INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

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

RESEARCH ARTICLE

ubiquity press

SEBASTIAN SIUDA
THOMAS SCHLÖSSER
DETLEF FETCHENHAUER

*Author affiliations can be found in the back matter of this article

ABSTRACT

Judgments about people's trustworthiness are made frequently and have important real-life consequences. However, the accuracy of these judgments is debated. We therefore systematically reviewed the current evidence for accurate trustworthiness detection in the literature. The overall evidence for accuracy is rather mixed; although we find only limited evidence for accurate trustworthiness detection from neutral photographs, trustworthiness detection becomes more accurate when the rater and target interact, when the target presentation resembles face-to-face contact, and when the target presentations contain cues or signals about the target's trustworthiness. We also find that the current literature lacks an overarching research agenda, which leads to a large heterogeneity in the extant studies' operationalizations. We address some of these operationalizations and suggest the following guidelines for future research: Studies should engage in stronger theory building, experimentally test moderators, strengthen generalizability by recruiting large target pools, and use appropriate methods for the analysis of nonindependent data.

CORRESPONDING AUTHOR:

Sebastian Siuda

University of Cologne, DE

research@sebastiansiuda.com

KEYWORDS:

trust; trustworthiness detection; person perception; impression formation; accuracy

TO CITE THIS ARTICLE:

Siuda, S., Schlösser, T., & Fetchenhauer, D. (2022). Do We Know Whom to Trust? A Review on Trustworthiness Detection Accuracy. *International Review of Social Psychology,* 35(1): 20, 1–16. DOI: https://doi.org/10.5334/irsp.623

> *There is a kernel of accuracy in trustworthiness perceptions that is of broad and substantial theoretical interest.* (Bonnefon et al., 2017b, 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), exhibit at least some shared stereotype on who appears 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).

Progress has been made on the formation of trustworthiness expectations, but 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. Distrust might be warranted if the other person were indeed untrustworthy. However, wrongfully withheld trust because of another person's appearance would be worrisome. Similarly, trust is neither good nor bad per se. Although trust is often a prerequisite for fruitful cooperation, placing trust in another person leaves oneself vulnerable to untrustworthy individuals. To solve this dilemma of trust as a double-edged sword, the most important task is knowing whom to trust.

Why has the debate on trustworthiness detection accuracy not yet been settled? A major reason for the ongoing debate is that the literature displays a large heterogeneity in its underlying theories and operationalizations, which leads to differences between studies in regard to independent variables (e.g., how target individuals are presented), dependent variables (e.g., which types of trust and trustworthiness are measured), and their correspondence (e.g., how detection accuracy is defined). This prevents study results from being meaningfully compared and also prevents adequate quantitative analyses because the studies' effect sizes relate to different operationalizations. The debate cannot yet be settled using quantitative approaches, but an important first step is to advance the debate using more narrative reviews. Therefore, 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.

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). Although 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.

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. These findings suggest that people can accurately detect trustworthiness via facial width, but the direct relationship between trust and trustworthiness was never directly established because the trustee's actual trustworthiness was measured only in the first but not the second study.

Second, the studies included must measure trust and trustworthiness *objectively*. 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, the studies included in this review were required to measure trust and trustworthiness behavior directly themselves using economic games in the laboratory. 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 (Levine & Schweitzer, 2015), the hidden action game (Charness & Dufwenberg, 2006) or the game of enthronement (Kiyonari & Yamagishi, 1999). The rely-or-verify game measures integrity-based trust by having trustors either rely on (trust) or verify (distrust) a trustee's statement. The hidden action game ensures the participants' anonymity by including a random component that reverses the participants' decisions in a small number of cases. The game of enthronement involves a binary trust decision (e.g., to keep or send 500 yen) and a continuous trustworthiness decision (e.g., how much of a resulting 900 yen to distribute between trustor and trustee). 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.

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 2,455 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 3,215 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).

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. Although 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).

	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)		–	–		–			–					–		–			–			–		–

(Contd.)

	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**
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	6	32	30	8	22	16	11	15	2	10	6	3	27	2	21	17	19	7	12	7	31	25	13
Frequency of accurate trustworthiness detection	5	11	12	4	6	10	5	9	1	1	4	1	10	1	9	7	6	3	7	5	11	11	5

Table 1 Overview of study conditions.

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

POTENTIAL MODERATORS OF TRUSTWORTHINESS DETECTION ACCURACY

Although the overall evidence for accurate trustworthiness detection is weak, this is not surprising for a task as diverse and complex as trustworthiness detection. Rather than drawing a conclusion on the accuracy of trustworthiness detection in general, the main task of this review is to identify under which conditions trustworthiness detection appears to be accurate and under which it does not. As we will illustrate, the literature is filled with a wide variety of studies from behavioral economics to facial symmetry research. This diversity allows us to compare trustworthiness detection accuracy across different settings and identify potential moderators of accuracy. However, the studies' diversity with respect to both dependent and independent variables also make a quantitative analysis of moderators using effect sizes of little use. We will therefore draw qualitative conclusions and focus only on the most notable dimensions on which studies differed. We will illustrate these dimensions with examples of the studies found in the literature, thereby giving some insight into each individual study included in this review. We will structure the differences alongside three broad categories, focusing on the following.

Differences in general:

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

Differences concerning targets:

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

Differences concerning raters:

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

1. Do the Rater and Target Interact? Studies on trustworthiness detection differ in the degree to which there is interaction between the rater and the target. Some studies build on findings from the cooperation detection literature that suggest detection accuracy is improved by personal interaction (DeSteno et al., 2012; Frank et al., 1993; Sparks et al., 2016). In one of these studies, Schilke and Huang (2018) directly investigated the influence of acquaintanceship on trustworthiness detection accuracy by comparing two interpersonal contact conditions. Before receiving information about an 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; although 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.

Studies without rater-target interactions, in contrast, 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, Bonnefon et al. (2013) asked participants to evaluate the trustworthiness of targets based on neutral black-and-white pictures of their faces. These pictures had been extracted and cropped to size from videos of a previous study in which the targets had played an anonymous trust game in the role of the trustee. Three interesting results emerged: First, raters accurately predicted the targets' trustworthiness from the pictures alone. Second, this detection accuracy was independent of general intelligence or cognitive load, suggesting that trustworthiness detection might be a modular process. Third, external features decreased accuracy in 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 see Jaeger et al. (2020) for opposing findings.

Did we find an overall trend across studies whether rater-target interaction improves accuracy? Yes, we did. Five of the 6 interactive study conditions reported accurate trustworthiness detection, whereas only 11 of the 32 noninteractive study conditions did so. Moreover, the fact that the study by Schilke and Huang (2018) was the only study that experimentally tested the moderating effect of personal interaction on accuracy within the same study design emphasizes this overall trend. This speaks against the idea that trustworthiness is a stable and easily observable trait and stresses the importance of personal interaction for detection accuracy.

2. 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., how a target behaves toward others in general) or as a target's *specific trustworthiness* (e.g., how a target behaves toward a specific person), 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. The study by De Neys et al. (2017) is an example of general trustworthiness detection. The raters in this study based their trust decisions on a subset of the same edited pictures previously used by Bonnefon et al. (2013) in which targets had played an anonymous trust game in which they neither knew nor saw their interaction partner. The results supported the original findings from Bonnefon et al. (2013) and further showed that the detection accuracy was also above chance level when the pictures were presented for as little as 100 milliseconds. Interestingly, however, the results were reversed when the pictures were only presented for 33 milliseconds; here, the participants trusted trustworthy targets significantly less than untrustworthy targets. Although this result can partly be explained by the fact that participants might have felt awkward assessing targets after an expose time at the brink of their conscious perception threshold, it nevertheless suggests that the overall detection effect is not very robust.

An example of a study on specific trustworthiness is that by Binzel and Fehr (2013). The researchers recruited pairs of friends from a Cairene slum and asked them to play a version of the trust game with each other, in which a random component insured the anonymity of each person's decision. Here, the participants could not accurately detect the trustworthiness of their friends. One potential explanation for the null finding could be the study's special setting in an Egyptian slum, in which inhabitants might more strongly rely on assurances than trust and therefore be less polished in assessing trustworthiness (see Yamagishi, 2011).

We take a look at the overall trend across studies, and 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 in theory it might 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. 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; although 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 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, whereas 16 study conditions tested accuracy via behavioral trust, of which 10 were significant. This pattern 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 et al. (2020) allowed us to directly test the accuracy for both types of trust and found no disparity in regard to accuracy.

4. When are the Targets Recorded? Following the discussion of more general differences, we now turn to the differences between the studies in regard to the targets. One potential moderator of trustworthiness detection might be timing of when targets are recorded. Depending on when they are recorded, targets might voluntarily or involuntarily exhibit emotional cues during their decision-making process (Verplaetse et al., 2007) or show emotional residues in their faces shortly after having 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.

A different timing was used in the 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, the 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. However, additional studies are needed for any substantial conclusion since only two study conditions, of which one was significant, have tested this idea so far.

5. How Are the Targets Presented? 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. Although these studies are closely related to the interactive studies discussed earlier, face-to-face type studies do not necessarily involve rater-target interactions. In a 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 or picture. Here, the raters usually predict the trustworthiness of targets who are recorded before, during or after a trust game. Zylbersztejn et al. (2020) tested trustworthiness detection accuracy across three conditions. First, before knowing about the upcoming trust game, the targets were photographed with a neutral expression and videorecorded reading a neutral text. Afterward, the targets learned about the upcoming trust game and were given the chance to deliver a video-recorded 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 the 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.

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.

Does the access to richer target cues improve trustworthiness detection accuracy? Overall, four of the six 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 one of the two study conditions if the targets were presented auditorily. Providing raters with more and richer target cues, thus, might indeed be a key to more accurate trustworthiness detection.

6. Are the 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 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 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 invest energy into appearing trustworthy only 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.

7. Are the 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 study by Hayashi and Yosano (2005), in which participants could get acquainted with each other in a 30-minute-long group discussion before they were informed 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 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. Overall, the raters' trustworthiness predictions and targets' actual back transfer rates were significantly associated. However, analyses revealed that this association was fully mediated by the information on whether a target had been trusted in the trust game. In fact, relying on the neutral targets photos decreased the overall detection accuracy for a number of raters. 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 for angry but not for 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. A potential explanation for the slightly more frequent significant results might be that the targets' natural facial expressions provide valuable cues for trustworthiness detection and that limiting access to these cues consequently decreases accuracy. However, we stress that this is just one of many possible interpretations that need to be systematically tested before any serious conclusions can be drawn.

8. When Do the Raters See the Targets? After discussing different operationalizations on the target side, 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 study by Eckel and Petrie (2011). The participants were photographed and then played trust games in which they either saw photographs of their interaction partners free of cost or had the opportunity to 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 be evaluated only based on previous small talk. An example of a study investigating this type of trustworthiness detection is the aforementioned study by Hayashi and Yosano (2005) in which the participants formed accurate expectations of their group members' trustworthiness even before knowing about the upcoming detection task.

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 an 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.

9. Are the 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 study by Bonnefon et al. (2013) offers an opportunity to test this idea. Although 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 two of the three 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: 11 of 25 study conditions with financial incentives reported accurate trustworthiness detection results, compared to 5 of 13 study conditions without financial incentives.

SUMMARY OF POTENTIAL MODERATORS

Across the studies reported in this review, three moderators emerged that appear to moderate trustworthiness detection accuracy. First, study conditions with rater-target interaction reported accurate trustworthiness detection more often than conditions without such an interaction. Importantly, this general trend across studies was also found in the study by Schilke and Huang (2018), which experimentally tested rater-target interaction within its study design. Thus, personal contact with another person may lead to accurate perceptions of that person's trustworthiness which would mirror the results of previous cooperation detection studies on the utility of strategic contact (Frank et al., 1993; Sparks et al., 2016).

Second, study 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 lead to more 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 like 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). 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.

The picture was less clear for the other potential moderators. The best case for an additional moderator could be made for the type of trust being measured as trustworthiness detection was more often accurate for behavioral than cognitive trust. The evidence is not clear-cut, however, because the results from Jaeger et al. (2020) did not find this trend within their study. Thus, future studies are needed to clearly disentangle the moderating impact of how trust is measured. All other potential moderators, although theoretically relevant, did not appear to independently influence the trustworthiness detection results. For example, although 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.

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.

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. Approximately half of all the study conditions regress trustworthiness ratings (or trust behavior) on the targets' trustworthiness, but 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 they 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. We therefore suggest that all future studies follow the analysis strategy that has already been most commonly used of regressing trustworthiness ratings (or trust behavior) on the targets' actual trustworthiness.

Second, some studies use improper methods to analyze nonindependent data. Many study designs generate multiple trustworthiness ratings for each rater, leading to data clustering, that is, 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 because it can lead to an overestimation of the true relationship between predicted and actual trustworthiness. Instead, we recommend the use of mixed-effect models to analyze nonindependent data. Over the last few years, these models have emerged as a useful method of analysis, and the R package *lme4* (Bates et al., 2015) has been most widely adopted.

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). Although 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). We therefore advise against aggregating ratings and encourage future studies to deal with the resulting nonindependence of data by using appropriate mixed-effect models.

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. Although 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). To investigate the accuracy of trustworthiness detection that is translatable to the real world, we therefore advise against altering the true prevalence of (un)trustworthy targets.

Second, some studies provide low generalizability by using the same pool of target pictures for a multitude of studies. Because 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. Given the relatively small target pools of 12–60 individuals, only a few easy-to-recognize targets would be sufficient for the small but better-than-chance detection accuracy that is usually found. The selection of stimuli persons is particularly critical because it affects all subsequent studies using the material from Bonnefon et al. (2013). As these studies make up a significant portion of the literature, the conclusion of any review (including meta-analyses) will likely be biased by this deliberatively selected stimulus set. We therefore strongly recommend recruiting new target persons for each study on trustworthiness detection. Moreover, the recruitment of new target persons should be done in an adequate quantity. It might be tempting to create statistical power by recruiting large (online) samples of raters who rate the trustworthiness of comparably few targets. However, increasing the sample size of raters is not very helpful because 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).

Third, many studies test trustworthiness detection in settings with limited ecological validity, for example, by only providing raters with neutral target pictures. Although 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 reminiscent 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. Trustworthiness detection from pictures certainly is an interesting research subject, but we argue that there has been an unwarranted focus on this specific subject. We therefore encourage future studies to explore trustworthiness detection beyond pictures and in more interactive settings closer related to how people assess other's trustworthiness in their real life.

Finally, we believe that most of the mentioned issues may arise because the field lacks a unified research agenda built on existing theory. Often, researchers seem to consider the accuracy of trustworthiness impressions as an interesting side note rather than the core focus of their studies which can result in studies being composed of seemingly random combinations of possible operationalizations. We urge researchers to more purposefully address trustworthiness detection accuracy in its own right and build on existing theory, for example, from person perception (Funder, 1995) or evolutionary psychology (Cosmides & Tooby, 1992; Frank, 2005). With specific hypotheses in mind, studies should then be consciously operationalized, which 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. An important part in this future research would also that more studies 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.

Systematically varying the levels of all 9 mentioned dimensions alone would already translate to 3,072 potential studies needed 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.

LIMITATIONS

As with any narrative review, there are obvious limitations to the extent to which we can draw conclusions regarding the overall accuracy and potential moderators of trustworthiness detection. We readily admit that a meta-analysis would have provided more satisfactory summary of the literature. As we have discussed, however, the current state of the literature prohibits any meaningful quantitative analyses that go beyond the simple vote-counting measures we have used. Although we believe that these measures are helpful in gaining a crude first impression of potential moderators, we urge the reader not to overinterpret our findings.

Moreover, the results and the interpretations in this review are limited to the detection of trustworthiness in trust games. The deliberate choice to only focus on trust games guarantees that the relationship between the participants' trust and trustworthiness behavior can be assessed directly without having to rely on self-reported or other-reported behavior, but for two reasons that limits the degree to which our findings can be applied to the real world. First, people's trustworthiness in trust games only represents one of many domains in the broader spectrum of trustworthiness (Wilson & Rule, 2017). Thus, we cannot draw confident conclusions about trustworthiness detection accuracy with regard to other trustworthiness domains like criminality, honesty, or infidelity. Second, research shows that people's behavior in economic games does not necessarily translate to real-life behavior outside of the laboratory (e.g., Galizzi & Navarro-Martinez, 2019). This should be kept in mind when considering how trustworthiness detection from economic games in laboratory settings might generalize to the real world.

CONCLUSION

Judgments about others' trustworthiness are made frequently and have important real-life consequences, yet their accuracy is still debated. We advanced 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 and test moderators within studies, strengthen generalizability with large target pools, and use appropriate methodology 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.

ACKNOWLEDGEMENTS

We thank Erik Hölzl, Corinna Michels, Daniel Ehlebracht, Esther Kang, Cora Linnartz, Carolina Dahlhaus, and Anne-Sophie Lang for comments on an earlier version of this article. We also thank Malin Beeh for her help with the literature search.

AUTHOR AFFILIATIONS

Sebastian Siuda orcid.org/0000-0003-1000-1755
University of Cologne, DE

Thomas Schlösser orcid.org/0000-0003-2685-6221
University of Cologne, DE

Detlef Fetchenhauer orcid.org/0000-0002-5469-471X
University of Cologne, DE

COMPETING INTERESTS

The authors have no competing interests to declare.

REFERENCES

Albohn, D. N., & **Adams, R. B.** (2020). Emotion residue in neutral faces: Implications for impression formation. *Social Psychological and Personality Science*, *12*(4), 479–486. DOI: https://doi.org/10.1177/1948550620923229

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. DOI: 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). DOI: https://doi.org/10.18637/jss.v067.i01

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

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

Bonnefon, J.-F., Hopfensitz, A., & **De Neys, W.** (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology: General*, *142*(1), 143–150. DOI: 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. DOI: 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. DOI: 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. DOI: https://doi.org/10.1017/S0140525X15002319

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

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. DOI: https://doi.org/10.1016/j.cogpsych.2010.03.001

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

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. DOI: https://doi.org/10.1177/0956797614523297

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

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. DOI: 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. DOI: 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. DOI: 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. DOI: https://doi.org/10.1177/0956797612448793

Dilger, A., Müller, J., & **Müller, M.** (2017). Is trustworthiness written on the face? *SSRN Electronic Journal*. Advance online publication. DOI: 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. DOI: https://doi.org/10.1037/a0036673

Dunning, D., Fetchenhauer, D., & **Schlösser, T.** (2019). Why people trust: Solved puzzles and open mysteries. *Current Directions in Psychological Science*, *28*(4), 366–371. DOI: https://doi.org/10.1177/0963721419838255

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

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

Fetchenhauer, D., & **Dunning, D.** (2009). Do people trust too much or too little? *Journal of Economic Psychology*, *30*(3), 263–276. DOI: https://doi.org/10.1016/j.joep.2008.04.006

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. DOI: https://doi.org/10.1098/rsos.181552

Frank, R. H. (2005). Altruists with green beards: Still kicking? *Analyse & Kritik*, *27*(1), 110. DOI: 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. DOI: https://doi.org/10.1016/0162-3095(93)90020-I

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. DOI: 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. DOI: 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. DOI: https://doi.org/10.1037//0022-3514.55.1.149

Galizzi, M. M., & **Navarro-Martinez, D.** (2019). On the external validity of social preference games: a systematic lab-field study. *Management Science*, *65*(3), 976–1002. DOI: https://doi.org/10.1287/mnsc.2017.2908

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. DOI: https://doi.org/10.11218/ojjams.20.59

Hox, J. J., Moerbeek, M., & **van de Schoot, R.** (2017). *Multilevel analysis: Techniques and applications* (Third edition). *Quantitative methodology series*. Routledge Taylor & Francis Group. DOI: https://doi.org/10.4324/9781315650982

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. DOI: https://doi.org/10.31234/osf.io/ayqeh

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. DOI: https://doi.org/10.1037/a0028347

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.

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. DOI: https://doi.org/10.1016/j.psyneuen.2013.10.015

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

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

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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: https://doi.org/10.1037/0022-3514.63.5.816

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

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. DOI: 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. DOI: https://doi.org/10.1016/j.cognition.2012.06.017

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. DOI: 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. DOI: https://doi.org/10.1037/apl0000321

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. DOI: https://doi.org/10.1146/annurev-psych-010418-102803

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 K. Gideon (Ed.), *Advances in Group Processes. Advances in Group Processes, 18*, 129–160. Emerald Group Publishing Limited. DOI: 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. DOI: https://doi.org/10.1016/j.evolhumbehav.2015.11.003

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

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press. DOI: https://doi.org/10.1515/9781400885725

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. DOI: 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. DOI: 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. DOI: 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. DOI: https://doi.org/10.1177/0956797614532474

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. DOI: 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. DOI: https://doi.org/10.1016/j.evolhumbehav.2007.04.006

Wells, G. L., & **Windschitl, P. D.** (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. DOI: 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. DOI: 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. DOI: 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. DOI: https://doi.org/10.1177/0963721416686211

Yamagishi, T. (2011). The Emancipation Theory of Trust. In: *Trust. The Science of the Mind*. Springer Publishing Company. DOI: https://doi.org/10.1007/978-4-431-53936-0_4

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. DOI: https://doi.org/10.1016/j.jebo.2019.12.009

TO CITE THIS ARTICLE:
Siuda, S., Schlösser, T., & Fetchenhauer, D. (2022). Do We Know Whom to Trust? A Review on Trustworthiness Detection Accuracy. *International Review of Social Psychology*, 35(1): 20, 1–16. DOI: https://doi.org/10.5334/irsp.623

Submitted: 24 June 2021 **Accepted:** 02 October 2022 **Published:** 14 November 2022

International Review of Social Psychology is a peer-reviewed open access journal published by Ubiquity Press.