Evolution and Human Behavior 37 (2016) 210–216

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

We can see inside: Accurate prediction of Prisoner's Dilemma decisions in announced games following a face-to-face interaction

Adam Sparks [a,b,*], Tyler Burleigh [a], Pat Barclay [a]

[a] Department of Psychology, University of Guelph, Guelph, ON N1G 2 W1, Canada
[b] Department of Anthropology and Center for Behavior, Evolution and Culture, University of California, Los Angeles, CA, 90095-1553, USA

ARTICLE INFO

Article history:
Initial receipt 28 August 2015
Final revision received 19 November 2015

Keywords:
Cooperation
Cheater detection
Thin slices
Impression formation
Personality judgment
Replication

ABSTRACT

Humans form impressions and make social judgments about others based on information that is quickly and easily available, such as facial and vocal traits. The evolutionary function of impression formation and social judgment mechanisms have received limited attention in psychology research; we argue that their function is to accurately forecast the behavior of others. There is some evidence for the predictive accuracy of social judgments, but much of it comes from situations where there is little incentive to deceive, which limits applicability to questions of the function of such mechanisms. A classic experiment that avoids this problem was conducted by R. H. Frank, T. Gilovich, and D. T. Regan (1993); their participants predicted each other's Prisoner's Dilemma Game decisions with above-chance accuracy after a short interaction period, knowing the game would follow. We report three original studies that replicate these aspects of the methods of Frank et al. (1993) and reanalyze data from all known replications. Our meta-analysis of these studies confirms the original report: humans can predict each other's Prisoner's Dilemma decisions after a brief interaction with people who have incentive to deceive.

1. Introduction

How do we assess our social partners? Psychology literatures on *impression formation* (Albright, Kenny, & Malloy, 1988; Eyal, Hoover, Fujita, & Nussbaum, 2011; Forgas, 2011; Neuberg & Fiske, 1987), *personality judgment* (Funder, 2012; Oh, Wang, & Mount, 2011), and *thin slicing* (Ambady, Conner, & Hallahan, 1999; Ambady & Gray, 2002; Ambady & Rosenthal, 1992; Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004; Willis & Todorov, 2006) have established that humans form impressions of others, that we do so quickly and with limited information, that our impressions match the impressions of others, and that these impressions can have predictive value.

However, why do we form these impressions? How do our assessments affect our social lives? What is the evolutionary function (Tinbergen, 1963) of these mechanisms, i.e., how do their effects on our social lives affect our reproductive success? Such questions have received less attention. For the most part, psychological research has focused on elucidating the mechanisms that produce judgments (Funder, 2012), giving less consideration to the adaptive significance of such judgments (e.g. Eyal et al., 2011; Forgas, 2011). Even when function is considered, as when evaluating the relationship between judge and target (discussed below), the discussion is often at the level of internal proximate mechanisms (e.g. resolving contradictory internal assessments; Neuberg & Fiske, 1987, pp 432-433) rather than on the level of solutions to strategic problems that may be favored by selection (e.g. the problem of predicting a potential social partner's behavior).

Function-focused research about assessing others should seek to characterize the relationship between a judge's impression of a target and the target's future behavior. Predicting people's future behavior is an important sub-problem of the broader task of making good social decisions. However, "accuracy" in the study of personality judgement and thin slices is most often operationalized as agreement between the judge's rating of a target and the target's self-rating or as agreement among multiple judges (e.g. Albright et al., 1988; Ambady & Rosenthal, 1993; Ambady & Gray, 2002); rarely is the "gold standard" of accurate behavior prediction employed (Funder, 2012). When the gold standard is used or discussed, it often takes the form of personality judgments correlating with behaviors of the targets (e.g. Fast & Funder, 2008; Hepler & Albarracín, 2013; Oh et al., 2011; Ozer & Benet-Martínez, 2006; Rule, Krendl, Ivcevic, & Ambady, 2013), rather than behavioral predictions about the targets corresponding to behaviors of the targets. In sum, it appears that little recent psychology research about how we make impressions of others explicitly considers whether we can and do use impressions to accurately predict people's behavior, and thus such research sheds little direct light on the evolutionary importance of such predictive ability.

Function-focused research should also approach this problem differently by considering the strategic interests of the judge and the target. Is it in the interests of the target to be accurately evaluated? How can the

* Corresponding author. Department of Anthropology and Center for Behavior, Evolution and Culture, University of California, Los Angeles, CA, 90095-1553, USA.
E-mail addresses: adspar@gmail.com (A. Sparks), tburleig@uoguelph.ca (T. Burleigh), barclayp@uoguelph.ca (P. Barclay).

http://dx.doi.org/10.1016/j.evolhumbehav.2015.11.003
1090-5138/

judge use an accurate judgment? Investigations of how judges' accuracy is influenced by motivation and outcome dependency (e.g. De Melo, Carnevale, Read, & Gratch, 2014; Neuberg & Fiske, 1987; Pelham & Neter, 1995) begin to consider strategic interests, but for the most part functional issues are seldom addressed. Efforts to situate assessment mechanisms in their social context, such as by considering how the relationship between the judge and the target affects the assessment, are rare (Fiske & Dépret, 1996). Indeed, a recent review paper on the accuracy of personality judgments (Funder, 2012) discusses what makes "good judges" without considering motivation or incentive to judge the target correctly, and what makes "good targets" without considering the target's strategic interest to inform or deceive the judge.

A great deal of evolutionary literature on the psychology of social behavior is devoted to the issue of how humans use information about past behavior to inform social decision-making (reviewed by Barclay, 2015). However how do people make judgments when such information is unavailable? Recent evolutionary psychology research has established that thin slice information can be used to accurately predict incentivized social behavior, such as vocal characteristics predicting aggression (Puts, Apicella, & Cárdenas, 2011; Sell et al., 2010) and facial features or expressions predicting cooperation and honesty (e.g. Brown, Palameta, & Moore, 2003; Fetchenhauer, Groothuis, & Pradel, 2010; Haselhuhn & Wong, 2011; Little et al., 2013; Oda, Naganawa, Yamauchi, Yamagata, & Matsumoto-Oda, 2009; Stirrat & Perrett, 2010; Tognetti, Berticat, Raymond, & Fauire, 2013; Verplaetse, Vanneste, & Braeckman, 2007; Verplaetse & Vanneste, 2010). These studies consider the strategic interests of the targets by assessing their behavior in the context of real social interactions, often operationalized as cooperate/defect decisions in economic games. Much of this literature is explicitly anchored by theoretical consideration of the functional significance of the mechanisms being investigated, such as harnessing the benefits of repeated mutual cooperation and avoiding the costs of interacting with non-cooperators, and related problems (Cosmides & Tooby, 1989).

However, we note two shortcomings of many these evolutionary studies. First, they involve presenting images or video of target individuals whose behavior will be predicted (see also: Kiyonari, 2010; Vogt, Efferson, & Fehr, 2013), rather than making predictions in the more ecologically-valid context of a face-to-face encounter. Second, many methods (see also: DeSteno et al., 2012; Manson, Gervais, & Kline, 2013) do not consider all strategic interests because they involve situations in which non-cooperators had little incentive to present themselves as cooperative because the games were unannounced (i.e. participants did not know a cooperation/defection decision would follow; Kiyonari, 2010). In reality, would-be defectors benefit from appearing cooperative when interacting with partners; accurate prediction under such circumstances is important to questions of function because many forms of human cooperation are explicitly negotiated.

In light of these concerns with both the social psychology literatures on impression formation, personality judgment, and thin slices, and the evolutionary literature on mechanisms for predicting social behavior, it is important to seek evidence for the accuracy of (1) incentivized predictions of (2) incentivized behavior (3) when the target has incentive to deceive the judge. Such evidence is crucial to any argument that social judgments are of practical consequence and crucial to evaluating models of the evolutionary origins of cooperation (Kiyonari, 2010).

Evidence meeting the criteria above was provided by Frank, Gilovich, and Regan (1993), who showed that previously unacquainted people can predict the decisions of others in a Prisoner's Dilemma Game (henceforth "PDG") after 30 minutes of face-to-face interaction during which they knew PDGs would follow. In a PDG, two players each have a choice of cooperating with their partner or defecting on her. Each player's individual payoff is higher if she chooses to defect, but choosing to cooperate improves the partner's payoff such that mutual cooperation results in the highest combined outcome. Thus when negotiating with a PDG partner prior to the game, would-be defectors have incentive to deceptively present themselves as cooperators in attempt to induce cooperation from the partner. Accordingly, would-be cooperators have incentive to detect deception to avoid the "sucker's payoff" (Table 1).

PDGs are widely used to model social interactions and the evolution of cooperation (Axelrod, 1984; Boyd & Lorberbaum, 1987; Brembs, 1996; Dawkins, 1976/2006; Klapwijk & Van Lange, 2009; Nowak & Sigmund, 1992, 1993; Wilson, 1971). Evidence about the psychology of this game is an important test of such models. The PDG prediction ability reported by Frank et al. (1993) has received modest but steady attention in the literature on cooperation (e.g. as of November 3, 2015 Google Scholar citation count is 276), especially the literature on cheater detection (e.g. DeSteno et al., 2012; Little, Jones, DeBruine, & Dunbar, 2013; Sylwester, Lyons, Buchanan, Nettle, & Roberts, 2012). Frank et al. (1993) have also been cited in studies of sociopathy (Mealy, 1995), morality and fairness (Baumard, Andre, & Sperber, 2013), determinants of cooperation (Fischer, 2009), interdependence theory (Rusbult & Van Lange, 2003), and the applicability of rational choice theory to understanding social interactions (Colman, 2003). As argued above, the empirical findings of Frank et al. are broadly relevant to studies of impression formation, personality judgment and thin slice judgments because they provide supportive evidence of a function of such mechanisms: to accurately forecast the behavior of our (potential) social partners.

Confidence in research findings is enhanced by independent replication, robust statistical analysis, and confirmatory meta-analysis. Brosig (2002) reported a successful replication of Frank et al. (1993) but did not analyze all relevant data and did not control for non-independence of observations. Reed, Zeglen, and Schmidt (2012) also reported a successful replication, but their report largely focussed on other results. In this paper we present our own research that replicates Frank et al. and controls for non-independent observations. We also reanalyze Brosig's data and present a meta-analysis of all available studies in which participants predicted each other's PDG decisions after an acquaintance period preceding an announced game. To be clear, our focus here is not on the specific theoretical ideas of Frank et al. (1993) regarding the evolution of one-shot cooperation (see Manson et al., 2013, for a nice review); rather, our focus is on evaluating evidence that people can accurately predict each other's PDG decisions despite incentives to deceive.

2. Methods

2.1. Overview

The first of three new studies we report here was designed as a pilot test of methods then intended to replicate and extend Frank et al. (1993). We used a slightly modified version of their methods as our control condition, expecting to observe similar predictive accuracy, and we expected to observe greater predictive accuracy in our experimental condition. We found no evidence of above-chance prediction in the control or experimental conditions; thus there was no direct test of our prediction and no replication of the original results. Wondering if minor method differences might explain the null replication, we designed the second study as a more precise replication. A third study was designed primarily to test whether PDG defection is correlated with outcomes of other tasks (results reported elsewhere; Sparks et al. in review), but we asked participants to predict each other's decisions to provide data relevant to the current investigation.

So, we conducted three studies that shared the following aspects of Frank et al. (1993), Brosig (2002) and Reed et al (2012): Student participants came to a laboratory, received an overview of the PDG, spent some time talking to their PDG partner(s), and then played a PDG with one or more partners and tried to predict the decisions of the partner(s). Our method details are publicly available (Sparks, Burleigh, & Barclay, 2015). The goal of this report is to evaluate data from all six of these studies about the accuracy of PDG predictions. Do the collective

Table 1
Methods discrepancies.

Discrepancy	Frank et al.	Brosig	Reed et al.	Study 1	Study 2	Study 3
Acquainted participants						
Identified?	Yes	Yes	Yes	Yes	No	No
By whom?	Participant	Participant	Participant	Experimenter	–	–
Anonymity						
PDG decisions revealed?	No	No	No	No	No	Yes
Confidence in prediction						
Measure included?	Yes	No	Yes	Yes	Yes	Yes
Scale	50 to 100	–	1 to 7	50 to 100	50 to 100	0 to 100
Sample size						
Number of participants	99	143	158	60	84	158
Predictions per player	2	3	1	2 or 3	2	1 or 2
predictions	198	394	158	168	167	283
Interaction structure						
Time (minutes)	30	10	10	25	30	10 to 20
Group size	3	2	2	3 or 4	3	2 or 3
PDG payoff parameters						
T	3	40	6	3	4	4
R	2	24	4	2	2.25	2.25
P	1	8	2	1	1.5	1.5
S	0	−8	0	0	0	0
Experimental condition	No	No	No	Yes	Yes	No

Summary of the differences in methods between studies, as discussed in main text and supplement. Sample size. Some participants' predictions were disqualified or missing, so total predictions is not always an exact multiple of participants and predictions per player. Interaction structure. Participants interacted before predicting each other's decisions. Brosig's participants had multiple paired interactions. Participants in other studies had a single interaction in groups of sizes 2, 3 or 4. PDG parameters. The payoff when a defector plays a cooperator is T ("Temptation to defect") for the defector and S ("Sucker's payoff") for the cooperator. Mutual cooperation pays R ("Reward for mutual cooperation") to each player. Mutual defection pays P ("Punishment for mutual defection") to each player. A Prisoner's Dilemma exists whenever payoffs are such that $T > R > P > S$ and $2R > T + S$. Units for Frank et al. and Reed et al. are USD. Units for Brosig are "lab dollars" convertible to DM at 4:1. Units for our studies are CAD.

results demonstrate that people can predict each other's PDG decisions after a brief interaction in which they have incentive to deceive each other?

2.2. Major methods differences

Some methods differences are inevitable in replication. Table 1 summarizes the most important differences among studies providing data on the accuracy of PDG predictions. These discrepancies should not obscure the fundamental similarity of these methods as described above. In the sections below, we discuss the relevance of major methodological differences to our analyses and interpretation of the prediction data. Minor differences are discussed in the supplementary material.

2.2.1. Previous acquaintance of participants

Frank et al. (1993) and Reed et al. (2012) ensured that their participants were previously unacquainted. Brosig (2002) collected data on acquainted participants but excluded their predictions from her reported analyses. Previous acquaintance was not relevant to the hypotheses our methods were originally designed to test, so we only systematically collected acquaintance data in study 1 (based on experimenter observation) and never screened participants on this basis. In study 1, 11.9% of games were played between acquainted participants; studies 2 and 3 drew from the same participant pool and we have no reason to suspect major differences. Thus, we expect that all of our studies included a minority of acquainted participants.

The challenge of making decisions involving novel partners is related to the challenge of making decisions involving known partners in novel settings—reputation information is either unavailable or of unclear value. As it is quite unlikely that acquainted participants in our studies have ever played a formal PDG with each other, they are likely in a novel social context. Obviously a data set that includes predictions of previously-acquainted participants has limited value for addressing the specific question investigated by Frank et al. (1993): can people distinguish cooperators from defectors based *only* on a brief interaction? But such data do provide information about a related and more general question: Can people predict social decisions of others in the absence of information about their past behavior in the same social context? The extent to which the latter question informs the former depends on whether previous acquaintance alters predictive accuracy. We see no clear *a priori* prediction on the matter. Recall the Godfather's advice: "keep your friends close but your enemies closer" (Coppola, F.F. (Producer, & Director), 1974), i.e. people who know each other well can apply that knowledge towards cooperative coordination or towards exploitative deception (see Sylwester et al., 2012).

In an effort to shed light on how previous acquaintance affects predictive accuracy, we report separate analyses of acquainted participants when possible—for our reanalysis of Brosig's data and for our study 1. We also re-computed our primary analyses with these acquainted sub-samples removed.

2.2.2. Anonymity

If a PDG is certain to be an isolated event—a one-shot interaction—defection is the dominant strategy and is favored by natural selection. However, if the outcome of one game can affect future social interactions (with the same partner or others) natural selection can favor the evolution of cooperative strategies (Delton, Krasnow, Tooby, & Cosmides, 2011). A common method for implementing one-shot games in experimental settings is ensuring anonymity such that participants' decisions cannot be known to other participants or to the experimenter with whom participants directly interact. But, it is unclear that participants regard such games as truly one-shot (Hagen & Hammerstein, 2006).

Experimenter guarantees of anonymity and one-shot interaction may be considered a *cue* that the current interaction is unlikely to have reputational consequence for any future interactions. However, such guarantees are not the only cues used to gauge reputational consequences. For example, a growing literature shows that an invalid cue to observation—images of eyes rather than real watching eyes— increase cooperativeness, even in effectively one-shot settings (reviewed by Sparks & Barclay, 2013 and Nettle et al., 2013). This suggests that such cues increase the estimated likelihood that the current interaction may have reputation-mediated consequences for future interactions, contrary to anonymity promises by experimenters. If exposure to images of eyes, or even images bearing only a crude resemblance to eyes (Rigdon, Ishii, Watabe, & Kitayama, 2009), can make otherwise anonymous participants behave as if their reputation is at stake, several minutes of face-to-face conversation between students at the same university could have a similar effect. Participants in Frank et al. (1993), Brosig (2002), Reed et al. (2012) and our studies 1 and 2 are likely to have assessed (perhaps implicitly) that there was a non-zero chance that their PDG decisions could influence their future social interactions, despite procedural guarantees to the contrary.

Thus, we argue that the difference between the experiences of participants in our study 3 and the other studies was not one of strictly public versus strictly anonymous decisions. We suggest that the decision-making mechanisms of our participants do not evaluate a dichotomy between anonymous and public, but a probabilistic continuum of likelihoods that the current social decision will affect future social outcomes. Compared to participants in the other studies, participants in study 3 faced social decisions with greater probabilities of influencing the outcomes of their future interactions. While this should be expected to increase cooperation rates (and therefore rates of cooperation predictions), we see no clear *a priori* reason to think that this difference would influence the accuracy of predictions (relative to chance), and thus no reason to exclude study 3 from our meta-analysis. Nevertheless, we report results of the meta-analysis with and without study 3.

2.2.3. Interaction structure

Table 1 summarizes differences in the structure of the interactions between participants, during which they presumably form impressions on which they base their predictions. Brosig's participants interacted in pairs. In Reed et al. (2012), each player had three consecutive paired interactions. Frank et al.'s participants interacted in groups of three before each player predicted the decisions of the other two. In our first study, interactions occurred in groups of size three or four; our second study included only groups of three; our third study included groups of three and pairs. The amount of time for the interactions varied as well.

Our designs reflected trade-offs between the quality of interaction provided to participants and the quantity of prediction data that can be collected under constraints. A reasonable hypothesis about the relationship between interaction structure and predictive accuracy is that accuracy will be greater when groups are smaller and when interaction time is longer, i.e. when the interactions are less diluted. Indeed, this was how Frank et al. explained the greater predictive accuracy in their study compared to a study involving much larger groups. That specific hypothesis can be tested in future work. Here, we simply assess predictive accuracy among people who interacted for 10 to 30 minutes in groups of four players or fewer.

2.3. Analysis

Our dependent variable, accuracy of a binary prediction, is dichotomous, so we use the odds ratio as our measure of effect size. Odds ratio values range between 0 and ∞. An OR = 1 would indicate no association between prediction and actual decision; whereas an OR > 1 would indicate above-chance prediction accuracy. For example, an OR = 2 would be interpreted to mean that the odds of a guesser predicting cooperation when playing with a cooperator was two times greater than the odds of a guesser predicting cooperation when playing with a defector.

Most PDG predictions were nonindependent because individuals provided multiple responses and also participated in groups. Thus, we use a generalized estimating equation (GEE) approach to account for the clustering of data. We report the adjusted odds ratios and 95% confidence intervals of our data and Brosig's (2002), and we also report a meta-analysis (Hunter & Schmidt, 2004) that combines the results of our studies with those of Frank et al. (1993), Brosig (2002), and Reed et al. (2012).

As in previous work, data from participants who promised to defect was eliminated. Such promises are highly credible, and in all known cases their partners correctly predicted defection. Including such data would raise estimates of accuracy.

2.3.1. Data clusters and model selection.

In our studies the prediction data were clustered at two nested levels: the level of the individual who made predictions, and the level of the group in which individuals participated. In groups of three, each individual would have made two predictions and two game decisions (one for each of the other players). Given this, we generated two alternative models where data were clustered by: (1) guessers within groups, and (2) players within groups. In Brosig's study the data were clustered at one level: the individual who played the game, which can be similarly specified in terms of guessers and players. We use the *quasilikelihood under the independence model criterion* (QIC; Pan, 2001) to select between models.

2.3.2. Study-level analyses and meta-analysis

2.3.2.1. Study. For each of our three studies, and also for our re-analysis of Brosig's data, we estimated the QIC for two models identified above, using the *qic* command in Stata (Cui, 2007; StataCorp, 2013). We select the model with the lowest value for parameter estimation. Next, we perform GEE on the selected models to estimate an odds ratio and confidence intervals describing prediction accuracy. For this we use the *xtgee* command in Stata; we specify the covariance structure as exchangeable and use the robust ("sandwich") variance estimator to obtain adjusted confidence intervals.

2.3.2.2. Meta-analysis. We use the *meta* command in Stata, entering the estimated log-odds and the standard errors of the log-odds from each study. We assign weights based on the inverse of the estimates' variance, such that greater weight is given to studies with less statistical noise, and the variance of the pooled effect estimate is minimized. We use a random-effects model; we test for the heterogeneity of studies to confirm or disconfirm this decision and report the Q and I^2 statistics (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

3. Results

Our data are publicly available (Sparks, Burleigh, & Barclay, 2015).

3.1. Study-level analyses

3.1.1. Model selection

The Player model is preferred in study 1, whereas the Guesser model is preferred in studies 2 and 3 and for Brosig's data (Table 2).

3.1.2. Predictive accuracy

Our studies 2 and 3, and Brosig's data provided evidence of above-chance predictive accuracy (study 2: OR = 2.19; CI = 1.034 to 4.637; Study 3: OR = 4.001; CI = 1.685 to 9.5; Brosig: OR = 2.582; CI = 1.668 to 3.999). Using rules-of-thumb (Chen, Cohen, & Chen, 2010), the

Table 2
QIC values for the data-cluster models.

Study	Data-cluster model	QIC value
Study 1	Guesser	208.94
	Player*	208.74
Study 2	Guesser*	206.30
	Player	209.50
Study 3	Guesser*	236.03
	Player	238.99
Brosig (2002)	Guesser*	529.63
	Player	533.29

* The selected model.

effects in study 2 and in Brosig (2002) were small (3.47 > OR > 1.68), whereas the effect in study 3 was large (OR > 3.47).

Study 1 did not provide any evidence of above-chance predictive accuracy (OR = 0.903; CI = 0.471 to 1.729).

3.1.3. Analysis of acquainted participants.

We conducted a separate analysis of acquainted participant subsamples from study 1 and Brosig (2002). These observations were not repeated and thus nonindependence was not an issue. The small sample sizes made calculations of the OR impossible, so we computed the percentage accuracy instead and compared it to chance accuracy, as in Frank et al. (1993) and Brosig (2002). We also re-computed the GEE models for each study with acquainted participants removed.

In study 1, we found no evidence of above-chance accuracy for acquainted participants (70% accuracy, CI = 49.4% to 90.6%; chance accuracy = 74.5%). This is consistent with the previous OR estimate for study 1. With these participants removed from the GEE model, the OR estimate changed only slightly (OR = 1.024, CI = 0.459 to 2.281); if anything the accuracy was slightly higher with the acquainted participants removed. This suggests that the above-chance accuracy observed in studies 2 and 3 was unlikely to be caused solely by the inclusion of acquainted participants. Our re-analysis of acquainted participants in Brosig (2002) revealed perfect accuracy (100%). However, because the subsample was a small minority of the overall sample, removing these participants from the GEE model reduced the OR estimate only slightly (OR = 2.343, CI = 1.508 to 3.64).

3.2. Meta-analysis

A test for the heterogeneity of studies was significant ($Q = 23.773$, $p < 0.001$; $I^2 = 78.968$, $p < 0.001$), which supports our decision to use a random effects model.

When the results of our three studies are combined with those obtained by Frank et al. (1993), Brosig (2002), and Reed et al. (2012), the effect for overall prediction accuracy is small and above-chance (OR = 2.187; CI = 1.341 to 3.568). This meta-analysis is summarized in Fig. 1. Removing study 3 from this analysis decreases the OR slightly (OR = 1.990, CI = 1.189 to 3.332).

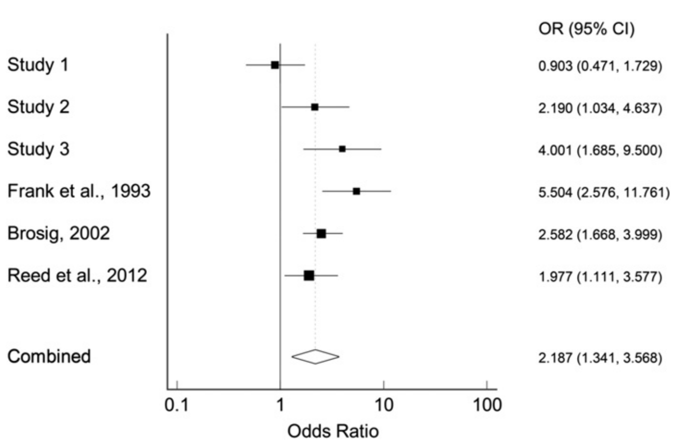

Fig. 1. Forest plot—marker size indicates a study's weight in the combined estimate.

4. Discussion

4.1. Summary

Data showing that people can explicitly predict the *behavior* of others is important for functional research into the mechanisms of impression formation and personality judgment, and meets an elusive "gold standard" for evidence of accuracy of personality judgments (Funder, 2012). We investigated the accuracy of predictions of behaviors of clear adaptive significance: incentivized cooperative social decisions, made by actors about whom we have limited information and who have incentive to deceive us.

Replicating underappreciated earlier work (Frank et al., 1993; Brosig, 2002; Reed et al, 2012), we conducted three studies in which participants were asked to predict the PDG decisions of other players following a brief face-to-face interaction. The combined results of all studies confirm the previously-reported results: people can accurately predict a social decision of a person who has incentive to mislead them. There is no evidence that this above-chance accuracy is driven by especially-accurate predictions of previously-acquainted participants, though this issue deserves further study. We also found some evidence that participants were more confident in their accurate predictions (Supplement). Can we figuratively "see inside" (Vogt et al., 2013) the minds of others? Yes.

4.2. Applications and future directions

What are the limits of our ability to see inside? Frank et al. (1993) suggested that predictions are likely to be more accurate after higher-quality interactions, such as when groups are smaller. Tentatively supporting this argument, the only study in our meta-analysis that did not provide evidence of predictive accuracy involved the largest groups (Table 1). Further, predictions about the decisions of economic game players seem to be less accurate when methods provide an unexpected game and/or when the "thin slice" is not a face-to-face encounter (Kiyonari, 2010; Manson et al., 2013; Vogt et al., 2013). This difference might be explained by such interactions being lower quality. Of course, real-life meetings between strangers often occur without a well-defined impending cooperation/defection opportunity. Perhaps people meeting in ambiguous contexts do form accurate impressions of each other, but on more salient dimensions than economic cooperativeness (as suggested by Manson et al., 2013), such as reproductive value or availability. Understanding how predictive accuracy is influenced by interaction quality and context could be applied towards the study of a variety of partner choice decisions of interest to social psychologists, such as employment and mating decisions.

Evidence of accurate prediction of social decisions under various conditions is a step towards understanding the function of impression formation mechanisms. A logical follow-up question is: How do predictions and impressions inform the social decisions of the predictor? A good social foraging strategy might be to attempt to identify non-cooperative partners and avoid interacting with them entirely. Reputation information can be used this way (Barclay & Willer, 2007; Sylwester & Roberts, 2010)—perhaps impressions from brief exposures are used similarly. We did not invite participants to refuse to play a PDG after their interactions, but in an open-ended debriefing question we did ask them which player they'd prefer to play an additional game with. The results of this exploratory non-incentivized partner choice measure are somewhat suggestive of a relationship between predictions and partner choice (Supplement), but future work could investigate the matter with more rigor.

We found that predictive accuracy is better than chance, but far from perfect. There was a bias to over-predict cooperation, which might reflect an error management problem whereby the costs of exposing oneself to a non-cooperative partner are (assumed to be) small relative to the gains of acquiring a new cooperative partner (Delton et al., 2011). Future research could investigate whether predictive accuracy and errors are influenced by these types of functional cost-benefit trade-offs. Improved understanding of when and why we can accurately predict behavior can in turn inform the study of memory or other down-stream social psychology phenomena. For example, when defectors are not initially detected or avoided, it is perhaps best to remember them and avoid further interaction (Barclay, 2008; Barclay & Lalumière, 2006).

Finally, the results of our model selection analysis may inform investigations of how predictive accuracy is influenced by cheater detection skill (i.e. good guessing) and effective signalling of cooperative intent (Manson et al., 2013; Reed et al., 2012). The only study in which a guesser-clustered model was not more informative was our study 1, where there was no evidence of predictive accuracy. This tentatively suggests that cheater detection based on cues of defection intent may be a more important factor than signaling cooperative intent.

Author Contributions

Sparks and Barclay designed the studies, obtained REB approval, and collected the data. Burleigh and Sparks analyzed the data. Sparks and Burleigh drafted the manuscript with Barclay's feedback. All authors approved the submitted manuscript.

Supplementary Materials

Supplementary data to this article can be found online at doi:10.1016/j.evolhumbehav.2015.11.003.

Acknowledgments

The authors thank Jeannette Brosig for sharing data and Kevin Kellenberger for help with data collection. We are grateful for comments from Kerry Kawakami, Daniël Lakens, Daniel Meegan, Stefan Linquist, Rob Kurzban, Elena Choleris and anonymous reviewers. The work was funded by an SSHRC grant to the third author.

References

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *55*, 387–395.

Ambady, N., Conner, B., & Hallahan, M. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology*, *77*, 538–547.

Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology*, *83*, 947–961.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*, 431–441.

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

Axelrod, R. (1984). *The evolution of cooperation.* New York: Basic Books.

Barclay, P. (2008). Enhanced recognition of defectors depends on their rarity. *Cognition*, *107*, 817–828.

Barclay, P. (2015). Reputation. In D. Buss (Ed.), *Handbook of evolutionary psychology* (pp. 810–828) (2nd ed.). Hoboken, NJ: J. Wiley & Sons.

Barclay, P., & Lalumière, M. L. (2006). Do people differentially remember cheaters? *Human Nature*, *17*, 98–113.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*, 749–753. http://dx.doi.org/10.1098/rspb.2006.0209.

Baumard, N., Andre, J. -B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *The Behavioral and Brain Sciences*, *36*, 59–78.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*, 599–614.

Boyd, R., & Lorberbaum, J. P. (1987). No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. *Nature*, *327*, 58–59.

Brembs, B. (1996). Chaos, cheating, and cooperation: Potential solutions to the Prisoner's Dilemma. *Oikos*, *76*(1), 14–24.

Brosig, J. (2002). Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior & Organization*, *47*, 275–290.

Brown, W. M., Palameta, B., & Moore, C. (2003). Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evolutionary Psychology*, *1*, 42–69.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation*, *39*(4), 860–864.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *The Behavioral and Brain Sciences*, *26*, 139–198.

Coppola, F. F. (1974). *The Godfather: Part II [Motion Picture].* United States: Paramount Pictures.

Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture, part II: Case study: A computational theory of social exchange. *Ethology and Sociobiology*, *10*, 51–97.

Cui, J. (2007). QIC program and model selection in GEE analyses. *Stata Journal*, *7*(2), 209.

Dawkins, R. (1976/2006). *The Selfish Gene (30th Anniversary Ed.).* Oxford Paperbacks.

De Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*, 73–88.

Delton, A. W., Krasnow, M. M., Tooby, J., & Cosmides, L. (2011). The evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(32), 13335–13340.

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

Eyal, T., Hoover, G. M., Fujita, K., & Nussbaum, S. (2011). The effect of distance-dependent construals on schema-driven impression formation. *Journal of Experimental Social Psychology*, *47*, 278–281.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, *94*, 334–346.

Fetchenhauer, D., Groothuis, T., & Pradel, J. (2010). Not only states but traits—Humans can identify permanent altruistic dispositions in 20 s. *Evolution and Human Behavior*, *31*, 80–86.

Fischer, I. (2009). Friend or foe: Subjective expected relative similarity as a determinant of cooperation. *Journal of Experimental Psychology. General*, *138*, 341–350.

Fiske, S. T., & Dépret, E. (1996). Control, interdependence and power: Understanding social cognition in its social context. *European Review of Social Psychology*, *7*, 31–61.

Forgas, J. P. (2011). Can negative affect eliminate the power of first impressions? Affective influences on primary and recency effects in impression formation. *Journal of Experimental Social Psychology*, *47*, 425–429.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology & Sociobiology*, *14*, 247–256.

Funder, D. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*, 177–182.

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology*, *69*, 339–348. http://dx.doi.org/10.1016/j.tpb.2005.09.005.

Haselhuhn, M. P., & Wong, E. M. (2011). Bad to the bone: Facial structure predicts unethical behaviour. *Proceedings of the Royal Society B*, *279*, 571–576. http://dx.doi.org/10.1098/rspb.2011.1193.

Hepler, J., & Albarracín, D. (2013). Attitudes without objects: Evidence for a dispositional attitude, its measurement, and its consequences. *Journal of Personality and Social Psychology*, *104*, 1060–1076.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, *11*(2), 193.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Sage.

Kiyonari, T. (2010). Detecting defectors when they have incentives to manipulate their impressions. *Letters on Evolutionary Behavioral Science*, *1*, 19–22.

Klapwijk, A., & Van Lange, P. A. M. (2009). Promoting cooperation and trust in noisy situations: The power of generosity. *Journal of Personality and Social Psychology*, *96*, 83–103.

Little, A. C., Jones, B. C., DeBruine, L. M., & Dunbar, R. I. M. (2013). Accuracy in discrimination of self-reported cooperators using static facial information. *Personality and Individual Differences*, *54*, 507–512.

Manson, J. H., Gervais, M. M., & Kline, M. A. (2013). Defectors cannot be detected during "small talk" with strangers. *PloS One*, *8*, e82531.

Mealy, L. (1995). The sociobiology of sociopathy: An integrated evolutionary model. *The Behavioral and Brain Sciences*, *18*, 523–541. http://dx.doi.org/10.1017/S0140525X00039595.

Nettle, D., Harper, Z., Kidson, A., Stone, R., Penton-Voak, I. S., & Bateson, M. (2013). The watching eyes effect in the dictator game: It's not how much you give, it's being seen to give something. *Evolution and Human Behavior*, *34*, 35–40.

Neuberg, S. L., & Fiske, S. T. (1987). Motivation influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*, 431–444.

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogenous populations. *Nature*, *355*, 250–253.

Nowak, M. A., & Sigmund, K. (1993). A strategy of win-stay, lost-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*, 56–58.

Oda, R., Naganawa, T., Yamauchi, S., Yamagata, N., & Matsumoto-Oda, A. (2009). Altruists are trusted based on non-verbal cues. *Biology Letters*, *5*, 752–754.

Oh, I., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *The Journal of Applied Psychology*, *96*, 762–773.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401–421.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120–125.

Pelham, B. W., & Neter, E. (1995). The effect of motivation of judgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology*, *68*, 581–594.

Puts, D. A., Apicella, C. L., & Cárdenas, R. A. (2011). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society B*, 601–609.

Reed, L. I., Zeglen, K. N., & Schmidt, K. L. (2012). Facial expressions as honest signals of cooperative intent in a one-shot anonymous Prisoner's Dilemma game. *Evolution and Human Behavior*, *33*, 200–209.
Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology*, *30*, 358–367. http://dx.doi.org/10.1016/j.joep.2009.02.002.
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*, 409–426.
Rusbult, C. E., & Van Lange, P. A. M. (2003). Interdependence, interaction, and relationships. *Annual Review of Psychology*, *54*, 351–375. http://dx.doi.org/10.1017/S0140525X00039595.
Sell, A., Bryant, G., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., ... Gurven, M. (2010). Adaptations in humans for assessing physical strength and fighting ability from the voice. *Proceedings of the Royal Society B*(277), 3509–3518.
Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior*, *34*, 317–322.
Sparks, A., Burleigh, T., & Barclay, P. (2015). *Accurate PDG predictions.* [osf.io/82rw6].
Sparks, A., Mishra, S., Rotella, A., & Barclay, P. (in review). Betting your reputation: Public (but not private) Prisoner's Dilemma defection is associated with behavioral risk-taking.
StataCorp (2013). *Stata Statistical Software: Release 13.* College Station, TX: StataCorp LP.
Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*, 349–354.
Sylwester, K., Lyons, M., Buchanan, C., Nettle, D., & Roberts, G. (2012). The role of theory of mind in assessing cooperative intentions. *Personality and Individual Differences*, *52*, 113–117. http://dx.doi.org/10.1016/j.paid.2011.09.005.
Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*, *6*, 659–662. http://dx.doi.org/10.1098/rsbl.2010.0209.
Tinbergen, N. (1963). On aims and methods in ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433.
Tognetti, A., Berticat, C., Raymond, M., & Fauire, C. (2013). Is cooperativeness readable in static facial features? An intercultural approach. *Evolution and Human Behavior*, *34*, 427–432.
Verplaetse, J., & Vanneste, S. (2010). Is cheater/cooperator detection an in-group phenomenon? Some preliminary findings. *Letters of Evolutionary Behavioral Science*, *1*, 10–14.
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*, 260–271.
Vogt, S., Efferson, C., & Fehr, E. (2013). Can we see inside? Predicting strategic behavior given limited information. *Evolution and Human Behavior*, *34*, 258–264. http://dx.doi.org/10.1016/j.evolhumbehav.2013.03.003.
Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*, 592–598.
Wilson, W. (1971). Reciprocation and other techniques for inducing cooperation in the Prisoner's Dilemma game. *Journal of Conflict Resolution*, *15*, 167–196.