Article

The Traits One Can Trust: Dissecting Reciprocity and Kindness as Determinants of Trustworthy Behavior

Personality and Social Psychology Bulletin
2015, Vol. 41(11) 1523–1536

DOI: 10.1177/0146167215600530
pspb.sagepub.com

Isabel Thielmann[1,2] and Benjamin E. Hilbig[1]

Abstract

Trustworthiness is a vital pillar of various social interactions hinging upon trust. However, the underlying determinants of trustworthiness—especially in terms of (basic) personality traits—are insufficiently understood. Specifically, three mechanisms underlying trustworthiness have been proposed: unconditional kindness, positive reciprocity, and negative reciprocity. The present research aims to disentangle these mechanisms using a trait-based approach, relying on the HEXACO (Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to Experience) model of personality. In three studies, participants acted as the trustee in the trust game. All studies revealed consistent support for the unconditional kindness mechanism, showing an exclusive link between Honesty-Humility and trustworthiness, irrespective of the level of prior trust. In turn, positive and negative reciprocity could not account for the pattern of results. In addition, our results reconcile the inconsistent evidence on the relation between Big Five-Agreeableness and trustworthiness: Unconditional kindness only refers to one component of the broad Big Five-Agreeableness factor (which subsumes various cooperative tendencies).

Keywords

trustworthiness, trust game, reciprocity, unconditional kindness, basic personality traits

Received March 6, 2015; revision accepted July 22, 2015

Trust is a vital pillar of various social interactions and society as a whole (e.g., Yamagishi, 2011). However, the benefits associated with trust are necessarily contingent upon others' trustworthiness—given that trust basically mirrors a risky choice to depend on another without being able to control the other's actions (Thielmann & Hilbig, in press). That is, trust—especially among strangers—is only defensible if one can expect the trusted person (the so-called *trustee*) to honor rather than betray the trust (Evans & Krueger, 2009). Despite this vital significance of trustworthiness, there is only a rudimentary understanding of trustworthy behavior, especially in terms of underlying personality traits.

Inspired by behavioral economics, recent research on trustworthiness has mostly relied on the *trust game* (Berg, Dickhaut, & McCabe, 1995). In this game, a trustor is asked to divide a certain endowment between herself and a trustee. The amount the trustor en*trusts* is multiplied (usually tripled) and transferred to the trustee who is then asked to decide how much to return to the trustor. By implication, the amount returned is considered a measure of behavioral trustworthiness, with high returns indicating high trustworthiness (cf. Johnson & Mislin, 2011). Note that, according to this conceptualization, we herein adopt a behavioral view of trust and trustworthiness.

As directly follows from the rules of the trust game, trustworthiness involves a *reaction*, that is, behavior contingent on another's (prior) trust. Correspondingly, trustworthiness in the trust game has typically been considered an expression of reciprocity (Berg et al., 1995), which can be defined as "a conditional behaviour aimed at reacting to a behaviour with another behaviour of the same valence" (Perugini & Gallucci, 2001, p. S20). Stated differently, reciprocity captures an individual's tendency to adjust her own behavior to an interaction partner's (previous) behavior. Depending on whether an individual rewards another's cooperative behavior or punishes another's uncooperative behavior, one can further distinguish between positive and negative reciprocity (e.g., Perugini, Gallucci, Presaghi, & Ercolani, 2003). Although recent research on the trust game almost exclusively interpreted trustworthiness in terms of positive reciprocity (e.g., Chaudhuri & Gangadharan, 2007; McCabe, Rigdon, & Smith, 2003), we consider both

[1]University of Koblenz-Landau, Germany
[2]University of Mannheim, Germany

Corresponding Author:
Isabel Thielmann, University of Koblenz-Landau, Fortstraße 7, 76829 Landau, Germany.
Email: thielmann@uni-landau.de

positive and negative reciprocity as potential factors accounting for trustworthy behavior.[1]

The idea that trustworthiness essentially mirrors reciprocity is largely based on evidence indicating that higher trust levels enhance trustees' willingness to behave trustworthily (see Johnson & Mislin, 2011, for a meta-analytic review). Likewise, if trustees highly (rather than only marginally) benefitted from the trustor's trust, trustworthiness increased (Malhotra, 2004). Vice versa, if trustees cannot rule out that trustors merely "trusted" out of strategic considerations rather than out of "true" trust or kindness, respectively, returns have been shown to decrease (Bauernschuster, Falck, & Große, 2013). Finally, trustee returns have been negatively related to trait negative reciprocity; however, for positive reciprocity a comparable (positive) link could not be corroborated (Yamagishi et al., 2012). Nevertheless, altogether, these findings suggest a mechanism of reciprocity underlying trustworthiness—implying that dispositions toward (positive or negative) reciprocity should determine individual differences in trustworthy behavior.

Besides reciprocity, it has also been argued—and empirically supported—that trustee returns are driven by individuals' unconditional kindness (e.g., Cox, 2004; Gambetta & Przepiorka, 2014), which has mostly been operationalized through giving in the dictator game[2] (Forsythe, Horowitz, Savin, & Sefton, 1994). Unconditional kindness implies that trustworthy behaviors are not perfectly contingent on the level of prior trust, but rather involve a relatively stable return. Correspondingly, meta-analytic evidence indicates that trustee returns decline less than proportionately with the multiplier of the entrusted amount (Johnson & Mislin, 2011). This implies relatively high returns if the amount is, for example, only doubled rather than tripled—a finding that is difficult to explain by reciprocity alone. Similarly, directly comparing (game-based) reciprocity and unconditional kindness revealed that the latter accounts for the majority of variance in trustworthiness (Ashraf, Bohnet, & Piankov, 2006). However, in another study, neither unconditional kindness nor reciprocity showed significant relations with trustworthiness (Ben-Ner & Halldorsson, 2010). In any case, an unconditional kindness mechanism would imply that dispositions toward altruistic and fair behavior can explain individual variation in trustworthiness.

Taken together, the extant evidence is inconclusive regarding the nature and underlying (trait) determinants of trustworthiness. Strikingly, though, almost all previous studies have exclusively relied on a game-theoretical approach, for example, by investigating behavioral tendencies across different games (or structural changes within one game). This common practice in behavioral economics is undoubtedly fruitful, but essentially misses out on more stable behavioral tendencies (i.e., traits) and may additionally yield caveats due to common-method variance and individuals' desire to respond consistently across games. Besides, given that a zero return in the trust game might reflect either negative reciprocity or a lack of positive reciprocity, distinguishing between positive and negative reciprocity is impossible using a purely game-based approach unless the game structure is changed considerably—which would, in turn, undermine comparability across studies.

As an alternative to the sole reliance on games, models of basic personality traits offer a promising avenue to identify the determinants underlying cooperation in general and trustworthiness in particular. Specifically, "broad and stable interpersonal traits can help explain behavioral heterogeneity across a range of games modeling social interactions" (Zhao & Smillie, 2015, p. 293). Regarding trustworthiness, most corresponding research focused on the widely accepted Five-Factor Model (FFM; Costa & McCrae, 1992; McCrae & Costa, 1985) and linked trustee behavior to Agreeableness (FFM-AG; see Zhao & Smillie, 2015, for a recent review). By definition, FFM-AG captures the tendency to cooperate in situations involving resource conflicts (Denissen & Penke, 2008). Correspondingly, some studies report a positive relation between FFM-AG and trustee returns (Becker, Deckers, Dohmen, Falk, & Kosse, 2012; Ben-Ner & Halldorsson, 2010). However, a similar number of studies could not corroborate said link (Evans & Revelle, 2008; Müller & Schwieren, 2012) or found that FFM-AG is only predictive in combination with other factors (low Neuroticism; Lönnqvist, Verkasalo, Wichardt, & Walkowitz, 2012). Summarized carefully, the evidence is currently inconclusive.

Plausibly, the inconsistent findings may not be due to a conceptual limitation of FFM-AG per se, but may be due to the broad nature of this factor capturing all kinds of cooperative tendencies (including unconditional kindness and positive/negative reciprocity; cf. Costa, McCrae, & Dye, 1991). Thus, the inconsistent evidence might actually suggest that only one of the proposed determinants explains trustworthy behavior. If, for example, only unconditional kindness accounts for trustworthiness whereas positive and negative reciprocity do not, the former mechanism would strengthen the association between FFM-AG and trustworthiness, whereas the latter would reduce it—leading to an inconsistent overall picture such as the one observed. In consequence, because FFM-AG covers all trait aspects potentially—but not necessarily—relevant for trustworthiness, a positive relation between FFM-AG and trustworthiness cannot help unravel which specific dispositional tendency is actually decisive.

Fortunately, the more recently proposed HEXACO (Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to Experience) model of personality (Ashton & Lee, 2007; Lee & Ashton, 2004) distinguishes between three trait dimensions accounting for individual variation in prosocial behavior: Honesty-Humility, Emotionality, and Agreeableness (Ashton, Lee, & De Vries, 2014). Whereas Emotionality involves a tendency toward kin altruism, Honesty-Humility (HEX-HH) and Agreeableness (HEX-AG) encompass complementary aspects of reciprocal

altruism. That is, HEX-HH is defined as "the tendency to be fair and genuine in dealing with others, in the sense of cooperating with others even when one might exploit them without suffering retaliation" (Ashton & Lee, 2007, p. 156). As such, high levels of HEX-HH imply sincerity, fairness, greed-avoidance, and modesty. HEX-AG, in turn, describes "the tendency to be forgiving and tolerant of others, in the sense of cooperating with others even when one might be suffering exploitation by them" (Ashton & Lee, 2007, p. 156). Thus, high levels of HEX-AG stand for forgiveness, gentleness, flexibility, and patience.

Corroborating the theoretical conceptualizations of both HEXACO factors with regard to prosocial behavior, HEX-HH has consistently (positively) been linked to active cooperation (i.e., non-exploitation) in social dilemmas (Hilbig, Zettler, & Heydasch, 2012; Zettler, Hilbig, & Heydasch, 2013) and—of particular interest for the issue at hand—to unconditional kindness in the dictator game (e.g., Baumert, Schlösser, & Schmitt, 2013; Hilbig, Thielmann, Hepp, Klein, & Zettler, 2015; Thielmann & Hilbig, 2014) as well as to positive reciprocity (Ackermann, Fleiß, & Murphy, in press; Perugini et al., 2003). For HEX-AG, in turn, studies point to negative associations with reactive cooperativeness (i.e., non-retaliation) in the ultimatum game (Hilbig, Zettler, Leist, & Heydasch, 2013; Thielmann, Hilbig, & Niedtfeld, 2014) and, specifically, negative reciprocity (Ackermann et al., in press; Perugini et al., 2003).

Due to this specific distinction between different cooperative tendencies captured by HEX-HH and HEX-AG, respectively, the HEXACO model allows for a particularly fine-grained analysis of dispositional tendencies underlying cooperative behavior. The model is thus especially useful whenever evidence based on the broader FFM is inconclusive (e.g., Hilbig, Glöckner, & Zettler, 2014). Correspondingly, for trustworthiness, specific predictions on the relation between trustworthy behavior and HEX-HH or HEX-AG, respectively, can be derived for each of the proposed potential mechanisms (as detailed below). Exactly this type of evidence—on the link between the HEXACO dimensions and trustworthiness (in the trust game)—is currently missing (Zhao & Smillie, 2015).

Consequently, the purpose of the present studies was to dissect the potential dispositional tendencies underlying trustworthiness based on the HEXACO model of personality. Given the theoretical conceptualizations of HEX-HH and HEX-AG—and corresponding evidence as sketched above—the following predictions can be derived: If trustworthiness is determined by unconditional kindness (alone), it should be positively linked to HEX-HH, but *not* linked to HEX-AG (as the latter refers to reactive cooperativeness which is, by definition, conditional). A similar main effect of HEX-HH would be compatible—but not necessary—if positive reciprocity is the responsible factor. Nonetheless, the unconditional kindness and positive reciprocity mechanisms make incompatible predictions on the presence of an interaction with the level of prior trust: By definition, unconditional kindness is unconditional and it should therefore drive trustworthiness independently of the trustor's level of trust (precluding an interaction). By contrast, positive reciprocity is inherently conditional and should thus drive trustworthiness contingent on the trustor's prior behavior (implying an interaction): The more is entrusted, the more strongly HEX-HH would have to predict trustworthiness. In summary, unconditional kindness would thus require a main effect of HEX-HH on trustworthiness, but none of HEX-AG, and no interaction of HEX-HH with prior trust. Positive reciprocity, in turn, would require said interaction; otherwise, a main effect of HEX-HH would be compatible, but not necessary.

Finally, if trustworthiness is determined by negative reciprocity, it should be (negatively) linked to HEX-AG. However, this relationship must also be a conditional one (given the conditional nature of reciprocity), implying an interaction between HEX-AG and prior trust (i.e., a stronger relation between HEX-AG and trustworthiness with decreasing levels of trust).[3] In turn, a main effect of HEX-AG on trustworthiness would not be required, but nonetheless be compatible with the negative reciprocity account.

To test the alternative mechanisms, we conducted three online studies on the link between basic personality traits and trustworthiness—all in close adherence to standards for Internet-based experimenting (e.g., Reips, 2002). As our primary goal was to disentangle the different potential determinants of trustworthy behavior, we exclusively focused on the relations between the HEXACO dimensions (particularly, HEX-HH and HEX-AG) and trustworthiness in Study 1. In Studies 2 and 3, we additionally aimed at clarifying whether the relatively weak link between FFM-AG and trustworthiness (Zhao & Smillie, 2015) can be attributed to the broader nature of FFM-AG (subsuming different cooperative tendencies).

Study 1

Method

Materials. To assess basic personality traits, we used the German 60-item version of the HEXACO Personality Inventory–Revised (HEXACO-60; Ashton & Lee, 2009; for psychometric properties of the German version, see Moshagen, Hilbig, & Zettler, 2014). The HEXACO-60 includes 10 items for each of the six HEXACO dimensions. Responses are given on a 5-point Likert-type scale ranging from *strongly disagree* to *strongly agree*.

To measure trustworthiness, we relied on the classical version of the trust game (Berg et al., 1995), placing participants in the role of the trustee.[4] During the game, participants could earn points, based on which they were later incentivized. Specifically, each participant was randomly matched to an unknown trustor. The combination of the participant's and the trustor's choices determined participants' point scores.

Table 1. Means, Standard Deviations (in Parentheses), and Bivariate Correlations (95% Confidence Intervals in Brackets) of All Focal Variables Assessed in Study 1, With Internal Consistency Reliabilities (Cronbach's α) in the Diagonal.

			Correlations	
Measure	Scale	*M* (*SD*)	HEX-HH	HEX-AG
HEX-HH	1-5	3.43 (0.64)	.80	
HEX-AG	1-5	3.08 (0.59)	.24** [.06, .41]	.78
Return 15 points (in %)	0-100	38.9 (23.6)	.35*** [.18, .51]	.00 [−.19, .19]
Return 30 points (in %)	0-100	40.3 (23.3)	.42*** [.26, .57]	.06 [−.13, .25]
Return 45 points (in %)	0-100	41.2 (21.1)	.39*** [.22, .54]	.08 [−.11, .26]
Return 60 points (in %)	0-100	42.8 (21.1)	.40*** [.22, .54]	.07 [−.12, .26]
Return 75 points (in %)	0-100	42.8 (22.0)	.34*** [.16, .50]	.05 [−.14, .24]
Return 90 points (in %)	0-100	43.8 (23.5)	.39*** [.22, .54]	.06 [−.13, .24]
M return (in %)	0-100	41.6 (20.5)	.42*** [.25, .56]	.06 [−.13, .24]

Note. HEX-HH = Honesty-Humility; HEX-AG = HEXACO-Agreeableness.
**$p \leq .01$. ***$p \leq .001$.

The 25% of participants with the highest final score received a 10€ (approximately US$12.60) gift voucher.

Initially, both trustors and trustees received an endowment of 30 points. Trustees (participants) were informed that the trustor (a randomly assigned unknown other, denoted as Player 1) could decide how much of this endowment (in 5-point increments) she wants to transfer to the trustee (denoted as Player 2). The transfer was tripled accordingly. Trustees' task was to indicate how much of the (tripled) transfer they wanted to return to the trustor. Corresponding to the widely accepted strategy method (Selten, 1967), participants were unaware of the trustor's actual transfer, but specified their return for each of the six potential (tripled) amounts (above 0) the trustor could transfer (i.e., between 15 and 90 points, in 15-point increments).

Procedure. After providing informed consent and demographic information, participants completed the HEXACO-60. Next, they were thoroughly introduced to the rules of the trust game and asked to indicate their return to the trustor for each potential transfer. Finally, participants answered a few control questions assessing their seriousness of participation and received individual feedback on their HEXACO scores. After completing data collection, each participant was randomly assigned to a trustor to determine their point scores (using pseudonymous codes preserving anonymity).

Participants. Participants were recruited via online social networks (e.g., Facebook) and university mailing lists. An a priori power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) revealed a required sample size of $N = 108$ to detect a medium-sized effect ($f^2 = .10$) in a two-tailed *t* test for single regression coefficients in a linear regression with a high power ($1 - \beta = .90$). Note that sample size calculations were based on the main effects of HEX-HH and HEX-AG, respectively, on trustworthiness, given that these mirror the most basic tests in our analyses. Corresponding to these calculations, we recruited $N = 108$ trustees (and the same number of trustors, see above), including 81 females and aged between 19 and 49 years ($M = 25.5$, $SD = 5.6$). The majority (72.2%) of participants were students, 17.6% were employees.

Results and Discussion

Table 1 summarizes the means, standard deviations, and zero-order correlations between all focal variables (for information on all variables, see Table S1 in the Supplemental Material). First off, trustee returns showed strong positive correlations with HEX-HH for all potential trust levels individually as well as for the average return across trust levels (all rs > .30, $p < .001$). For HEX-AG, in turn, no (or only very weak) correlations with trustworthiness could be observed (all rs < .08, $p > .43$).

To statistically test this pattern (i.e., an influence of HEX-HH, but not of HEX-AG, on trustworthiness), we analyzed the average return across trust levels using a three-step analytical approach (cf. Hilbig et al., 2014). First, we used an approximation of the Bayesian Information Criterion (BIC) from R^2 (Raftery, 1995, Equation 26; Wagenmakers, 2007). From the BIC, we calculated Bayes Factors (BF). Following Wagenmakers (2007), we refer to BF_{01}, relating the probability of the null hypothesis to the probability of the alternative hypothesis. Thus, $BF_{01} < 1$ indicates evidence in favor of the alternative hypothesis, whereas $BF_{01} > 1$ indicates evidence in favor of the null hypothesis. For HEX-HH, a $BF_{01} = 0.0003$ indicated that the alternative hypothesis (a meaningful correlation between HEX-HH and trustworthiness) was more than 3,000 times as likely as the null hypothesis given the data. By contrast, for HEX-AG, a $BF_{01} = 8.75$ indicated the opposite, with the null being almost 9 times as likely as the alternative hypothesis. Second, we compared the size of the two correlation coefficients (i.e., $r = .42$ for HEX-HH

vs. $r = .06$ for HEX-AG) using a z test for dependent correlations (Meng, Rosenthal, & Rubin, 1992). As implied by a significant difference, $z = 3.16$, $p = .002$, the correlation between trustworthiness and HEX-HH was indeed larger than its counterpart for HEX-AG. Finally, concurrently regressing the average return on both HEX-HH and HEX-AG revealed a unique impact of HEX-HH, $\beta = .43$, $p < .001$, 95% CI = [.25, .61], but not of HEX-AG, $\beta = -.05$, $p = .592$, 95% CI = [−.23, .13]. In sum, all three analyses consistently supported an influence of HEX-HH, but not of HEX-AG, which, in turn, corresponds to the unconditional kindness mechanism and is also compatible with the positive reciprocity mechanism.

However, as detailed above, an unconditional kindness mechanism also requires that the relation between HEX-HH and trustworthiness is independent of the level of trust, thus prohibiting an interaction with prior trust. By contrast, the positive reciprocity mechanism specifically necessitates said interaction. To hence test the interaction between HEX-HH and prior trust, we used a linear mixed model, regressing trustworthiness on HEX-HH (between-level predictor), trust (within-level predictor), and their interaction.[5] In line with the unconditional kindness mechanism, the model revealed a significant main effect of HEX-HH, $B = 13.47$, $p < .001$, 95% CI = [7.92, 19.01], but no interaction between HEX-HH and trust, $B = -0.02$, $p = .759$, 95% CI = [−0.16, 0.12]. To test this null interaction more conclusively, we approximated the BIC values for the two regression models (i.e., the main-effects model without the interaction term and the interaction model including the interaction term) based on their respective log-likelihood (Wagenmakers, 2007; Equation 9) and calculated the BF_{01} for the difference between the two BIC values (Wagenmakers, 2007; Equation 10). As indicated by $BF_{01} = 9.91$ (based on $\Delta BIC_{10} = 4.59$), the probability of the main-effects model given the data was almost 10 times greater than the probability of the interaction model. Overall, this pattern implies that positive reciprocity cannot account for trustworthy behavior. As displayed in Figure 1, the (positive) relation between HEX-HH and trustworthiness was indeed equivalent across trust levels.

To finally test the negative reciprocity mechanism, the same linear mixed model was used, including HEX-AG, trust, and their interaction as predictors. Contradicting the negative reciprocity mechanism, the model did not reveal an interaction between HEX-AG and trust, $B = 0.06$, $p = .481$, 95% CI = [−0.10, 0.21] (and—mirroring the results for the average return—no main effect of HEX-AG, $B = 1.96$, $p = .558$, 95% CI = [−4.64, 8.56]). Correspondingly, comparing the HEX-AG main-effects model with the interaction model revealed $BF_{01} = 8.10$ ($\Delta BIC_{10} = 4.18$), thus indicating that the probability for the main-effects model was about 8 times greater than the corresponding probability for the interaction model. Hence, the data contradicted negative reciprocity as underlying determinant of trustworthiness.

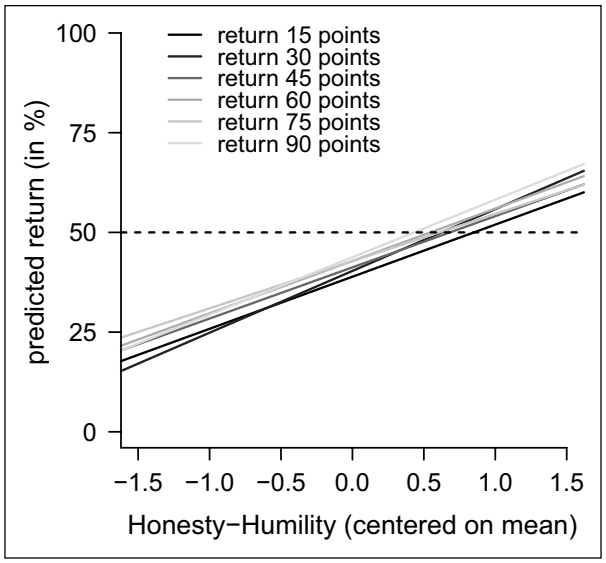

Figure 1. Predicted trustee returns for all levels of prior trust (between 15 and 90 points) depending on individual Honesty-Humility scores (centered on mean) in Study 1.

In summary, our analyses yielded a positive and consistent influence of HEX-HH on trustee returns across different trust levels. This suggests that trustworthiness is driven by dispositional tendencies of unconditional kindness. By contrast, there was no evidence favoring the positive or negative reciprocity mechanisms as we found no interactions of HEX-HH or HEX-AG, respectively, with prior trust.

However, a potential limitation of our study was that participants distributed points (rather than money) in the trust game and that only 25% of participants were incentivized. Although meta-analytic evidence on the trust game suggests that trustee behavior should be unaffected by the actual rate of incentivization (Johnson & Mislin, 2011), it nonetheless seemed prudent to replicate the above results providing monetary behavior-contingent incentives for all participants. In addition, given that the majority of participants in Study 1 were students and that 75% were female, we aimed at critically testing the results in a more heterogeneous (non-student) sample. Therefore, participants in Study 2 were recruited by an independent professional panel provider. Furthermore, we considered it important to rule out demand effects of participants' personality self-reports on subsequent behavior in the trust game. Therefore, in Study 2 we implemented a longitudinal design, separating the personality assessment from the assessment of trustworthiness in time. Finally, as outlined above, Study 2 aimed at clarifying the mixed extant evidence on the relation between FFM-AG and trustworthiness. To this end, we additionally tested which aspects of FFM-AG actually link to (or show no relation with) trustworthiness.

Table 2. Means, Standard Deviations (in Parentheses), and Bivariate Correlations (95% Confidence Intervals in Brackets) of All Focal Variables Assessed in Study 2, With Internal Consistency Reliabilities (Cronbach's α) in the Diagonal.

			Correlations		
Measure	Scale	*M* (*SD*)	HEX-HH	HEX-AG	FFM-AG
HEX-HH	1-5	3.51 (0.59)	.73		
HEX-AG	1-5	3.13 (0.52)	.45*** [.29, .58]	.77	
FFM-AG	1-5	3.62 (0.45)	.49*** [.34, .62]	.60*** [.47, .71]	.75
Return 1.50€ (in %)	0-100	49.2 (26.4)	.28** [.11, .44]	.12 [−.07, .29]	.21* [.03, .38]
Return 3.00€ (in %)	0-100	47.9 (23.6)	.21* [.03, .38]	.07 [−.11, .25]	.15 [−.04, .32]
Return 4.50€ (in %)	0-100	45.9 (22.7)	.24** [.06, .40]	.15 [−.03, .32]	.10 [−.08, .28]
Return 6.00€ (in %)	0-100	46.3 (23.5)	.26** [.09, .42]	.15 [−.04, .32]	.16 [−.02, .33]
Return 7.50€ (in %)	0-100	48.5 (22.7)	.23* [.05, .40]	.13 [−.05, .31]	.16 [−.02, .33]
Return 9.00€ (in %)	0-100	49.8 (24.0)	.27** [.09, .43]	.14 [−.04, .31]	.18* [.00, .35]
M return (in %)	0-100	47.9 (20.8)	.29** [.11, .44]	.14 [−.04, .32]	.18* [.00, .35]

Note. HEX-HH = Honesty-Humility; HEX-AG = HEXACO-Agreeableness; FFM-AG = FFM-Agreeableness.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

Study 2

Method

Materials. As in Study 1, we used the German version of the HEXACO-60 to assess the six HEXACO dimensions. In addition, the FFM factors were measured via the German 60-item NEO Five-Factor Inventory (NEO-FFI; Borkenau & Ostendorf, 2008), including 12 items for each personality factor. In both questionnaires, participants' responses were collected on 5-point Likert-type scales, ranging from *strongly disagree* to *strongly agree*.

To measure trustworthiness, we again relied on the classical trust game, with participants acting in the role of the trustee.[6] However, in contrast to Study 1, participants now played for real money (rather than points). That is, participants (as well as trustors) received an initial endowment of 3.00€ (approximately US$3.80) and were—again corresponding to the strategy method—asked to decide how much they wanted to return to the trustor for each potential (tripled) transfer (between 1.50€ and 9.00€, in 1.50€ increments). In addition, we slightly changed the response format to allow maximum transparency, such that participants received explicit information on the outcomes for themselves and the unknown trustor for all potential returns.

Procedure. Study 2 was again conducted via the Internet. Yet, to further strengthen our data compared to Study 1, we implemented a longitudinal design, separating the personality assessment from the trust game. At both measurement occasions, participants first provided informed consent and demographic information. At Time 1, they completed the HEXACO-60 and the NEO-FFI, followed by other measures not pertinent to the current investigation. At Time 2 (about 5 months later), a random subsample of participants were re-invited to a follow-up study. In this study, participants received detailed information on the rules of the trust game and indicated their returns for each potential (tripled) trust transfer as a trustee. After completing data collection, participants were randomly matched to a trustor (assessed in a separate study) to determine individual payoffs. Incentive payment (consisting of a flat fee for participation and payoffs earned in the trust game) was handled entirely (and anonymously) by the panel provider.

Participants. Following the power analysis reported in Study 1, the subsample recruited for Time 2 comprised $N = 118$ participants (51 female). Supporting the heterogeneous composition of the sample, participants' ages covered a broad range (20-66 years), with a relatively high average age ($M = 42.0$, $SD = 12.4$). Only 5.1% of participants were students, whereas about two thirds (68.6%) were in employment. Also, there was a substantial diversity in educational levels.

Results and Discussion

Unconditional kindness versus (positive/negative) reciprocity. Table 2 reports the means, standard deviations, and correlations between all variables of interest (for information on all variables, see Table S2 in the Supplemental Material). Similar to Study 1, HEX-HH showed significant (positive) correlations with trustworthiness for all trust levels individually as well as for the average return (all rs > .20, $p < .025$). For HEX-AG, in turn, no noteworthy associations with trustworthiness were apparent (all rs ≤ .15, $p > .10$). Altogether, these zero-order correlations corroborate those observed in Study 1, although effect sizes were slightly different.

To test this pattern of results statistically, we relied on the same three-step approach as in Study 1 (based on the average return across trust levels). First, we approximated the BIC and corresponding BF_{01} based on R^2 (predicting trustworthiness with HEX-HH and HEX-AG, respectively; Raftery, 1995; Wagenmakers, 2007). For HEX-HH, $BF_{01} = 0.07$

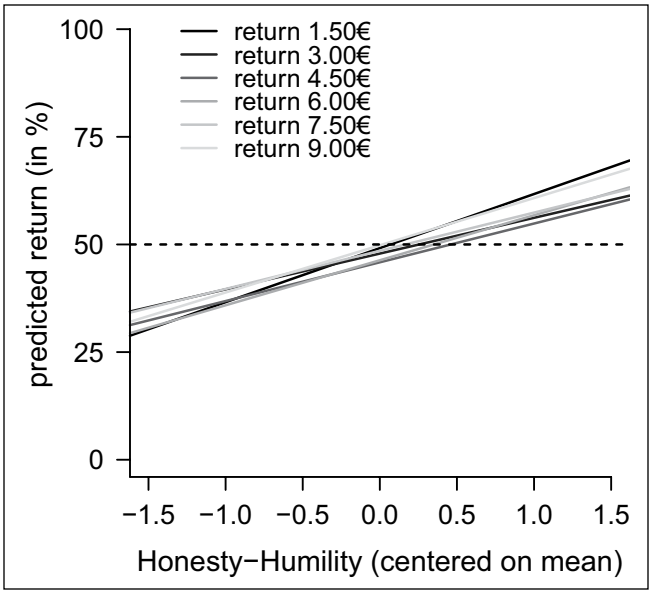

Figure 2. Predicted trustee returns for all levels of prior trust (between 1.50€ and 9.00€) depending on individual Honesty-Humility scores (centered on mean) in Study 2.

yielded substantial evidence in favor of the alternative hypothesis, being more than 14 times as likely as the null hypothesis given the data. For HEX-AG, by contrast, BF_{01} = 3.11 suggested the null hypothesis to be about 3 times as likely as the alternative hypothesis. Second, we compared the zero-order correlations between trustworthiness and HEX-HH (r = .29) and HEX-AG (r = .14), respectively (Meng et al., 1992). Unlike in Study 1, this test did not reach a conventional level of statistical significance (z = 1.49, p = .137). Nevertheless, when finally regressing the average return on both HEX-HH and HEX-AG concurrently, only HEX-HH predicted unique variance in trustworthiness, β = .28, p = .006, 95% CI = [.08, .47], whereas HEX-AG did not, β = .02, p = .830, 95% CI = [−.18, .22]. Taken together, these results largely replicate the findings of Study 1, providing evidence for the unconditional kindness and compatible with the positive reciprocity mechanism.

To further test whether the link between HEX-HH and trustworthiness is unconditional or conditional on the level of prior trust, we once more used a linear mixed model, first regressing trustworthiness on HEX-HH, trust, and their interaction (cf. Study 1). Consistent with Study 1, and in line with unconditional kindness, we found a significant main effect of HEX-HH, B = 10.03, p = .002, 95% CI = [3.92, 16.13], but no interaction between HEX-HH and trust, B = −0.28, p = .774, 95% CI = [−2.17, 1.61]. Correspondingly, comparing the main-effects model with the interaction model (cf. Study 1) yielded BF_{01} = 10.43 (based on ΔBIC_{10} = 4.69), thus indicating the main-effects model to be about 10 times as likely as the interaction model given the data. Indeed, the relation between HEX-HH and trustworthiness was again comparable for all potential trust levels (Figure 2), thus contradicting the positive reciprocity mechanism.

Finally, aiming to test the negative reciprocity mechanism, we reran the linear mixed model including HEX-AG as between-level predictor. However, the model showed no interaction between HEX-AG and prior trust in predicting trustworthiness, B = 0.60, p = .584, 95% CI = [−1.55, 2.75] (and, corroborating the results for the average return, no main effect of HEX-AG, B = 5.77, p = .114, 95% CI = [−1.40, 12.93]). Correspondingly, the BF_{01} = 9.35 (ΔBIC_{10} = 4.47) for the model comparison revealed that the main-effects model was more than 9 times as likely as the interaction model. Overall, Study 2 hence replicated the findings observed in Study 1, corroborating that trustworthy behavior can be considered an expression of unconditional kindness rather than of positive/negative reciprocity.

FFM-AG and trustworthiness. Regarding the relation between FFM-AG and trustee behavior, our results mirror previous findings in that FFM-AG showed small to medium-sized (Cohen, 1988) correlations with trustworthiness, with significant effects for some trust levels, but not for others (and r = .18, p = .045, for the average return; cf. Table 2). As reasoned above, a potential explanation for this relatively weak effect might be that only very specific aspects of FFM-AG—namely those driving unconditional kindness, as implied by the above results—actually account for trustworthiness. To test this assumption, we first regressed trustworthiness (average return) on both FFM-AG and HEX-HH, given that HEX-HH captures unconditional kindness. Indeed, only HEX-HH predicted unique variance in trustworthiness, β = .26, p = .013, 95% CI = [.05, .46], whereas FFM-AG did not, β = .06, p = .579, 95% CI = [−.15, .26]. This suggests that those aspects linking FFM-AG to trustworthiness are the ones covered by HEX-HH.[7] To test this interpretation more thoroughly, we used a mediation approach (Preacher & Kelley, 2011) in Mplus (Muthén & Muthén, 2012). Indirect effects ($a \times b$) refer to the standardized solution with the corresponding bootstrap confidence interval based on B = 1,000 bootstrap samples. As the analyses revealed, HEX-HH indeed mediated the link between FFM-AG and trustworthiness, $a \times b$ = 0.128, p = .043, 95% CI = [0.004, 0.252], but not vice versa (for FFM-AG as mediator), $a \times b$ = 0.028, p = .592, 95% CI = [−0.075, 0.132]. Overall, the results hence suggest that the relation between FFM-AG and trustworthiness can be attributed to aspects of active cooperativeness (including unconditional kindness) as covered by HEX-HH. However, as these aspects only constitute one component of FFM-AG (among several others), FFM-AG seems somewhat too broad to predict a specific behavior like trustworthiness in a satisfactory manner.

Taken together, Study 2 successfully replicated the results of Study 1 in a heterogeneous (non-student) sample with monetary behavior-contingent incentives for all participants,

using a longitudinal design. Thus, the findings support the conclusion that trustworthiness is mainly driven by unconditional kindness. Moreover, Study 2 provides clarification concerning the mixed evidence linking FFM-AG to trustworthiness: Aspects mirroring unconditional kindness are only a relatively minor component of the broad FFM-AG factor, but are well captured by HEX-HH.

Still, a limitation of Studies 1 and 2 refers to our exclusive reliance on the strategy method to assess trustworthiness. Although the strategy method has the inherent advantage of providing as much data as possible for each individual (cf. Brandts & Charness, 2011), responses are partially hypothetical in nature and do not necessarily mirror actual reactions toward another's trust. Consequently, the influence of reciprocity might be suppressed to some extent. To hence rule out that the strategy method undermined the positive/negative reciprocity mechanisms, participants in Study 3 indicated their return for one specific trust level only. Moreover, so as to provide more direct evidence on the proposed mechanisms underlying trustworthiness, we tested whether participants' expectations and evaluations regarding another's trust account for trustworthiness and further explicitly assessed participants justifications for their return decision.

Study 3

Method

Materials. Similar to Study 2, we assessed individuals' trait levels on the six HEXACO dimensions (using the German version of the HEXACO-60) and the FFM traits (using the 60-item NEO-FFI). Responses were given on 5-point Likert-type scales, ranging from *strongly disagree* to *strongly agree*.

Trustworthiness was again assessed via the classical trust game with participants acting as the trustees. However, unlike Studies 1 and 2, we now relied on the direct-response method. That is, participants only indicated their return for one specific trust level a trustor could transfer from her 6.00€ (approximately US$6.80) endowment. To ensure that all potential trust levels were almost equally covered in our data, we implemented a hypothetical design without "real" trustors and money involved. However, note that evidence supports the equivalence of trustee behavior across hypothetical and real scenarios (Holm & Nystedt, 2008). As an advantage, this procedure allowed us to systematically manipulate the trust level between participants (from 1.00€ to 6.00€, in 1.00€ increments). Correspondingly, participants were asked to imagine having received a specific transfer (tripled trust level) by an unknown other and to indicate how much they want to return.

Besides trustworthiness, we created two ad hoc scales to measure participants' evaluation of the specific trust level as well as their justification for their return decision (see Supplemental Material for individual items). The "Evaluation scale" consisted of 10 items (adjectives) in total, comprising 4 positive attributes (e.g., kind), 4 negative attributes (e.g., uncooperative), and 2 rationality-related attributes (e.g., understandable). The "Justification scale" comprised 6 items in total, with 2 items referring to each proposed mechanism (i.e., unconditional kindness, positive reciprocity, negative reciprocity). In both questionnaires, responses were given on 5-point Likert-type scales, ranging from *strongly disagree* to *strongly agree*. Analyses were based on the means of the three subscales for each questionnaire.

Procedure. Similar to Study 2, we again implemented a longitudinal design in Study 3, separating the personality assessment from the trust game. Specifically, another random subsample of participants taking part in the "pre-study" (Time 1) for Study 2 (in which participants completed the HEXACO-60 and the NEO-FFI) was re-invited to an online follow-up study (about 11 months later) by the same panel provider (excluding participants from Study 2). After providing informed consent for this follow-up study and demographic information, participants received detailed information on the rules of the trust game (as trustee). Next, they indicated the level of trust they would expect from an unknown other (between 1.00€ and 6.00€). Thereafter, participants received information on the actual (hypothetical) trust level, evaluated this trust level (using the 10-adjective Evaluation scale), and indicated how much of the tripled amount they wanted to return. Finally, participants provided reasons for their return decision (using the 6-item Justification scale). A flat fee for participation was paid out anonymously by the panel provider.

Participants. To further strengthen our conclusions, we recruited a slightly larger sample compared with Studies 1 and 2. Thus, the subsample recruited for the second measurement occasion of Study 3 comprised $N = 177$ participants (81 female). Overall, the composition of the sample was comparable to Study 2, covering a broad diversity in age (19 to 66 years, $M = 41.5$, $SD = 12.7$) and educational level. Only 7.9% of participants were students; 65.0% were in employment. Participants were almost equally distributed across trust levels (ranging between $n = 27$ and $n = 31$)

Results and Discussion

Unconditional kindness versus (positive/negative) reciprocity. Table 3 reports the means, standard deviations, and correlations between all focal variables (for information on all variables, see Table S3 in the Supplemental Material). As before, HEX-HH showed a positive (albeit weaker) relation to trustworthiness ($r = .17$, $p = .025$), which now referred to trustees' return (percentage of tripled transfer) in response to one specific trust level. For HEX-AG, a corresponding link was again absent ($r = .04$, $p = .611$). The zero-order correlations hence largely corroborated the results of Studies 1 and 2.

Table 3. Means, Standard Deviations (in Parentheses), and Bivariate Correlations (95% Confidence Intervals in Brackets) of All Focal Variables Assessed in Study 3, With Internal Consistency Reliabilities (Cronbach's α) in the Diagonal.

			Correlations			
Measure	Scale	*M* (*SD*)	HEX-HH	HEX-AG	FFM-AG	Return (in %)
HEX-HH	1-5	3.54 (0.74)	.84			
HEX-AG	1-5	3.13 (0.46)	.32*** [.18, .44]	.66		
FFM-AG	1-5	3.64 (0.45)	.56*** [.45, .66]	.53*** [.41, .63]	.73	
Return (in %)	0-100	50.5 (24.7)	.17* [.02, .31]	.04 [−.11, .18]	.05 [−.10, .20]	—
Expected trust	0-6	2.83 (1.60)	.10 [−.05, .24]	.03 [−.12, .18]	−.04 [−.18, .11]	.14 [.00, .28]
Trust evaluation: Positive	1-5	3.55 (0.91)	.08 [−.07, .22]	.19* [.04, .33]	.15* [.00, .29]	.04 [−.10, .19]
Trust evaluation: Negative	1-5	2.04 (0.96)	−.14 [−.29, .00]	−.18* [−.31, −.03]	−.22** [−.36, −.08]	−.09 [−.24, .05]
Trust evaluation: Rational	1-5	3.27 (0.88)	−.07 [−.22, .08]	.09 [−.05, .24]	.08 [−.06, .23]	−.12 [−.26, .03]
Justification: Unconditional kindness	1-5	3.53 (0.84)	.28*** [.13, .41]	.21** [.07, .35]	.17* [.03, .31]	.33*** [.19, .45]
Justification: Positive reciprocity	1-5	3.71 (0.89)	.08 [−.07, .22]	.14 [−.01, .28]	.18* [.03, .32]	.19* [.04, .33]
Justification: Negative reciprocity	1-5	1.75 (0.97)	−.18* [−.31, −.03]	−.13 [−.27, .02]	−.21** [−.35, −.06]	−.19* [−.33, −.04]

Note. HEX-HH = Honesty-Humility; HEX-AG = HEXACO-Agreeableness; FFM-AG = FFM-Agreeableness.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

To test this pattern of results statistically, we used the same three-step analytical approach as described above (now based on the return for a specific trust level). First, the approximation of the BIC and corresponding BF_{01} based on R^2 (Raftery, 1995; Wagenmakers, 2007) revealed $BF_{01} = 1.03$ for the link between HEX-HH and trustworthiness, and thus—unlike Studies 1 and 2—only inconclusive evidence (neither for nor against the alternative hypothesis). For HEX-AG, by contrast, $BF_{01} = 11.67$ indicated strong evidence in favor of the null, being almost 12 times as likely as the alternative hypothesis given the data. Second, comparing the zero-order correlations (Meng et al., 1992) between trustworthiness and HEX-HH ($r = .17$) and HEX-AG ($r = .04$), respectively, failed to reveal a significant difference, $z = 1.48$, $p = .139$. Nevertheless, in a multiple regression including both HEX-HH and HEX-AG as predictors of trustworthiness, HEX-HH predicted unique variance, $\beta = .17$, $p = .029$, 95% CI = [.02, .33], whereas HEX-AG did not, $\beta = -.02$, $p = .832$, 95% CI = [−.17, .14]. In sum, these results replicated the findings of Studies 1 and 2 only partially. Nonetheless, they still align with the unconditional kindness mechanism and are also compatible with the positive reciprocity mechanism.

To further test whether unconditional kindness or positive reciprocity accounts for the (albeit weak) link between HEX-HH and trustworthiness, we regressed trustworthiness (one return per participant) on HEX-HH, trust (between-participants), and their interaction in a multiple regression analysis. In line with unconditional kindness, we found a (one-tailed significant) main effect of HEX-HH, $\beta = .15$, $p = .053$, 95% CI = [.00, .30], but no interaction between HEX-HH and prior trust, $\beta = .00$, $p = .916$, 95% CI = [−.16, .14]. Correspondingly, comparing the main-effects-regression model with the interaction model revealed $BF_{01} = 13.23$ (based on $\Delta BIC_{10} = 5.16$), thus rendering the former 13 times as likely as the latter. Altogether, evidence once more corresponded better to the unconditional kindness mechanism than to the positive reciprocity mechanism.

Regarding the negative reciprocity mechanism, we reran the multiple regression analysis from above, now including HEX-AG as trait-based predictor. However, corroborating the results obtained with the strategy method (Studies 1 and 2), the interaction between HEX-AG and trust once more failed to explain significant variance in trustworthiness, $\beta = .02$, $p = .689$, 95% CI = [−.12, .18] (as did the main effect of HEX-AG, $\beta = .03$, $p = .672$, 95% CI = [−.12, .18]). In turn, comparing the main-effects model with the interaction model yielded $BF_{01} = 12.25$ ($\Delta BIC_{10} = 5.01$), thus implying a 12 times higher probability for the main-effects model given the data. Summing up, the trait-based evidence hence rendered the unconditional kindness mechanism most likely, thus essentially corroborating the conclusions from Studies 1 and 2.

Finally, analyses of the additional variables (i.e., expected trust, evaluations of trust, and justifications of return) provided further evidence for the unconditional kindness mechanism. First, expectations toward another's trust should reasonably influence one's interpretation of another's cooperativeness (cf. Gallucci & Perugini, 2000; see also Note 1)—as also observable in our data (Table S3 in the Supplemental Material)—and, according to a reciprocity account, affect the willingness to behave trustworthily. However, the relation between trustworthiness and expected trust level was only small (Cohen, 1988) and failed to reach statistical significance ($r = .14$, $p = .058$; Table 3). The corresponding $BF_{01} = 2.13$ implied the null hypothesis to be twice as likely as the alternative hypothesis given the data. Even smaller effect sizes emerged for individuals' evaluations of the specific trust level. That is, trustee returns were not contingent on whether participants evaluated the trust

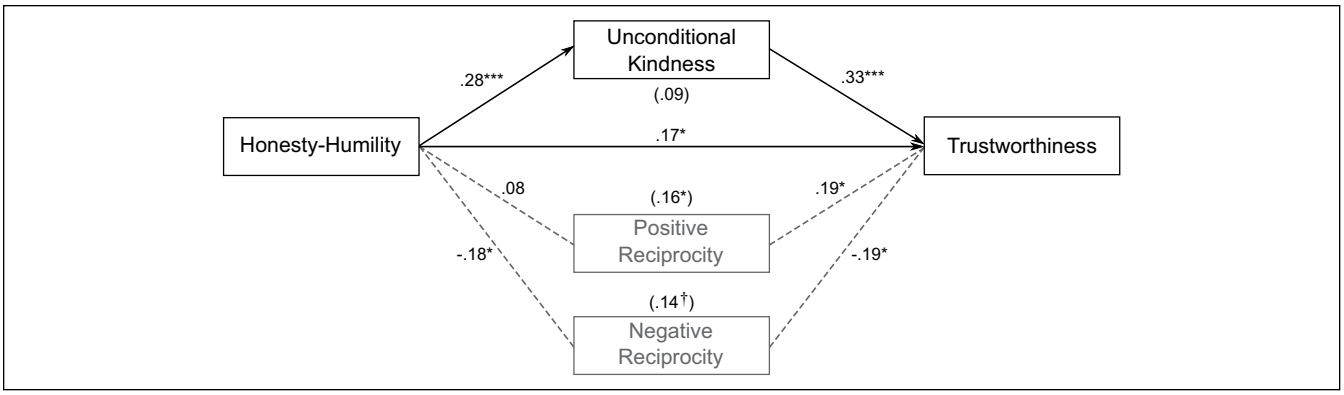

Figure 3. Mediation model of Honesty-Humility to trustworthiness with corresponding correlation and partial correlation coefficients (the latter controlling for the respective justification as mediator).
Note. Solid connections depict significant indirect effects; dashed relations depict a lack of significant indirect effects.
†$p \leq .10$. *$p \leq .05$. ***$p \leq .001$.

level positively ($r = .04$, $p = .561$, $BF_{01} = 11.21$), negatively ($r = -.09$, $p = .215$, $BF_{01} = 11.21$), or as being rational ($r = -.12$, $p = .116$, $BF_{01} = 3.81$). Moreover, participants' justification of their return decision showed the strongest correlation between trustworthiness and an unconditional kindness justification ($r = .32$, $p < .001$, $BF_{01} = 0.0006$), compared with the positive and negative reciprocity justification ($r = \pm.19$, $p = .012$, $BF_{01} = 0.54$). Correspondingly, the unconditional kindness justification was the sole significant predictor in a multiple regression including all three justification scales, $\beta = .30$, $p < .001$, 95% CI = [.15, .45]. Also, as depicted in Figure 3, the unconditional kindness justification mediated the link between HEX-HH and trustworthiness, $a \times b = 0.084$, $p = .008$, 95% CI = [0.022, 0.147], whereas both the positive ($a \times b = 0.014$, $p = .548$, 95% CI = [−0.031, 0.058]) and negative ($a \times b = 0.029$, $p = .156$, 95% CI = [−0.011, 0.070]) reciprocity justifications did not.[8] Overall, analyses of our complementary variables fit in well with the trait-based evidence from above, further supporting that trustworthiness is an expression of unconditional kindness.

FFM-AG and trustworthiness. Unlike Study 2—but converging with the mixed extant evidence—the relation between FFM-AG and trustee returns did not reach statistical significance in Study 3 ($r = .05$, $p = .505$). Correspondingly, in a multiple regression analysis concurrently considering HEX-HH and FFM-AG as predictors of trustworthiness, only HEX-HH predicted unique variance, $\beta = .21$, $p = .024$, 95% CI = [.03, .38], whereas FFM-AG did not, $\beta = -.07$, $p = .467$, 95% CI = [−.24, .11]. As such, results once more suggest that FFM-AG is somewhat too broad to account for sufficient variance in trustworthiness as a specific type of prosocial behavior.

General Discussion

The vital importance of trustworthiness for well-functioning social interactions is necessarily implied by the corresponding significance of trust. Surprisingly, however, the determinants of trustworthiness—especially in terms of (basic) personality traits—are insufficiently understood. Most prominently, two mechanisms to explain trustworthy behavior have been proposed: reciprocity (including positive and negative reciprocity) and unconditional kindness. However, purely game-theoretical evidence is mixed, and the empirical picture has remained inconclusive. The same holds for the few studies referring to basic personality traits (as conceptualized within the FFM). Most consistently, they suggest a positive, but weak and unreliable relation between trustworthiness and Agreeableness (FFM-AG). In any case, this link is insufficient to clarify the determinants underlying trustworthiness, given the broad nature of FFM-AG covering different aspects of cooperative tendencies (including positive/negative reciprocity and unconditional kindness).

To provide an enhanced understanding of the dispositional determinants of trustworthiness, we investigated the relation between trustworthiness and the HEXACO personality factors. In particular, HEX-HH and HEX-AG have consistently been linked to distinct aspects of cooperative tendencies, namely active cooperativeness (including unconditional kindness and positive reciprocity) versus reactive cooperativeness (including negative reciprocity). In turn, specific predictions on the to-be-expected relations between these two trait dimensions and trustworthy behavior can be derived for the proposed mechanisms to trustworthiness—thus disentangling unconditional kindness, positive reciprocity, and negative reciprocity through distinct hypotheses.

In three online studies, we assessed participants' trustworthiness (trustee behavior) in incentivized and hypothetical versions of the classical trust game, using either the strategy method (i.e., asking participants to indicate their trustworthiness for all potential trust levels; Studies 1 and 2) or the direct-response method (i.e., asking participants to respond to one specific trust level; Study 3). As implied by a mechanism of unconditional kindness, all studies revealed a positive link

between trustworthiness and HEX-HH, irrespective of the level of trust. Across studies, this resulted in a medium-sized average effect of $r = .28$ (sample-size weighted average correlation; Field, 2001). That is, HEX-HH showed a main effect on trustworthiness, but no interaction with prior trust—thus contradicting the positive reciprocity mechanism which inherently predicts such an interaction. HEX-AG, in turn, showed no relation to trustworthiness whatsoever (meta-analytic $r = .07$) and no interaction with prior trust, thus further contradicting that trustworthiness is determined by negative reciprocity. This interpretation was further supported by the finding that participants' expectations regarding another's trust as well as their evaluations of trust did not account for trustworthiness. In turn, participants' justification for their return decision corroborated the idea that trustworthiness is an expression of unconditional kindness. Altogether, our results are hence in line with the unconditional kindness mechanism, but cannot be reconciled with the positive or negative reciprocity mechanisms. Nevertheless, it might be worthwhile for future research to uncover potential situation-specific moderators that might render trustworthiness a conditional (reciprocal) behavior.

Overall, our findings support previous research implying that unconditional kindness is a prime determinant of trustworthiness—primarily based on a positive relation between trustworthiness and dictator game altruism (Ashraf et al., 2006; Cox, 2004; Gambetta & Przepiorka, 2014) and charitable giving (Fehrler & Przepiorka, 2013). Specifically, our studies extend the extant literature by using a trait-based approach, thus overcoming some of the inherent limitations associated with purely game-based approaches (e.g., common-method variance, desire for consistent responding). Moreover, the results—especially those clashing with the negative reciprocity mechanism—further support that trust does not correspond to a social norm which other people expect to be upheld (Bicchieri, Xiao, & Muldoon, 2011; Dunning, Anderson, Schlösser, Ehlebracht, & Fetchenhauer, 2014).

In addition, the results of Studies 2 and 3 replicated previous research in that they only revealed a weak link between FFM-AG and trustworthiness (meta-analytic $r = .10$). However, our results offer a reasonable explanation for this pattern: Unlike in HEX-HH, trait aspects predicting unconditional kindness are only marginally represented in FFM-AG. Correspondingly, HEX-HH mediated the link between FFM-AG and trustworthiness, but not vice versa. Similarly, HEX-HH predicted unique variance in trustworthiness beyond FFM-AG. Overall, the weak link between FFM-AG and trustworthiness thus seems to result from the misfit between the rather specific nature of trustworthiness (mainly incorporating unconditional kindness) and the broad nature of FFM-AG.

Regarding the HEXACO model in particular, the findings provide first evidence on the relation between the HEXACO dimensions and (incentivized) trust game behavior—thereby filling a gap identified in a recent meta-analytic review (Zhao & Smillie, 2015). In other words, they extend previous research linking HEX-HH and non-exploitation in the dictator game (e.g., Baumert et al., 2013; Hilbig et al., 2015) to situations in which the to-be-divided endowment is provided by another person rather than the investigator. Thereby, the results essentially corroborate the notion that HEX-HH should, by definition, drive trustworthiness (Thielmann & Hilbig, 2014).

Nonetheless, some limitations of the present studies should be acknowledged. First, the degree of interpersonal contact is obviously minimized in web-based studies. Thus, our web-based procedure might have diminished the feeling that one's own behavior is consequential for another's outcome. Although evidence suggests a high comparability of trust game behavior across web-based and lab-based studies (Holm & Nystedt, 2008), future studies might consider replicating our findings in lab-based settings. Second, due to our focus on basic and broad personality traits, we did not incorporate more specific trait scales of unconditional kindness and positive/negative reciprocity. This might be a worthwhile extension for future research. Finally, the trust game obviously represents only one specific situation eliciting trust and trustworthiness. Future studies might hence critically test the generalizability of our findings to other trust contexts.

In conclusion, the present studies provide trait-based evidence that trustworthiness is mainly an expression of unconditional kindness rather than positive or negative reciprocity. Thus, the findings contribute to the understanding of trustworthy behavior in terms of underlying personality traits and provide a valuable starting point for future research. Also, they point to the specific usefulness of HEX-HH and the higher resolution afforded by the HEXACO model to explain individual variation in trustworthiness and prosocial behavior more generally.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work reported herein was supported by grants to the second author from the Baden-Württemberg Foundation (Germany) and the German Research Foundation (HI 1600/1-1).

Notes

1. Indeed, from a game-theoretical perspective—assuming that individuals are rational utility-maximizers and thus untrustworthy—any trust transfer above zero should basically be perceived as cooperative. In turn, a trustworthy reaction necessarily implies positive reciprocity, whereas an untrustworthy reaction would simply mirror rationality (rather than implying negative reciprocity). However, whether a transfer is *actually* perceived as cooperative or uncooperative will lie in the eye of

the beholder and, for example, depend on the level of trust a trustee expects an interaction partner to place in her (Gallucci & Perugini, 2000). Correspondingly, Berg, Dickhaut, and McCabe (1995) themselves noted that trustees returning nothing "may not have *interpreted* the [trustor's] decisions as placing a trust" (p. 137, emphasis added).

2. Note that giving in the dictator game—as well as other forms of charitable giving—can also be considered in terms of fairness (cf. Forsythe, Horowitz, Savin, & Sefton, 1994) and might thus not provide an optimal measure of *pure* unconditional kindness.
3. On closer inspection, two HEXACO-Agreeableness items (items 3 and 27 of the herein used HEXACO-60; Ashton & Lee, 2009) are already conditional in nature. According to the negative reciprocity mechanism, these items should thus show an unconditional relation to trustworthiness (i.e., a simple main effect). Correspondingly, we have repeated all analyses for the two-item parcel—which basically yielded similar conclusions as will be reported below.
4. More precisely, participants were randomly assigned to the role of the trustor or the trustee. However, in what follows, we will exclusively refer to participants acting as the trustee. Specifically, trustors' responses mainly served the purpose of making the game "real" (without requiring deception).
5. The model specified the repeated trustworthiness measurements (Level 1) nested within participants (Level 2) and was estimated using the lm function of the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in R. All variables were centered on the global mean. Model statistics are based on maximum-likelihood estimates. However, all models reported here and in the following were also fitted based on restricted maximum-likelihood (REML), which did not result in any noteworthy differences.
6. The trustors, to whom participants were randomly matched, were assessed in a separate study. In what follows, we will only refer to the trustees.
7. To check whether our data (based on a moderate sample size) might have over- or underestimated the differential predictive power of Five-Factor Model-Agreeableness (FFM-AG) and Honesty-Humility (HEX-HH) on trustworthiness, we compared the correlation between FFM-AG and HEX-HH ($r = .49$) with a larger (and representative) German sample ($N = 2{,}027$). However, supporting the validity of the current findings, the corresponding effect size in the large sample ($r = .39$) still fell in the 95% CI = [.34, .62] for the observed correlation.
8. Results remain similar when including the trust level as between-participants covariate in the mediation models (for details on the mediation approach, see Study 2). We refrained from calculating the same analyses for HEX-AG due to the observed null relation between HEX-AG and trustworthiness.

Supplemental Material

The online supplemental material is available at http://pspb.sagepub.com/supplemental.

References

Ackermann, K. A., Fleiß, J., & Murphy, R. O. (in press). Reciprocity as an individual difference. *Journal of Conflict Resolution*. doi: 10.1177/0022002714541854

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

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*, 150-166. doi: 10.1177/1088868306294907

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*, 340-345. doi: 10.1080/00223890902935878

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*, 139-152. doi: 10.1177/1088868314523838

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package* (Version 1.1-7). Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

Bauernschuster, S., Falck, O., & Große, N. (2013). When trustors compete for the favour of a trustee—A laboratory experiment. *Journal of Economic Psychology*, *34*, 133-147. doi: 10.1016/j.joep.2012.09.004

Baumert, A., Schlösser, T. M., & Schmitt, M. (2013). Economic games: A performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment*. doi: 10.1027/1015-5759/a000183

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*, 453-478. doi: 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*, 64-79. doi: 10.1016/j.joep.2009.10.001

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

Bicchieri, C., Xiao, E., & Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy & Economics*, *10*, 170-187. http://doi.org/10.1177/1470594X10387260

Borkenau, P., & Ostendorf, F. (2008). *NEO-Fünf-Faktoren Inventar nach Costa und McCrae (NEO-FFI). Manual (2., neu normierte und vollständig überarbeitete Auflage)* [NEO Five-Factor Inventory by Costa and McCrae (NEO-FFI). Manual (2nd ed.)]. Göttingen, Germany: Hogrefe.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, *14*, 375-398. doi: 10.1007/s10683-011-9272-x

Chaudhuri, A., & Gangadharan, L. (2007). An experimental analysis of trust and trustworthiness. *Southern Economic Journal*, *73*, 959-985.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI–R) and NEO Five–Factor Inventory (NEO–FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences*, *12*, 887-898. doi: 10.1016/0191-8869(91)90177-D

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*, 260-281. doi: 10.1016/S0899-8256(03)00119-2

Denissen, J. J. A., & Penke, L. (2008). Motivational individual reaction norms underlying the Five-Factor Model of personality: First steps towards a theory-based conceptual framework. *Journal of Research in Personality*, *42*, 1285-1302. doi: 10.1016/j.jrp.2008.04.002

Dunning, D. A., Anderson, J. E., Schlösser, T. M., 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*, 122-141. doi: 10.1037/a0036673

Evans, A. M., & Krueger, J. I. (2009). The psychology (and economics) of trust. *Social & Personality Psychology Compass*, *3*, 1003-1017. doi: 10.1111/j.1751-9004.2009.00232.x

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, *42*, 1585-1593. doi: 10.1016/j.jrp.2008.07.011

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160. doi: 10.3758/BRM.41.4.1149

Fehrler, S., & Przepiorka, W. (2013). Charitable giving as a signal of trustworthiness: Disentangling the signaling benefits of altruistic acts. *Evolution & Human Behavior*, *34*, 139-145. doi: 10.1016/j.evolhumbehav.2012.11.005

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*, 161-180.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*, 347-369.

Gallucci, M., & Perugini, M. (2000). An experimental test of a game-theoretical model of reciprocity. *Journal of Behavioral Decision Making*, *13*, 367-389.

Gambetta, D., & Przepiorka, W. (2014). Natural and strategic generosity as signals of trustworthiness. *PLoS ONE*, *9*(5), e97533.

Hilbig, B. E., Glöckner, A., & Zettler, I. (2014). Personality and pro-social behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology*, *107*, 529-539. doi: 10.1037/a0036074

Hilbig, B. E., Thielmann, I., Hepp, J., Klein, S., & Zettler, I. (2015). From personality to altruistic behavior (and back): Evidence from a double-blind dictator game. *Journal of Research in Personality*. doi: 10.1016/j.jrp.2014.12.004

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional Honesty-Humility. *European Journal of Personality*, *26*, 245-254. doi: 10.1002/per.830

Hilbig, B. E., Zettler, I., Leist, F., & Heydasch, T. (2013). It takes two: Honesty–Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, *54*, 598-603. doi: 10.1016/j.paid.2012.11.008

Holm, H. J., & Nystedt, P. (2008). Trust in surveys and games—A methodological contribution on the influence of money and location. *Journal of Economic Psychology*, *29*, 522-542. doi: 10.1016/j.joep.2007.07.010

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

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*, 329-358. doi: 10.1207/s15327906mbr3902_8

Lönnqvist, J.-E., Verkasalo, M., Wichardt, P. C., & Walkowitz, G. (2012). Personality disorder categories as combinations of dimensions: Translating cooperative behavior in borderline personality disorder into the five-factor framework. *Journal of Personality Disorders*, *26*, 298-304. doi:10.1521/pedi.2012.26.2.298

Malhotra, D. (2004). Trust and reciprocity decisions: The differing perspectives of trustors and trusted parties. *Organizational Behavior and Human Decision Processes*, *94*, 61-73. doi: 10.1016/j.obhdp.2004.03.001

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, *52*, 267-275. doi: 10.1016/S0167-2681(03)00003-9

McCrae, R. R., & Costa, P. T. (1985). Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology*, *49*, 710-721. doi: 10.1037/0022-3514.49.3.710

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172-175. doi: 10.1037/0033-2909.111.1.172

Moshagen, M., Hilbig, B. E., & Zettler, I. (2014). Faktorstruktur, psychometrische Eigenschaften und Messinvarianz der deutschen Version des 60-Item HEXACO Persönlichkeitsinventars [Factor structure, psychometric properties, and measurement invariance of the German-language version of the 60-item HEXACO personality inventory]. *Diagnostica*, *60*(2), 86-97.

Müller, J., & Schwieren, C. (2012). *What can the Big Five personality factors contribute to explain small-scale economic behavior?* (Tinbergen Institute Discussion Paper No. 12-028.1). Amsterdam, The Netherlands: Tinbergen Institute.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Perugini, M., & Gallucci, M. (2001). Individual differences and social norms: The distinction between reciprocators and prosocials. *European Journal of Personality*, *15*(S1), S19-S35. doi: 10.1002/per.419

Perugini, M., Gallucci, M., Presaghi, F., & Ercolani, A. P. (2003). The personal norm of reciprocity. *European Journal of Personality*, *17*, 251-283.doi: 10.1002/per.474

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93-115. doi: 10.1037/a0022658

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163. doi: 10.2307/271063

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*, 243-256. doi: 10.1026//1618-3169.49.4.243

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes [The strategy method as a tool to analyze bounded rationality in oligopoly experiments]. In H. Sauermann

(Ed.), *Beiträge zur Experimentellen Wirtschaftsforschung* [Contributions in experimental economics] (pp. 136–168). Tübingen, Germany: J. C. B. Mohr.

Thielmann, I., & Hilbig, B. E. (2014). Trust in me, trust in you: A social projection account of the link between personality, cooperativeness, and trustworthiness expectations. *Journal of Research in Personality*, *50*(3), 61-65. doi: 10.1016/j.jrp.2014.03.006

Thielmann, I., & Hilbig, B. E. (in press). Trust: An integrative review from a person-situation perspective. *Review of General Psychology*. doi: 10.1037/gpr0000046

Thielmann, I., Hilbig, B. E., & Niedtfeld, I. (2014). Willing to give but not to forgive: Borderline personality features and cooperative behavior. *Journal of Personality Disorders*, *28*, 778-795. doi: 10.1521/pedi_2014_28_135

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804. doi: 10.3758/BF03194105

Yamagishi, T. (2011). *Trust: The evolutionary game of mind and society*. Tokyo, Japan: Springer. doi:10.1007/978-4-431-53936-0

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., . . . Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 20364-20368. doi: 10.1073/pnas.1212126109

Zettler, I., Hilbig, B. E., & Heydasch, T. (2013). Two sides of one coin: Honesty-Humility and situational factors mutually shape social dilemma decision making. *Journal of Research in Personality*, *47*, 286-295. doi: 10.1016/j.jrp.2013.01.012

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*, 277-302. doi: 10.1177/1088868314553709