https://doi.org/10.1108/PR-08-2019-0444

Smither, J. W., & Walker, A. G. (2004). Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time? *The Journal of Applied Psychology, 89*(3), 575–581. https://doi.org/10.1037/0021-9010.89.3.575

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology, 71*(3), 299–333. https://doi.org/10.1111/peps.12263

Speer, A. B. (2020). Judging performance – General mental ability and the convergence of operational performance ratings. *Journal of Personnel Psychology, 19*(1), 44–48. https://doi.org/10.1027/1866-5888/a000244

Spence, J. R., & Keeping, L. (2011). Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review, 21*(2), 85–95. https://doi.org/10.1016/j.hrmr.2010.09.013

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test–retest reliability of employee job performance ratings. *Journal of Applied Psychology, 90*(2), 269–283. https://doi.org/10.1037/0021-9010.90.2.269

Sturman, M. C., & Trevor, C. O. (2001). The implications of linking the dynamic performance and turnover literatures. *Journal of Applied Psychology, 86*(4), 684–696. https://doi.org/10.1037/0021-9010.86.4.684

Sundvik, L., & Lindeman, M. (1998). Performance rating accuracy: Convergence between supervisor assessment and sales productivity. *International Journal of Selection and Assessment, 6*(1), 9–15. https://doi.org/10.1111/1468-2389.00067

Vanhove, A. J., Gibbons, A. M., & Kedharnath, U. (2016). Rater agreement, accuracy, and experienced cognitive load: Comparison of distributional and traditional assessment approaches to rating performance. *Human Performance, 29*(5), 378–393. https://doi.org/10.1080/08959285.2016.1192632

Varma, A., Pichler, S., & Srinivas, E. S. (2005). The role of interpersonal affect in performance appraisal: Evidence from two samples – the US and India. *The International Journal of Human Resource Management, 16*(11), 2029–2044. https://doi.org/10.1080/09585190500314904

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557–574. https://doi.org/10.1037/0021-9010.81.5.557

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science, 49*(4), 400–415. https://doi.org/10.1287/mnsc.49.4.400.14430

Wilson, K. Y. (2010). An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations, 63*(12), 1903–1933. https://doi.org/10.1177/0018726710369396

Woods, A. (2012). Subjective adjustments to objective performance measures: The influence of prior performance. *Accounting, Organizations and Society, 37*(6), 403–425. https://doi.org/10.1016/j.aos.2012.06.001

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment*, *13*(2), 97–107. https://doi.org/10.1111/j.0965-075X.2005.00304.x