# Essays on Interpersonal Trust and Trustworthiness Detection

Inauguraldissertation

zur

Erlangung des Doktorgrades

der

Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

2021

vorgelegt

von

Sebastian Alexander Siuda

aus

München

| | |
|---|---|
| Referent: | Prof. Dr. Detlef Fetchenhauer |
| Korreferenten: | PD Dr. Thomas Schlösser |
| | Prof. Dr. Erik Hölzl |

Tag der Promotion: 23. November 2021

# Acknowledgments

I would like to thank Detlef Fetchenhauer, my primary research supervisor, and Thomas Schlösser, my co-supervisor and second member of my thesis committee, for their patient guidance and support, their advice and critical comments, and their openness and sincerity. It is unmistakably connected with them that I have learned not to confine human behavior to 2x2 ANOVA designs but to cherish the complexities of life outside the comfort of one's own laboratory. I am further grateful to Erik Hölzl for posting as third member of my thesis committee and giving me more hope and strength than he realizes with his openhearted, empathetic, and helpful advice.

Furthermore, I would like to thank Anne-Sophie Lang for the many discussions, private and research-related, that have shaped my perspective on human decision-making. I am equally grateful to Daniel Ehlebracht and Carolina Dahlhaus who have enriched my time at the University of Cologne with their invaluable feedback and their easy-going nature. My special thanks also go to David Dunning and Corinna Michels who have substantially improved my work with their helpful comments and remarks. Moreover, I would like to thank all my colleagues who have invested their time and energy into reading and critiquing my work during the many research seminars. This dissertation would not exist without their efforts. Additionally, my thank goes to Ingrid Kampkötter for ensuring that my studies' funding never run out even during a world-wide pandemic – and to all friends and colleagues at the Cologne Graduate School where I spent my first three and a half years of my PhD life.

I am deeply grateful for the unconditional love and support from my parents, Anneliese and Sigismund Siuda, who have given me the strength and security to pursue my hopes and dreams. I am aware of the great privilege of having them as my parents. My special thanks also go to my brother Fabian Siuda who I have been able to lean on during my whole life. Lastly, I want to thank all my friends for the support throughout the years.

Finally, I am forever grateful to my soulmate and partner Lisa Müsse. It is incredibly rare to find a partner who combines warmth, charm, intelligence, and beauty and I am endlessly thankful for having you in my life. Thank you for giving me strength and confidence throughout the past four years. This dissertation is dedicated to you.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Opening Remarks

## 1.1 Introduction

Trust is a vital factor for the functioning of society. It aids cooperation between individuals (Simpson, 2007), guarantees the success and health of organizations (Kramer, 1998), and improves the economic growth of nations (Fetchenhauer & van der Vegt, 2001). Trust even influences how much people are willing to adhere to health care recommendations such as vaccine uptake (Larson et al., 2018). The current coronavirus pandemic currently illustrates the important role of trust. As of September 2021, vaccination rates have only slowly progressed in most developed countries even though vaccines are readily available at zero cost for the individual. Vaccination rates appear to be particularly low in countries characterized by low trust in their countries' healthcare system; whereas 60.5% of the German population is fully vaccinated, only 24.8% of the Russian population has received full vaccination (Bloomberg, 2021). Although there are many contributing factors to vaccine uptake, it seems likely that trust in the healthcare system, which has traditionally been lower in the Russian (43%) than in the German (62%) population (Edelman, 2021), might contribute to the slow vaccine uptake.

Focusing on interpersonal trust, the decision to trust another person (e.g., a healthcare provider) does not appear to follow the same rationale as non-social decisions such as risky financial decisions. This is true for at least two reasons. First, trust behavior is largely principled and less strongly influenced by objective probabilities. People playing the trust game (Berg et al., 1995) regularly trust more than they objectively should, given their pessimistic trustworthiness expectations (Fetchenhauer & Dunning, 2009). Moreover, changing the probability of being matched with a trustworthy interaction partner influences trust behavior to a much smaller extend than changing the win probability of an otherwise identical non-social lottery (Fetchenhauer & Dunning, 2012). Thus, rather than being the product of a rational analysis of potential costs and benefits, trust behavior appears to be rather principled and motivated by moral norms not to offend another person's character. Second, trust behavior is largely driven by emotional reactions to potential violations of the aforementioned moral norms. Especially the feeling of agitation at the thought of distrusting another person appears to explain why people are more risk seeking in trust games versus lottery games (Schlösser et al., 2016). Taken together, trust behavior appears not only to be influenced by a rational weighing of

risks but also by feelings invoked at the thought of breaking a moral norm not to question another person's character.

In addition to normative and emotional characteristics of the trustor, trust is influenced by how a particular trustee is perceived. That is, people trust another person more or less depending on how trustworthy that person appears. People readily form congruent impressions of others' trustworthiness (Todorov et al., 2009) and these impressions have been shown to influence criminal sentencing decisions (Wilson & Rule, 2015, 2016) and economic cooperation (Chang et al., 2010). There are two alternatives for why people agree on who appears trustworthy. On the one hand, people might agree *because* trustworthiness is accurately detectable even after minimal exposure. If this were the case, exclusions from economic transactions might actually be reasonable. On the other hand, people might have congruent trustworthiness impressions *even though* these impressions are inaccurate. Such false but congruent perceptions could, for example, be caused by emotion overgeneralization effects (Todorov et al., 2008) or self-fulfilling effects of facial impressions (Hong et al., 2021; Todorov, Olivola, et al., 2015). Unfortunately, while progress has been made in understanding how trustworthiness impressions arise, the picture is much less clear regarding the accuracy of these impressions (Bonnefon et al., 2015; Todorov, Funk, & Olivola, 2015; Wilson & Rule, 2017).

## 1.2 Outline of the Thesis

It is at this stage of the research on interpersonal trust and trustworthiness detection that the joint research work together with Detlef Fetchenhauer, Thomas Schlösser, David Dunning, and Anne-Sophie Lang began. My dissertation guides through the joint work and consists of three independent studies regarding the potential effect of foreign language on trust behavior (Chapter 2), the overall evidence in the extant literature on the accuracy of trustworthiness impressions (Chapter 3), and an empirical investigation when and through which mechanism trustworthiness detection is accurate (Chapter 4). The dissertation ends with a short integrative discussion of the results presented in Chapters 2 to 4 as well as implications for future research (Chapter 5).

Chapter 2 is joint work with Anne-Sophie Lang, Detlef Fetchenhauer, and David Dunning. It focuses on trust decisions toward an unknown and unobservable interaction

partner. As already mentioned, trust decisions in these situations are largely principled, rely on people's injunctive moral norms and should therefore be stable over time and not be easily influenced. However, trust behavior is also driven by emotional reactions to potential norm violations (e.g., agitation at the thought of distrusting), which indicates that changes in the emotional processing of norm violations might influence trust behavior. One potential way to influence the emotional processing during trust situations is to present the trust situation either in a native or in a foreign language. The use of foreign language has been shown to decrease the intensity of emotional reactions (Dewaele, 2004) so that childhood reprimands and swearwords in a foreign language are perceived as less emotionally arousing than the same words in one's native language (Caldwell-Harris, 2015; Harris et al., 2003). Moreover, these foreign-language effects have been shown to influence individuals' judgment and decision-making, for example in the area of moral judgment (Costa, Foucart, Hayakawa, et al., 2014), loss aversion (Keysar et al., 2012), and certain norm violations (Geipel et al., 2015b). As trust behavior is also largely driven by immediate emotional reactions to potential norm violations, foreign language might attenuate these reactions and thereby lead to a decrease in people's overall trust rate. On the other hand, it is not the emotional reactions per se that lead people to trust. Instead, people trust because they try to avoid violating a moral norm not to insult another person's character. As these norms are internalized and not easily influenced, it could also be that trust behavior does not fall prey to a foreign-language effect. In Chapter 2, we directly tested these two accounts, which we called the *language sensitivity account* and the *principled trustfulness account*. Using a binary version of the trust game (Berg et al., 1995) and two versions of the trolley dilemma (Thomson, 1985) we found that the use of foreign language influenced choices in both moral dilemmas but not in the trust game. This is consistent with the principled trustfulness account and emphasizes the principled nature of trust toward unknown and unobservable interaction partners in the trust game.

Chapter 3 is joint work with Thomas Schlösser and Detlef Fetchenhauer. It focuses on trust decisions toward observable individuals and systematically reviews the current literature on the accuracy of trust behavior and trustworthiness impressions. As already alluded to, there is an ongoing debate on whether people can accurately detect others' trustworthiness (Bonnefon et al., 2015; Todorov, Funk, & Olivola, 2015). As the ambiguity of the results in the literature may be influenced by different conceptual

definitions and methodological approaches (Wilson & Rule, 2017), the review also critically examined the potential moderators and the different methodological approaches used thus far. The overall evidence for trustworthiness detection accuracy in the extant literature was rather mixed. Whereas the detection accuracy from neutral photographs of individuals is limited at best, trustworthiness detection becomes more accurate when individuals can interact, observe each other face-to-face, and provide cues or signals about their own trustworthiness. On a conceptual level, we found that studies' operationalizations show a high heterogeneity, which we attribute to a lack of an overall research agenda. We therefore suggest that new studies should more strongly follow common theoretical assumptions and experimentally test potential moderators for trustworthiness detection. On a methodological level, we found that a number of studies used the same or similar participant pools for a variety of studies and that older studies failed to control for multiple trustworthiness predictions within the same trustors. We therefore suggest that future research should recruit new and larger sets of participants in the role of trustees for each study and use appropriate methods for the analysis of nonindependent data.

Chapter 4 is joint work with Thomas Schlösser and Detlef Fetchenhauer. It builds upon the findings of Chapter 3 and empirically investigates the accuracy of trust behavior and cognitive trustworthiness predictions toward individuals after face-to-face contact. Over a total of four studies, Chapter 4 provides evidence that people accurately detect others' trustworthiness if a) they are asked to predict their interaction partners' specific trustworthiness toward them and b) they are acquainted with their interaction partners. It also shows that people know which of their trustworthiness predictions are particularly accurate. This suggests that the accuracy of trustworthiness detection in the real-world, where people largely self-select how frequently they interact with certain interaction partners, might be more accurate than previously assumed. Regarding potential mechanisms, Chapter 4 finds evidence that trustworthiness detection accuracy is the result of people using their relationship quality toward acquaintances as a one-clever-cue heuristic (Gigerenzer & Gaissmaier, 2011). The relevance of this *relation-as-information* heuristic for detection accuracy is further strengthened by the finding that people accurately predicted whether their acquaintances would be trustworthy specifically toward them but not whether the same acquaintances would be generally trustworthy toward

others. Thus, while people might not know whether another acquainted person is *generally* trustworthy, they appear to know whether they can *personally* trust that person.

## 1.3 Contributions of the Authors

Chapter 2 is a research article together with Anne-Sophie Lang, Detlef Fetchenhauer, and David Dunning. Anne-Sophie Lang and I designed and conducted the study, and she is the main author of the manuscript. Detlef Fetchenhauer provided advice on the study design and on the preparation of the manuscript. David Dunning provided advice on revising the manuscript.

Chapter 3 is a review article together with Thomas Schlösser and Detlef Fetchenhauer. I designed and conducted the literature search and wrote the manuscript. Both coauthors provided advice on the structure and the preparation of the manuscript.

Chapter 4 is a research article together with Thomas Schlösser and Detlef Fetchenhauer. Both coauthors and I designed and conducted the studies, and I analyzed the data and wrote the manuscript. Both coauthors provided advice on the preparation of the manuscript. Thomas Schlösser also provided advice on the data analysis and came up with the clever name *"relation-as-information"* for the heuristic identified in the article.

## 1.4 Funding

The data collection for Study 4 of Chapter 4 was financially supported by a research grant from the Center for Social and Economic Behavior (C-SEB) at the University of Cologne. The organization did not exert influence on the choice of study design, the analysis, or the interpretation of the data.

# Chapter 2

# Foreign Language Influences Choices in Moral Dilemmas but not in Trust Situations

# 2.1 Introduction

When do we trust? From a rational choice perspective, decisions whether to trust can simply be seen as lotteries in which the source of risk lies in another person (e.g., Coleman, 1990, Chapter 5). However, trust decisions are unlike other risk-taking decisions (Ashraf et al., 2006; Eckel & Wilson, 2004; Houser et al., 2010). One difference is that they do not merely depend on calculative, analytical processes but also on emotional reactions. To show trust, that is, to give others the benefit of the doubt, has a normative component, and a substantial share of people feel negative emotions at the thought of violating that norm by signaling distrust (Dunning et al., 2019; Schlösser et al., 2013; Schlösser et al., 2015; Schlösser et al., 2016).

This feature of trust decisions might make them susceptible to so-called foreign-language effects: Foreign language has been shown to alter people's judgment and decision-making (e.g., Hayakawa et al., 2016). It may alter emotional involvement (e.g., Dewaele, 2004), particularly involving negative emotions (Sheikh & Titone, 2016; Wu & Thierry, 2012), which is a possible reason for these kinds of effects (e.g., Costa, Foucart, Hayakawa, et al., 2014; Hadjichristidis et al., 2019; Hayakawa et al., 2017; Keysar et al., 2012). Another, related possible reason is a reduced access to normative knowledge (Geipel et al., 2015b, 2015a). Other (also related) explanations focus on increased deliberation (Bereby-Meyer et al., 2020; Cipolletti et al., 2016) or increased cognitive load (Costa et al., 2017; Frey & Gamond, 2015). Since the explanations supported best by empirical evidence touch on factors that are also relevant for trust decisions, it makes sense to ask whether these decisions might be language-susceptible as well. We study this question, asking whether trust decisions couched in a foreign language, rather than in a person's first-language, reduce the emotionality connected to trust, and thus the rate of trust and trustworthiness itself. In doing so, we link different strands of research from moral psychology, judgment and decision-making, and trust research, to examine whether common mechanisms underlie these disparate types of behavior.

## 2.1.1 Foreign-Language Effects

A broad variety of foreign-language effects on judgment and decision-making has been found (for a short review on effects on choices see Hayakawa et al., 2016, for a

more recent and exhaustive review on effects and explanations see Hadjichristidis et al., 2019). Foreign language may, for instance, reduce loss aversion and framing effects and increase risk-taking (Costa, Foucart, Arnon, et al., 2014; Keysar et al., 2012). It may also reduce the impact of other heuristic biases like the self-bias (Ivaz et al., 2016), the causality bias (Díaz-Lago & Matute, 2019), or the "hot-hand" fallacy (Gao et al., 2015). It may reduce mental imagery (Hayakawa & Keysar, 2018) as well as foster honesty as opposed to lying in a dice-roll task (Bereby-Meyer et al., 2020). In more applied domains, foreign language may reduce fear conditioning (García-Palacios et al., 2018), and it has been shown to make people more willing to consume food that is sustainable yet commonly perceived as unpleasant, such as insect-based cookies, by attenuating disgust (Geipel et al., 2018).

Most relevant for our concerns, foreign language has been shown to alter judgment in moral dilemmas: In several studies, the use of foreign language caused a shift from deontological towards utilitarian responding – that means from responding based on absolute views about what kind of actions are morally right or wrong towards responding based on the evaluation of a specific action's consequences. This occurred mainly in dilemmas encompassing emotional conflict (e.g., Costa, Foucart, Hayakawa, et al., 2014; Geipel et al., 2015b, 2015a; see also Hadjichristidis et al., 2019).

Most often, these studies made use of the so-called trolley dilemma (Foot, 1967; Thomson, 1985). This hypothetical scenario exists in different versions. In what we call the footbridge version of the dilemma, the respondent has to decide whether to push another person off a bridge and onto train tracks in order to save the lives of five people. The person's body would stop an approaching train (originally, a trolley) which would otherwise run over the five people. So the decision is between sacrificing one life in order to save five lives. In this situation, pushing the person represents a utilitarian choice: In pure numbers, one lost life is less than five lost lives. Not sacrificing the person represents a deontological choice: To actively kill someone is perceived as morally wrong. In this high-conflict situation, foreign language leads to less deontological and more utilitarian choices (e.g., Cipolletti et al., 2016; Costa, Foucart, Hayakawa, et al., 2014; Geipel et al., 2015b, 2015a; Hayakawa et al., 2017; Shin & Kim, 2017).

However, there is typically no such effect of foreign language in another version of the dilemma. In the switch (or button) version, the respondent does not have to push

someone but only has to pull a switch (or push a button) in order to divert the train to another track on which only one person will be killed. The numbers (1 versus 5) are the same, but the emotional quality of the dilemma is different (e.g., Costa, Foucart, Hayakawa, et al., 2014): Measures of neural activity have shown that emotion-related brain areas are more strongly activated in the footbridge dilemma than in the switch dilemma (Greene et al., 2001; Greene et al., 2004). Also, damage of the ventromedial prefrontal cortex that processes social emotions only affects responses in the footbridge dilemma (Koenigs et al., 2007).

## 2.1.2 Explanations for Foreign-Language Effects

The literature on foreign-language effects is nascent and has produced different explanations. These accounts are mostly not necessarily exclusive but overlap and may play different roles in explaining different kinds of foreign-language effects.

### 2.1.2.1 Emotionality

The fact that foreign-language effects in the moral domain seem more common in emotionally charged situations (like the footbridge dilemma) suggests that foreign language could attenuate emotions. This explanation is supported by research on foreign language and emotions. For example, case studies from psychoanalysis and psychotherapy describe patients becoming more detached and easier able to talk about distressing topics in a language learned later in life (Pavlenko, 2005). Outside of a therapy context, there is also ample evidence for a reduced emotionality in foreign languages (Dewaele, 2004; Pavlenko, 2005). For example, the participants of Harris et al. (2003) showed higher skin conductance when being exposed to swearwords and childhood reprimands in their native language rather than in a language they learned later in life. When Hsu et al. (2015) fMRI scanned subjects who were reading excerpts from Harry Potter, they found that the emotional experience of reading emotionally charged written texts is weaker and less differentiated in a foreign language. Iacozza et al. (2017) measured larger pupil sizes when participants read aloud emotional sentences in their native than in a foreign language, indicating larger emotional arousal.

A possible reason for the reduced emotionality is the environment in which a foreign language is usually learned: In a classroom setting, language acquisition does not come with the same emotional richness as the experiences attached to this process in

one's early childhood. Instead of linking words to concepts embodied by sensory representation or autobiographic memories, students learn by defining and translating them, i.e., by explicit rather than implicit memory. At the same time, they often focus on structure rather than on meaning. These differences in the learning processes of a foreign language compared to the native language lead to a lower degree of involvement of emotion-processing structures such as the amygdala, and thereby to a disembodiment of language (Pavlenko, 2005).

#### 2.1.2.2 Dual-Process Accounts

Greene (2008, 2014) and colleagues have linked findings on the heightened emotionality of high-conflict dilemmas like the footbridge dilemma with dual-process theories. These theories distinguish between heuristic or intuitive modes of reasoning on the one hand and controlled or deliberate modes of reasoning on the other hand (e.g., J. S. B. T. Evans, 1984). Greene's dual-process theory of moral judgment postulates that deontological judgment is associated with the first kind (with automatic emotional responses) and utilitarian judgment is associated with the latter kind (with controlled cognitive processes).

In studies on the moral foreign language effect, it has been suggested that foreign language triggers controlled cognitive processes and thereby, in the case of the trolley dilemmas, utilitarian responding (Bereby-Meyer et al., 2020; Geipel et al., 2015b, 2015a). Disfluency has been mentioned as a driving factor for this (Hayakawa et al., 2016). However, recent research challenged this notion by showing that foreign language does not increase deliberate reasoning in reasoning tasks (Mækelæ & Pfuhl, 2019), and that it reduces rather than increases sensitivity to consequences of dilemmas. This was shown by Białek et al. (2019) who compared responses across different kinds of dilemmas to model the impact of sensitivity to consequences, sensitivity to norms, and general inertia on responses (the CNI model, Gawronski et al., 2017).

#### 2.1.2.3 Normative Knowledge

Another mechanism that has been proposed is a reduced access to normative knowledge. Based on findings that people judge moral transgressions less harshly in a foreign language, Geipel et al. (2015a, 2015b) argue that automatic processes may still be active but mental accessibility of moral and social rules may be reduced (see also

Hadjichristidis et al., 2019). The notion of a reduced sensitivity to norms is backed by Białek et al. (2019), who deem their result consistent with reduced emotionality.

#### 2.1.2.4 Other Explanations

Explanations related to dual-process accounts used the concept of psychological distance (Costa, Foucart, Hayakawa, et al., 2014), a concept that, when applied, however, requires elaboration as to what exactly is meant by it (Hadjichristidis et al., 2019). Furthermore, the factor cognitive load has been mentioned (Costa et al., 2017; Volk et al., 2014). But while it is intuitive that thinking in a foreign language is cognitively more demanding than thinking in the native language, several documented foreign-language effects run counter to what a mere effect of cognitive load would be expected to look like. Triggering automatic responding, cognitive load would, for instance, be expected to increase rather than decrease risk aversion (Hadjichristidis et al., 2019).

### 2.1.3 Trust

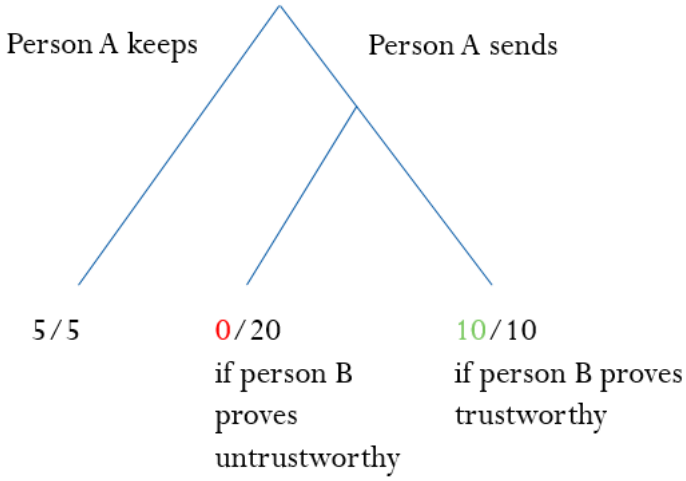
Both the reduced emotionality and the reduced access to normative knowledge explanation give reason to suspect that foreign language could also influence trust behavior. Both emotionality and moral norms are important components for trust decisions. At least in Western cultures, trusting behavior has been shown to represent an internalized moral norm: People tend to perceive trust as something they *should* do even though they do not necessarily *want* to do it (Dunning et al., 2014; Dunning et al., 2019). In a multitude of studies, participants have been overly cynical about other people's trustworthiness. Given these overly pessimistic expectations, for the majority extending trust does not represent the rational option in terms of expected utility. Yet, participants still show trusting behavior at high rates even though many think their action will produce a negative return (e.g., Fetchenhauer & Dunning, 2009).

Studies showing these high levels of trust have usually used a binary version of the so-called "trust game" (e.g., Snijders & Keren, 1999) (originally introduced by Berg et al., 1995, as the "investment game"). This game can be played with various payoff structures; here we describe the one we used in this study (see Figure 1 for illustration): A trustor receives € 10. The trustor then has to decide whether to keep the money (thereby behaving distrustfully) or whether to send it to an anonymous trustee (thereby behaving trustfully). Given the trustor sends the money, the experimenters triple the amount. The

trustee then has to decide whether to keep the € 30 (behaving untrustworthily) or whether to send € 15 back (behaving trustworthily). As the trustor has to decide whether to make themselves vulnerable to the trustee, it can be argued that this game captures the essence of trust (Dunning & Fetchenhauer, 2010).

**Figure 1**

*The Binary Trust Game*




For many people, trust behavior is associated with emotions such as anxiety and guilt at the thought of showing distrust instead of trust (Dunning et al., 2012; Dunning et al., 2014; Dunning et al., 2017; Schlösser et al., 2013; Schlösser et al., 2015). That is, people feel bad at the thought of not granting others the benefit of the doubt with regard to their moral character. They appear to follow a norm of respect, even in anonymous one-shot interactions with strangers (Dunning et al., 2014; Dunning et al., 2019). This norm is likely acquired in early socialization: People tend to state that showing trust would be something their parents would approve of rather than their peers, which is indicative of an internalized norm (Dunning et al., 2014, Studies 1 and 2). Furthermore, trust decisions have proven to be largely uninfluenced by whether they are made in private or in public (Dunning et al., 2014, Study 4), which provides evidence for a high degree of internalization.

We can deduce two opposing accounts from these prior findings when it comes to the question whether trust may be susceptible to foreign language. On the one hand, for many people, being forced to make a trust decision poses a dilemma and an emotional conflict: If they hold pessimistic expectations about their peers' trustworthiness (which many people do) trusting does not lie in their self-interest, but they may also feel obliged

to trust and experience negative emotions at the thought of not doing so (Dunning et al., 2014). The foreign-language effect in moral decision-making appears primarily in emotionally engaging dilemmas, and processing information in a foreign language reduces primarily negative emotions. If in trust decisions the negative emotions at the thought of distrusting (such as anxiety and guilt) were reduced, this might lead to a lower probability of trust behavior being shown. A reduced access to normative knowledge could also lead to the same outcome, since the negative emotions arise at the thought of a norm violation. We call this (preregistered) account of a foreign-language effect on trust the *language-sensitivity account*.[1]

That said, on the other hand, trust behavior may not be influenced by foreign language. This is because to a certain degree it is principled. As discussed previously, people often show trust despite pessimistic expectations of others' trustworthiness – in other words, they show trust even though from a rational choice perspective they should consider this behavior to be a bad investment (e.g., Fetchenhauer & Dunning, 2009). Further, Fetchenhauer and Dunning (2012) found that the share of people who trusted in a trust game was not influenced by whether the probability of being paired with a trustworthy partner was set to 46 % or 80 %, respectively (while the share of people who gambled in a lottery was significantly influenced by the same change in the probability of winning).

This suggests that trust behavior is less influenced by the specific features of a situation than other types of risky behavior. For many people, this implies a general tendency to trust others, whilst for others it may also imply a general tendency to distrust (see also Lönnqvist et al., 2015). This tendency could stem from internalized normative knowledge on trust behavior acquired in early socialization (Dunning et al., 2014). For internalized norms, it makes sense to assume that their impact should not be strongly influenced by the language of the situation, nor by differences in emotionality provoked by those languages: To use a simple but extreme example, one would not expect a committed vegetarian to opt for a steak when being given the menu in a foreign language.

We call this account which opposes the language-sensitivity account the *principled trustfulness account*, in that choices are produced by already crystallized normative principles

[1] Regarding the other explanations of foreign-language effects, it is not clear whether a dual-process explanation would speak for such an effect on trust. Be it increased deliberation or the opposite effect (due to cognitive load) – trustful and distrustful responses cannot be associated with one of either processing styles per se. In this regard, trust decisions cannot be compared with moral dilemma decisions.

associated with the decision rather than its emotionality. To be sure, facing the decision may evoke strong emotions, but those are not determinative in of themselves of how the person behaves. They might mark that a social norm is present but might carry little to no causal weight. This possible principled nature of trust could make it immune to foreign-language influences.

The present study tested the language-sensitivity account and the principled trustfulness account against each other by investigating whether or not there is a foreign-language effect on trust. To our knowledge, research on the moral foreign-language effect and research on trust have not been linked in previous studies. To address our research question, we let German-speaking participants play a binary trust game once with an unknown anonymous interaction partner either in German or in English. They were also presented with the footbridge version of the trolley dilemma and a version analogous to the switch version (the button version). This was done to ensure that the foreign-language effect on moral judgment would replicate, that is, that our foreign language manipulation was sufficient to produce the usual language effect on moral reasoning.

## 2.2 Method

### 2.2.1 Participants

As preregistered, a power analysis based on typical (rather small) sizes of foreign-language effects of moral judgment yielded a required sample size of roughly 400 participants ($\omega = 0.18$, power = 95 %). Ultimately, 408 German-speaking participants between 17 and 32 years ($M = 21.60$, $SD = 2.86$) were recruited on the campus of a large German university for a study on decision-making. Sample size was determined before any data analysis. Participants were randomly assigned to one of two language conditions: German or English.

The sample's gender distribution was approximately even: 217 participants were female (53.2 %) and 191 participants were male (46.8 %). Participants studied a wide variety of subjects at the faculty of economics, business administration and social sciences (27.5 %), the faculty of human sciences (31.1 %), the faculty of mathematics and natural sciences (23.0 %), the faculty of arts and humanities (8.3 %), the faculty of law (7.8 %),

and the faculty of medicine (1.0 %; the remaining 0.7 % indicated another university or faculty).

The language pairing German and English was chosen because German university students can generally be expected to have a sufficiently sound knowledge of the English language: It is usually the first foreign language they learn within the German education system. In our study, the average age at which participants in the English condition had started to learn the language was 8.68 years ($SD$ = 1.90). On scales from 1 (= *very little knowledge*) to 7 (= *very good knowledge*), they rated their reading skills as $M$ = 5.55 ($SD$ = 1.03), hearing skills as $M$ = 5.02 ($SD$ = 1.40), writing skills as $M$ = 4.58 ($SD$ = 1.28), and speaking skills as $M$ = 4.75 ($SD$ = 1.34). These scores support, on average, good knowledge of English. Asked whether they had lived in an English-speaking country for over a month, 24.5 % ($n$ = 50) of participants in the English condition indicated this had been the case. Excluding these participants did not change the overall pattern of results in a meaningful way. Neither did excluding participants with a high degree of proficiency in the English language or excluding two participants who indicated their native language as English in the English condition.[2] In general, English proficiency did not correlate with our dependent variables, e.g., for an index of self-rated English skills (Cronbach's alpha = .84) and trust behavior, $r$ = -.09, $p$ = .18. Results are reported for all participants without any exclusions.

## 2.2.2 Materials and Procedure

As they entered the laboratory, participants were seated in front of one of eight computers, separated by opaque dividers. Depending on condition, they were presented with either the German or the English version of the study. We report all measures and manipulations. The English version was a translation of the German version and had been checked and back-translated by native speakers in order to ensure correctness and comparability (Brislin, 1970). It had also been pre-tested in order to ensure understandability and an appropriate processing time ($N$ = 25 for trust game, $N$ = 32 for moral dilemmas).

[2] Due to an error in the questionnaire, we lack information on the language skills of participants in the German condition. This also means we were not able to detect participants with a native language other than German in this condition. However, since these participants represented less than 5 % of the sample in the English condition, we consider it very unlikely that this would impair our quite clear overall results.

After seeing a welcome screen with general information on the study, participants generated a code word enabling them to receive their payoff anonymously after they completed the study. They were then presented with four situations: a binary trust game, two trolley dilemmas (the footbridge and the button version), and a coin-flip that was not of theoretical interest to us but was mainly included to facilitate payment.

The trust game was described as an anonymous interaction between Person A and Person B, using neutral wording. We made use of the strategy method, so every participant decided in both roles, Person A and B. In the role of Person A, they decided whether to keep € 10 or whether to send them to an anonymous Person B. In the role of Person B, they decided whether to keep € 30 or whether to send € 15 back if Person A trusts. Participants were told that their interaction partner was a student of the same university who had already made their decision at a previous session of the experiment. This was to prevent participants in the English condition from assuming they might be interacting with outgroup members (Geipel et al., 2015a) since this could have negatively influenced trust and trustworthiness rates.

In the two trolley dilemmas, participants hypothetically decided whether to push the person off the bridge (footbridge version), and whether to push a button to divert the train (button version, analogous to the switch version). The word "button" was used rather than the word "switch" because in a pre-test, the English word "switch" had caused problems of understanding in this context. (For the same reason, "train" was used rather than "trolley".)

The order of the trust game and the block of the two moral dilemmas was permutated, as well as the order of the moral dilemmas within their block. For each situation, participants made moral judgments about which of the options they considered the morally correct option on a 7-point scale: A value of 1 indicated that *not choosing* the option – of sending money (back), pushing the person, or pushing the button – was morally correct, 4 was the neutral middle category, and 7 indicated that *choosing* the option was morally correct. After this judgment, participants made their actual binary decision, and filled in comprehension checks. In the trust game, participants additionally stated the percentage of participants in the role of Person B they estimated to send money back (i.e., their expectations of other participants' trustworthiness).

After participants made all their decisions, the language of the questionnaire switched to German in both conditions. Participants filled in demographic information, as well as a short measure of guilt proneness for exploratory purposes (a translated version of the GP-5; Cohen et al., 2014). In the English condition, participants were also asked to translate some key words and phrases used in the decision situations and to provide detailed information on their English skills, while participants in the German condition filled in some further exploratory measures (moral foundations, moral attentiveness, sleepiness and hunger, another guilt proneness measure, a moral reputation concern item and a generalized trust item). After completing a short second, unrelated study, participants were thanked and led to a separate room to obtain their payment.

Payment worked as follows: Participants had been asked to make a decision in a simple coin-flip in which they could either keep € 10 or stake this with the chance of either doubling it (€ 20) or the risk of losing it (€ 0). In the study, they had been told that one of their decisions (either as Person A or as Person B, or the one in the coin-flip) would be conducted for actual money. In the separate room, participants were told that this decision was to be the coin-flip. If they had chosen to participate, the experimenter flipped a coin and paid them according to the outcome; otherwise they were handed € 10. Participants were then dismissed.

### 2.2.3 Comprehensibility of Materials

In studies implementing foreign language, a crucial point is to ensure that participants are able to comprehend instructions. Participants understood the trust game well in both conditions: 87 % answered all control questions correctly. The number of errors was slightly higher in the English condition ($M$ = 0.26, $SD$ = 0.69) than in the German condition ($M$ = 0.17, $SD$ = 0.65) but this difference was not significant, $t(403.82) = 1.48$, $p = .14$, $d = .15$. In the moral dilemmas, more than 95 % of participants answered the control questions correctly in both the German and the English condition. The translation task also indicated a generally good understanding of the situations in the English condition: Most participants (73.0 %) translated all 10 key words and phrases accurately ($M$ = 9.64, $SD$ = 0.70, *Range* = 6–10). Accordingly, participants rated the English part of the questionnaire as easy to understand ($M$ = 4.38, $SD$ = 0.86, on a scale from 1 = *very hard to work on* to 5 = *very easy to work on*).

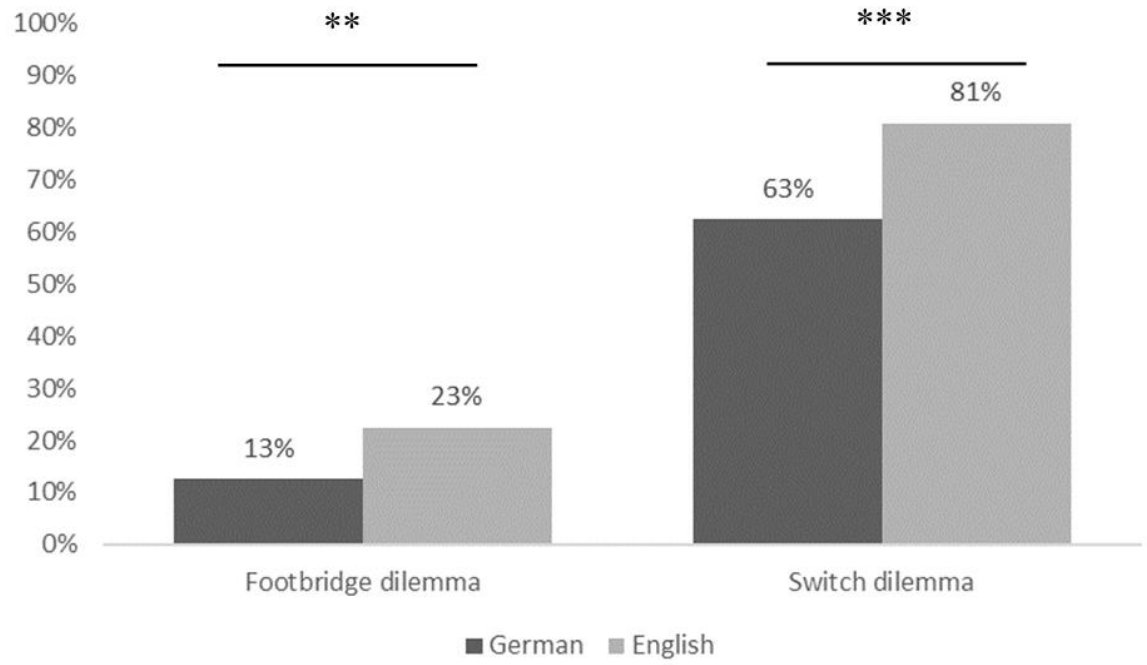
## 2.3 Results

### 2.3.1 Moral Dilemmas

We were able to replicate the moral foreign-language effect in the footbridge scenario (e.g., Costa, Foucart, Hayakawa, et al., 2014): 22.5 % of participants in the English condition decided to push the person while only 12.7 % in the German condition decided to do so. This means that foreign language increased utilitarian choices, $\chi^2(1) = 6.75$, $p = .01$, $\varphi = .13$. Unlike prior studies using the switch version, we also found such an effect in the button version. In the English condition, 80.9 % of participants decided to push the button, while in the German condition, only 62.5 % decided to do so, $\chi^2$ (1) = 12.75, $p < .001$, $\varphi = .18$ (see Figure 2 for both dilemmas).

**Figure 2**

*Percentage of Utilitarian Decisions (Push Person/Button) in the Trolley Dilemmas per Language Condition*




*Note.* ** $p < .01$. *** $p < .001$.

In terms of moral judgment, participants judged pushing the button to be more morally appropriate in the foreign language condition than in the native language one (English: $M = 4.78$, $SD = 1.54$, German: $M = 4.40$, $SD = 1.70$, $t(406) = 2.35$, $p = .02$, $d = .23$). However, in the footbridge version, the difference in moral judgment did not reach

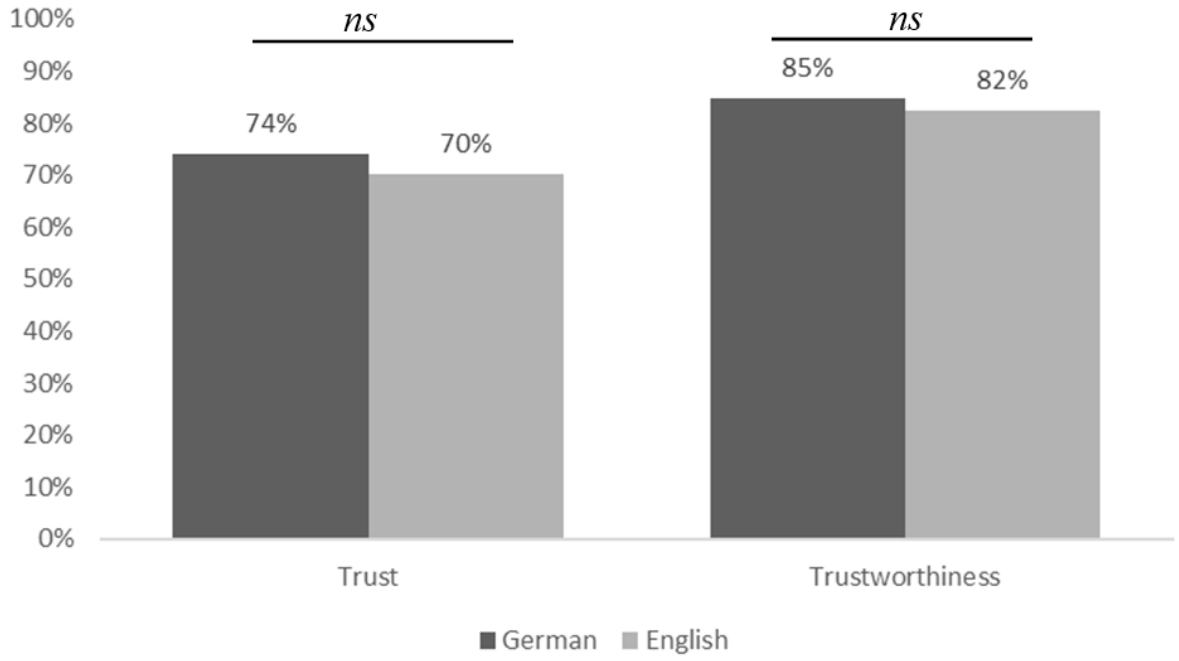
statistical significance (English: $M$ = 3.31, $SD$ = 1.92, German: $M$ = 3.14, $SD$ = 1.71, $t$(400.91) = 0.95, $p$ = .34, $d$ = .09). The averages of the 7-point Likert scale ratings indicate that subjects were clearly conflicted about the moral correctness of the options, with a tendency towards utilitarian judgment (to push is correct) in the button dilemma and a tendency towards deontological judgment (to refrain from pushing is correct) in the footbridge dilemma.

### 2.3.2 Trust Game

Unlike the moral dilemmas, differences in trust and trustworthiness rates were not significant across language conditions. In the German condition, 74.0 % of participants trusted (i.e., decided to send € 10 to Person B), in the English condition, 70.1 % did, $\chi^2(1) = 0.78$, $p$ = .38, $\varphi$ = -.04. Similarly, 84.8 % behaved trustworthily (i.e., decided to send € 15 back to Person A) in the German condition; in the English condition, 82.4 % behaved trustworthily, $\chi^2(1) = 0.45$, $p$ = .50, $\varphi$ = -.03 (see Figure 3).

**Figure 3**

*Percentage of Trust and Trustworthiness Decisions in the Trust Game per Language Condition*




*Note. ns* $p > .05$.

Moral judgments for trust did not significantly differ between language conditions either (German: $M$ = 5.27, $SD$ = 1.37, English: $M$ = 5.17, $SD$ = 1.38, $t$(406) = 0.72, $p$ = .47, $d$ = .07), nor did they for trustworthiness (German: $M$ = 6.67, $SD$ = 0.82, English: $M$

= 6.50, *SD* = 1.18, *t*(361.47) = 1.71, *p* = .09, *d* = .17). The absolute means of the ratings indicate that in both languages, trust and trustworthiness were perceived to be the morally correct options, even though for trustworthiness this perception was more unanimous than for trust.

Participants' estimates of the percentage of participants who would choose the trustworthy option in the role of Person B did not differ between conditions (German: *M* = 52.17, *SD* = 23.88; English: *M* = 51.38, *SD* = 23.98), *t*(406) = 0.33, *p* = .74, *d* = .03. This indicated that language did not influence the perception of the interaction partner in the trust game. Participants' averaged trustworthiness expectations were once again too pessimistic, in that a large majority (83.6 %) of participants split the money evenly when in the role of Person B. Also, once again, participants as Person A trusted too often given these pessimistic expectations: From a rational point of view, a risk-neutral Person A should have been indifferent between trusting and not trusting when expecting two out of three trustees to prove trustworthy (not trusting provided them with a sure outcome of € 10, but to reach an expected value of € 10 for trusting, a probability of 66.7 % to receive € 15 was necessary). However, although only 32.8 % of participants indicated at least such a level of expectations, 72.1 % of participants, more than twice as many, chose to trust.

## 2.4 Discussion

Does trust behavior follow the same psychological logic as moral behavior? Our study replicated the moral foreign-language effect but did not show a foreign-language effect of trust. Hence, the *principled trustfulness account* was supported over one based on emotional processes or salience of norms. It seems that the internalized nature of trust largely prevents it from being susceptible to language influences, even if language does reduce emotionality and/or the access to normative knowledge.

This result can be seen in line with other results regarding the relative insusceptibility of trust rates to changes in situational features, such as changing odds. For example, trust behavior varies less for different levels of risk than decisions to participate in a lottery (Fetchenhauer & Dunning, 2012). In general, decisions whether to trust strangers in a trust game do not show a very high elasticity towards situational circumstances. For example, people still show trust when they have to hand over their

own money rather than money given to them by the experimenter (Schlösser et al., 2015), and they show trust facing different stake sizes (Johnson & Mislin, 2011).

The notion that trust behavior is relatively stable is also supported by individual difference work. For example, Lönnqvist et al. (2015) found a high test-retest stability of trust behavior. In a recent study, trustors were asked to predict the trustworthiness of 45 video stimulus persons and played trust games with them. For trustworthiness predictions, only 15 % of the variance lay within the trustors, but it was 58 % for actual trust behavior (Siuda et al., 2019, unpublished data). This means that while trust on a cognitive level was much more situation specific, actual trust behavior was mainly influenced by the trustor's attributes. This suggests that principles play an important role for trust behavior even when the trustee is no longer an anonymous other. Such individual differences in tendencies to show trust or distrust are possibly influenced by normative knowledge passed on from parents to children (Dunning et al., 2014, Study 2). They may also be promoted by past reinforcement history, in that trust will persist in stable and benign environments where it pays off, such as in well-functioning societies (Jordan et al., 2015; Stavrova & Ehlebracht, 2016).

Trust may be more impervious to language effects because of the frequency of engagement with the issue. Occasions requiring trust decisions are frequent in our daily lives, not only regarding friends and family but also regarding strangers: Do I leave my valuables at my table in the café when I go to the toilet? Do I hand my phone to someone who asks to make a quick call? Do I order a used item on-line and trust the buyer to truthfully describe its state as well as to actually ship it to me? Of course, people use a variety of cues to make these decisions, but if an underlying disposition to trust or not is involved as well, it makes sense that such a cue should not be something as peripheral as the language of the situation.

Conversely, moral dilemmas like the trolley dilemmas are not something people encounter on a daily basis. This might mean that for most people the decisions they make do not reflect a well-learned, well-rehearsed, pre-formed disposition. As such, moral dilemmas may be more ad hoc and improvised at the time of the decision. Decisions are more impromptu, and thus more responsive to emotional and normative pressures that arise at the moment of decision, with those pressures depending in part on the language used to describe the choice.

Further, two additional issues should be raised. First, one issue with many studies on foreign-language effects is that choices are hypothetical (exemptions include Geipel et al., 2018, and Urbig et al., 2016). In our study, participants knew they were going to make actual – incentivized – trust and trustworthiness decisions themselves. This situation is very different from rating an action possibly performed by an equally hypothetical person in the abstract. Prior research has shown that hypotheticality reduces trust (Fetchenhauer & Dunning, 2009), an explanation being that people do not actually have to disrespect someone by withholding trust when they decide merely hypothetically (Dunning et al., 2019). Thus, it is possible – and could be tested in future studies – that abstract moral evaluations of trust situations that lack an actual decision context might be influenced by foreign language like other moral evaluations (e.g., of moral transgressions in Geipel et al., 2015a). At the same time, more research investigating actual choices is needed, preferably involving different types of norms to enable a better understanding of when and how their influence on behavior is affected by foreign language.

Second, Li (2017) found considerably lower trust rates among Hong Kong participants making trust decisions in English than among those making them in Chinese. This is an interesting finding since his participants learned both languages from early childhood. For a high level of proficiency and immersion in a second language, however, a genuine foreign-language effect is not to be expected (Čavar and Tytus, 2017; but see Białek and Fugelsang, 2019). For example, Dylman and Champoux-Larsson (2020) only found foreign-language effects in the Asian disease problem and the footbridge dilemma for Swedish/French but not for Swedish/English. They argue that this is because English is highly culturally influential in Sweden, and therefore also acquired in emotional situations such as watching movies. With regard to Li's study, bearing in mind the English language's special and prominent role in Hong Kong, e.g., as the language used in economic life, we consider it likely that he measured a culturally idiosyncratic effect rooted in sociolinguistic factors (Pavlenko, 2005) attached to the two languages.

Thinking about proficiency and immersion, it is of course conceivable that our participants, like Dylman and Champoux-Larsson (2020), might have been too immersed in the English language for an effect to reveal itself (most of them indicated that they have learned the language at least to some degree not only in classroom but also in natural settings). However, countering this line of reasoning, German students can be considered

somewhat less immersed in the English language than Scandinavian students – and, in our study, the moral foreign-language effect did show up in both moral dilemmas for the behavioral measure (i.e., the binary decision) and in the button dilemma for the moral judgment measure. To be sure, the pattern of results was interesting because the considerable decrease in deontological responding in the button dilemma is at odds with earlier studies in which the effect was limited to the footbridge dilemma (e.g., Costa, Foucart, Hayakawa, et al., 2014; Geipel et al., 2015b). Notwithstanding this difference, the foreign language English clearly influenced our participants' moral choices. Still, we did not find a foreign-language effect in the domain of trust, despite having run a laboratory study with a large sample size. This does not rule out influences of foreign language – for example, emotionality might still have been reduced – but speaks against an important role of any of these influences for trust decisions. It should be noted, however, that even with our results it is still possible that a foreign-language effect of trust exists that might be very small or subject to moderators. For example, Urbig et al. (2016) found that foreign language increased free riding in a public good setting, but almost only for less conscientious students. Since our experiment was not designed to look for interactions, we could only speculate in this regard.

In regard to research on foreign-language effects in general, and to moral decisions more specifically, our work was able to strengthen confidence in the existing literature, which is still very recent and usually based on small effect sizes (see Białek & Fugelsang, 2019). We have also added to knowledge about the contexts in which foreign language may or may not affect choices, seeing that in our study trust remained unaffected by it. The knowledge in this domain is currently limited (Costa et al., 2019), particularly for interpersonal contexts (Hadjichristidis et al., 2019), so that more research would be desirable.

## 2.5 Concluding Remarks

We showed that trust behavior is generally stable and resistant to the influence of foreign language. Seeing how important trust is for the functioning of societies – facilitating all kinds of interactions (e.g., Fukuyama, 1995; Luhmann, 1968) – this observation is a valuable insight adding to the understanding of which factors determine it and which do not. At the same time, we were able to replicate an existing foreign-

language effect on moral choices, and to identify valuable opportunities for future research. As a closing remark, we may be able to provide multinational organizations with good news. Although it is possible that operating in a foreign language may influence some aspects of their employees' work, when it comes to the important asset of trust, our evidence suggests that this will not be affected, at least not by the use of a foreign language per se.

## 2.6 Open Practices

The preregistration of the experiment is available at https://osf.io/g6vzp/?view_only=6439b862c2354f4f93e36335f461d546.

The materials and data are available at https://osf.io/atdzk/files/?view_only=aec8e8685d0a47dbaa74290e0fd675ff.

# Chapter 3

# Do We Know Whom to Trust? A Review on Trustworthiness Detection Accuracy

# 3.1 Introduction

> *There is a kernel of accuracy in trustworthiness perceptions that is of broad and substantial theoretical interest.* (Bonnefon et al., 2017b, p. 24)

> *The modern models of visualizing first impressions are mathematical maps of our appearance stereotypes, not of reality.* (Todorov, 2017, p. 268)

People automatically evaluate strangers' trustworthiness with little time and effort. As little as 34 milliseconds are sufficient to form stable expectations of another person's trustworthiness (Todorov et al., 2009), and most people, even young children (Cogsdill et al., 2014), agree on which people seem trustworthy. Importantly, these trustworthiness expectations have real-life consequences; in comparison to their trustworthy-appearing counterparts, untrustworthy-appearing individuals are remembered better (Rule et al., 2012), receive harsher criminal penalties (Wilson & Rule, 2015, 2016) and are more often excluded from economic exchanges (Chang et al., 2010).

While progress has been made on the formation of trustworthiness expectations, the accuracy of these expectations is still debated (Bonnefon et al., 2015; Todorov, Funk, & Olivola, 2015; Wilson & Rule, 2017). Resolving this debate is critical because agreement about who appears trustworthy could serve as useful or harmful, depending on the accuracy of these expectations. Excluding individuals who appear untrustworthy from cooperation might be warranted if those individuals were indeed untrustworthy. However, excluding trustworthy individuals wrongly because of their appearance would be worrisome. The same is true from the perspective of the trustor. Trusting another person is neither good nor bad per se but critically depends on that person's trustworthiness. Wrongfully withheld trust hinders fruitful cooperation; however, placing trust in another person leaves the trustor vulnerable to untrustworthy individuals. To solve this dilemma of trust as a double-edged sword, the most important task is knowing whom to trust.

To advance the debate on whether people can accurately detect others' trustworthiness, this article systematically reviews the current state of the literature and is structured as follows: First, we discuss which studies can meaningfully contribute to the question of accuracy and outline our literature search. Second, we review the overall evidence for accurate trustworthiness detection and explore potential moderators. Third, we critically address some of the current methodological and conceptual operationalizations and suggest guidelines for future research.

## 3.2 Literature Search

Over the last decade, an increasing number of studies have focused on the accuracy of trustworthiness impressions – a quick search on "Google Scholar" reveals more than 15,000 hits. What exactly is meant by the term accuracy in these studies, however, depends on the particular research question at hand. An often-encountered definition of accuracy, for example, is the consensus between people about who *appears* trustworthy (e.g., Lambert et al., 2014). While these studies are illustrative of how people form uniform trustworthiness impressions, they are uninformative regarding the actual validity of this consensus. We, therefore, developed three critical requirements for studies to be included in our review.

### 3.2.1 Which Studies can Meaningfully Advance the Debate?

First, the studies included must investigate the *direct* relationship between a trustor's trust (or a trustor's expectation of a trustee's trustworthiness) and the trustee's actual trustworthiness so that accuracy could be defined as the correspondence between these two measures. Studies that did not measure their correspondence directly were not included in this review. To provide an illustrative example, Stirrat and Perrett (2010) showed that men's facial width was related to trustworthiness in an economic game and that in a subsequent study, men with wider (compared to narrower) faces were generally trusted less. While these findings suggest that people can accurately detect trustworthiness via facial width, the direct relationship between trust and trustworthiness was never directly established because the trustee's actual trustworthiness was only measured in the first but not the second study.

Second, the studies included must measure trust and trustworthiness *objectively* via economic games. Trustworthiness detection is often used as a catchall term for detecting different positive behaviors ranging from economic behavior to infidelity to crime (Wilson & Rule, 2017). Regarding infidelity, accuracy is usually defined as the correspondence between a rater's prediction of a behavior and targets' self-reports that might or might not be honest (Foo et al., 2019). For criminality, it is defined as the correspondence between a rater's general trustworthiness judgment of a target and that target's criminal record (Rule et al., 2013). However, self-reported behavior and criminal records are subject to personal or systemic biases that limit the criterion's objectivity. In contrast, economic games offer the possibility to objectively measure trust and trustworthiness. In this way, the "gold standard" of comparing a rater's prediction of a specific behavior with that target's actual behavior can be used for accuracy (Funder, 2012). Moreover, economic games offer the advantage that participants are often financially motivated to accurately predict others' trustworthiness, and the rules of the games (including their anonymity) make it comparatively acceptable to distrust (Bonnefon et al., 2017a).

Third, the studies included must distinguish the roles of trustors and trustees and measure their behaviors (or expectations) *separately*. Note that this requirement excludes studies using cooperation games such as prisoner's dilemma (Luce & Raiffa, 1957). In the prisoner's dilemma, two actors decide simultaneously whether to cooperate or defect. If both cooperate, they receive payoffs larger than their original endowments. However, cooperators risk receiving a "sucker's payoff" if their interaction partner defects (who, in this case, receives more compensation than that received for mutual cooperation). In this setup, the roles of both actors are interchangeable, and the actors' actions are influenced by trust *and* trustworthiness simultaneously. Actors may defect because they are untrustworthy themselves but also because they fear being exploited by their interaction partner (Hayashi & Yosano, 2005). This confounding of trust and trustworthiness is resolved in the trust game (Berg et al., 1995). Here, a trustor first decides how much of an original endowment to send to a trustee who may then send back some of that (now increased) money. Like in the prisoner's dilemma, both parties receive larger payoffs if they cooperate. However, different from the prisoner's dilemma, only trustors but not trustees decide under uncertainty so that the trustee's behavior is not motivated by a fear

of exploitation. Thus, the trust game creates different roles for trustors and trustees and conceptually separates their behavior into undiluted measures of trust and trustworthiness (Snijders & Keren, 1999). This is also true for structurally similar games, such as the rely-or-verify game (E. E. Levine & Schweitzer, 2015), the hidden action game (Charness & Dufwenberg, 2006) or the game of enthronement (Kiyonari & Yamagishi, 1999). For the sake of simplicity, we will refer to all of these games as trust games.

Taken together, we required the studies reviewed to *objectively* investigate the *direct* relationship between a trustor's trust (or the expectation of a trustee's trustworthiness) and the trustee's actual trustworthiness using *trust games*.

### 3.2.2 Identification of Studies

Systematic reviews should include all relevant published and unpublished works to limit bias toward studies with significant findings (Siddaway et al., 2019). We therefore used a variety of databases in our literature search. First, we searched the Web of Science database for published articles that included the terms "trustworthiness" or "trust" or "cooperation" in the title and "detection" or "accuracy" or "ratings" or "judgment" in the text. This resulted in a total of 2455 articles, which we scanned for our inclusion requirements. Altogether, 105 articles remained after the initial screening. Second, we searched for (yet) unpublished manuscripts and dissertations using the databases of the Social Science Research Network (SSRN), EconPapers, PsyArXiv, and ProQuest using the search terms "trustworthiness" or "trust" or "cooperation" and "detection" or "accuracy" or "ratings" or "judgment". The resulting 3215 manuscripts were scanned for our inclusion requirements, leading to a total of 64 manuscripts after the initial screening. Third, we searched the reference lists of all thus far included articles for additional manuscripts on the topic to ensure that no papers were missed. This produced additional 35 articles. Next, we assessed all 204 published and unpublished articles on a full-text basis for our inclusion criteria. At this stage, only articles that objectively measured the direct relationship between trust and trustworthiness using trust games remained in the literature pool. This resulted in a grand total of 19 research articles (excluding 3 reviews or opinion articles), providing 38 individual study conditions (see Table 1 for an overview).

## 3.3 Evidence for Accurate Trustworthiness Detection

What evidence for accurate trustworthiness detection could we find among the research articles? Overall, the evidence was rather mixed; across all 38 study conditions, 16 study conditions reported accurate trustworthiness detection, whereas 22 study conditions did not. While simple vote-counting measures should not be overinterpreted (Bushman & Wang, 1994), the fact that less than half of all the study conditions found accurate trustworthiness detection suggests that it is, at the very least, a noisy endeavor. It is also worth mentioning that the number of nonsignificant findings might be underreported due to publication bias: None of the four (yet) unpublished study conditions reported better than chance accuracy, and it is not unlikely that similar studies ended up in the file drawer (Rosenthal, 1979).

While the evidence for accurate trustworthiness detection overall is not exactly overwhelming, we did not necessarily expect it to be for a task as diverse as trustworthiness detection. As we will illustrate, the literature is filled with a wide variety of studies from behavioral economics to facial symmetry research that investigate accuracy across different settings. The main task of this review is rather to identify under which conditions trustworthiness detection appears to be accurate and under which it does not. We first focus on the most likely moderators that have already been tested *within* comparably few studies before we turn to additional possible moderators that might be identified by comparing the larger number of individual studies with each other.

### 3.3.1 Moderators Within Studies

Studies that experimentally vary the setting for trustworthiness detection offer a good opportunity to identify moderators because other confounding variables are kept constant. There are two studies in the literature that identified moderators within their study design.

Schilke and Huang (2018) tested trustworthiness detection accuracy across two interpersonal contact conditions. Before receiving information about the upcoming trust game, participants received the name of their future interaction partner and were either given the chance to briefly interact with that person or not. Then, the participants were introduced to the trust game and privately made their decisions regarding their partner. The results indicated that the accuracy results were significantly higher in the contact

condition than in the no-contact condition. This effect was extended in another experiment with four conditions in which the level of interpersonal contact varied. The participants either received their partner's name or photograph or interacted with their partner via a short phone call or face-to-face conversation before receiving information about and playing the trust game. Again, the accuracy results improved with interpersonal contact; while the accuracy was above chance in the phone and face-to-face conditions, it was only at chance levels under the name and photograph conditions. These findings suggest that even short interactions of up to five minutes enable people to accurately predict another person's trustworthiness toward them. Interestingly, trustworthiness detection was accurate even though the participants had not been informed about the upcoming game and did not know what to look for when becoming acquainted with their partner.

Zylbersztejn et al. (2020) tested trustworthiness detection accuracy across three conditions that varied in terms of the strategic content and richness of the target cues made available to the raters. A first set of participants was recruited as targets, photographed with a neutral expression, and videorecorded reading a neutral text. Afterwards, the targets learned they would be playing the role of trustee in a trust game and were given the chance to deliver a videorecorded statement to potential trustors about why they could be trusted. For the critical trustworthiness detection task, another set of participants was recruited as raters; the raters were presented neutral pictures, neutral videos, or strategic videos and asked to predict the behavior of each target. The results indicated that trustworthiness detection was accurate only for the strategic videos but not the neutral videos or photographs. One reason for the improved accuracy in the strategic condition seemed to be that the raters accurately detected strategic signals (e.g., promises to be trustworthy) sent by the trustworthy targets.

Taken together, these two studies point to the following three potential moderators for accuracy: interpersonal contact, the possibility of detecting strategic content, and the richness of target cues. Trustworthiness detection was more accurate when the raters had interacted with or seen strategic messages from targets. These findings are consistent with previous studies from cooperation detection on the utility of strategic contact (Frank et al., 1993; Sparks et al., 2016). Moreover, the richness of target cues appeared as another moderator: accuracy improved as the richness of the available

target information increased from seeing the names, neutral pictures, or neutral videos of targets to seeing unscripted target videos or seeing the targets face-to-face. This is consistent with person perception theories pointing to "good information" (both in terms of quality and quantity) as a main moderator of accuracy (Funder, 1995) and the idea that trustworthiness detection "depends on the 'bandwidth' of the signaling stage of the game" (Bacharach & Gambetta, 2001, p. 172) because face-to-face encounters offer more opportunity to signal and detect trustworthiness than less information-laden exchanges. It also fits with evidence from the cooperation detection literature that finds higher accuracy after people interacted face-to-face in comparison to interacting via virtual chats (DeSteno et al., 2012).

### 3.3.2 Moderators Between Studies

In addition to identifying moderators within the extant studies, we also searched for different operationalizations of trustworthiness detection between studies that might moderate accuracy. We identified 22 nontrivial dimensions (e.g., the ratio of (un)trustworthy targets presented to raters) on which studies differed. To make these dimensions more tangible, we will illustrate the 9 most notable dimensions with examples of the studies found in the literature, thereby also giving some insight into each individual study included in this review. Note that we will draw rather qualitative conclusions on the impact of each dimension because the conflation of the different operationalizations (with respect to both dependent and independent variables) in the studies made any quantitative comparison using effect sizes of little use. We will structure the differences alongside three broad categories, focusing on…

…differences in general:

1. Is general or specific trustworthiness measured?
2. Is cognitive or behavioral trust measured?
3. Do the rater and target interact?

…differences concerning targets:

4. How are the targets presented?
5. When are the targets recorded?
6. Are the targets incentivized to appear trustworthy?

7. Are the targets instructed how to act?

…and differences concerning raters:

8. When do the raters see the targets?
9. Are the raters incentivized to provide an accurate judgement?

#### 3.3.2.1 Is General or Specific Trustworthiness Measured?

Although rarely given much attention, a possible moderator of accuracy could be the type of trustworthiness being measured. Trustworthiness can be defined as a target's *general trustworthiness* (e.g., toward an unknown individual) or as a target's *specific trustworthiness* (e.g., toward the rater), and it is unclear a priori whether both types lead to the same accuracy. It might be, for example, that a person's specific trustworthiness toward oneself is easier to predict than that person's general trustworthiness because one can take the specific relationship with that person into account.

In one study on the accuracy of general trustworthiness by Bonnefon et al. (2013), the raters played trust games in which they could base their decisions only on neutral target pictures. These pictures were extracted from videos of a previous study in which the targets had played an anonymous trust game in the role of trustee. The results indicated that the raters could accurately predict the targets' general trustworthiness when given cropped black-and-white versions of target pictures. The effect was also found to be independent of general intelligence or cognitive load, suggesting that the detection of general trustworthiness might be a modular process. This idea is supported by the fact that accuracy was only at chance level when the raters were given the target pictures in an unedited color version that included the targets' hairstyles and clothes (but also see Jaeger, Oud, et al. (2020) for opposing findings).

An example of specific trustworthiness detection is the aforementioned study by Schilke and Huang (2018), in which raters predicted targets' specific trustworthiness toward them after seeing the name or picture of the target or after engaging in a phone or face-to-face interaction. As mentioned, trustworthiness detection was more accurate in the latter than in the former two conditions.

Overall, 30 study conditions investigated general trustworthiness, of which 12 were significant, whereas 8 study conditions investigated specific trustworthiness, of which 4

were significant. Thus, although it might, in theory, be easier to predict a person's specific trustworthiness toward oneself than that person's general trustworthiness, we do not find clear evidence that supports this idea.

#### 3.3.2.2 Is Cognitive or Behavioral Trust Measured?

Another potential moderator could be the type of trust being measured. The extant studies differ in regard to whether trust is measured via cognitive judgments of trustworthiness (cognitive trust) or via actual behavior in the trust game (behavioral trust). This distinction is important because trust rates on the cognitive level and on the behavioral level differ; while cognitive trust is guided by rather cynical views of trustees (Dunning et al., 2019), trust behavior is often guided by normative principles to respect trustees' moral character (Dunning et al., 2014). It is not unlikely that these differences may also lead to differences in accuracy.

As an example of *cognitive* trust, Okubo et al. (2018) presented raters with target pictures and asked them to rate each target's trustworthiness on a seven-point scale. In these pictures, targets were photographed slightly from the right- and left-hand sides with posed happy and angry expressions before completing several trust games. The results indicated that the cognitive trust ratings were accurate for angry faces viewed from the right side but inaccurate for the other three combinations.

In contrast, De Neys et al. (2015) investigated the accuracy of actual trust *behavior*. Raters were shown a subset of the same edited target pictures used by Bonnefon et al. (2013) and asked to play a trust game with each target. The results replicated the above chance accuracy and showed that the result held true for raters as young as 13 years of age.

Different again, Jaeger, Oud, et al. (2020) investigated the accuracy of both cognitive and behavioral trust. Raters saw photographs of targets who had already made their trust game decisions and indicated whether they wanted to send money to each target (behavioral trust) and how much money they expected back from each target (cognitive trust). Here, there was no significant relationship between the targets' trustworthiness and the raters' cognitive or behavioral trust. Moreover, this null result was independent of whether full-sized or cropped versions of target photographs were used.

Overall, 22 study conditions tested accuracy via cognitive trust, of which 6 were significant, while 16 study conditions tested accuracy via behavioral trust, of which 10

were significant. This pattern, thus, seems to suggest that trust behavior might be more accurate than cognitive trust, which would echo previous findings from anonymous trust games (Fetchenhauer & Dunning, 2009). However, also note that the study by Jaeger, Oud, et al. (2020) allowed us to directly test the accuracy for both types of trust and found no disparity in regard to accuracy.

#### 3.3.2.3 Do Rater and Target Interact?

As already mentioned, Schilke and Huang (2018) found that interpersonal contact improved the detection accuracy. Did we find a similar pattern across studies? Most studies with rater-target interactions assume that trustworthiness might be detectable via a "sympathetic manner" (p. 249) if individuals become sufficiently acquainted with each other (Frank et al., 1993). In one of these studies, Hayashi and Yosano (2005) gave participants 30 minutes to become acquainted during a group discussion before informing them about the upcoming trust game. The participants privately indicated their behavior toward one of their group members (who had yet to be randomly decided) and then rated the trustworthiness of each group member. The results indicated that the participants' actual trustworthiness in the game was significantly correlated with the group members' aggregated trustworthiness ratings. In another study, Binzel and Fehr (2013) recruited pairs of friends from a Cairene slum in Egypt to play a trust game. Here, the participants were unable to accurately detect the trustworthiness of their friends.

Studies without rater-target interactions, on the other hand, test whether trustworthiness is a stable feature that is detectable by outside observers. Some of these studies build on the idea that trustworthiness can be objectively measured via facial width (Stirrat & Perrett, 2010) and assume that a person's trustworthiness can be detected from viewing neutral photographs. In one of these studies, De Neys et al. (2017) found accurate trustworthiness detection for a subset of the same edited pictures previously used by Bonnefon et al. (2013). They further showed that the accuracy results were above chance when the pictures were presented for as little as 100 milliseconds. However, the results also showed that the effect was reversed when the pictures were only presented for 33 milliseconds; here, the participants trusted trustworthy targets significantly less than untrustworthy targets.

Does the overall trend across studies mirror the results from Schilke and Huang (2018) that interpersonal contact improves accuracy? Yes, it does. Five of the 6 interactive

study conditions reported accurate trustworthiness detection. It is curious in this light that Binzel and Fehr (2013) did not find accurate trustworthiness detection among friends. While this could be due to the study's special setting in an Egyptian slum, it raises the interesting possibility that too much previous interaction could also be detrimental to trustworthiness detection. Conversely, only 11 of the 32 noninteractive study conditions reported accurate trustworthiness detection. This speaks against the idea that trustworthiness is a stable and physically observable trait that can be detected without previous interaction.

#### 3.3.2.4 How are the Targets Presented?

Following the discussion of more general differences, we now turn to the differences between the studies in regard to the targets. The most obvious dimension on which studies differ is how many (and which) target cues are observable for raters. On the one side of the spectrum, raters are given access to numerous cues when observing targets face-to-face. While these studies are closely related to the interactive studies discussed earlier, face-to-face type studies do not necessarily involve rater-target interactions. In the study by Snijders and Keren (2001), the participants sat in opposing rows and privately played trust games with each opposing participant, with whom they had no previous interaction. They were also asked to privately predict each other's trustworthiness. Limited to information based on physical appearance, the participants were unable to accurately detect each other's trustworthiness.

Fewer target cues are available in the studies that present targets via video. Here, the raters usually predict the trustworthiness of targets who are recorded before, during or after a trust game. In the aforementioned study by Zylbersztejn et al. (2020), the trustworthiness detection results were accurate for the videos of targets attempting to convince the potential trustors of their trustworthiness but were inaccurate for the neutral target videos.

Even fewer target cues are available in the studies that present the targets via pictures. As mentioned earlier, raters in an additional condition of Zylbersztejn et al. (2020) could only base their trust game decisions on neutral target photographs. Similar to the neutral videos, the raters were unable to distinguish (un)trustworthy targets.

Finally, at the other end of the spectrum, some studies limit the observable cues to the voices of the targets. Schild et al. (2020) tested trustworthiness detection accuracy via

men's voice pitch. The targets played an anonymous trust game, and their voices were recorded while reading a pre-established text. The raters then listened to these recordings and predicted each target's trustworthiness. A lower voice pitch was linked to higher perceived trustworthiness but was unrelated to the targets' actual trustworthiness so that the overall detection accuracy was not better than chance.

We previously suggested that the richness of target cues could be a moderator of accuracy. Did we find evidence across studies to support this idea? Again, 4 of the 6 study conditions in which the targets engaged in face-to-face interactions reported accurate trustworthiness detection results. Conversely, only 1 of 3 and 10 of 25 study conditions reported accurate trustworthiness detection for videos or pictures, respectively. Moreover, the accuracy results were above chance for 1 of the 2 study conditions if the targets were presented auditorily. Providing raters with more and richer target cues might, thus, be a key to accurate trustworthiness detection.

#### 3.3.2.5 When are the Targets Recorded?

Recall that Zylbersztejn et al. (2020) reported accurate trustworthiness detection results for videos recorded before but not after the targets knew about the upcoming trust game. Could the timing of target recording be a moderator of trustworthiness detection? After all, if raters were able to correctly identify voluntary trustworthiness signals sent by targets, they might also pick up on involuntary trustworthiness cues, such as emotional expressions during (Verplaetse et al., 2007) the decision-making process or emotional residues after the targets made their decisions (Albohn & Adams, 2020).

Verplaetse and Vanneste (2010) investigated this question by having raters observe targets during their trust game decisions. The targets were filmed so that short videos taken at the moment of their trustee decision could be shown to the raters who then predicted which target had sent money back. Here, the trustworthiness detection results were accurate, which suggests that viewing people's emotional reactions during their decision-making process could indeed improve accuracy. On the other hand, the participants in the aforementioned study by Snijders and Keren (2001) were unable to distinguish (un)trustworthy targets, although they were able to observe each other face-to-face during their trust games. However, the raters in this study were also the targets of their fellow participants and might have been too busy with their own trust game decisions to adequately focus on reading others' emotional reactions.

A different timing was used in the aforementioned study by Ask et al. (2020), who videorecorded targets expressing why they could be trusted after they had already made their trustworthiness decision. Viewing these videos, raters were unable to distinguish (un)trustworthy targets. Thus, untrustworthy individuals might be able to mask their intentions if given adequate time to emotionally distance themselves from the decision. Interesting in this regard is also the study by Okubo et al. (2012) in which male targets completed a series of trust games before having their pictures taken with posed happy and angry facial expressions. Raters then saw a subset of these pictures and rated each target's trustworthiness. The results indicated that the trustworthiness detection results were only accurate for angry but not happy expressions. Although speculative, happy expressions might, thus, be better suited to conceal one's trustworthy intentions than angry expressions.

The general trend across studies supports the idea that raters make inaccurate judgements when targets have already made their decisions; only 1 of 10 study conditions reported accurate trustworthiness detection in this case. However, accuracy might improve when the targets are unaware of the upcoming game or when targets are aware of the game but have not yet made their decisions. Here, 5 of 11 and 9 of 15 study conditions reported accurate trustworthiness detection results, respectively. It also seems possible that observing targets' emotional reactions during their decision could lead to accurate trustworthiness detection. Obviously, additional studies are needed for any substantial conclusion since only the 2 study conditions mentioned above, of which 1 was significant, have so far tested this idea.

#### 3.3.2.6 Are Targets Incentivized to Appear Trustworthy?

Another difference between the studies that might moderate the trustworthiness detection accuracy is whether the targets had financial incentives to appear trustworthy. We categorized the study conditions as providing an incentive if the targets knew that their recordings would later be used to predict their trustworthiness and if those predictions had consequences for their own trust game payoffs.

An example is the aforementioned study by Ask et al. (2020), in which the targets tried to convince the potential trustors of their trustworthiness via video messages. As already mentioned, the raters were unable to accurately detect the targets' actual trustworthiness. However, there are also studies in which the trustworthiness detection

was accurate for targets who had financial incentives. De Neys et al. (2013) used a subset of the same edited target pictures as Bonnefon et al. (2013), which featured targets trying to convince potential trustors of their trustworthiness. Upon viewing these pictures, the raters accurately distinguished (un)trustworthy targets.

In contrast to the two studies above, Dilger et al. (2017) did not financially incentivize targets to appear trustworthy. Here, the targets had already played a trust game with an anonymous interaction partner and knew they would be paid according to this trust game before having their pictures taken. Only later were the pictures shown to raters who were unable to accurately predict the targets' trustworthiness. Similar studies, however, found accurate trustworthiness detection results. As in the previous study, the targets in the aforementioned study by Verplaetse and Vanneste (2010) knew their trust game partner would not see their recordings and, thus, had no financial incentive to appear trustworthy. However, unlike in the previous study, the raters were able to accurately predict the targets' trustworthiness.

One argument for the incentivization of targets is that untrustworthy targets might only invest energy into appearing trustworthy if given financial incentives to do so. As a result, trustworthiness detection should be less accurate when targets are incentivized to appear trustworthy. Did we find evidence in support of this argument? No. There was no clear difference in the ratio of significant study conditions between studies with (9 of 21 significant conditions) or without (7 of 17 significant conditions) target incentivization. This suggests that giving targets financial incentives to appear trustworthy is not as critical as often assumed.

#### 3.3.2.7 Are Targets Instructed how to Act?

Another potential moderator could be whether targets were instructed how to act while they were being recorded. Studies vary in this regard mainly because of differing assumptions about trustworthiness detection. The studies that do not restrict the targets' appearance usually assume that trustworthiness detection is dependent on situational cues or signals. An example is the aforementioned study by Hayashi and Yosano (2005), in which participants took part in a group discussion before playing their trust games. As already reported, the group members accurately detected each other's trustworthiness. Another example is the earlier mentioned study by Ask et al. (2020), in which targets

recorded video messages to convince potential trustors of their trustworthiness. Here, watching these videos did not lead to accurate trustworthiness predictions.

In contrast, the studies that specifically instruct targets to act neutrally in their recordings test the assumption that trustworthiness is a stable feature of a person that can be detected from a neutral appearance. In one of these studies, Efferson and Vogt (2013) photographed male targets with neutral expressions after they had played a version of the trust game that allowed them to send back money even when they had not been trusted. Later, the raters were presented with these neutral target photos alongside the information on whether each target had been trusted by their trustor. While the raters accurately predicted that the targets would act reciprocally if trusted, they could not use pictures of the targets to further improve this accuracy. This once more speaks against the notion that trustworthiness detection is accurate after viewing neutral faces.

Another set of studies instructs targets to make specific emotional expressions when posing for their photographs. These studies assume that trustworthiness can be masked by posed emotional expressions. In the aforementioned study by Okubo et al. (2012), targets were instructed to feign happy and angry expressions when posing for their photographs. As reported, the raters accurately predicted target trustworthiness only for angry but not happy expressions, indicating that trustworthiness detection may be more accurate for some emotional expressions than others.

Across studies, trustworthiness detection was accurate in comparably few studies when targets had been instructed to act neutrally (6 of 19 significant conditions) or emotionally (3 of 7 significant conditions). In contrast, 7 of 12 study conditions reported accurate trustworthiness detection results when targets had not been instructed on how to act. Although the differences in accuracy between studies with vs. without instructions provided to targets were small, it might be possible that the targets' natural facial expressions provide valuable cues for trustworthiness detection and that limiting access to these cues consequently decreases accuracy.

#### 3.3.2.8 When Do Raters See Targets?

After discussing how different operationalizations on the target side influenced the detection accuracy results, we now turn to the differences between the studies on the rater side. An important difference between studies is whether raters see targets before or after they know about their upcoming detection task. Why is this important? Again, the

different operationalizations result from opposing assumptions about what constitutes trustworthiness detection in the real world.

On the one hand, people frequently enter trust situations knowingly (e.g., when buying a used car) in which they can strategically look for trustworthiness cues or signals shown by their interaction partner. An example of a study considering these dynamics is the aforementioned study by Zylbersztejn et al. (2020). Raters were fully informed about the trust game and could accurately detect the targets' trustworthiness by picking up signals of trustworthiness sent by targets. In other studies, however, fully informed raters were unable to accurately detect targets' trustworthiness. Eckel and Petrie (2011) photographed participants and had them play trust games in which they either saw photographs of their interaction partners free of cost or could buy them. The results showed that participants were willing to pay at least some money for target pictures but were unable to use them to their advantage.

On the other hand, there are many social situations in which people need to assess trustworthiness from past observations. A new neighbor might, for example, ask to borrow an expensive tool, and their trustworthiness can only be evaluated based on previous small talk. This type of trustworthiness detection was tested in the aforementioned study by Hayashi and Yosano (2005), in which participants formed accurate expectations of their group members' trustworthiness even before knowing about the upcoming detection task. As we have already mentioned, however, trustworthiness detection among naïve individuals is not always accurate – even when these individuals are friends (Binzel & Fehr, 2013).

Did the ratio of significant studies vary depending on whether the raters knew about the upcoming detection task? Taken together, 11 of 31 study conditions with informed raters reported accurate trustworthiness detection results, whereas 5 of 7 study conditions with naïve raters reported accurate trustworthiness detection results. This pattern might be viewed as evidence for the rather counterintuitive conclusion that people who are naïve about any upcoming trustworthiness detection task achieve higher accuracy than people who are consciously looking for potential cues or signals of trustworthiness. However, we caution against any overinterpretation, as the pattern could simply be because the ratio of interactive studies is higher with naïve raters than informed raters.

#### 3.3.2.9 Are Raters Incentivized to Be Accurate?

All else being equal, it could be assumed that financial incentives motivate raters to be more accurate with their predictions. Did rater incentivization moderate accuracy across studies? The aforementioned study by Bonnefon et al. (2013) offers an opportunity to test this idea. While the raters in most study conditions were paid according to one randomly chosen trust game they played, the raters in another condition rated trustworthiness on a seven-point scale without financial incentives for accurate judgements. Whereas trustworthiness detection was accurate in 2 of the 3 incentivized conditions, it was inaccurate in the unincentivized condition.

However, this does not indicate that trustworthiness detection is accurate only when raters are incentivized. Okubo et al. (2017) photographed targets who were instructed to appear as trustworthy as possible before having them play a series of trust games. Raters later viewed these photographs and rated the targets' trustworthiness on a seven-point scale. Even though the raters had no incentives to provide accurate ratings, trustworthy targets were rated as more trustworthy than untrustworthy targets.

Overall, we found no clear trend to support the idea that providing raters with financial incentives increased accuracy. Eleven of 25 study conditions with financial incentives reported accurate trustworthiness detection results, compared to 5 of 13 study conditions without financial incentives.

### 3.3.3 Summary of Potential Moderators

We set out the following three probable moderators of trustworthiness accuracy that had been experimentally manipulated within studies: interpersonal contact, the richness of target cues, and the possibility of detecting strategic content. Did we find evidence in support of these moderators across studies? Indeed, we did. First, study conditions with rater-target interaction reported accurate trustworthiness detection more often than conditions without interaction. This supports the idea that personal contact with another person leads to accurate perceptions of that person's trustworthiness. While we can only speculate as to why this is the case, it seems plausible that, on average, people may indeed detect trustworthiness in personal interactions via a "sympathetic manner" (Frank et al., 1993, p. 249).

Second, conditions that included rich target cues more often reported accurate trustworthiness detection than conditions with limited target cues. Thus, increasing the richness of target cues (e.g., by observing another person face-to-face) may indeed lead to accurate trustworthiness detection results. In contrast, we found only mixed evidence for accuracy in information-poor contexts. For example, the evidence for accurate trustworthiness detection from neutral faces was limited to studies using the target pool created by Bonnefon et al. (2013) with some studies even using only a subset of the previously most diagnostic faces (e.g., De Neys et al., 2015).

Third, trustworthiness detection was more often accurate when strategically relevant content was observable, either because targets were not limited in how to act or because they were recorded just before or during their trust game. Thus, trustworthiness detection appears more accurate when raters have access to situational cues, such as the targets' emotional expressions. In contrast, accuracy was lowest when targets were recorded after their decisions or when situational cues were masked by specific instructions (e.g., on how to pose for a picture).

Did we find additional moderators between the study conditions? The answer is a resounding maybe. As can be appreciated from our discussion of the studies, the nonindependence of operationalizations (especially those regarding the dependent variable) prevents meta-analytical comparisons using effect sizes. We therefore relied on simple vote-counting, even though this approach obviously only allows for less conclusive results. The best case for an additional moderator could be made for the type of trust being measured because trustworthiness detection was more often accurate for behavioral than cognitive trust. This is unlikely to be simply due to trust behavior being more often incentivized than cognitive trust because neither rater nor target incentivization themselves appeared to influence accuracy. The evidence is not clear-cut, however, because the results from Jaeger, Oud, et al. (2020) did not support this trend within their study. Thus, future studies are needed to clearly disentangle the moderating impact of how trust is measured. All other differences between studies, although theoretically relevant, did not appear to independently influence the trustworthiness detection results. For example, while raters appeared to be more accurate when observing targets before as opposed to after being informed about the upcoming detection task, this difference could likely be because only the conditions with rater-target interaction involved naïve raters.

## 3.4 Toward Unified Research on Trustworthiness Detection

After summarizing the current evidence under which conditions trustworthiness detection appears to be accurate, we now turn to a more overarching issue. During our literature review, we discovered that studies strongly varied in their methodological and conceptual designs. In an ideal world, this diversity would have enabled us to compare accuracy across a rich field of different situations and identify potential moderators. In the real world, however, the absence of similar research methods (e.g., how accuracy is defined and analyzed) made it difficult to meaningfully compare the findings across studies. Part of the problem, we believe, is that the field lacks a unified research agenda with common research practices. We, therefore, decided to address some of the current methodological and conceptual practices and offer suggestions regarding how to improve the comparability of future research and open up the possibility of more quantitative analyses in the future.

### 3.4.1 Methodological Designs of Studies

There are three methodological practices that most prevent results from being comparable across studies. First, the studies use different and sometimes misleading definitions of accuracy. While approximately half of all the study conditions regress trustworthiness ratings (or trust behavior) on the targets' trustworthiness, there are some departures from this procedure. Schilke and Huang (2018), for example, coded ratings as 1 if rater trust and target trustworthiness corresponded, and 0 otherwise, and compared these scores across conditions. This procedure might be problematic, however, because the participants' overall trust and trustworthiness rates also differed across conditions. Note that the trust games in this study were not played under anonymity, and the participants might have felt a stronger obligation to both trust and be trustworthy in conditions with more intensive interpersonal contact. As a result, higher accuracy in more interactive conditions could simply be due to different base rates and not because raters in interactive conditions more successfully detected untrustworthy targets than in the less interactive conditions. To show this quantitatively, by simply trusting everyone, participants would have reached 89% accuracy in the face-to-face condition but only 54%

accuracy in the no-contact condition. This also illustrates that while interpersonal contact might be a promising candidate for a moderator of trustworthiness detection, further studies with different measures of accuracy are needed to assess this conclusively.

Second, some studies use improper methods to analyze nonindependent data. Many study designs generate multiple trustworthiness ratings for each rater, leading to data clustering, i.e., an underestimation of standard errors and an increase in type I errors if left unaccounted (Hox et al., 2017). In particular, older studies suffer from a lack of corresponding data analysis because adequate methods were not as widespread as those available currently. For example, Hayashi and Yosano (2005) collected up to five ratings from every participant and analyzed these data using traditional test statistics. This approach does not meet current practice standards, as it can lead to an overestimation of the true relationship between predicted and actual trustworthiness.

Third, some studies aggregate ratings over raters or over targets before testing for accuracy. In the first procedure, ratings are aggregated for each target so that a target's actual trustworthiness can be compared to the average predicted trustworthiness of that target (e.g., Dilger et al., 2017). While this procedure provides results regarding the detection accuracy of groups as a whole, it systematically overestimates trustworthiness detection accuracy at the individual level because idiosyncratic rater biases are evened out (Efferson & Vogt, 2013). Moreover, the results of such analyses may not replicate when other raters are used (Judd et al., 2012). In the second method, ratings are aggregated for each rater so that a rater's average rating of trustworthy targets can be compared with that rater's average rating of untrustworthy targets (e.g., Okubo et al., 2017). This procedure creates accurate estimates of the differences between (un)trustworthy targets but limits the generalizability of these differences to the targets used in the experiment. As the goal is usually to generalize results to the general population, aggregating over targets should therefore be avoided (Judd et al., 2012).

### 3.4.2 Conceptual Designs of Studies

Apart from methodological issues, some conceptual procedures also need to be addressed. First, some studies test trustworthiness detection with nonrepresentative subsets of targets in which the ratio of (un)trustworthy targets is either artificially set to 50:50 or equally distributed between genders. While this creates a clean benchmark for

measuring better than chance accuracy, it impairs the external validity of the results if trustworthiness is not also set at this very specific ratio in the real world (Todorov, Funk, & Olivola, 2015). Moreover, conducting studies in the vacuum of balanced trustworthiness ratios could lead to an artificial increase in the detection accuracy because base rates do not have to be considered (Olivola & Todorov, 2010).

Second, some studies provide low generalizability by using the same pool of target pictures for a multitude of studies. As people largely agree on who appears trustworthy (Todorov, Olivola, et al., 2015), it is not surprising that accurate trustworthiness detection in an initial study is repeated in subsequent studies. This is even less surprising when considering that subsequent studies often used subsets of the previously most diagnostic pictures. Moreover, given the relatively small target pools of 12 to 60 individuals, only a few easy-to-recognize targets would be sufficient for the small but better-than-chance detection accuracy that is usually found.

Third, many studies test trustworthiness detection in settings with limited ecological validity, for example, by only providing raters with neutral target pictures. While previous research suggests that trustworthiness detection could be accurate in ecologically valid settings, for example, after interpersonal contact (Frank et al., 1993) or among acquainted participants (Funder & Colvin, 1988; Paulhus & Bruce, 1992), the idea that trustworthiness is readable in static faces is reminiscing of past physiognomic beliefs that have been largely refuted (Todorov, Olivola, et al., 2015). As an illustration of the limited value of photographs, Todorov and Porter (2014) showed that pictures of the same target were perceived differently depending on slight changes in their facial expression. The trustworthiness ratings varied so much that any target could be ranked as the most or least trustworthy-looking individual depending on which pictures were chosen.

### 3.4.3 Guidelines for Future Research

While some of the current research practices make it difficult to conclude under which conditions trustworthiness detection is accurate, we are hopeful that the debate can ultimately be settled. To further advance the debate, we suggest four guidelines for future research.

First, researchers need to purposefully address trustworthiness detection accuracy instead of solely considering it a byproduct of their own study design. Too many studies seem to emerge from the thought "I wonder if these trustworthiness judgments that we measure are actually accurate" and are thus composed of a seemingly random combination of possible operationalizations. Instead, studies should build on existing theory, for example, from person perception (Funder, 1995) or evolutionary psychology (Cosmides & Tooby, 1992; Frank, 2005), and consciously operationalize how to measure trustworthiness detection. This includes knowing which type of trustworthiness (general vs. specific) or trust (cognitive or behavioral) is relevant or how much interpersonal contact, cue richness, strategic content or acquaintanceship should be adequate for accurate trustworthiness detection.

Second, studies should try to find moderators of detection accuracy within their own study design. As we illustrated, there are numerous operationalizations for trustworthiness detection research that could independently influence accuracy. Even if we limit ourselves to the 9 most notable differences above, they still translate to 3,072 potential studies that would need to be conducted before all operationalizations were systematically varied. Thus, experimental conditions within studies appear to be the most fruitful approach to identify under which conditions trustworthiness detection is accurate.

Third, studies need to ensure the generalizability and validity of the results by recruiting as many target persons as possible without altering the targets' true trustworthiness prevalence. It might be tempting to create statistical power by recruiting large (online) sample sizes of raters who rate the trustworthiness of comparably few targets. However, increasing the sample size of raters is not very helpful, as people generally agree about who appears trustworthy, and adding further raters only consolidates the same overall findings. Instead, generalizability and construct validity can be improved by increasing the sample size of targets because it decreases outlier effects from particularly easy (or difficult) to detect targets (Wells & Windschitl, 1999).

Fourth, researchers need to use an appropriate methodology to test accuracy. This entails analyzing data at the rating level instead of rater or target level, controlling for dependencies in the data, and using an appropriate definition of accuracy. Over the last few years, mixed-effect models have emerged as a useful method to analyze the

corresponding data, and the R package *lme4* (Bates et al., 2015) has been most widely adopted.

## 3.5 Conclusion

Judgments about others' trustworthiness are made frequently and have important real-life consequences, yet their accuracy is still debated. We advance this current debate in two ways. First, we identified the following three moderators of trustworthiness detection: interpersonal contact, the richness of target cues, and the possibility of detecting strategic content. Second, we addressed some current research methods and developed the following guidelines for future research: studies should engage in stronger theory building, test moderators within studies, strengthen generalizability with large target pools, and use appropriate analyses for nonindependent data.

With these promising moderators and guidelines, we call on future studies to investigate trustworthiness detection accuracy more systematically. People in their everyday life are constantly engaged in trustworthiness detection tasks, for example, when thinking about leaving their laptop on the café table while they go to the bathroom or when buying a used car. It is worthwhile to uncover the mysteries behind these everyday challenges.

**Table 1**

*Overview of Study Conditions*

| | 1 | | 2 | | 3 | | 4 | | | | 5 | | | | 6 | | 7 | | | 8 | | 9 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Study Condition | a | b | a | b | a | b | a | b | c | d | a | b | c | d | a | b | a | b | c | a | b | a | b |
| Ask et al. (2020) | - | | - | | | - | | - | | | | | | - | - | | | | - | | - | | - |
| Binzel & Fehr (2013) | | - | - | | - | | - | | | | - | | | | | - | | | - | - | | - | |
| Bonnefon et al. (2013): Study 1 | + | | | + | | + | | | + | | | + | | | + | | + | | | | + | + | |
| Bonnefon et al. (2013): Study 2 | + | | | + | | + | | | + | | | + | | | + | | + | | | | + | + | |
| Bonnefon et al. (2013): Study 3 | - | | | - | | - | | | - | | | - | | | - | | - | | | | - | - | |
| Bonnefon et al. (2013): Study 4 | - | | - | | | - | | | - | | | - | | | - | | - | | | - | | | - |
| De Neys et al. (2013) | + | | | + | | + | | | + | | | + | | | + | | + | | | | + | + | |
| De Neys et al. (2015) | + | | | + | | + | | | + | | | + | | | + | | + | | | | + | | + |
| De Neys et al. (2017): Study 1 | + | | | + | | + | | | + | | | + | | | + | | + | | | | + | + | |
| De Neys et al. (2017): Study 2 (> 33 ms) | + | | | + | | + | | | + | | | + | | | + | | + | | | | + | + | |
| De Neys et al. (2017): Study 2 (< 33 ms) | - | | | - | | - | | | - | | | - | | | - | | - | | | | - | - | |
| Dilger et al. (2017) | - | | - | | | - | | | - | | | | | - | | - | - | | | - | | | - |
| Eckel & Petrie (2011): Condition 2 | | - | | - | | - | | | - | | - | | | | | - | - | | | | - | - | |
| Eckel & Petrie (2011): Condition 3 | - | | | - | | - | | | - | | - | | | | | - | - | | | | - | - | |
| Efferson & Vogt (2013) | - | | - | | | - | | | - | | | | | - | | - | - | | | | - | - | |
| Hayashi & Yosano (2005) | + | | + | | + | | + | | | | + | | | | | + | | | + | + | | | + |
| Jaeger et al. (2020): Study 1 (behavioral trust) | - | | | - | | - | | | - | | | | | - | - | | - | | | | - | - | |
| Jaeger et al. (2020): Study 1 (cognitive trust) | - | | - | | | - | | | - | | | | | - | - | | - | | | | - | - | |
| Jaeger et al. (2020): Study 2 (cropped photos) | - | | - | | | - | | | - | | | | | - | - | | - | | | | - | - | |
| Jaeger et al. (2020): Study 2 (uncropped photos) | - | | - | | | - | | | - | | | | | - | - | | - | | | | - | - | |
| Okubo et al. (2018): Angry faces (right) | + | | + | | | + | | | + | | | + | | | + | | | + | | | + | | + |
| Okubo et al. (2018): Angry faces (left) | - | | - | | | - | | | - | | | - | | | - | | | - | | | - | | - |
| Okubo et al. (2018): Happy faces (right) | - | | - | | | - | | | - | | | - | | | - | | | - | | | - | | - |
| Okubo et al. (2018): Happy faces (left) | - | | - | | | - | | | - | | | - | | | - | | | - | | | - | | - |
| Okubo et al. (2017) | + | | + | | | + | | | + | | | + | | | + | | | + | | | + | | + |
| Okubo et al. (2012): Angry faces | + | | + | | | + | | | + | | | | | + | | + | | + | | | + | | + |
| Okubo et al. (2012): Happy faces | - | | - | | | - | | | - | | | | | - | | - | | - | | | - | | - |
| Schild et al. (2020) | - | | - | | | - | | | | - | | | | - | | - | | | - | | - | | - |
| Schilke & Huang (2018): Study 1 | | + | | + | + | | + | | | | + | | | | | + | | | + | + | | + | |
| Schilke & Huang (2018): Study 2 | | + | | + | + | | + | | | | + | | | | | + | | | + | + | | + | |
| Schilke & Huang (2018): Study 3 (photo) | | - | | - | | - | | | - | | - | | | | | - | | | - | - | | - | |
| Schilke & Huang (2018): Study 3 (phone) | | + | | + | + | | | | | + | + | | | | | + | | | + | + | | + | |
| Schilke & Huang (2018): Study 3 (face-to-face) | | + | | + | + | | + | | | | + | | | | | + | | | + | + | | + | |
| Snijders & Keren (2001) | | - | - | | | - | - | | | | | | - | | - | | | | - | | - | - | |
| Verplaetse & Vanneste (2010) | + | | + | | | + | | | + | | | | + | | | + | | | + | | + | + | |
| Zylbersztejn et al. (2020): (strategic video) | + | | + | | | + | | + | | | | + | | | + | | | | + | | + | + | |
| Zylbersztejn et al. (2020): (neutral video) | - | | - | | | - | | - | | | - | | | | | - | - | | | | - | - | |
| Zylbersztejn et al. (2020): (neutral photo) | - | | - | | | - | | | - | | - | | | | | - | - | | | | - | - | |
| Frequency of occurrence | 30 | 8 | 22 | 16 | 6 | 32 | 6 | 3 | 27 | 2 | 11 | 15 | 2 | 10 | 21 | 17 | 19 | 7 | 12 | 7 | 31 | 25 | 13 |
| Frequency of accurate trustworthiness detection | 12 | 4 | 6 | 10 | 5 | 11 | 4 | 1 | 10 | 1 | 5 | 9 | 1 | 1 | 9 | 7 | 6 | 3 | 7 | 5 | 11 | 11 | 5 |

*Note*. + represents accurate and - represents inaccurate trustworthiness detection; categories are listed in the same order as in the text: 1a: general trustworthiness, 1b: specific trustworthiness; 2a: cognitive trust, 2b: behavioral trust; 3a: interaction, 3b: no interaction; 4a: face-to-face, 4b: video, 4c: picture, 4d: voice; 5a: before information was given, 5b: after information was given but before a decision was made, 5c: during decision making, 5d: after the decision was made; 6a: target incentive, 6b: no target incentive; 7a: neutral instructions, 7b: emotional instructions, 7c: no instructions; 8a: before information was given, 8b: after information was given; 9a: rater incentive, 9b: no rater incentive.

# Chapter 4

# People Accurately Detect Acquaintances' Specific but not General Trustworthiness by Using Relationship Quality as Information

# 4.1 Introduction

We often try to assess people's trustworthiness. Will the other person at the beach take my belongings if I go for a swim? Do I believe the used car dealer that the car is still in good condition? Can I trust my neighbor with the house key when I am on vacation? Assessments such as these are made frequently in our daily lives toward strangers, acquaintances, and friends. However, how accurate are we at detecting trustworthiness?

Past research shows that people readily form stable impressions of others' trustworthiness (Todorov et al., 2009) and that these impressions have important real-life consequences (Chang et al., 2010; Wilson & Rule, 2015, 2016). However, the evidence for the accuracy of these impressions is weak (Todorov, Olivola, et al., 2015; Wilson & Rule, 2017). For every paper finding accurate trustworthiness detection (Bonnefon et al., 2013), a similar paper finding no such effect (Jaeger, Oud, et al., 2020) can usually be found. Moreover, those studies that do find accurate trustworthiness detection often report such small effects that reasonable generalizations into the real world are inappropriate (Todorov, Funk, & Olivola, 2015).

These conclusions raise a mystery: How do people successfully navigate their social lives if they are unable to distinguish (un)trustworthy individuals? We believe that the apparent lack of evidence for accurate trustworthiness detection in the current literature can be attributed to two misunderstandings. The first misunderstanding is the lingering sentiment, once popular with physiognomists, that character traits can be read from neutral faces. As physiognomy has largely been discarded, it is no surprise that evidence for accurate trustworthiness detection from neutral photographs is particularly weak (Todorov, Olivola, et al., 2015). The second, more opaque, misunderstanding is the treatment of trustworthiness as a stable character trait. To date, most studies have asked participants to predict trustees' general trustworthiness. However, in this paper, we argue that trustworthiness also includes an interpersonal component, and accurate trustworthiness detection critically depends on access to this component. As a result, previous research has underestimated people's trustworthiness detection abilities. In other words, we argue that the most relevant question to ask participants is not "is this a trustworthy person?" but "is this person trustworthy toward you?"

## 4.2 Trustworthiness

It is interesting that a clear definition of trustworthiness in the literature is missing. Rather, trustworthiness is used as a catch-all term for different kinds of prosocial behavior (Wilson & Rule, 2017). Trust, on the other hand, has received more conceptual attention and is often defined as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or the behavior of another" (Rousseau et al., 1998, p. 395). In line with this definition, trustworthiness can be regarded as the intention or the behavior not to take advantage of this vulnerability of another person.

Depending on the research field, the operationalization and measurement of trustworthiness can be quite different. Research on criminality operationalizes trustworthiness as the tendency not to engage in criminal behavior, measured via criminal records (Rule et al., 2013), whereas research on infidelity focuses on cheating behavior, measured via self-reported transgressions (Foo et al., 2019). Unfortunately, the personal and systematic biases that accompany self-reports or criminal records can undermine the objective measurement of trustworthiness. Similarly, research on organizational trustworthiness generally measures organizations' or managers' perceived but not actual trustworthiness using employee questionnaires (Caldwell & Clapham, 2003; Ferrin et al., 2008). To resolve this issue, trustworthiness is often measured objectively using economic games such as the trust game (Berg et al., 1995). In the trust game, a trustor first decides how much of an original endowment to send to a trustee, who may then send back some of that (now increased) money. The game's sequential nature thus creates distinct roles for the trustor and the trustee. Trust is measured via the trustor's decision to send the endowment and make herself vulnerable, while trustworthiness is measured via the trustee's decision not to take advantage of this vulnerability. While this conceptualization of trustworthiness does not capture (arguably related) concepts such as criminal or aggressive behavior, its objective measurement ensures that the "gold standard" of accuracy measurement via behavioral prediction is met (Funder, 2012). The trust game also offers the advantage that participants are financially incentivized to accurately predict the trustee's trustworthiness. In this paper, we will therefore follow the operationalization of trustworthiness via the trust game.

As can be appreciated from the operationalization of trustworthiness as the trustee's reaction to the trustor's trust, trustworthiness is *relational* and thus features both a personality component and an interpersonal component. The personality component of trustworthiness refers to a trustee's *general* trustworthiness and relates to how a trustee acts across different situations and interaction partners. For example, a particular trustee may generally be more or less trustworthy than others. Research has shown that people differ in general trustworthiness and that these differences are associated with personality traits (Zhao & Smillie, 2015). Regarding the Big Five personality traits, Ben-Ner and Halldorsson (2010) found that trustees high in agreeableness were more trustworthy than less agreeable trustees. Becker et al. (2012) showed that both openness and agreeableness were associated with trustworthiness, although the link was rather weak. Thielmann and Hilbig (2015) largely replicated these results for the HEXACO personality traits, finding an association between honesty-humility and trustworthiness. Thus, general trustworthiness appears to be related to personality traits such as honesty-humility (HEXACO-Model) or agreeableness (Big-5-Model), although the overall relationship appears to be rather weak.

In addition to the trustee's general trustworthiness, there is also a more interpersonal component of trustworthiness in that a trustee may be more or less trustworthy toward a specific trustor. This *specific* trustworthiness relates to how a trustee acts toward particular interaction partners. For example, although a trustee may be untrustworthy in general, she may be trustworthy toward a particular trustor with whom she has a close relationship. Previous research demonstrates the importance of considering interpersonal components for behavior related to trustworthiness. Columbus et al. (2021) found that although cooperation in everyday life was common, there was substantial within-person variation depending on the interaction partner and their level of interdependence. Similarly, Weiss et al. (2020) reported overall high trust rates in everyday life that varied strongly depending on the specific trustee and the individual relationship with that trustee. Whereas only 16.2% of the variance in trust was attributable to stable trustor effects, 62.8% of the variance was attributable to differing trustees. Although the literature lacks comparable studies on how much a person's trustworthiness varies in everyday life, it seems reasonable to assume that trustworthiness also contains a flexible interpersonal component that is influenced by relational aspects.

Taken together, we suggest that trustworthiness contains both a personality component in that some trustees are *generally* more trustworthy than others and an interpersonal component in that trustees are more or less trustworthy *specifically* toward particular trustors. How a trustee acts toward another specific trustor depends on both components. That is, a trustee may generally be more or less trustworthy but deviate from this trustworthiness depending on the relationship with a specific trustor.

## 4.3 Trustworthiness Detection

In contrast to our conceptual distinction between general and specific trustworthiness, most studies on the detection of trustworthiness have – at least implicitly – treated trustworthiness as a stable personality trait. In a typical study, trustees are photographed with neutral facial expressions before being asked to play a trust game. For the critical detection task, the trustor is then presented the photographs and asked to predict which trustee acted (un)trustworthily. Trustors are thereby asked to predict the general trustworthiness of a yet unknown interaction partner. For example, Bonnefon et al. (2017a) showed in a series of studies (using the same stimulus material) that trustors could indeed distinguish (un)trustworthy trustees, although the detection accuracy was only slightly better than chance. However, studies using the exact same (Jaeger, Oud, et al., 2020) or similar setups (Dilger et al., 2017; Eckel & Petrie, 2011; Efferson & Vogt, 2013) have failed to find accurate trustworthiness detection. Overall, the evidence for the accurate detection of trustworthiness, as defined as a stable personality trait, is weak, and the answer to the question "do people know who is trustworthy?" appears to be "no".

Rather than trying to add more nuance to this question, we suggest that it is not the adequate question to ask. People in their everyday life generally do not face the question of whether a potential trustee is trustworthy per se but whether that potential trustee is trustworthy toward them. Therefore, the more important question to ask is "do people know who is trustworthy *toward them*?". To answer this question, trustors need to be able to accurately predict the trustee's specific rather than general trustworthiness. Since the specific trustworthiness depends on the relationship between the trustor and the trustee, studies must give them an opportunity to form at least some minimal relationship with each other. Interactive studies from cooperation detection serve as useful examples (DeSteno et al., 2012; Frank et al., 1993; Sparks et al., 2016). In their seminal paper, Frank

et al. (1993) showed that participants accurately distinguished cooperators from defectors in a prisoner's dilemma game after sufficiently long strategic interaction in groups of three. Cooperation detection accuracy was inaccurate, however, when participants had not had enough time to get acquainted.

Only one study has thus far investigated trustworthiness detection accuracy in a similarly interactive setting. Schilke and Huang (2018) compared the accuracy of trustworthiness judgments across four conditions that varied in the degree of interpersonal contact between trustor and trustee. The trustors either received the trustee's name or photograph or swiftly talked to the trustee via phone call or face-to-face interaction before predicting the trustee's specific trustworthiness toward them. The results mirrored those from cooperation detection: Whereas trustworthiness detection was inaccurate when only a name or photograph was given, it was better than chance when trustors had interacted with the trustee. Thus, people appear to accurately detect another person's specific trustworthiness toward them after having established some previous relationship. Unfortunately, the study used a somewhat misleading definition of detection accuracy. Different from most studies in the field, trustworthiness judgments were coded as accurate if trust and trustworthiness corresponded (and inaccurate otherwise), and these overall accuracy rates were then compared across conditions. The comparison of these overall accuracy scores is problematic, however, when the overall trust and trustworthiness rates also differ across conditions. As the trust games in the study were not anonymous, participants likely felt a stronger obligation to both trust and be trustworthy in conditions with interpersonal contact. As a result, the apparent higher overall accuracy in conditions with interpersonal contact could be an artifact from the higher base rates of trust and trustworthiness in these conditions. To put things into numbers, by simply trusting everyone, trustors would have reached 89% accuracy in the face-to-face condition but only 54% accuracy in the name condition.

Therefore, there is currently no conclusive evidence regarding whether people can accurately detect the *specific* trustworthiness of others toward them. Most studies have limited themselves to the detection of *general* trustworthiness as a stable trait or used a misleading definition of accuracy. In this paper, we aim to close this gap. We investigate the type of trustworthiness detection that people more regularly face in their everyday life ("can *I* trust this person"?) and allow participants to predict their interaction partner's

trustworthiness in the trust game after having interacted with that person in a face-to-face context. Moreover, we introduce a round-robin design that allows us to compare the detection accuracy of general and specific trustworthiness. More specifically, Study 1 tests trustworthiness detection accuracy among previously unacquainted participants after a short group task. Study 2 tests trustworthiness detection accuracy among acquainted participants after a multiday seminar. Study 3 compares the detection accuracy of specific and general trustworthiness and identifies a potential mechanism for accurate trustworthiness detection. Study 4 replicates and extends these results.

## 4.4 Study 1

The cooperation detection literature and the results of Schilke and Huang (2018) suggest that the detection of specific trustworthiness might be accurate after relatively short interpersonal contact. We tested this idea among previously unacquainted participants following a group task. As previous research highlights the importance of distinguishing expectations of trustworthiness (cognitive trust) and actual trust behavior (behavioral trust), we measured and investigated the accuracy of both types of trust separately (Dunning et al., 2019).

### 4.4.1 Method

#### 4.4.1.1 Participants

We recruited 144 students between 16 and 42 years ($M$ = 22.67, $SD$ = 4.20; 45.8% female, 54.2% male) at a large German university for a study on decision-making. The participants were recruited individually and given specific appointments to ensure that groups would mostly consist of strangers. Altogether, 24 groups consisting of 6 participants each were recruited in this way, which translated to an expected total of 720 individual trust interactions.

#### 4.4.1.2 Procedure

Upon arriving at the laboratory, the participants were seated around a large desk and informed about the group task. In this task, each group had to build the largest possible paper tower within 15 minutes using only the materials provided (paper, scissors, and glue). The group task was originally developed for assessment centers and gave group

members a short opportunity to become acquainted (Heilmann, 2002). If their group won the task, each participant could earn an additional €20 for participating in the study.

After the group task, each participant individually and silently filled out a questionnaire containing a binary version of the trust game that they played with each group member. In this version of the trust game, the participants decided whether to keep or send an original endowment of €3 in the role of trustor and, if they were trusted by their counterpart, decided whether to return half or none of a resulting €10 in the role of trustee. The participants knew their choices were anonymous and that one of their behavioral decisions would later be carried out with real money. The participants indicated their behavior both in the roles of the trustor and trustee toward each of the other five group members and indicated what they believed each group member would do in the role of the trustee toward them. The answers served as measures of the participants' trust behavior, trustworthiness behavior and cognitive trust. All answers were recorded using the strategy method, meaning that the participants indicated their choices before knowing whether they would ultimately participate as trustor or trustee.

After filling out their questionnaires, the participants were randomly matched with one of their group members, randomly assigned either the role of the trustor or trustee and paid their individual payout in a sealed envelope. To match participants to their envelopes without breaching anonymity, each envelope had a code word written on it, which the participants had generated in their questionnaires. In this way, neither the participants nor the experimenters could track individual choices. Last, the participants of the winning group were contacted and invited to collect their €20 bonus. No deception was used throughout the study.

#### 4.4.1.3 Analysis Strategy

To investigate trustworthiness detection accuracy at the interaction level, all analyses were based on the 720 trust game interactions between participants. After excluding one interaction because of missing data, 719 individual trust game interactions remained for the analyses. Because there were multiple observations within groups, trustors, and trustees, observations were nonindependent. To account for this clustering in the data, we used the *lme4* package (Bates et al., 2015) to estimate separate multilevel regression models for cognitive and behavioral trust that included random intercepts for

groups, trustors, and trustees[3]. Following the suggestion by Peugh (2010), the predictors were tested via likelihood ratio tests of nested models that either did or did not include the relevant predictor variable.

### 4.4.2 Results

In the role of trustor, the participants predicted their colleagues to behave trustworthily toward them in 81.5% of all cases (cognitive trust) and sent their endowment to the other person in 83.0% (behavioral trust). In the role of trustee, the participants behaved trustworthily 81.9% of the time. With regard to accuracy, trustworthy trustees were predicted to behave trustworthily 82.2% of the time and were trusted 83.4% of the time, whereas untrustworthy trustees were predicted to behave trustworthily 78.5% of the time and were trusted only 81.5% of the time.

To test if these differences were statistically significant, we proceeded with two separate multilevel regression analyses in which the trustors' cognitive and behavioral trust were each regressed on the trustee's actual trustworthiness. Model comparisons between the empty models, which only included the random intercepts, and the saturated models, which also included the trustees' actual trustworthiness as a predictor, showed that the trustees' actual trustworthiness was not significantly related to the trustors' cognitive trust ($\chi^2(1) = 1.55$, $p = .21$, $OR = 1.47$, 95% CI [0.80, 2.68], see Table 2) or trust behavior ($\chi^2(1) = 1.24$, $p = .26$, $OR = 3.00$, 95% CI [0.44, 20.35], see Table 3). Thus, the participants were unable to accurately detect their group members' trustworthiness toward them after getting to know each other in the group task.

### 4.4.3 Discussion

Contrary to Schilke and Huang (2018), we did not find evidence for the accurate detection of specific trustworthiness (both on a cognitive and behavioral level) after short interpersonal contact. This could be due to a few factors. First, participants in our study did not interact in dyads but in groups of six. While this round-robin design allowed us to gather many trust interactions from comparably few participants, it likely limited participants' chance to become sufficiently acquainted. Second, the group task might not

[3] Across all four studies reported in this paper, some models only converged after removing one or two random intercepts. Models with maximal and simplified random intercept structure resulted in very similar estimates and significance levels. We therefore report the results of models with the maximum number of random intercepts throughout the paper.

have provided a sufficient "diagnostic situation" for trustworthiness detection. According to interdependence theory, people accurately identify others' attitudes toward themselves by observing how they behave in conflict-of-interest situations (Columbus et al., 2021). While the 15-minute group task provided some conflict (e.g., deciding on a plan under time pressure), it might not have been diagnostic enough. Third, although the group task gave the participants an opportunity to become acquainted, they were likely focused on successfully completing the group task rather than getting to know their group members.

Taken together, the setting of Study 1 likely prevented participants from developing the kind of relationships that are needed for person-specific trustworthiness to arise. As a result, decisions to trust or be trustworthy could have been driven by rather principled injunctive norms to be generally trusting and trustworthy (e.g., see Dunning et al., 2014). Indeed, an ICC analysis of the empty multilevel models revealed that 87.7% of the variance in trust behavior and 86.7% of the variance in trustworthiness behavior could be attributed to stable between-person differences. Thus, the detection of previously unacquainted strangers after only limited personal contact in groups resembled rather general trustworthiness detection. Unsurprisingly, then, trustworthiness judgments were inaccurate.

## 4.5 Study 2

The previous results suggested that short interpersonal contact in groups is insufficient for accurate trustworthiness detection to arise. We therefore tested trustworthiness detection accuracy in a more natural setting among acquaintances. This aimed to provide participants with diagnostic information about their real-life relationships that had naturally developed over time.

### 4.5.1 Method

#### 4.5.1.1 Participants

To ensure that the participants were acquainted, we conducted Study 2 at the end of a three-day university seminar that was held at a youth hostel in Germany. At that time,

the students[4] had spent days together, taking part in joint work sessions in various group constellations during the daytime and sleeping in shared dormitories during the nighttime. While this assured that the participants were well acquainted at the end of the seminar, it also led to a comparably small sample size of 18 participants. However, recent data simulations show that the sample size itself is not as important for detection tasks as the number of targets and total ratings (T. R. Levine et al., 2021). Since the group size was larger than that in Study 2, the procedure resulted in a total of 18 targets and 306 individual trust interactions.

#### 4.5.1.2 Procedure

On the last day of the seminar, the participants sat in a circle, were informed about the trust game, and privately recorded their decisions. In the role of trustor, the participants chose between keeping or sending an original endowment of €10. In the role of trustee, they chose whether to return half or none of a resulting €40 if they had been trusted. All answers were recorded using the strategy method, meaning that the participants indicated their choices before knowing whether they would ultimately participate as trustor or trustee. The participants indicated their behavior as the trustor and trustee toward each classmate and indicated how they expected each classmate to behave toward them in the role of trustee. These answers again served as measures of participants' trust behavior, trustworthiness behavior and cognitive trust. As in Study 1, the participants knew that their choices would remain anonymous, as only one of their trust interactions was randomly chosen to be carried out with real money. After all trust game interactions had been randomly determined, the participants were handed their individual payment in sealed envelopes, which they opened in private. No deception was used.

#### 4.5.1.3 Analysis Strategy

We followed the same analysis strategy as in Study 1 with the only exception that we included random intercepts for the trustors and trustees but not the groups because there was just one group. All analyses were based on the 306 trust game interactions

[4] To guarantee the participants' complete anonymity, we chose not to include any personal information in the questionnaire that might later be used to identify any student. Because of this, we do not report on any demographic information about participants in this study and all following studies.

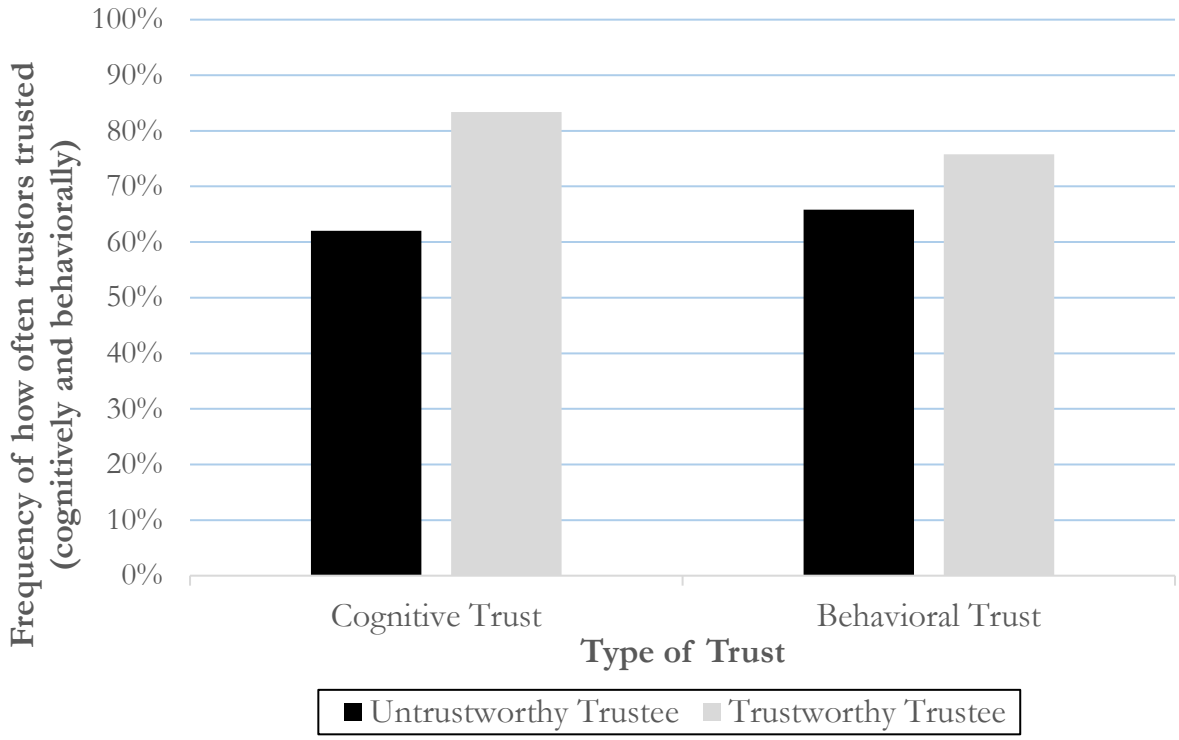
between participants, and after excluding 4 interactions because of missing data, a total of 302 individual trust game interactions remained for the analyses.

## 4.5.2 Results

In the role of trustor, the participants predicted their classmates to behave trustworthily toward them in 77.8% of all cases and sent their endowment to the other person in 73.2%. In the role of trustee, the participants behaved trustworthily 73.8% of the time. A descriptive look at the data revealed remarkable accuracy of both cognitive and behavioral trust. Trustworthy trustees were predicted to behave trustworthily 83.4% of the time and were trusted 75.8% of the time, whereas untrustworthy trustees were predicted to behave trustworthily only 62.0% of the time and were trusted only 65.8% of the time (see Figure 4).

**Figure 4**

*Frequency of how Often Trustors in Study 2 Predicted Trustees to be Trustworthy (Cognitive Trust) and Sent Money to Trustees (Behavioral Trust), Depending on Trustees' Actual Trustworthiness.*




To confirm that this detection accuracy was significant, we proceeded with two separate multilevel regression analyses, in which the trustors' cognitive and behavioral trust were each regressed on the trustees' actual trustworthiness. Model comparisons

showed that the trustees' actual trustworthiness was significantly related to both the trustors' cognitive trust ($\chi^2(1) = 16.82$, $p < .001$, $OR = 5.16$, 95% CI [2.39, 11.14], see Table 4) and trust behavior ($\chi^2(1) = 4.04$, $p = .04$, $OR = 2.39$, 95% CI [1.02, 5.58], see Table 4). Thus, the participants accurately detected their classmates' trustworthiness toward them at the end of the seminar.

### 4.5.3 Discussion

Study 2 provided the first evidence for the accurate detection of specific trustworthiness, although a few issues limit the scope of this finding. First, while we used over 300 trust interactions for the analyses, these interactions were based on only one seminar group of eighteen participants. Any effects could therefore be due to participants' idiosyncrasies and are thus limited in their informative value. Second, we only focused on the detection of specific but not general trustworthiness, which prevents any direct comparisons between the two. We cannot rule out the possibility that participants would have accurately detected their classmates' general trustworthiness. Third, we did not measure whether participants took their relationships into account when evaluating their classmates' specific trustworthiness. While it is not unlikely that participants used their relationships as information to distinguish (un)trustworthy classmates, there are various alternative explanations for the accurate trustworthiness detection observed (e.g., the higher financial stakes than in Study 1).

## 4.6 Study 3

To address these weaknesses, the purpose of Study 3 was threefold. First, it aimed to replicate and generalize the results of Study 2 by recruiting a larger sample who were not university students. Second, it directly compared the accuracy of specific and general trustworthiness detection. If information about the trustor-trustee relationship was necessary for accurate trustworthiness detection, we would expect the detection of the *specific* but not the *general* trustworthiness to be accurate. Third, Study 3 tested this idea more directly by considering relationship quality as a potential mechanism for trustworthiness detection. We assumed that when evaluating another person's specific trustworthiness, participants might use their relationship with that person as information. This "relation-as-information" heuristic could thus serve as a one-clever-cue-heuristic

(Gigerenzer & Gaissmaier, 2011) to accurately evaluate another person's benevolence and thus trustworthiness toward oneself.

## 4.6.1 Method

### 4.6.1.1 Participants

We recruited three groups of 11, 16, and 24 participants[4] from workshops of a German banking organization. The participants in each workshop were well acquainted, as they were part of one-year training programs. Altogether, 51 participants were recruited for Study 3, which follows the recommended number of at least 20 raters and 50 targets for detection experiments (T. R. Levine et al., 2021) and translated to an expected total of 902 individual trust interactions.

### 4.6.1.2 Procedure

The procedure was the same as in Study 2, except for four changes. First, before being informed about the trust game, participants rated their relationship quality with each trustee (*How would you describe your relationship with Person X?*) on a scale from 1 (*very bad*) to 7 (*very good*). This single item served as a fast and frugal measurement of the idiosyncratic relationship with each trustee. Second, participants played a slightly altered trust game in which the trustor's original €10 endowment was increased to only €30 instead of €40, if sent. This change effectively increased the trustor's risk (i.e., the ratio of cost over benefit), which has been shown to decrease overall trust rates (A. M. Evans & Krueger, 2014), and aimed to counteract anticipated ceiling effects for trust among well-acquainted participants. Third, after each trustworthiness prediction, trustors additionally indicated their confidence in that prediction on a scale ranging from 1 (*not confident at all*) to 7 (*very confident*). This tested whether participants had an accurate assessment of their own trustworthiness detection abilities. Fourth, after predicting whether a trustee would behave trustworthily toward them (specific trustworthiness), trustors were asked to predict how often a trustee would behave trustworthily toward all other trustors (general trustworthiness). Answers could range from zero to $n$ - 2 for a group with $n$ participants and were later transformed into percentages to make them comparable across the three differently sized groups.

#### 4.6.1.3 Analysis Strategy

To investigate trustworthiness detection accuracy at the interaction level, analyses were based on 902 trust game interactions between participants. After excluding four interactions because of missing data, 898 individual trust game interactions remained for the analyses. We followed the same analysis strategy as in previous studies, this time estimating separate multilevel regression models for general cognitive trust, specific cognitive trust, and specific behavioral trust. All continuous predictors were group mean centered prior to the analyses except for trustees' general trustworthiness toward the group, which was an aggregated level two variable and therefore centered around the grand mean.

### 4.6.2 Results

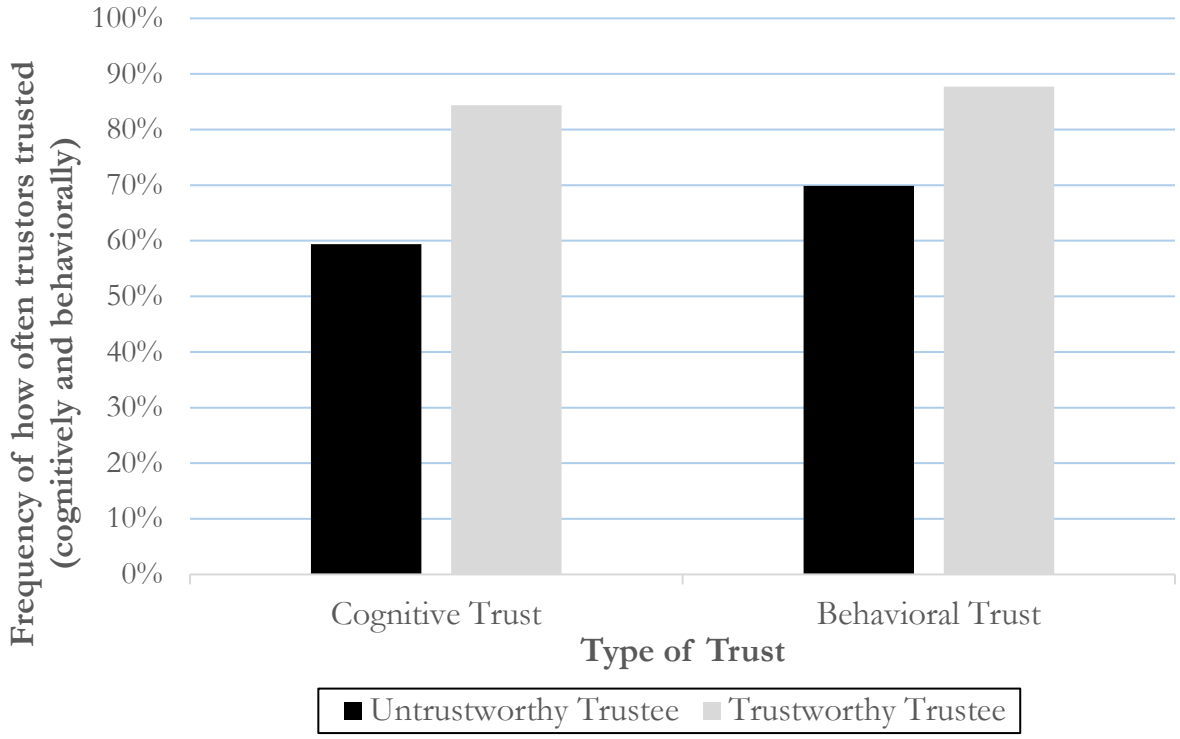
In the role of trustor, the participants predicted their colleagues to behave trustworthily toward them in 80.7% of all cases and sent their endowment to the other person in 85.1%. In the role of trustee, the participants behaved trustworthily 85.2% of the time.

#### 4.6.2.1 Detection of Specific Trustworthiness

Mirroring the results of Study 2, a descriptive look at the data revealed remarkable accuracy of both cognitive and behavioral trust. Trustworthy trustees were predicted to behave trustworthily 84.4% of the time and were trusted 87.7% of the time, whereas untrustworthy trustees were predicted to behave trustworthily only 59.4% of the time and were trusted only 69.9% of the time (see Figure 5).

**Figure 5**

*Frequency of how Often Trustors in Study 3 Predicted Trustees to be Trustworthy (Cognitive Trust) and Sent Money to Trustees (Behavioral Trust), Depending on Trustees' Actual Trustworthiness.*




To confirm that this detection accuracy was significant, we proceeded with two separate multilevel regression analyses, in which the trustors' cognitive and behavioral trust were each regressed on the trustees' actual trustworthiness. Model comparisons between the empty and saturated models showed that the trustee's actual trustworthiness was significantly related to both the trustors' cognitive trust ($\chi^2(1) = 34.48$, $p < .001$, $OR = 5.83$, 95% CI [3.16, 10.76], see Table 5) and trust behavior ($\chi^2(1) = 17.85$, $p < .001$, $OR = 3.36$, 95% CI [1.92, 5.88], see Table 6). Thus, we replicated the results of Study 2. The participants accurately detected their acquaintances' specific trustworthiness toward them. Moreover, the accuracy of any particular trustworthiness prediction was positively related to the confidence in that prediction ($\chi^2(1) = 8.47$, $p < .01$, $OR = 1.92$, 95% CI [1.23, 2.99], see Table 7). In other words, the participants knew which of their predictions was more accurate than others.

### 4.6.2.2 Relationship Quality

Did participants use their relationships as valid information for trustworthiness detection? To test this hypothesis, we again conducted two separate multilevel regression

analyses for cognitive and behavioral trust. After controlling for relationship quality, the association between the trustees' actual trustworthiness and the trustors' trust behavior ($\chi^2(1) = 3.20$, $p = .07$, $OR = 1.80$, 95% CI [0.95, 3.41], see Table 6) was no longer significant. In contrast, the association between the trustees' actual trustworthiness and the trustors' cognitive trust was weakened but remained significant even after controlling for relationship quality ($\chi^2(1) = 15.77$, $p < .001$, $OR = 3.62$, 95% CI [1.90, 6.91], see Table 5). Thus, we found mixed evidence that the participants used relationship quality as a one-clever-cue-heuristic to accurately detect others' trustworthiness toward them. While relationship quality indeed predicted trustors' trust behavior, it only partly explained the accuracy of trustors' cognitive trust.

#### 4.6.2.3 Detection of General Trustworthiness

Did participants also detect trustees' general trustworthiness toward other group members? To answer this question, we estimated a multilevel regression model in which general trustworthiness predictions were regressed onto trustees' actual trustworthiness toward all other group members. Trustees' general trustworthiness was not significantly related to the trustors' general trustworthiness predictions ($\chi^2(1) = 2.36$, $p = .12$, $B = 0.11$, 95% CI [-0.03, 0.25], see Table 8), indicating that the participants were unable to accurately detect trustees' general trustworthiness.

### 4.6.3 Discussion

Again, the participants successfully detected others' specific trustworthiness toward them and knew which of their trustworthiness predictions were particularly accurate. More importantly, however, Study 3 allowed us to compare the detection accuracy of specific and general trustworthiness. As predicted, the participants accurately detected their peers' specific trustworthiness toward them but failed to do so for their peers' general trustworthiness toward others. This was true at both the cognitive and behavioral levels and indicates that trustworthiness detection critically depends on which type of trustworthiness is actually measured. Why was this the case? We assumed that the difference in accuracy might be due to the degree to which people can rely on their relationship with another person as valid information about that person's trustworthiness. Did we find evidence for this idea? At least partly. Relationship quality fully explained whether participants accurately sent their money to reciprocal peers. Thus, the

participants indeed relied on a relation-as-information heuristic when deciding whom to trust. However, this relationship was weaker for the cognitive trustworthiness predictions. Here, relationship quality only partly mediated whether participants accurately predicted their peers' reciprocity on a cognitive level. It is unclear why relationship quality fully explains the accuracy of behavioral but not cognitive trust. One reason might simply be that the participants' trust behavior was less accurate than their cognitive predictions. By this logic, relationship quality might have equally mediated behavioral and cognitive trust, but the (originally higher) accuracy for cognitive trust might nevertheless have remained better than chance. Another reason might be that trust behavior is more closely driven by relational aspects than cognitive trustworthiness predictions. Although trustors might have certain expectations of a trustee's trustworthiness, their actual trust behavior could still be largely influenced by their idiosyncratic relationship with that trustee. As a result, controlling for relationship quality would more strongly weaken behavioral but not cognitive trust accuracy. It would be interesting to further illuminate this question using multilevel mediation analyses. However, current statistical methods are based on data nested within two levels (Hayes & Rockwood, 2020; Yu & Li, 2020) and do not correspond to the data observed in our studies, which are nested simultaneously across multiple levels (groups, trustors, and trustees).

## 4.7 Study 4

In our first three studies, participants accurately detected others' specific trustworthiness toward them when they were well acquainted. Moreover, relationship quality fully mediated behavioral trust accuracy and party mediated cognitive trust accuracy. In contrast, participants had been unable to detect others' general trustworthiness. Study 4 aimed to strengthen the robustness and generalizability of these findings. First, we added an element of temptation to the trust game to test whether accuracy would change with lower overall trust and trustworthiness rates. Recall that the level of trust and trustworthiness was above 70% in all studies thus far. Could trustworthiness detection be limited to high trust environments? To answer this question, we sought to test trustworthiness detection accuracy in a lower trust environment.

Second, although we sought to test trustworthiness detection among acquaintances, Studies 2 and 3 included the possibility that the participants were friends

rather than acquaintances. Our accuracy results could therefore be inflated and limited to trustworthiness detection among friends. To test whether the results would hold under a stricter definition of acquaintanceship, we tested detection accuracy among previously unacquainted individuals after they had become acquainted for only one week.

## 4.7.1 Method

### 4.7.1.1 Participants & Analysis Strategy

We recruited four classes of 10, 11, 20, and 21 students at a German vocational school at the end of their first week of school to participate in a study on decision-making. This setup ensured that most students had known each other for only one week, as German vocational school students transitioned into new schools at the beginning of their 11th grade. All 62 students[4] attended 11th grade and were between 16 and 17 years of age. Altogether, our setup translated to an expected total of 1,000 individual trust interactions. Since some students had already known each other at the beginning of the school year, we chose to exclude those 67 interactions from the analyses. After excluding another 67 interactions because of missing data, a total of 866 individual trust interactions remained for the analyses. We followed the same analysis strategy as in Study 3.

### 4.7.1.2 Procedure

The procedure was the same as in Study 3, except for two changes. First, to test trustworthiness detection in a lower trust environment, we included an element of temptation to be untrustworthy in the trust game. Whereas the trustors still faced the same choice as in Study 3 (i.e., to either keep or send €10), the trustees decided between keeping €15 and sending back €15 (the trustworthy option) and keeping €50 and sending back €0 (the untrustworthy option). We chose this specific mechanism because it has been shown to reliably decrease overall rates for trust and trustworthiness (A. M. Evans & Krueger, 2014). Second, as mentioned, we asked the participants to list any classmates they had known before their transition to the current school, which enabled us to later exclude those interactions from the analyses.

## 4.7.2 Results

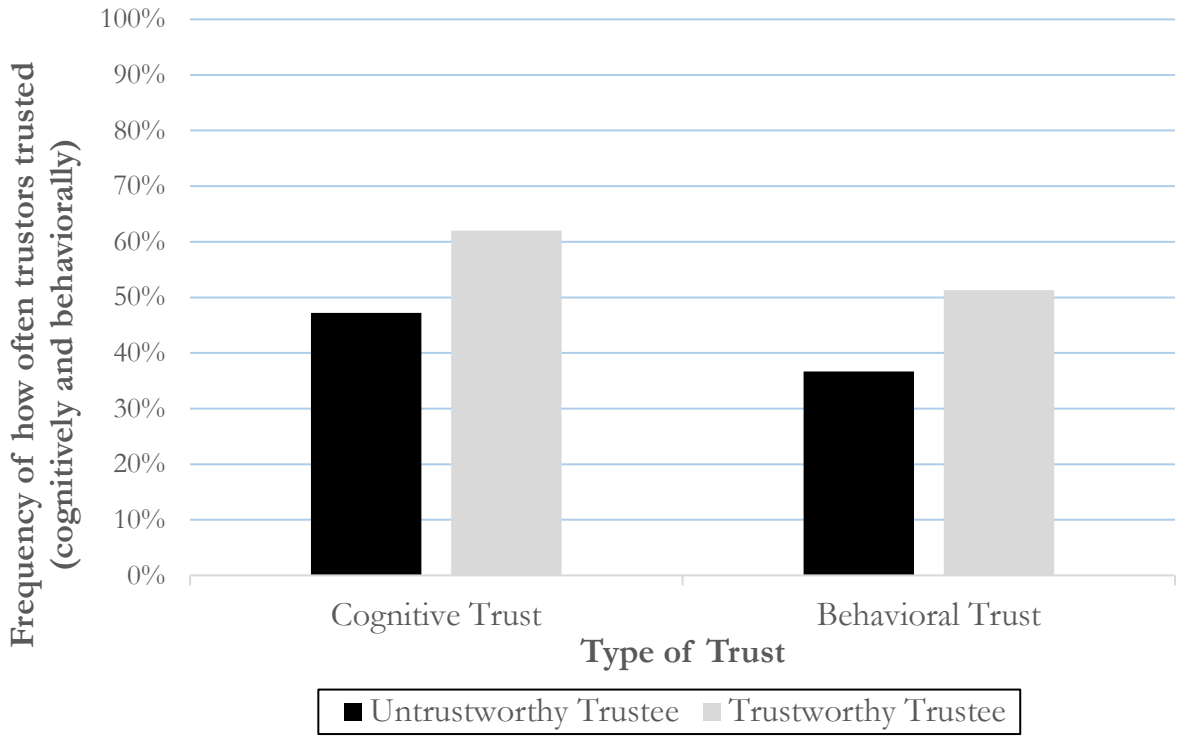
As intended, the inclusion of temptation into the trust game appeared to decrease the overall levels of trust and trustworthiness. In the role of trustor, the participants predicted their classmates to behave trustworthily toward them in 56.8% of all cases and sent their endowment to the other person in 46.2%. In the role of trustee, the participants behaved trustworthily 64.8% of the time.

### 4.7.2.1 Detection of Specific Trustworthiness

While the levels of cognitive and behavioral trust were smaller than those in Study 3, the descriptive data showed a similar pattern for accuracy. Trustworthy trustees were predicted to behave trustworthily 62.0% of the time and were trusted 51.3% of the time, whereas untrustworthy trustees were predicted to behave trustworthily 47.2% of the time and were trusted 36.7% of the time (see Figure 6).

**Figure 6**

*Frequency of how Often Trustors in Study 4 Predicted Trustees to be Trustworthy (Cognitive Trust) and Sent Money to Trustees (Behavioral Trust), Depending on Trustees' Actual Trustworthiness.*




We proceeded with separate multilevel regression analyses to test whether the detection accuracy was statistically significant. Model comparisons between the empty and

saturated models replicated the results from Study 3: trustees' actual trustworthiness was significantly related to both the trustors' cognitive trust ($\chi^2(1) = 6.50$, $p = .01$, $OR = 1.66$, 95% CI [1.14, 2.41], see Table 9) and trust behavior ($\chi^2(1) = 5.84$, $p = .02$, $OR = 1.78$, 95% CI [1.15, 2.75], see Table 10). Moreover, the accuracy of any particular trustworthiness prediction was, again, positively related to the confidence in that prediction ($\chi^2(1) = 7.37$, $p < .01$, $OR = 1.56$, 95% CI [1.13, 2.15], see Table 11). As in Study 3, the participants could accurately distinguish between (un)trustworthy individuals and knew which of their trustworthiness predictions was more accurate than others.

#### 4.7.2.2 Relationship Quality

Study 3 provided the first evidence that relationship quality mediated trustworthiness detection. Did we find support of a mediating effect of the relation-as-information heuristic? Indeed, we did; after controlling for relationship quality, the associations between the trustees' actual trustworthiness and the trustors' cognitive trust ($\chi^2(1) = 1.70$, $p = .19$, $OR = 1.32$, 95% CI [0.88, 1.97], see Table 9) and behavioral trust ($\chi^2(1) = 1.87$, $p = .17$, $OR = 1.40$, 95% CI [0.89, 2.23], see Table 10) were no longer significant. Thus, participants used their relationships as information when deciding whom to trust. This time, relationship quality fully mediated both cognitive and behavioral trust accuracy. The relation-as-information heuristic could therefore be seen as a one-clever-cue-heuristic for the valid detection of specific trustworthiness.

#### 4.7.2.3 Detection of General Trustworthiness

We next tested whether trustors could accurately detect trustees' general trustworthiness toward all other classmates. Recall that we hypothesized that the detection of general trustworthiness would be inaccurate because participants had insufficient insight into the unique relationships among classmates. Did the data support this idea? Yes, it did. As in Study 3, trustees' general trustworthiness was not significantly related to the trustors' general trustworthiness predictions ($\chi^2(1) = 2.95$, $p = .09$, $B = 0.05$, 95% CI [-0.01, 0.10], see Table 12). Thus, the participants were once again unable to accurately detect the trustees' general trustworthiness toward all other classmates.

### 4.7.3 Discussion

Study 4 replicated the previous results among recently acquainted individuals after a comparably short acquaintance period of one week and in a lower trust environment. As in Study 3, the participants accurately detected their new classmates' specific trustworthiness toward them (and knew which of their predictions were particularly accurate) but failed to detect their general trustworthiness toward other classmates. Moreover, trustworthiness detection does not appear to be limited to high trust environments. Interestingly, in contrast to Study 3, the relation-as-information heuristic mediated trustworthiness detection for both cognitive and behavioral trust. While we can only speculate why this was the case, it could be that the trustworthiness predictions after only a short acquaintance period were less accurate than in Study 3, so it was easier for relationship quality to explain detection accuracy. However, it could also be that both trust and trustworthiness among the students were driven more strongly by personal relationships than they were among the participants of banking workshops in Study 3. Thus, younger and less economically trained individuals might be less principled in their economic decisions and more likely to (accurately) base their cognitive trustworthiness predictions on their relationship quality with another person.

## 4.8 General Discussion

We started out with two questions on the detection of trustworthiness: "do people know who is generally trustworthy?" and "do people know who is trustworthy toward them?". The answer to the first question appears to be "no". In none of our four studies could the participants accurately detect who in their group was generally more or less trustworthy than others. The answer to the second question appears to be "yes". When the participants had sufficient time to become acquainted, they accurately predicted who would act trustworthily toward them and sent their money more often to those individuals. It is important to note that the difference in answers to the two questions is not trivial. In fact, it could be argued that specific trustworthiness detection was the more difficult task in our studies. This is because the trustors predicted one single behavior of a trustee toward them for the prediction of specific trustworthiness, whereas they predicted a trustee's average behavior across many interactions for the prediction of general

trustworthiness. If both tasks were equally difficult, then general trustworthiness predictions should have been more accurate because trustors' random prediction errors for each interaction could be evened out[5].

How can the discrepancy between general and specific trustworthiness detection be explained? We proposed that trustworthiness is not a stable personality trait but also has an interpersonal component. The interclass correlation coefficients from our data support this idea. Across the four studies, 8.5%, 17.4%, 36.9%, and 23.4% of the variance in trustworthiness behavior was attributable to the interpersonal dyad level. These numbers were strikingly similar to those for trust behavior, where 2.4%, 27.2%, 35.3%, and 17.9% of the variance was at the dyad level. This illustrates two points. First, at least in our studies, the participants were often not consistently (un)trustworthy but adjusted their behavior depending on their interaction partner. For example, 27 of the 51 participants in Study 3 and 38 of the 62 participants in Study 4 were not (un)trustworthy per se but tailored their trustworthiness behavior to their interaction partner. Second, this adjustment was at similar levels for trust and trustworthiness behavior. Here, 23 of the 51 participants in Study 3 and 35 of the 62 participants in Study 4 adjusted their trust behavior to their interaction partner. Taken together, while the participants did have somewhat stable tendencies to be more or less trusting and trustworthy than others in their group, they were *conditionally* rather than principally trusting and trustworthy. This conditional or specific trustworthiness thus resembled a moving target for trustors to detect. While they could do so accurately for the specific interaction between themselves and any particular trustee, they were unable to do so for the other interactions.

### 4.8.1 Relation-as-Information

How did trustors accurately detect whether a specific trustee would be trustworthy toward them? We argued that a trustee's trustworthiness toward a trustor depends on their unique relationship and that trustors would use their relationship quality as valid information for trustworthiness detection. As relationships develop over time, the detection of specific trustworthiness would be inaccurate among strangers but accurate among acquaintances and friends. The results across the four studies seem to support this

[5] In another previous study, we recorded trustees' general trustworthiness with one single behavior in an anonymous trust game in which the trustees' interaction partner was yet unknown. Here, general trustworthiness detection accuracy was also far from reaching significance.

view. The specific trustworthiness detection was inaccurate when the participants had limited time to get acquainted (and thus limited diagnostic information) but was accurate when the participants had gotten to know each other for at least a few days. When we controlled for the relationship quality that participants had developed over time, trustworthiness detection became inaccurate for behavioral trust in Study 3 and behavioral and cognitive trust in Study 4.

Thus, the accuracy of specific trustworthiness detection appears to mostly rely on the relatively simple heuristic "do I have a good relationship with this person?". Whereas people can rely on this relation-as-information heuristic for the detection of a trustee's specific trustworthiness toward them, the heuristic is inappropriate for the detection of general trustworthiness. Although we did not test this directly, it also seems difficult to use the relation-as-information-heuristic to detect the specific trustworthiness among third parties. Recall that the participants in our studies predicted each trustee's general trustworthiness by indicating how often that trustee would act trustworthily toward all other group members. Thus, general trustworthiness detection could have been achieved by accurately assessing group members' relationship quality. As general trustworthiness detection was inaccurate, however, the participants appeared to be unable to do so. This is understandable from an evolutionary perspective. Humans regularly faced evolutionary pressures to determine who would behave trustworthily *toward them* and likely developed correspondingly specific adaptations to keep track of potential cheaters (Cosmides et al., 2010). It seems less likely that humans developed additional cognitive modules to predict trustworthy behavior toward *third parties,* as there would not have been much evolutionary pressure to do so.

Why do we think that the relation-as-information-heuristic is so uniquely adept at detecting trustworthiness? Anthropological data suggest that during most of evolution, humans lived in comparably small tribes (Marlowe, 2005). With only a limited number of potential interaction partners, any trustworthiness detection ability should therefore be tailored toward assessing the trustworthiness of individuals with whom there is at least some (positive or negative) relationship. This idea is in line with arguments that relational information such as a sympathetic manner might serve as trustworthiness signals that are difficult (although not entirely impossible) to fake (Frank et al., 1993). Thus, the relation-as-information heuristic serves as a fast and frugal mechanism to accurately detect

ecologically valid trustworthiness signals that would have been prevalent during most human evolution. This distinguishes our approach from those that have suggested a human ability to accurately detect trustworthiness from nonecological information, such as cropped black-and-white photographs of strangers (e.g., see Bonnefon et al., 2013).

Moreover, the relation-as-information heuristic is useful not only because of its ecological adaptiveness but also because of its simplicity. Rather than collecting and weighing all potential cues, heuristics such as "relation-as-information" are simple and cost-effective bottom-up judgments that ignore part of the information. In a complex and uncertain world, fast and frugal heuristics have been shown to not only be efficient but also often lead to more accurate judgments than more complex procedures (Gigerenzer & Gaissmaier, 2011). For trustworthiness detection, the relation-as-information heuristic decreases decision complexity because trustors do not need to assess trustees' overall trustworthiness. The simple concept "good relationship equals trustworthiness" is sufficient to arrive at accurate trustworthiness detection. In contrast, a potential mechanism that incorporates the trustee's general trustworthiness would need to use a variety (e.g., the trustee's behavior toward various others) of rather noisy (e.g., because contextual information is missing) cues and then adjust this general trustworthiness assessment depending on the relationship with the trustee. Such a mechanism would be cognitively demanding and might lead to noisy detection results. The relation-as-information heuristic offers a simple and adaptive solution to this complexity.

### 4.8.2 Knowing When to Trust

Another finding of our studies was that the accuracy of the specific trustworthiness predictions increased with certainty in those predictions, which suggests that people have some "meta" accuracy about their detection abilities. Knowing *when* to trust one's abilities is particularly important in everyday life, where people largely self-select how frequently and trustfully they interact with interaction partners. Therefore, trustworthiness detection might not have evolved to detect everyone's trustworthiness equally but to detect those interaction partners that are "a sure bet". We would not necessarily expect accurate trustworthiness detection among strangers after short exposure, which, as Study 1 shows, was not the case. What is rather necessary is that people have an accurate assessment of *when* they can rely on their detection abilities. This

is what we found in our studies. The differences in accuracy between (un)certain trustworthiness predictions are best illustrated in Study 4, in which accuracy was exclusively driven by predictions that scored above the median level of certainty. Trustworthy targets were predicted to be trustworthy 62.2% of the time, whereas untrustworthy targets were predicted to be trustworthy only 35.3% of the time. For uncertain predictions, however, the trustworthiness predictions for trustworthy (61.6%) and untrustworthy (59.6%) targets were virtually equal. This can be regarded as evidence in favor of the hypothesis that trustworthiness detection should be accurate if individuals are acquainted and can largely self-select their interaction partners (Frank, 2005).

### 4.8.3 Limitations and Future Research

There are also limitations to our findings and their generalizability. First, the sample sizes of 144, 18, 51, and 62 participants were not particularly large. The observed detection accuracy might therefore be artificially augmented if some of the participants' idiosyncrasies rendered them more easily assessable than people are in general. However, this concern is mitigated by the fact that the initial results of Study 2 were replicated consistently across Studies 3 and 4 and that the participants accurately detected others' specific but not general trustworthiness. If the participants' trustworthiness was particularly easy to detect, trustors should have been able to accurately detect both types of trustworthiness.

Second, although we suggest that the detection of specific trustworthiness becomes more accurate as acquaintanceship increases, we did not experimentally test this hypothesis. While trustworthiness detection was inaccurate after a short group task in Study 1 but accurate after longer acquaintanceship periods in Studies 2 to 4, this increase in accuracy could be due to numerous other factors. We also cannot rule out the possibility that causality between acquaintanceship and detection accuracy is reversed. It might after all be the case that individuals develop stronger relationships with those whose intentions they can accurately assess. In future studies, it would be reasonable to experimentally manipulate the degree of acquaintance to test whether accuracy indeed increases with acquaintanceship.

Third, although it would be desirable to statistically explore the mediating effects of the relation-as-information heuristic, current methods for multilevel mediation analyses

are based on data nested within one or two levels (Hayes & Rockwood, 2020; Yu & Li, 2020). Unlike these simpler data structures, the trustworthiness detection judgments in our study are nested simultaneously at multiple levels (groups, trustors, and trustees).

### 4.8.4 Conclusion

The results across four studies show that people know whom to trust if they are sufficiently well acquainted to have developed relationships, which they use as a valid heuristic. We think that these findings advance the ongoing debate on trustworthiness detection accuracy. Most studies have treated trustworthiness as a stable personality trait and only found weak evidence for accurate trustworthiness detection. However, we suggest that studies on the detection of trustworthiness should keep the distinction between general and specific trustworthiness in mind when investigating whether and when people accurately know whom to trust. As our studies show that trustees are (un)trustworthy conditional on their interaction partner, it is unsurprising that the answer to the question "do people know who is trustworthy?" is "no". Our results illustrate that the more adequate question to ask is "do people know who is trustworthy toward them?" to which the answer appears to be "yes".

# Supplementary Material

**Table 2**

*Summary of Study 1's Multilevel Regression Models for Variables Predicting Specific Cognitive Trust (n = 719)*

| | Empty Model | | Predictor Model | |
|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 7.36*** | 4.73 – 11.44 | 5.47*** | 2.95 – 10.12 |
| Actual Trustworthiness | | | 1.47 | 0.80 – 2.68 |
| Random effects | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 1.70 | | 1.75 |
| $\tau_{00}$ (Trustee) | | 0.18 | | 0.19 |
| $\tau_{00}$ (Group) | | 0.01 | | 0.00 |
| Model fit | | | | |
| n (Trustor) | | 144 | | 144 |
| n (Trustee) | | 144 | | 144 |
| n (Group) | | 24 | | 24 |
| Observations | | 719 | | 719 |
| Deviance | | 651.69 | | 650.14 |
| Likelihood ratio tests | | $\chi^2(1) = 1.55, p = .21$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 3**

*Summary of Study 1's Multilevel Regression Models for Variables Predicting Specific Behavioral Trust (n = 719)*

| | Empty model | | Predictor model | |
|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 27468.07*** | 1239.75 – 608584.26 | 10881.39*** | 467.83 – 253093.88 |
| Actual trustworthiness | | | 3.00 | 0.44 – 20.35 |
| Random effects | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 121.10 | | 120.59 |
| $\tau_{00}$ (Trustee) | | 13.65 | | 13.21 |
| $\tau_{00}$ (Group) | | 0.00 | | 0.00 |
| Model fit | | | | |
| n (Trustor) | | 144 | | 144 |
| n (Trustee) | | 144 | | 144 |
| n (Group) | | 24 | | 24 |
| Observations | | 719 | | 719 |
| Deviance | | 449.38 | | 448.13 |
| Likelihood ratio tests | | $\chi^2(1) = 1.24, p = .26$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 4**

*Summary of Study 2's Multilevel Regression Models for Variables Predicting Specific Cognitive and Behavioral Trust (n = 302)*

| | Cognitive trust | | | | Behavioral trust | | | |
|---|---|---|---|---|---|---|---|---|
| | Empty model | | Predictor model | | Empty model | | Predictor model | |
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 9.94*** | 2.67 – 37.08 | 3.19 | 0.84 – 12.17 | 9.04** | 1.68 – 48.58 | 4.84 | 0.83 – 28.05 |
| Actual trustworthiness | | | 5.16*** | 2.39 – 11.14 | | | 2.39* | 1.02 – 5.58 |
| Random effects | | | | | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 4.82 | | 4.98 | | 8.59 | | 8.64 |
| $\tau_{00}$ (Trustee) | | 0.32 | | 0.00 | | 0.24 | | 0.16 |
| Model fit | | | | | | | | |
| n (Trustor) | | 18 | | 18 | | 18 | | 18 |
| n (Trustee) | | 18 | | 18 | | 18 | | 18 |
| Observations | | 302 | | 302 | | 302 | | 302 |
| Deviance | | 255.22 | | 238.40 | | 246.10 | | 242.06 |
| Likelihood ratio tests | $\chi^2(1) = 16.82, p < .001$ | | | | $\chi^2(1) = 4.04, p = .04$ | | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 5**

*Summary of Study 3's Multilevel Regression Models for Variables Predicting Specific Cognitive Trust (n = 898)*

| | Empty model | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 10.58*** | 5.17 – 21.66 | 2.48* | 1.17 – 5.28 | 14.73*** | 6.51 – 33.31 | 4.83*** | 1.93 – 12.10 |
| Actual trustworthiness | | | 5.83*** | 3.16 – 10.76 | | | 3.62*** | 1.90 – 6.91 |
| Relationship quality | | | | | 2.50*** | 1.96 – 3.18 | 2.28*** | 1.78 – 2.92 |
| Random effects | | | | | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 2.26 | | 2.48 | | 3.12 | | 3.15 |
| $\tau_{00}$ (Trustee) | | 0.79 | | 0.81 | | 0.80 | | 0.79 |
| $\tau_{00}$ (Group) | | 0.06 | | 0.00 | | 0.08 | | 0.02 |
| Model fit | | | | | | | | |
| n (Trustor) | | 51 | | 51 | | 51 | | 51 |
| n (Trustee) | | 51 | | 51 | | 51 | | 51 |
| n (Group) | | 3 | | 3 | | 3 | | 3 |
| Observations | | 898 | | 898 | | 898 | | 898 |
| Deviance | | 756.73 | | 722.25 | | 686.78 | | 671.01 |
| Likelihood ratio tests | | | $\chi^2(1) = 34.48, p < .001$ | | | $\chi^2(1) = 15.77, p < .001$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 6**

*Summary of Study 3's Multilevel Regression Models for Variables Predicting Specific Behavioral Trust (n = 898)*

| | Empty model | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 29.04*** | 9.92 – 85.05 | 10.73*** | 3.45 – 33.31 | 98.89*** | 20.84 – 469.13 | 57.47*** | 11.32 – 291.67 |
| Actual trustworthiness | | | 3.36*** | 1.92 – 5.88 | | | 1.80 | 0.95 – 3.41 |
| Relationship quality | | | | | 4.20*** | 2.98 – 5.92 | 3.99*** | 2.82 – 5.64 |
| Random effects | | | | | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 6.01 | | 6.00 | | 11.02 | | 10.68 |
| $\tau_{00}$ (Trustee) | | 0.02 | | 0.00 | | 0.00 | | 0.00 |
| $\tau_{00}$ (Group) | | 0.00 | | 0.00 | | 0.00 | | 0.00 |
| Model fit | | | | | | | | |
| n (Trustor) | | 51 | | 51 | | 51 | | 51 |
| n (Trustee) | | 51 | | 51 | | 51 | | 51 |
| n (Group) | | 3 | | 3 | | 3 | | 3 |
| Observations | | 898 | | 898 | | 898 | | 898 |
| Deviance | | 590.31 | | 572.46 | | 483.85 | | 480.64 |
| Likelihood ratio tests | $\chi^2(1) = 17.85, p < .001$ | | | | $\chi^2(1) = 3.20, p = .07$ | | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 7**

*Association in Study 3 Between Confidence and Actual Trustworthiness for Predicting Specific Cognitive Trust (n = 898)*

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 3.37** | 1.49 – 7.61 | 2.87* | 1.26 – 6.56 |
| Actual trustworthiness | 5.17*** | 2.74 – 9.73 | 6.45*** | 3.35 – 12.40 |
| Confidence | 1.73*** | 1.43 – 2.10 | 1.05 | 0.72 – 1.54 |
| Actual trustworthiness x confidence | | | 1.92** | 1.23 – 2.99 |
| Random effects | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 2.93 | | 3.06 |
| $\tau_{00}$ (Trustee) | | 0.90 | | 0.90 |
| $\tau_{00}$ (Group) | | 0.00 | | 0.00 |
| Model fit | | | | |
| n (Trustor) | | 51 | | 51 |
| n (Trustee) | | 51 | | 51 |
| n (Group) | | 3 | | 3 |
| Observations | | 898 | | 898 |
| Deviance | | 686.99 | | 678.52 |
| Likelihood ratio tests | | $\chi^2(1) = 8.47, p < .01$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 8**

*Summary of Study 3's Multilevel Regression Models for Variables Predicting General Cognitive Trust (n = 898)*

| | Empty model | | Model 1 | |
|---|---|---|---|---|
| Fixed effects | *B* | *95% CI* | *B* | *95% CI* |
| Intercept | 77.45*** | 68.82 – 86.08 | 76.20*** | 70.39 – 82.01 |
| General trustworthiness | | | 0.11 | -0.03 – 0.25 |
| Random effects | | | | |
| $\sigma^2$ (Interaction) | | 482.37 | | 481.87 |
| $\tau_{00}$ (Trustor) | | 340.90 | | 353.95 |
| $\tau_{00}$ (Trustee) | | 63.47 | | 63.32 |
| $\tau_{00}$ (Group) | | 30.94 | | 0.00 |
| Model fit | | | | |
| n (Trustor) | | 51 | | 51 |
| n (Trustee) | | 51 | | 51 |
| n (Group) | | 3 | | 3 |
| Observations | | 898 | | 898 |
| Deviance | | 8285.3 | | 8282.9 |
| Likelihood ratio tests | | $\chi^2(1) = 2.36, p = .12$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 9**

*Summary of Study 4's Multilevel Regression Models for Variables Predicting Specific Cognitive Trust (n = 866)*

| | Empty model | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 4.29 | 0.85 – 21.56 | 2.96 | 0.61 – 14.27 | 4.69 | 0.86 – 25.53 | 3.80 | 0.71 – 20.44 |
| Actual trustworthiness | | | 1.66** | 1.14 – 2.41 | | | 1.32 | 0.88 – 1.97 |
| Relationship quality | | | | | 1.58*** | 1.37 – 1.81 | 1.54*** | 1.34 – 1.78 |
| Random effects | | | | | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 2.46 | | 2.47 | | 2.87 | | 2.85 |
| $\tau_{00}$ (Trustee) | | 0.14 | | 0.09 | | 0.15 | | 0.12 |
| $\tau_{00}$ (Group) | | 2.41 | | 2.20 | | 2.64 | | 2.51 |
| Model fit | | | | | | | | |
| n (Trustor) | | 60 | | 60 | | 60 | | 60 |
| n (Trustee) | | 62 | | 62 | | 62 | | 62 |
| n (Group) | | 4 | | 4 | | 4 | | 4 |
| Observations | | 866 | | 866 | | 866 | | 866 |
| Deviance | | 924.55 | | 918.05 | | 880.33 | | 878.63 |
| Likelihood ratio tests | | $\chi^2(1) = 6.50, p = .01$ | | | | $\chi^2(1) = 1.70, p = .19$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 10**

*Summary of Study 4's Multilevel Regression Models for Variables Predicting Specific Behavioral Trust (n = 866)*

| | Empty model | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 3.38 | 0.27 – 42.72 | 2.21 | 0.18 – 26.61 | 3.67 | 0.24 – 54.87 | 2.85 | 0.19 – 41.98 |
| Actual trustworthiness | | | 1.78** | 1.15 – 2.75 | | | 1.40 | 0.89 – 2.23 |
| Relationship quality | | | | | 1.67*** | 1.42 – 1.98 | 1.64*** | 1.38 – 1.94 |
| Random effects | | | | | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 9.17 | | 9.15 | | 10.49 | | 10.52 |
| $\tau_{00}$ (Trustee) | | 0.06 | | 0.00 | | 0.00 | | 0.00 |
| $\tau_{00}$ (Group) | | 5.86 | | 5.50 | | 6.68 | | 6.46 |
| Model fit | | | | | | | | |
| n (Trustor) | | 60 | | 60 | | 60 | | 60 |
| n (Trustee) | | 62 | | 62 | | 62 | | 62 |
| n (Group) | | 4 | | 4 | | 4 | | 4 |
| Observations | | 866 | | 866 | | 866 | | 866 |
| Deviance | | 683.23 | | 677.39 | | 645.93 | | 644.06 |
| Likelihood ratio tests | | $\chi^2(1) = 5.84$, $p = .02$ | | | | $\chi^2(1) = 1.87$, $p = .17$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 11**

*Association in Study 4 Between Confidence and Actual Trustworthiness for Predicting Specific Cognitive Trust (n = 860)*

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| Fixed effects | *OR* | *95% CI* | *OR* | *95% CI* |
| Intercept | 2.99 | 0.61 – 14.61 | 2.89 | 0.60 – 13.94 |
| Actual trustworthiness | 1.62* | 1.11 – 2.37 | 1.66** | 1.14 – 2.43 |
| Confidence | 0.95 | 0.82 – 1.10 | 0.72* | 0.56 – 0.93 |
| Actual trustworthiness x confidence | | | 1.56** | 1.13 – 2.15 |
| Random effects | | | | |
| $\sigma^2$ (Interaction) | | 3.29 | | 3.29 |
| $\tau_{00}$ (Trustor) | | 2.48 | | 2.49 |
| $\tau_{00}$ (Trustee) | | 0.10 | | 0.08 |
| $\tau_{00}$ (Group) | | 2.23 | | 2.19 |
| Model fit | | | | |
| n (Trustor) | | 60 | | 60 |
| n (Trustee) | | 62 | | 62 |
| n (Group) | | 4 | | 4 |
| Observations | | 860 | | 860 |
| Deviance | | 909.96 | | 902.59 |
| Likelihood ratio tests | | $\chi^2(1) = 7.37, p < .01$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

**Table 12**

*Summary of Study 4's Multilevel Regression Models for Variables Predicting General Cognitive Trust (n = 866)*

| | Empty model | | Model 1 | |
|---|---|---|---|---|
| Fixed effects | *B* | *95% CI* | *B* | *95% CI* |
| Intercept | 62.81*** | 42.41 – 83.21 | 62.47*** | 42.67 – 82.27 |
| General trustworthiness | | | 0.05 | -0.01 – 0.10 |
| Random effects | | | | |
| $\sigma^2$ (Interaction) | | 536.76 | | 536.88 |
| $\tau_{00}$ (Trustor) | | 530.52 | | 533.58 |
| $\tau_{00}$ (Trustee) | | 11.89 | | 9.99 |
| $\tau_{00}$ (Group) | | 390.14 | | 364.71 |
| Model fit | | | | |
| n (Trustor) | | 60 | | 60 |
| n (Trustee) | | 62 | | 62 |
| n (Group) | | 4 | | 4 |
| Observations | | 866 | | 866 |
| Deviance | | 8084.4 | | 8081.4 |
| Likelihood ratio tests | | $\chi^2(1) = 2.95, p = .09$ | | |

*Note.* *** $p < .001$, ** $p < 0.01$, * $p < 0.05$

# Chapter 5

# Integrative Discussion

## 5.1 Key Message

The Chapters of this dissertation illustrate that trust is an elusive concept that presents itself differently depending on how it is observed. On the one hand, trust toward an unknown and unobservable trustee is largely principled (Dunning et al., 2014). The participants in Chapter 2 trusted at similar levels independent of whether they were confronted with the trust situation in their native or in a foreign language. Moreover, we observed the same following phenomenon as in previous native-language studies (e.g., Fetchenhauer & Dunning, 2009): although most participants were (unjustifiably) skeptical of the trustees' overall trustworthiness, a large proportion still decided to trust. Taken together, the use of foreign language did not significantly change people's trust, which contrasted with the results in the two moral dilemmas. Here, the use of foreign language increased the share of participants who chose a consequentialist over a deontological (hypothetical) moral decision.

A potential explanation for this difference might be that participants faced a decision with real consequences in the trust game but only a hypothetical decision in the moral dilemmas. Bostyn et al. (2018) showed that the choices in moral dilemmas with real consequences that involved painful electric shocks for either one or five laboratory mice were different from the choices in hypothetical but otherwise identical moral dilemmas. Moreover, only hypothetical but not consequential choices were related to participants' choices in trust dilemmas such as the trolley dilemmas used in Chapter 2. For future studies, it would therefore be worthwhile to investigate a) whether choices in consequential moral dilemmas might be immune to foreign language effects and b) whether choices in hypothetical trust situations could be influenced by foreign language. Another potential explanation for the different choices in the trust game and the moral dilemmas might be the frequency with which people normally experience both situations. Whereas people regularly find themselves in trust situations and might have developed principled stances on trust behavior, they are less often confronted with the specific situations presented in moral dilemmas. Choices in moral dilemmas might consequently be more malleable by situational factors, while choices in trust situations might be made more principally on the basis of internalized norms.

However, this dissertation also demonstrates that, depending on the situation, trust may not be entirely principled. When the participants in Chapter 4 made their trust decisions toward specific and observable trustees, most of their trust behavior was *conditional* on the trustees' perceived trustworthiness rather than *on principle*. Moreover, trust toward specific observable trustees was not characterized by the phenomenon that individuals trusted too much given their (unjustifiably) pessimistic trustworthiness expectations. These findings suggest that perceptions of trustworthiness are especially influential when potential trustees can be observed, which, in turn, raises the question whether these perceptions are actually accurate. As mentioned in the introduction of this dissertation, a reliance on trustworthiness perceptions might be good or bad depending on the accuracy of these perceptions.

Let us start with the good news. Chapter 4 indicates that people accurately detect whether another person will be trustworthy toward them when they are sufficiently acquainted with that person. When we asked participants to rate their confidence in the trustworthiness assessments, we further discovered that they knew *which* of their assessments were particularly accurate. This suggests that trust behavior in everyday life might be even more accurate than in our studies because such a "meta" accuracy about one's detection abilities should be particularly useful in situations in which a trustor can choose between multiple potential trustees (for similar arguments, see Frank, 2005). When deciding whom to select as the godfather of one's child, for example, it is sufficient to know of *one* particular person worthy of one's trust. Studies 3 and 4 of Chapter 4 further identified how people arrive at accurate trustworthiness perceptions. The participants in our studies accurately detected their acquaintances' trustworthiness toward them by taking their relationship quality into account. By using the simple heuristic "do I have a good relationship with this person?" – which my co-author Thomas Schlösser fittingly named *relation-as-information* – they were able to determine which of their acquaintances would be (un)trustworthy toward them.

Unfortunately, assessments about others' trustworthiness are not always accurate. When we asked the participants in our studies to assess their acquaintances' *general* trustworthiness toward others instead of their *specific* trustworthiness toward them, detection accuracy dropped to chance levels. For one, this illustrates the importance of dyadic decision-making processes like the relation-as-information heuristic for

trustworthiness detection. For another, it shows that people should be cautious when extrapolating from one's own experience with a person to how that person will behave in general or toward third parties. Study 1 of Chapter 4 revealed an additional limitation to the accuracy of trustworthiness detection. When the participants assessed the trustworthiness of *strangers* toward them after only a short group task, they failed to distinguish between (un)trustworthy individuals. Although speculative, one potential explanation for the inaccuracy might be that the participants had not had enough time to get acquainted, which would have prevented them from using the relation-as-information heuristic for accurate dyadic trustworthiness assessments. For future studies, it would therefore be worthwhile to a) experimentally test whether detection accuracy indeed increases with acquaintanceship and b) measure whether people increasingly rely on the relation-as-information heuristic as acquaintance grows.

Overall, Chapter 4 warns us not to be overconfident when assessing acquaintances' general trustworthiness or when assessing people's trustworthiness at little acquaintance. The literature review in Chapter 3 underpins this notion. While face-to-face interactions with another person might provide useful cues or signals and improve detection accuracy, people appear to be unable to accurately detect strangers' trustworthiness when limited to little information (e.g., neutral photographs of these strangers). This inaccuracy might seem surprising when considering the congruency of people's trustworthiness assessments (Todorov et al., 2008). Why do people agree who appears trustworthy if these assessments are inaccurate? One potential explanation might be that people systematically try to identify character traits from momentary snapshots of another person (temporal extension) and overgeneralize apparent emotional resemblances in otherwise neutral expressions (Todorov, Olivola, et al., 2015). A person who appears to slightly smile in a photograph (e.g., because of unique facial structures) might, therefore, be systematically rated as more friendly, warm, or trustworthy than another person who exhibits a fully neutral expression. Emotion overgeneralization might thus be an example of an evolutionary mismatch (N. P. Li et al., 2018) in that cognitive processes involved in normally useful inferences about the emotional state of another person in face-to-face interactions become falsely activated when observing neutral photographs. A photograph of a slightly smiling person might thus be interpreted as if that person was slightly smiling toward oneself, which might then activate cognitive processes akin to the relation-as-

information heuristic reported above. The importance of even minimal emotional expressions for impression formation was recently underscored in a study showing that the emotion resemblances in neutral faces were most indicative of whether a face was perceived as (un)trustworthy (Jaeger & Jones, 2021). The effects of first impressions might also be reinforced by self-fulfilling effects (Todorov, Olivola, et al., 2015). Hong et al. (2021) showed that individuals whose photographs were perceived as untrustworthy were approached more often with offers of unethical behavior (i.e., accepting a bribe) than their trustworthy-appearing counterparts. Consequently, untrustworthy-appearing individuals ended up with a larger number of accepted bribes, even though the relative chance of accepting a bribe was equally distributed between (un)trustworthy-appearing individuals.

The Chapters of this dissertation can be condensed to the following key messages. Trust toward an unknown and unobservable person appears to be largely principled and immune to situational influences such as the use of a foreign language. Trust toward an observable person, however, appears to be less principled and more strongly influenced by how that person is perceived. These perceptions appear to be accurate when assessing how an acquainted person will behave toward oneself but inaccurate when assessing how that same person will behave toward others in general. Moreover, trustworthiness detection at zero (or little) acquaintance appears to be inaccurate, especially when people are unable to interact with the other person and the information about the other person is limited. As a result, people in their everyday lives should cautiously follow their trustworthiness assessments toward acquainted others when they are confident in their assessments but be skeptical of trustworthiness impressions at first sight.

## 5.2 The Elusiveness of Trust

One feature of trust that is particularly worthy of discussion is its elusiveness. Trust is an exceptionally tricky concept to measure as its dynamics change depending on how or in which situation it is studied. For example, this dissertation illustrates the importance to distinguish trust toward anonymous and unobservable trustees (which is mostly principled in nature) from trust toward specific and observable trustees (which appears to also be influenced by characteristics of the trustees and the trustor-trustee relationship). A potential explanation for the different trust dynamics could stem from the

access to potential trustworthiness cues. That is, people might simply follow these cues whenever they have the opportunity to do so. When we investigated trustworthiness detection from short videos of trustees, we found that trustor characteristics only accounted for 57.8% of the variance in trust behavior and 14.9% of the variance in trustworthiness expectations (Siuda et al., 2019, unpublished data). To be sure, trust behavior was still influenced by trustors' general stances on whether to trust but presenting visual cues of the potential trustees appeared to render trust less strongly principled and more strongly reliant on potential cues to trustees' trustworthiness that could be extracted from the videos. This finding is congruent with a recent study showing that trust decisions are influenced by the (visual) presentation of potential trustworthiness cues which may lead trustors to underemphasize more important aspects of the trust situation (Jaeger et al., 2019). People even appear to follow visual trustworthiness cues when they are explicitly told that these trustworthiness cues do not hold much validity. For example, Jaeger, Todorov, et al. (2020) reported that educating participants on the inaccuracy of facial inferences reduced explicit trustworthiness impressions but did not ultimately affect participants' behavior. Taken together, people appear to react to potential cues to the trustee's trustworthiness when such cues are available. This, in turn, may lead trust to be less principled and more strongly reliant on the cues accessible in the specific trust situation.

The less principled and more strongly trustee-dependent nature of trust toward observable (vs. unobservable) trustees could also help explain the comparably high correspondence between trustors' cognitive trustworthiness expectations and their actual trust behavior that we found in Chapter 4. Although speculative, people might more strongly follow their trustworthiness expectations when those are informed by trustworthiness cues. It is interesting to note that we observed high correspondences between trustors' cognitive and behavioral trust rates in the aforementioned video study as well as in the studies presented in Chapter 4. Thus, the participants in our studies seemed to follow both accurate (Studies 2-4 of Chapter 4) and inaccurate (Video Study & Study 1 of Chapter 4) trustworthiness expectations, which suggests that the observed correspondence between cognitive and behavioral trust might largely be caused by the mere presence (but not validity) of trustworthiness cues. For example, the participants in the video study appeared to falsely take the sex of potential trustees into consideration.

Although the return rates of male (39.8%) and female trustees (40.2%) did not significantly differ, female trustees were trusted more often (in 74.9% of all cases) than male trustees (in 60.8% of all cases). Unfortunately, we do not fully know why the participants took trustee sex into consideration and whether the participants knew of the unreliableness of such cues. In this regard, it would be interesting to explore trustors' confidence in their trustworthiness expectations more systematically in future work: Does confidence in one's trustworthiness expectations increase with the mere presence of trustworthiness cues (e.g., a visual presentation of the trustee) or with the actual validity of these cues (e.g., the quality of the available cues)?

Another difference between trust toward observable (vs. unobservable) trustees is that trust does not appear to be characterized by overly pessimistic expectations of trustworthiness when trustors see with whom they interact. Across all of the 2904 trust interactions in Chapter 4, trustworthiness expectations (73.0%), trust behavior (70.2%), and actual trustworthiness (76.5%) were at similar levels. Recall, moreover, that the overall correspondence between trust and trustworthiness rates did not automatically lead to accurate trustworthiness expectations. Although all three variables were at similar levels in Study 1 of Chapter 4, the overall accuracy of trustworthiness expectations and trust behavior was not better than chance. Thus, we did not observe too little cognitive and too much behavioral trust (see Fetchenhauer & Dunning, 2009) because trustors had accurate assessments of trustees' trustworthiness but because the dynamics of trust appeared to have changed.

There are also differences in the trust dynamics within the same trust paradigm. The results of Chapter 4 indicate that trust behavior was more strongly principled than cognitive trust. Characteristics of the trustors accounted for 87.7% (Study 1), 70.8% (Study 2), 64.5% (Study 3), and 49.9% (Study 4) of the variance in trust behavior but only for 32.8% (Study 1), 57.2% (Study 2), 35.3% (Study 3), and 29.7% (Study 4) of the variance in cognitive trustworthiness expectations. The less principled nature of cognitive trust also extended to trustworthiness expectations about trustees' general trustworthiness for which trustor characteristics accounted for 37.1% (Study 3) and 36.1% (Study 4) of the variance. As already mentioned, we found similar differences in the principledness of cognitive and behavioral trust in the aforementioned video study. Therefore, while people might readily differentiate between whom they *perceive* as trustworthy, they might be more

reluctant to differentiate between whom to *trust* when there are actual consequences for both themselves and the other person. This demonstrates once again the importance to distinguish cognitive from behavioral trust and suggests that, compared to their trustworthiness expectations, people do act rather principally even when they can see (and personally know) the trustee.

## 5.3 Trustworthiness and its Detection

Chapters 3 and 4 of this dissertation covered whether and when people accurately detect others' trustworthiness. The work that contributed to both chapters demonstrates that the answers to these questions are more complex than we originally appreciated. When we first started the empirical investigation of trustworthiness detection, we had adopted (at least implicitly) the apparent general consensus in the literature that trustworthiness was strongly person-dependent and not much unlike a personality trait. In fact, most studies on trustworthiness detection accuracy design and operationalize their experiments in accordance with this notion. Studies trying to identify trustworthiness cues from neutral voice recordings (Schild et al., 2020) or neutral pictures (Bonnefon et al., 2017a) measure trustworthiness via anonymous trust games analogous to those we used in Chapter 2. This single measurement of a person's general trustworthiness automatically presumes that trustworthiness is not a situation- or relation-specific but a person-specific and therefore stable character trait.

It is not unlikely that the attempts to identify valid trustworthiness cues grew out of a desire to explain the congruency of facial trustworthiness impressions (Todorov, Olivola, et al., 2015). In contrast to studies that explain people's congruency in terms of cognitive biases such as emotion overgeneralization (Todorov et al., 2008) or consequentialist approaches such as self-fulfilling prophecy effects (Hong et al., 2021), some studies have suggested that people largely agree on who appears trustworthy because trustworthiness impressions are simply accurate (Bonnefon et al., 2015). For example, stable facial characteristics such as people's facial-width-to-height ratio, have been proposed as a potential mechanism to identify which people behave (un)trustworthily in trust games (Stirrat & Perrett, 2010) or (un)cooperatively in public-goods games (Stirrat & Perrett, 2012). These studies generally build on research showing that a variety of personal attributes can be accurately assessed based on thin slices of

behavior (Ambady & Rosenthal, 1992). Among other things, people have been shown to accurately predict strangers' sexual orientation (Rule et al., 2009; Rule & Ambady, 2008), political ideology (Samochowiec et al., 2010) as well as intelligence and personality traits (Carney et al., 2007) after watching short silent video clips of those strangers. Therefore, it would not have been unreasonable to assume that trustworthiness would also be detectable from little information – if trustworthiness similarly were a stable character trait of a person.

As it turned out, however, the implicit assumption that people are consistently (un)trustworthy independent of whom they are interacting with did not hold true. When we first recorded people's trust game behavior toward each other in small seminar groups, a large proportion of them behaved (un)trustworthily *conditional* on their interaction partner. The conditional nature of trustworthiness can also be understood from its definition as "the intention or the behavior not to take advantage of the vulnerability of another person" given in Chapter 4. Thus, people usually behave (un)trustworthily *toward another person* which differentiates trustworthiness from person-specific characteristics such as sexual orientation, political ideology, intelligence, or personality traits. This relational aspect of trustworthiness is consequential for its detection. Whereas person-specific characteristics of a person should be detectable from another person independent of any relational aspects, relational aspect are important for situation- or relation-specific characteristics such as trustworthiness. In other words, people might be able to accurately predict whether another person is generally extraverted or introverted but fail to accurately predict whether that person is trustworthy without knowing the identity of the person's interaction partner.

To be sure, people appear to be somewhat principled in their trustworthiness behavior. Between 62.1% and 86.7% of the participants' trustworthiness behavior was attributable to stable trustee characteristics in the studies presented in Chapter 4. Moreover, trustworthiness as measured in anonymous trust games appears to be weakly to moderately associated with the Honesty-Humility trait of the HEXACO model of personality (Thielmann & Hilbig, 2015). It might therefore be possible that people assess another person's prosocial tendencies and infer that person's general trustworthiness from this information somewhat accurately. After all, the participants' predictions of acquaintances' general trustworthiness trended toward marginal significance in Study 3

and were marginally significant in Study 4. It is not unlikely that a study with a different set of participants might show accurate detection of trustees' general trustworthiness.

It is therefore important to note that we do not rule out that people may be able to accurately detect others' general trustworthiness. We argue, however, that the detection of general trustworthiness is a) not the type of trustworthiness detection task people regularly find themselves in and b) different (and less accurate) than the detection of specific trustworthiness because relational information cannot directly be used for assessments. In addition to the lack of relational information, general trustworthiness detection also appears to be difficult according to the Realistic Accuracy Model (Funder, 1995) because less visible traits such as deceptiveness are more difficult to detect than more visible traits such as extraversion or talkativeness (Funder, 2012). Thus, it is still an open question whether people's general trustworthiness (or the trustworthiness toward specific third parties) can accurately be detected. Research from person perception suggests that acquaintanceship might be a contributing factor to the accurate detection of general trustworthiness. Early work has already established that personality inferences become more accurate as acquaintanceship increases (Funder et al., 1995; Funder & Colvin, 1988; Paulhus & Bruce, 1992; Paunonen, 1989) and Lee and Ashton (2017) more recently found that the less visible traits such as conscientiousness, agreeableness and honesty-humility appear to especially profit from this acquaintanceship effect. For future work it would therefore be interesting to investigate whether general trustworthiness detection becomes accurate as acquaintanceship increases.

Acquaintanceship is also likely to be an underlying requirement for the accurate detection of trustees' *specific* trustworthiness. The participants in Chapter 4 relied on inferences about their relationship with a potential trustee when deciding whether to trust. From a theoretical perspective, the participants therefore needed at least some minimal relationship with a potential trustee for the detection of that trustee's specific trustworthiness. After we had conducted a series of studies, a pattern indeed seemed to emerge that supports this view. Whereas the participants were unable to accurately detect strangers' specific trustworthiness after short interaction in Study 1, the participants in Studies 2 and 3, who were already acquainted, accurately detected their acquaintances' specific trustworthiness. We therefore sought to experimentally test the influence of acquaintanceship on detection accuracy in Study 4. We had originally planned to compare

the detection accuracy of a first group of participants during their first week of school with that of a second group of participants at the end of their first semester of school. Although we successfully conducted the experiment at the start of the school year, the coronavirus pandemic unfortunately prevented us from conducting the second part of the experiment. Future research is therefore needed to experimentally test the hypothesis that acquaintanceship improves trustworthiness detection accuracy. In this regard it would be valuable to also explore whether the detection of general trustworthiness reaches better than chance levels with sufficient acquaintanceship. Future studies should keep in mind, however, that we already found an accurate detection of specific trustworthiness among students who knew each other for only one week. Although overall accuracy might have increased even further with each passing week, it might be useful to focus on the first emergence of acquaintanceship for trustworthiness detection.

## 5.3 Directions for Future Research

A variety of directions for future research have emerged during the integrative discussion of this dissertation. Regarding foreign language effects, we found that trustworthiness expectations and trust behavior in a consequential trust game appeared to be immune to foreign language whereas choices in hypothetical moral dilemmas were influenced by foreign language. It would be interesting to more systematically investigate whether the hypotheticality of a situation may be an underlying moderator for foreign language effects. For trust research, it may be worthwhile to investigate whether choices in a hypothetical trust situation akin to the moral dilemma vignettes may be influenced by foreign language. For research on moral decision-making, it would be equally (if not more) relevant to experimentally test whether foreign language effects on moral behavior hold up in the real world.

Regarding trustworthiness detection, our results suggest that acquaintanceship might be a prerequisite for accuracy. Since we have not yet experimentally tested this idea, however, future work is needed to confirm this hypothesis. As already mentioned, however, future work should keep in mind that accurate trustworthiness detection in our studies emerged in a matter of days rather than weeks. Thus, future studies should test the acquaintanceship effect early in the getting-to-know-you process. Regarding the potential mechanisms for trustworthiness detection, it would also be interesting to explore whether

the use of the relation-as-information heuristic increases with accuracy. The participants in our studies had a surprisingly good understanding of their detection abilities which suggests that people might more strongly rely on valid trustworthiness cues (and ignore invalid trustworthiness cues) as they become available. People might therefore more strongly rely on their relationships as information for trustworthiness detection when these relationships have (positively or negatively) crystallized themselves.

In contrast to studies that investigated people's trustworthiness detection abilities from little information (e.g., neutral photographs), we found that trustors did not always blindly follow all available trustworthiness cues. While the participants in the aforementioned video-study falsely used trustee sex as an apparent cue to trustworthiness, the participants in Chapter 4 (who had more relevant cues available) did not show such a sex bias. In this regard, it would be interesting to explore more systematically which trustworthiness cues trustors themselves think are valid and whether these cues actually are valid. Moreover, it would be interesting to investigate whether trustors' confidence in their trustworthiness expectations increases with the mere presence of cues or with the true validity of these cues. Future studies might, for example, compare trust confidence between anonymous trust games, trust games with little information about the trustee (e.g., neutral photographs or videos), and trust games with previous trustor-trustee interaction.

# References


Albohn, D. N., & Adams, R. B. (2020). Emotion residue in neutral faces: Implications for impression formation. *Social Psychological and Personality Science*, 194855062092322. https://doi.org/10.1177/1948550620923229

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, *9*(3), 193–208. https://doi.org/10.1007/s10683-006-9122-4

Ask, K., Calderon, S., & Mac Giolla, E. (2020). Human lie-detection performance: Does random assignment vs. self-selection of liars and truth-tellers matter? *Journal of Applied Research in Memory and Cognition*, *9*(1), 128–136. https://doi.org/10.1016/j.jarmac.2019.10.002

Bacharach, M., & Gambetta, D. (2001). Trust in signs. In K. S. Cook (Ed.), *Trust in society* (pp. 148–184). Russel Sage Foundation.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Becker, A., Deckers, T., Dohmen, T., Falk, A., & Kosse, F. (2012). The relationship between economic preferences and psychological personality measures. *Annual Review of Economics*, *4*(1), 453–478. https://doi.org/10.1146/annurev-economics-080511-110922

Ben-Ner, A., & Halldorsson, F. (2010). Trusting and trustworthiness: What are they, how to measure them, and what affects them. *Journal of Economic Psychology*, *31*(1), 64–79. https://doi.org/10.1016/j.joep.2009.10.001

Bereby-Meyer, Y., Hayakawa, S., Shalvi, S., Corey, J. D., Costa, A., & Keysar, B. (2020). Honesty Speaks a Second Language. *Topics in Cognitive Science*, *12*(2), 632–643. https://doi.org/10.1111/tops.12360

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Białek, M., & Fugelsang, J. (2019). No evidence for decreased foreign language effect in highly proficient and acculturated bilinguals: a commentary on Čavar and Tytus (2018). *Journal of Multilingual and Multicultural Development*, *40*(8), 679–686. https://doi.org/10.1080/01434632.2018.1547072

Białek, M., Paruzel-Czachura, M., & Gawronski, B. (2019). Foreign language effects on moral dilemma judgments: An analysis using the CNI model. *Journal of Experimental Social Psychology*, *85*, 103855. https://doi.org/10.1016/j.jesp.2019.103855

Binzel, C., & Fehr, D. (2013). Social distance and trust: Experimental evidence from a slum in Cairo. *Journal of Development Economics*, *103*, 99–106. https://doi.org/10.1016/j.jdeveco.2013.01.009

Bloomberg (2021, September 20). Bevölkerungsanteil mit COVID-19-Impfung nach ausgewählten Ländern weltweit [Proportion of population with COVID-19 vaccination by selected countries worldwide]. *Statista*. https://de.statista.com/statistik/daten/studie/1203308/umfrage/impfstoffabdeckung-der-bevoelkerung-gegen-das-coronavirus-nach-laendern/

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology: General*, *142*(1), 143–150. https://doi.org/10.1037/a0028930

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences*, *19*(8), 421–422. https://doi.org/10.1016/j.tics.2015.05.002

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2017a). Can we detect cooperators by looking at their face? *Current Directions in Psychological Science*, *26*(3), 276–281. https://doi.org/10.1177/0963721417693352

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2017b). Trustworthiness perception at zero acquaintance: Consensus, accuracy, and prejudice. *Behavioral and Brain Sciences*, *40*, E4. https://doi.org/10.1017/S0140525X15002319

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas. *Psychological Science*, *29*(7), 1084–1093. https://doi.org/10.1177/0956797617752640

Brislin, R. W. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Bushman, B. J., & Wang, M. C. (1994). Vote-counting procedures in meta-analysis. *The Handbook of Research Synthesis*, *236*, 193–213.

Caldwell, C., & Clapham, S. E. (2003). Organizational Trustworthiness: An International Perspective. *Journal of Business Ethics*, *47*(4), 349–364. https://doi.org/10.1023/A:1027370104302

Caldwell-Harris, C. L. (2015). Emotionality Differences Between a Native and Foreign Language. *Current Directions in Psychological Science*, *24*(3), 214–219. https://doi.org/10.1177/0963721414566268

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*(5), 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004

Čavar, F., & Tytus, A. E. (2017). Moral judgement and foreign language effect: When the foreign language becomes the second language. *Journal of Multilingual and Multicultural Development*, *39*(1), 17–28. https://doi.org/10.1080/01434632.2017.1304397

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, *61*(2), 87–105. https://doi.org/10.1016/j.cogpsych.2010.03.001

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601. https://doi.org/10.1111/j.1468-0262.2006.00719.x

Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The Moral Foreign-Language Effect. *Philosophical Psychology*, *29*(1), 23–40. https://doi.org/10.1080/09515089.2014.993063

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science*, *25*(5), 1132–1139. https://doi.org/10.1177/0956797614523297

Cohen, T. R., Kim, Y., & Panter, A. T. (2014). *Five-Item Guilt Proneness Scale (GP-5).* https://doi.org/10.13140/RG.2.1.2847.2167

Coleman, J. S. (1990). *Foundations of social theory*. Belknap Press.

Columbus, S., Molho, C., Righetti, F., & Balliet, D. (2021). Interdependence and cooperation in daily life. *Journal of Personality and Social Psychology*, *120*(3), 626–650. https://doi.org/10.1037/pspi0000253

Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 9007–9014. https://doi.org/10.1073/pnas.0914623107

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, *163*, 163–228.

Costa, A., Duñabeitia, J. A., & Keysar, B. (2019). Language context and decision-making: Challenges and advances. *Quarterly Journal of Experimental Psychology*, *72*(1), 1–2. https://doi.org/10.1177/1747021818789799

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). "Piensa" twice: On the foreign language effect in decision making. *Cognition*, *130*(2), 236–254. https://doi.org/10.1016/j.cognition.2013.11.010

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS ONE*, *9*(4), e94842.

Costa, A., Vives, M.-L., & Corey, J. D. (2017). On Language Processing Shaping Decision Making. *Current Directions in Psychological Science*, *26*(2), 146–151. https://doi.org/10.1177/0963721416680263

De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2013). Low second-to-fourth digit ratio predicts indiscriminate social suspicion, not improved trustworthiness detection. *Biology Letters*, *9*(2), 20130037. https://doi.org/10.1098/rsbl.2013.0037

De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2015). Adolescents gradually improve at detecting trustworthiness from the facial features of unknown adults. *Journal of Economic Psychology*, *47*, 17–22. https://doi.org/10.1016/j.joep.2015.01.002

De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2017). Split-second trustworthiness detection from faces in an economic game. *Experimental Psychology*, *64*(4), 231–239. https://doi.org/10.1027/1618-3169/a000367

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., & Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, *23*(12), 1549–1556. https://doi.org/10.1177/0956797612448793

Dewaele, J.-M. (2004). The Emotional Force of Swearwords and Taboo Words in the Speech of Multilinguals. *Journal of Multilingual and Multicultural Development*, *25*(2-3), 204–222. https://doi.org/10.1080/01434630408666529

Díaz-Lago, M., & Matute, H. (2019). Thinking in a Foreign language reduces the causality bias. *Quarterly Journal of Experimental Psychology*, *72*(1), 41–51. https://doi.org/10.1177/1747021818755326

Dilger, A., Müller, J., & Müller, M. (2017). Is trustworthiness written on the face? *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.2930064

Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology*, *107*(1), 122–141. https://doi.org/10.1037/a0036673

Dunning, D., & Fetchenhauer, D. (2010). Trust as an expressive rather than an instrumental act. In S. R. Thye & E. J. Lawler (Eds.), *Advances in Group Processes. Advances in Group Processes* (Vol. 27, pp. 97–127). Emerald Group Publishing Limited. https://doi.org/10.1108/S0882-6145(2010)0000027007

Dunning, D., Fetchenhauer, D., & Schlösser, T. M. (2012). Trust as a social and emotional act: Noneconomic considerations in trust behavior. *Journal of Economic Psychology*, *33*(3), 686–694. https://doi.org/10.1016/j.joep.2011.09.005

Dunning, D., Fetchenhauer, D., & Schlösser, T. (2017). The varying roles played by emotion in economic decision making. *Current Opinion in Behavioral Sciences*, *15*, 33–38. https://doi.org/10.1016/j.cobeha.2017.05.006

Dunning, D., Fetchenhauer, D., & Schlösser, T. (2019). Why people trust: Solved puzzles and open mysteries. *Current Directions in Psychological Science*, *3*, 096372141983825. https://doi.org/10.1177/0963721419838255

Dylman, A. S., & Champoux-Larsson, M.-F. (2020). It's (not) all Greek to me: Boundaries of the foreign language effect. *Cognition*, *196*, 104148. https://doi.org/10.1016/j.cognition.2019.104148

Eckel, C. C., & Petrie, R. (2011). Face value. *American Economic Review*, *101*(4), 1497–1513. https://doi.org/10.1257/aer.101.4.1497

Eckel, C. C., & Wilson, R. K. (2004). Is trust a risky decision? *Journal of Economic Behavior & Organization*, *55*(4), 447–465. https://doi.org/10.1016/j.jebo.2003.11.003

Edelman (2021, May 24). Percentage of persons with trust in healthcare in 2021, by country. *Statista*. https://www.statista.com/statistics/1071027/trust-levels-towards-healthcare-in-select-countries/

Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports*, *3*, 1047. https://doi.org/10.1038/srep01047

Evans, A. M., & Krueger, J. I. (2014). Outcomes and expectations in dilemmas of trust. *Judgment & Decision Making*, *9*(2), 90–103.

Evans, J. S. B. T. (1984). Heuristic and analytic processes in reasoning *British Journal of Psychology*, *75*(4), 451–468. https://doi.org/10.1111/j.2044-8295.1984.tb01915.x

Ferrin, D. L., Bligh, M. C., & Kohles, J. C. (2008). It takes two to tango: An interdependence analysis of the spiraling of perceived trustworthiness and cooperation in interpersonal and intergroup relationships. *Organizational Behavior and Human Decision Processes*, *107*(2), 161–178. https://doi.org/10.1016/j.obhdp.2008.02.012

Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology*, *30*(3), 263–276.

Fetchenhauer, D., & Dunning, D. (2012). Betrayal aversion versus principled trustfulness—How to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization*, *81*(2), 534–541. https://doi.org/10.1016/j.jebo.2011.07.017

Fetchenhauer, D., & van der Vegt, G. (2001). Honesty, Trust and Economic Growth. *Zeitschrift Für Sozialpsychologie*, *32*(3), 189–200. https://doi.org/10.1024//0044-3514.32.3.189

Foo, Y. Z., Loncarevic, A., Simmons, L. W., Sutherland, C. A. M., & Rhodes, G. (2019). Sexual unfaithfulness can be judged with some accuracy from men's but not women's faces. *Royal Society Open Science*, *6*(4), 181552. https://doi.org/10.1098/rsos.181552

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 143–155. https://doi.org/10.1093/0199252866.003.0002

Frank, R. H. (2005). Altruists with green beards: Still kicking? *Analyse & Kritik*, *27*(1), 110. https://doi.org/10.1515/auk-2005-0104

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, *14*(4), 247–256. https://doi.org/10.1016/0162-3095(93)90020-I

Frey, D. P., & Gamond, L. (2015). Second Language Feedback Reduces the Hot Hand Fallacy, But Why? *Journal of Neuroscience*, *35*(34), 11766–11768. https://doi.org/10.1523/JNEUROSCI.2295-15.2015

Fukuyama, F. (1995). *Trust: The social virtues and the creation of prosperity*. Free Press.

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., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology*, *55*(1), 149–158. https://doi.org/10.1037//0022-3514.55.1.149

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, *69*(4), 656–672. https://doi.org/10.1037/0022-3514.69.4.656

Gao, S., Zika, O., Rogers, R. D., & Thierry, G. (2015). Second language feedback abolishes the "hot hand" effect during even-probability gambling. *Journal of Neuroscience*, *35*(15), 5983–5989. https://doi.org/10.1523/jneurosci.3622-14.2015

García-Palacios, A., Costa, A., Castilla, D., Del Río, E., Casaponsa, A., & Duñabeitia, J. A. (2018). The effect of foreign language in fear acquisition. *Scientific Reports*, *8*(1), 1157. https://doi.org/10.1038/s41598-018-19352-8

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

Geipel, J., Hadjichristidis, C., & Klesse, A.-K. (2018). Barriers to sustainable consumption attenuated by foreign language use. *Nature Sustainability*, *1*(1), 31–33. https://doi.org/10.1038/s41893-017-0005-9

Geipel, J., Hadjichristidis, C., & Surian, L. (2015a). The Foreign Language Effect on Moral Judgment: The Role of Emotions and Norms. *PLoS ONE*, *10*(7), e0131529. https://doi.org/10.1371/journal.pone.0131529

Geipel, J., Hadjichristidis, C., & Surian, L. (2015b). How foreign language shapes moral judgment. *Journal of Experimental Social Psychology*, *59*, 8–17. https://doi.org/10.1016/j.jesp.2015.02.001

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

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (pp. 35–80). MIT Press.

Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D., Nystrom, L. E [Leigh E.], Engell, A. D., Darley, J. M [John M.], & Cohen, J. D [Jonathan D.] (2004). The neural bases of cognitive conflict and control in

moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E [L. E.], Darley, J. M [J. M.], & Cohen, J. D [J. D.] (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Hadjichristidis, C., Geipel, J., & Keysar, B. (2019). The influence of native language in shaping judgment and choice. *Progress in Brain Research*, *247*, 253–272. https://doi.org/10.1016/bs.pbr.2019.02.003

Harris, C. L., Ayçíçeğí, A., & Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics*, *24*(4), 561–579. https://doi.org/10.1017/S0142716403000286

Hayakawa, S., Costa, A., Foucart, A., & Keysar, B. (2016). Using a foreign language changes our choices. *Trends in Cognitive Sciences*, *20*(11), 791–793. https://doi.org/10.1016/j.tics.2016.08.004

Hayakawa, S., & Keysar, B. (2018). Using a foreign language reduces mental imagery. *Cognition*, *173*, 8–15. https://doi.org/10.1016/j.cognition.2017.12.010

Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking more or feeling less? Explaining the foreign-language effect on moral judgment. *Psychological Science*, *28*(10), 1387–1397. https://doi.org/10.1177/0956797617720944

Hayashi, N., & Yosano, A. (2005). Trust and belief about others: Focusing on judgment accuracy of others' trustworthiness. *Sociological Theory and Methods*, *20*(1), 59–80. https://doi.org/10.11218/ojjams.20.59

Hayes, A. F., & Rockwood, N. J. (2020). Conditional Process Analysis: Concepts, Computation, and Advances in the Modeling of the Contingencies of Mechanisms. *American Behavioral Scientist*, *64*(1), 19–54. https://doi.org/10.1177/0002764219859633

Heilmann, K. (2002). Die Konstruktionsübung: Eine besondere Übung im Assessment-Center. In E. Fay (Ed.), *Das Assessment-Center in der Praxis: Konzepte, Erfahrungen, Innovationen* (pp. 103–129). Vandenhoeck & Ruprecht.

Hong, S., Suk, H. W., Choi, Y., & Na, J. (2021). Face-Based Judgments: Accuracy, Validity, and a Potential Underlying Mechanism. *Psychological Science*, 095679762110003. https://doi.org/10.1177/09567976211000308

Houser, D., Schunk, D., & Winter, J. (2010). Distinguishing trust from risk: An anatomy of the investment game. *Journal of Economic Behavior & Organization*, *74*(1-2), 72–81. https://doi.org/10.1016/j.jebo.2010.01.002

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (Third edition). *Quantitative methodology series*. Routledge Taylor & Francis Group.

Hsu, C.-T., Jacobs, A. M., & Conrad, M. (2015). Can Harry Potter still put a spell on us in a second language? An fMRI study on reading emotion-laden literature in late bilinguals. *Cortex*, *63*, 282–295. https://doi.org/10.1016/j.cortex.2014.09.002

Iacozza, S., Costa, A., & Duñabeitia, J. A. (2017). What do your eyes reveal about your foreign language? Reading emotional sentences in a native and foreign language. *PLoS ONE*, *12*(10), e0186027. https://doi.org/10.1371/journal.pone.0186027

Ivaz, L., Costa, A., & Duñabeitia, J. A. (2016). The emotional impact of being myself: Emotions and foreign-language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(3), 489–496. https://doi.org/10.1037/xlm0000179

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). Explaining the persistent influence of facial cues in social decision-making. *Journal of Experimental Psychology. General*, *148*(6), 1008–1021. https://doi.org/10.1037/xge0000591

Jaeger, B., & Jones, A. L. (2021). Which Facial Features Are Central in Impression Formation? *Social Psychological and Personality Science*, 194855062110349. https://doi.org/10.1177/19485506211034979

Jaeger, B., Oud, B., Williams, T., Krumhuber, E., Fehr, E., & Engelmann, J. B. (2020). Trustworthiness detection from faces: Does reliance on facial impressions pay off? Advance online publication. https://doi.org/10.31234/osf.io/ayqeh

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, *90*, 104004. https://doi.org/10.1016/j.jesp.2020.104004

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889. https://doi.org/10.1016/j.joep.2011.05.007

Jordan, J., Peysakhovich, A., & Rand, D. G. (2015). Why we cooperate. In J. Decety & T. Wheatley (Eds.), *The moral brain: A multidisciplinary perspective.* MIT Press.

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

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science*, *23*(6), 661–668. https://doi.org/10.1177/0956797611432178

Kiyonari, T., & Yamagishi, T. (1999). A comparative study of trust and trustworthiness using the game of enthronement. *The Japanese Journal of Social Psychology*, *15*(2), 100–109.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911. https://doi.org/10.1038/nature05631

Kramer, R. M. (1998). Paranoid cognition in social systems: Thinking and acting in the shadow of doubt. *Personality and Social Psychology Review*, *2*(4), 251–275. https://doi.org/10.1207/s15327957pspr0204_3

Lambert, B., Declerck, C. H., & Boone, C. (2014). Oxytocin does not make a face appear more trustworthy but improves the accuracy of trustworthiness judgments. *Psychoneuroendocrinology*, *40*, 60–68. https://doi.org/10.1016/j.psyneuen.2013.10.015

Larson, H. J., Clarke, R. M., Jarrett, C., Eckersberger, E., Levine, Z., Schulz, W. S., & Paterson, P. (2018). Measuring trust in vaccination: A systematic review. *Human Vaccines & Immunotherapeutics*, *14*(7), 1599–1609. https://doi.org/10.1080/21645515.2018.1459252

Lee, K., & Ashton, M. C. (2017). Acquaintanceship and self/observer agreement in personality judgment. *Journal of Research in Personality*, *70*, 1–5. https://doi.org/10.1016/j.jrp.2017.05.001

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88–106. https://doi.org/10.1016/j.obhdp.2014.10.007

Levine, T. R., Daiku, Y., & Masip, J. (2021). The number of senders and total judgments matter more than sample size in deception-detection experiments. *Perspectives on Psychological Science.* Advance online publication. https://doi.org/10.1177/1745691621990369

Li, K. K. (2017). How does language affect decision-making in social interactions and decision biases? *Journal of Economic Psychology*, *61*, 15–28. https://doi.org/10.1016/j.joep.2017.03.003

Li, N. P., van Vugt, M., & Colarelli, S. M. (2018). The Evolutionary Mismatch Hypothesis: Implications for Psychological Science. *Current Directions in Psychological Science*, *27*(1), 38–44. https://doi.org/10.1177/0963721417731378

Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., & Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior & Organization*, *119*, 254–266. https://doi.org/10.1016/j.jebo.2015.08.003

Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. Wiley.

Luhmann, N. *Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität* [Trust. A mechanism of the reduction of social complexity] (5th ed.). UVK. (Original work published 1968)

Mækelæ, M. J., & Pfuhl, G. (2019). Deliberate reasoning is not affected by language. *PLoS ONE*, *14*(1), e0211428. https://doi.org/10.1371/journal.pone.0211428

Marlowe, F. W. (2005). Hunter-gatherers and human evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *14*(2), 54–67. https://doi.org/10.1002/evan.20046

Okubo, M., Ishikawa, K., & Kobayashi, A. (2018). The cheek of a cheater: Effects of posing the left and right hemiface on the perception of trustworthiness. *Laterality: Asymmetries of Body, Brain and Cognition*, *23*(2), 209–227. https://doi.org/10.1080/1357650X.2017.1351449

Okubo, M., Ishikawa, K., Kobayashi, A., & Suzuki, H. (2017). Can I trust you? Laterality of facial trustworthiness in an economic game. *Journal of Nonverbal Behavior*, *41*(1), 21–34. https://doi.org/10.1007/s10919-016-0242-z

Okubo, M., Kobayashi, A., & Ishikawa, K. (2012). A fake smile thwarts cheater detection. *Journal of Nonverbal Behavior*, *36*(3), 217–225. https://doi.org/10.1007/s10919-012-0134-9

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*(2), 315–324. https://doi.org/10.1016/j.jesp.2009.12.002

Paulhus, D. L., & Bruce, M. N. (1992). The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. *Journal of Personality and Social Psychology*, *63*(5), 816–824. https://doi.org/10.1037/0022-3514.63.5.816

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology*, *56*(5), 823–833. https://doi.org/10.1037/0022-3514.56.5.823

Pavlenko, A. (2005). *Emotions and multilingualism*. Cambridge University Press.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

Rule, N. O., & Ambady, N. (2008). Brief exposures: Male sexual orientation is accurately perceived at 50 ms. *Journal of Experimental Social Psychology*, *44*(4), 1100–1105. https://doi.org/10.1016/j.jesp.2007.12.001

Rule, N. O., Ambady, N., & Hallett, K. C. (2009). Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal of Experimental Social Psychology*, *45*(6), 1245–1251. https://doi.org/10.1016/j.jesp.2009.07.010

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*(3), 409–426. https://doi.org/10.1037/a0031050

Rule, N. O., Slepian, M. L., & Ambady, N. (2012). A memory advantage for untrustworthy faces. *Cognition*, *125*(2), 207–218. https://doi.org/10.1016/j.cognition.2012.06.017

Samochowiec, J., Wänke, M., & Fiedler, K. (2010). Political Ideology at Face Value. *Social Psychological and Personality Science*, *1*(3), 206–213. https://doi.org/10.1177/1948550610372145

Schild, C., Stern, J., Zettler, I., & Barrett, L. (2020). Linking men's voice pitch to actual and perceived trustworthiness across domains. *Behavioral Ecology*, *31*(1), 164–175. https://doi.org/10.1093/beheco/arz173

Schilke, O., & Huang, L. (2018). Worthy of swift trust? How brief interpersonal contact affects trust accuracy. *Journal of Applied Psychology*, *103*(11), 1181–1197. https://doi.org/10.1037/apl0000321

Schlösser, T., Dunning, D., & Fetchenhauer, D. (2013). What a Feeling: The Role of Immediate and Anticipated Emotions in Risky Decisions. *Journal of Behavioral Decision Making*, *26*(1), 13–30. https://doi.org/10.1002/bdm.757

Schlösser, T., Fetchenhauer, D., & Dunning, D. (2016). Trust against all odds? Emotional dynamics in trust behavior. *Decision*, *3*(3), 216–230. https://doi.org/10.1037/dec0000048

Schlösser, T., Mensching, O., Dunning, D., & Fetchenhauer, D. (2015). Trust and rationality: Shifting normative analyses of risks involving other people versus nature. *Social Cognition*, *33*(5), 459–482.

Sheikh, N. A., & Titone, D. (2016). The embodiment of emotional words in a second language: An eye-movement study. *Cognition & Emotion*, *30*(3), 488–500. https://doi.org/10.1080/02699931.2015.1018144

Shin, H. I., & Kim, J. (2017). Foreign Language Effect and Psychological Distance. *Journal of Psycholinguistic Research*, *46*(6), 1339–1352. https://doi.org/10.1007/s10936-017-9498-7

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Simpson, J. A. (2007). Psychological Foundations of Trust. *Current Directions in Psychological Science*, *16*(5), 264–268. https://doi.org/10.1111/j.1467-8721.2007.00517.x

Siuda, S., Schlösser, T., & Fetchenhauer, D. (2019) [*Unpublished raw data on the accuracy of trustworthiness detection]*. University of Cologne.

Snijders, C., & Keren, G. (1999). Determinants of trust. In D. V. Budescu, I. Erev, & R. Zwick (Eds.), *Games and human behavior* (pp. 355–385). Erlbaum.

Snijders, C., & Keren, G. (2001). Do you trust? Whom do you trust? When do you trust? In Keren Gideon (Ed.), *Advances in Group Processes. Advances in Group Processes* (Vol. 18, pp. 129–160). Emerald Group Publishing Limited. https://doi.org/10.1016/S0882-6145(01)18006-9

Sparks, A., Burleigh, T., & Barclay, P. (2016). We can see inside: Accurate prediction of Prisoner's Dilemma decisions in announced games following a face-to-face interaction. *Evolution and Human Behavior*, *37*(3), 210–216. https://doi.org/10.1016/j.evolhumbehav.2015.11.003

Stavrova, O., & Ehlebracht, D. (2016). Cynical beliefs about human nature and income: Longitudinal and cross-cultural analyses. *Journal of Personality and Social Psychology*, *110*(1), 116–132. https://doi.org/10.1037/pspp0000050

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354. https://doi.org/10.1177/0956797610362647

Stirrat, M., & Perrett, D. I. (2012). Face structure predicts cooperation: Men with wider faces are more generous to their in-group when out-group competition is salient. *Psychological Science*, *23*(7), 718–722. https://doi.org/10.1177/0956797611435133

Thielmann, I., & Hilbig, B. E. (2015). The Traits One Can Trust: Dissecting Reciprocity and Kindness as Determinants of Trustworthy Behavior. *Personality & Social Psychology Bulletin*, *41*(11), 1523–1536. https://doi.org/10.1177/0146167215600530

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395. https://doi.org/10.2307/796133

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

Todorov, A., Funk, F., & Olivola, C. Y. (2015). Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences. *Trends in Cognitive Sciences*, *19*(8), 422–423. https://doi.org/10.1016/j.tics.2015.05.013

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813–833. https://doi.org/10.1521/soco.2009.27.6.813

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*(7), 1404–1417. https://doi.org/10.1177/0956797614532474

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Urbig, D., Terjesen, S., Procher, V., Muehlfeld, K., & van Witteloostuijn, A. (2016). Come on and Take a Free Ride: Contributing to Public Goods in Native and Foreign Language Settings. *Academy of Management Learning & Education*, *15*(2), 268–286. https://doi.org/10.5465/amle.2014.0338

Verplaetse, J., & Vanneste, S. (2010). Is cheater/cooperator detection an in-group phenomenon? Some preliminary findings. *Letters on Evolutionary Behavioral Science*, *1*(1), 10–14. https://doi.org/10.5178/lebs.2010.3

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel. A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28*(4), 260–271. https://doi.org/10.1016/j.evolhumbehav.2007.04.006

Volk, S., Köhler, T., & Pudelko, M. (2014). Brain drain: The cognitive neuroscience of foreign language processing in multinational corporations. *Journal of International Business Studies*, *45*(7), 862–885. https://doi.org/10.1057/jibs.2014.26

Weiss, A., Michels, C., Burgmer, P., Mussweiler, T., Ockenfels, A., & Hofmann, W. (2020). Trust in everyday life. *Journal of Personality and Social Psychology.* Advance online publication. https://doi.org/10.1037/pspi0000334

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. https://doi.org/10.1177/01461672992512005

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*(8), 1325–1331. https://doi.org/10.1177/0956797615590992

Wilson, J. P., & Rule, N. O. (2016). Hypothetical sentencing decisions are associated with actual capital punishment outcomes: The role of facial trustworthiness. *Social Psychological and Personality Science*, *7*(4), 331–338. https://doi.org/10.1177/1948550615624142

Wilson, J. P., & Rule, N. O. (2017). Advances in understanding the detectability of trustworthiness from the face: Toward a taxonomy of a multifaceted construct. *Current Directions in Psychological Science*, *26*(4), 396–400. https://doi.org/10.1177/0963721416686211

Wu, Y. J., & Thierry, G. (2012). How reading in a second language protects your heart. *Journal of Neuroscience*, *32*(19), 6485–6489. https://doi.org/10.1523/jneurosci.6119-11.2012

Yu, Q., & Li, B. (2020). Third-variable effect analysis with multilevel additive models. *PLoS ONE*, *15*(10), e0241072. https://doi.org/10.1371/journal.pone.0241072

Zhao, K., & Smillie, L. D. (2015). The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review*, *19*(3), 277–302. https://doi.org/10.1177/1088868314553709

Zylbersztejn, A., Babutsidze, Z., & Hanaki, N. (2020). Preferences for observable information in a strategic setting: An experiment. *Journal of Economic Behavior & Organization*, *170*, 268–285. https://doi.org/10.1016/j.jebo.2019.12.009