7-2020

# Trust, Reciprocity, and Social History: New Pathways of Learning When Max U (own reward) Fails Decisively

Vernon L. Smith

# Trust, Reciprocity, and Social History: New Pathways of Learning When Max U (own reward) Fails Decisively

## Comments

ESI Working Paper 20-28

# TRUST, RECIPROCITY, AND SOCIAL HISTORY

## New Pathways of Learning When Max U (own reward) Fails Decisively

Vernon L. Smith[1]

Chapman University

In 1995, Joyce Berg, John Dickhaut, and Kevin McCabe (hereafter BDM) inaugurated entirely new directions of experimental investigation, business research, and human sociability in their study of trust and trustworthiness (BDM called it trust and reciprocity). Building on research results from the study of ultimatum and dictator games, BDM ignited widespread research interest—Google Scholar indicates 5242 citations through September, 2019. Johnson and Mislin (2011) offer a meta-analysis of BDM using data from 162 replications across 35 countries, and 23,000 subjects, but replications and extensions have continued unabated since these data were assembled. Few experiments have approached that of BDM in launching such extensive further investigation.

What accounts for the incredible scholarly popularity of the BDM protocol and its many derivative studies?

Subjects in the BDM protocol chose unexpectedly high levels of cooperative other-regarding action, under conditions of strict privacy and anonymity that invited self-interested action under a heavy cloak of secrecy. Their findings seemed at odds with dictator games in which similar conditions of anonymity and secrecy had greatly reduced dictator game "generosity." After all, BDM was "merely" a two stage dictator game. The observations also appeared to be at complete odds with the own-regarding actions that had dominated in market experiments beginning a half century earlier. (Chamberlin, 1948; Smith, 1962; see Holt, 2019, pp 1-35 for a historical summary)[2] The BDM investment game challenged the beliefs underlying economic modelling, altered research directions, and ignited a search for understanding—for reconciling disparate bodies of data, each highly replicable and coming from people in the same sampling populations. Can these distinct and contradictory patterns be reconciled in one underlying theory, or are we stuck with a two-regime theoretical justification?

[1] I am grateful to Andreas Ortmann for his careful reading of two earlier drafts of this paper and providing extensive comment, while not absolving myself from responsibility for and errors that remain.

[2] As we shall see below these appearances were not correct; both own-regarding and other-regarding human action are consistent with strictly self-interested preferences in Smith (1759; 1853). The path-breaking work of BDM has helped immeasurably in enabling Smith's classical work to be rehabilitated.

This evaluation begins with the BDM protocol—itself a methodological contribution—and the experimental findings. The question of the replicability and robustness of these unexpected results is addressed next in a summary of two subsequent experimental papers. We follow with a discussion of two attempts to explain qua understand the BDM findings; both, however, have methological deficiencies—Reciprocity and Social Preference explanations. Finally, we offer a brief on Adam Smith's (1759; 1853; hereafter in the text, *Sentiments*) model of human sociability, based on strictly self-interested actors, that culminates in propositions that (1) account for trust game choices, and (2) predict action in new variations on trust game designs that, in the absence of Adam Smith's model, would be neither natural or well-motivated.

## THE BDM PROTOCOL: FIRST RESULTS AND "SOCIAL HISTORY" REPLICATION

### Introduction

Across three sessions BDM recruited 32 pairs of subjects. In each session half the individuals were recruited for room A, and half for room B.[3] Each received ten $1 bills as an upfront payment for showing up on time—an intentional form of earned compensation that belongs to the individual. Each person in room A is free to select from their money payment any number, from 0 to 10 one-dollar bills to be sent to their anonymous and randomly paired counterpart in room B. In route, the sum is tripled before delivery to the person's counterpart in room B. BDM implemented a double-anonymity protocol wherein each pair is anonymous with respect to each other and to any and all third parties including the experimenters. No one can know who sent whom how much money.

This procedure was a departure from commonly practiced protocols and generated some entirely appropriate controversy that we will discuss as part of our reexamination and review.[4] Methodological challenges—what does it mean to test a theory? —are part of daily life in any and all experimental sciences. That meaning emerges out of the personal experiences and conversation of experimentalists whose "knowledge in science is not made but discovered, and as such it claims to establish contact with reality beyond the clues on which it relies….For we live in it as in the garment of our own skin." (Polanyi, 1964, p 64;

[3] When the subjects were recruited, they were told to come to either room A or B, making it credible that there really were two rooms (no deception); a monitor was chosen in each room to carry envelops to the other room, further making it evident that real people were paired with real people and no deception was possible. Also, by not first meeting in one room, there was additional social control in support of the concept of paired "strangers."

[4] Every experimental science depends on an immense body of "experimental knowledge," a specialized form of human capital based on practice and the ongoing evaluation, and re-evaluation, of the state of that knowledge; this is the life blood of any experimental science. (Mayo, 1996; Smith, 2008, chapter 13)

also see Mayo, 1996 on the role of experimental methodology in reducing belief error.)

## The Dictator Game as a Precursor of the Investment Trust Game

The dictator game (DG) evolved from the study of the ultimatum game (UG) as part of explorations designed to better understand the unexpected findings by Guth, Schmittberger and Schwarze (1982) who originated the UG. In this game the Proposer offers to split M one-dollar bills with the Responder, yielding (Proposer payoff, Responder payoff) = (M—X, X). M is commonly $10 or $20. Play then passes to the Responder who either accepts the offer, in which case the imputation is (M—X, X), or rejects the offer, in which case the outcome is (0, 0). If the players are each strictly self-interested and always choosing dominant own payoff outcomes, the predicted equilibrium offer is (M—1, 1), since $1 clearly dominates 0, and the Responder is predicted to accept. On average, Proposers offer about 0.45M, and Responders accept almost all offers. Responders routinely tend to reject infrequent offers of $1 or $2 and even of $3.

The predominance of offers of $4 and $5 led to the *ex post hoc* explanation that people had a strong preference for "fairness" or an equal-outcome division of M. This interpretation was challenged by Forsythe, et al. (1994). They argued that if the results were driven by a strong preference for equal split "fairness" then the results would not be effected by eliminating the Responder's right to veto the Proposer's offer. Hence, they compare UG treatments with and without the responder being allowed to veto the offer, and report mean offers of 0.47M when Responder can veto, and (significantly less) 0.24M when Responder cannot veto (See Camerer, 2003, Tables 2.2 and 2.3, pp 50-55; and 2.4, pp 57-58; he conveniently reports data from all the early studies). The no-veto treatment quickly became known as the DG and took on an experimental life of its own.

Next—in this scenario from UG to DG to BDM—enters Hoffman et al. (1994; hereafter HMSS) who report many treatment variations on the UG and the DG.[5] In particular, although DG offers are significantly less than UG offers reported in Forsythe et al. (1994), HMSS were intrigued and impressed by the fact that dictators are nevertheless giving away twenty-four percent of their endowments. To stress the boundary of these unexpected, but persistent DG results, they

[5] The reader should note that the four papers in this scenario were all published in 1994-1995 as the various authors were all in contact with each other and working from pre-publication drafts of their respective papers.

introduce a "double-blind" procedure to see if DG generosity is materially reduced, or stubbornly resistant, to this treatment protocol.[6]

## Legitimacy of the Double-Blind Treatment Component

The double-blind treatment procedures used by HMSS and by BDM, has been criticized as representing an illegitimate experimenter-demand effect "by too clearly indicating the goals of the experimenter." (Kagel and Roth, 1995, p 303) This constitutes a misunderstanding of the purpose and objectives behind this protocol. BDM and HMSS, in controlling for reputation and other social effects, intentionally sought to invite and encourage strictly self-serving action by making it surveillance-safe and transparent, that it is ok not to send money, and ok to keep any money received. (Smith, 1982, refers to such explorations as boundary experiments) HMSS found that the procedure substantially (as well as significantly) lowered dictator giving (on average from 0.24M to 0.10M, with the percent giving nothing rising from 20% to 60%). Similarly, in BDM, which is merely a two stage DG, we have a doubling of the opportunity to secretly give nothing. If cooperation fails, we have evidence of the power of self-interested motivation—Max-U(own)—to be expressed under the cloak of secrecy as a control for social value and influence. If cooperation persists, we meaningfully expand the range of conditions where the standard "strangers" model fails. In exploring the boundaries of that persistence, the BDM experiment either expands the range of self-interested action, or launches us into explorations of why robust cooperation trumps the temptation to serve private advantage?

We were not to be disappointed in this polarizing stress test, for BDM find that the dramatic effect found by HMSS in reduced DG giving does not carry over to the trust game. Moreover, far from reducing cooperation it is substantially increased. Hence, the interpretation that people in the BDM trust game see it in a completely different way than they see the DG. The trust game is indeed much different than a sequential DG—the tripling of any amount sent implies gains from trust/trustee interaction, a synergy that is absent in DG, and it is this leveraging of the reward stakes that seems to invite a much different experiential response, such as the BDM emphasis on "reciprocity".

The important learning from these experiments is that other-regarding action trumps and robustly survives instructional treatments designed strongly to encourage self-interested action. This powerful finding demonstrates the strength of human sociability, and robustly falsifies the traditional economic and game-

[6] The term "double-blind" is used here in the sense that subject identity is protected (1) between and among the subjects participating in a ("single-blind") experiment, but is also protected (2) between the subject, the experimenter, and any other potential observer.

theoretic modeling based on self-interested action. Methodologies that preclude such boundary experiments because of unexamined hypothetical experimenter-demand effects fail to afford opportunities for identifying the edges of validity of new and unexpected findings—or establish that there are no edges.

Beyond the moral imperative that subjects be treated with respect, dignity, payment for their earnest service, and strict adherence to the principle that the experimenter shalt not bear false witness (don't lie to the subjects or anyone else), experimental methods must be free and open to new means of learning.[7]

For a non-cooperative equilibrium of the game, sufficient conditions are that (1) all are strictly self-interested, (2) this is common knowledge, and (3) each chooses to maximize their own utilitarian outcome. It follows that individuals in room A are predicted to send nothing; those in room B return nothing if any money is sent. This prediction does badly even under the supposed favorable condition where no one, not even the experimenter, can know the identity of any individual actor. The primary implication was that it was a good idea for researchers to seek better ways of thinking about two-person connectedness. Massive prediction failure ought to motivate re-evaluation and new learning on a similar scale. As we aim to show here, that failure was not newsworthy within the framework of *Sentiments*, published over two hundred fifty years before BDM. Moreover, this classical contribution to social psychology expands the range of new experimental designs and prediction. Our only excuse was that we were either ignorant of *Sentiments* or did not understand its message for embracing BDM and their aftermath.

## BDM Results

On average, individuals in room A sent $5.16, but the average amount returned was $4.66; two subjects sent zero and five sent $10; twenty-eight of thirty-two people in room A sent more than $1. Since sending money yielded an overall loss, senders' beliefs in the game appeared mistaken. Hence, BDM followed with their "social history" treatment in which new subjects, informed by a summary of the first experiment results, could adapt and correct their beliefs. The social history summary treatment reported the number of subjects sending each amount from $1.00 to $10.00, the average amount returned and average profit of the sender; the only net profitable amounts sent were $5 and $10.

However, the BDM conjecture—that subjects would correct their belief error—was not supported: Now, the average amount sent increases slightly to $5.36, but

[7] Psychological research traditions are not self-bound by any such moral imperative although the latter is rooted in the experimental research of the psychologist, Sydney Siegel, one of the early founders of experimental economics. (Smith, 2017)

the average returned increases to $6.46. The baseline norm, "be generous in sending" does not unravel in the social history treatment, while the trustworthy norm, "be generous in rewarding trust" is enhanced; 3 of 28 send nothing, half (14) of those in room A send $5 or $10, with only one recipient in room B keeping all that was sent.

## SKEPTICS CHALLENGE THE BDM RESULTS, FIND ONLY CONFIRMING EVIDENCE, AND SIGNIFICANTLY EXTEND THE DOMAIN OF BDM APPLICABILITY

In the large subsequent literature, two studies, both by scholars skeptical of the robustness of these remarkable findings, continued to observe results inconsistent with Max-U (own payoff) rationality in experiments motivated by the BDM findings. In the first, Andreas Ortmann, John Fitzgerald, and Carl Boeing (2000; hereafter OFB) comprehensively replicated and reexamined the BDM experiments adding new treatments that they hypothesized would change the findings.

They study five treatments:

> First, a baseline "No History" treatment which replicated the original BDM experiments.
>
> Second, a replication of the BDM "Social History" treatment by presenting the results from the first baseline treatment, precisely as did BDM, by simply presenting the values of previous investments and returns in a table.[8]
>
> Third, a "Social History" treatment framing the previous experimental results in terms of the portion that room B participants returned to A, clearly showing room A participants that the returns were not equitable.
>
> Fourth, a second "baseline No History treatment" characterized, however, by several key modifications.

[8] Experimental economists often encounter journal editorial resistance to publishing "mere replications," especially in the leading journals that seek to pioneer new and innovative work, while personally placing high priority on the scientific importance of replications. The solution to this challenge for many has taken the form of combining replication with new treatment variations on the original motivating study. (Smith, 1994, p 128)

Specifically, OFB included a questionnaire for the room A participants which they were to complete prior to their decision. Specifically, this questionnaire had two purposes. First, it was to ensure that room A subjects understood the design and considered their decisions carefully before making them. Second, it was to help subjects determine how much to invest by encouraging them to think carefully (prompt strategic reasoning)[9] about the consequences of their decisions before they made them. The subjects were asked the following four questions:

1. How much money do you think you will send?

2. How much money will your room B counterpart receive if you send this much?

3. How much money do you think will be returned to you?

4. How much money would you return if you were in room B?

The authors hypothesized that by changing the presentation format in Treatment Three, and prompting strategic reasoning as in Treatment Four it would cause significant drops in both the amounts sent to room B from A, and consequently the amounts returned to room A from B. As we shall see, however, these modifications had no effect.

Fifth, Treatment Five and Five R each applied the combined modifications of Treatments Three and Four, with Five R a replication of Treatment Five designed to further test the statistical significance finding in Treatment Five. "When Berg et al. used their social history treatment, contributions did not change much. The median remained at $5 and only 3 out of 28 subjects sent zero…none of our treatments led to significantly different results. This means that neither the way information is presented (BDM presentation, OFB presentation) nor strategic reasoning prompts (the questionnaire) matter statistically to our subject pool. In fact, as the results for treatment Five and its replication show, nor do these two modifications to the original design matter jointly if we pool the data." (Ortmann, et al., 2000, pp 85-6) In their abstract OFB express the unexpectedness of their findings: "To our

[9] Hoffman, McCabe, and Smith (2000) used instructions to "prompt" subjects to think about what their paired counterpart would do in the UG, but it simply focused them on the prospect that Responder might veto the proposal (not on the strategic idea that $1 was better than nothing, which ought to be acceptable), and their offers became more generous.

> surprise, none of our various treatments led to a reduction in the amount invested."

A second skeptical examination of BDM substantially alters the BDM framework, while polarizing the potential outcome depending upon how the subjects' respond. (McCabe and Smith, 2000; hereafter MS) Senders in the BDM game can choose any of eleven amounts from zero to ten dollars to send to their counterpart; if X dollars are sent ($0 \leq X \leq 10$), receivers can return any amounts from zero up to and including 3X dollars.

MS dichotomize the choices for each of the players so that each can choose only two starkly contrasting actions. The MS payoffs are motivated by BDM, but the BDM context—two people matched in a sending, tripling and returning money relationship—is stripped out of the MS narrative. One of only two actions by each of the players provides the largest self-interested outcome, the other a "fair" equal split of the joint gains. Thus, Player 1 can choose to send nothing—the self-interested "best" outcome, yielding the payoff: (Player 1= \$10; Player 2 = \$10). Or, alternatively, Player 1 chooses to send the entire ten dollars, which is tripled to thirty dollars. Player 2 can only respond with either of two actions: split the thirty dollars equally with Player 1, yielding the payoff (Player 1 = \$15; Player 2 = \$25 = \$10 + \$15), or take all the money resulting in the payoff (Player 1= \$0; Player 2 = \$40 =\$10 + \$ 30). In the first option, Player 1 receives a 50 percent larger amount than if nothing is sent. However, Player 2 receives an increase that is 150 percent larger than if Player 1 sent nothing. Clearly, Player 2 is made strictly and asymmetrically better off. In the second option, however, Player 1 can end with nothing.

By removing all context and starkly focusing on the hazards to any Player 1 who passes to Player 2, MS intentionally probe the boundary of validity of the original BDM results hypothesizing that this will discourage cooperative play.

Remarkably and surprisingly the frequencies with which subjects offered and accepted the cooperative chose actions that were trusting and trustworthy were high enough that, on average, the earnings of both players increased relative to the self-interested equilibrium payoffs. Of twenty-four undergraduates, Twelve Player 1s' (50 percent) passed play to their counterpart Player 2 and nine of the twelve responded cooperatively (75 %), only three taking all the money (25 %).

Also MS report the data for twenty-eight graduate PhD students who play the same game twice, knowing that they will retain the same pairing.

> First Play: Twenty-one (75%) pass to their counterpart Player 2, of which 16 (76%) choose to cooperate, and only five ((24%) take all the money. Second

Play: Fourteen (50%) of Player 1s pass to their Player 2s of which nine (64%) cooperate, and only five (36%) defect.

Far from failing this simplified, alleged self-interest promoting test, the MS subjects were even more other-regarding in rejecting non-cooperative action and tended to earn more money than in the BDM game. Thus, did MS introduce a simplified design that enabled subjects to better coordinate actions designed to achieve their cooperative intentions.

How are we to explain these findings, so bizarre by accustomed economic standards?

## RECIPROCITY AND SOCIAL PREFERNCE EXPLANATIONS

### Reciprocity Explanation

As indicated in their title, reciprocity as an explanation of trusting and trustworthy behavior was very much part of how BDM thought about their discovery. Their title was not "Trust, Trustworthiness and Social History;" trustworthiness was identified with reciprocity in a non-market exchange.

By sending money the first mover in the BDM game is offering to cooperate; by returning money the second mover is accepting the offer in an exchange; "reciprocity" is simply a word for describing those two actions.[10] How can a description of what transpires be an explanation of why we observe the behavior? The argument is circular. (Smith and Wilson, 2019)[11]

### Social preference Explanation

Social preference theory had its origins in the proposition that other-regarding action is a direct consequence of other-regarding preference or utility functions. (Fehr and Fischbacher, 2002) What seemed to fail in BDM was the neoclassical assumption that people cared only about their own payoff. Thus—for generations of economists brought up on utility as the cause of all action—it was natural to

[10] Many of us fell into that pattern. Smith (1998) elevates reciprocity to the key to an understanding of the connection between Adam Smith's (1759, 1776) two books. This is neither wrong, nor very deep, as an explanation of the results in BDM and the large subsequent literature demonstrating the robustness of their results.

[11] The same circular reasoning has been accepted to explain the strong tendency toward equal-split outcomes in the UG. The preference for "fair" outcomes is said to be the explanation and the cause. However, equal-split "fair" outcomes also constitute what is observed, which itself cannot serve as an explanation and a cause. It is correct to state that "[t]he (UG) data falsify the assumption that players maximize their own payoffs as clearly as experimental data can" which recognizes the contradiction between prediction and observation. But it is leading and questionable to add: "Since the equilibria are so simple to compute,… the ultimatum game is a crisp way to measure social preferences rather than a deep test of strategic thinking." (Camerer, 2005, p 43) This inference follows only if all action is a direct consequence of preference, but it takes only one non-preference-based model of action to negate it. *Sentiments* performs that function.

explain the new findings with a utility of the form U(own payoff, other payoff) in which actions reflected the actors concern about other as well as own payoff in the trust game. But if social preference is to be the predictor of action, we need to know the form of the Utility function in advance. For example, suppose Player 1 has social presences such that they want to transfer money to Player 2. Suppose Player 1 passes to player 2, and Player 1 defects; that mat be a better outcome than if Player 2 chooses the cooperative outcome. Social preference theory cannot assume that defection hurts Player 1. If Player 1 is given the opportunity to punish defection by Player 2, some do, but most do not. The methodology is that of retrofitting utility to actions discovered empirically, then looking for "epi-cycle" parameters that accommodate the observation.[12]

Attempts to solve the puzzle—How do we explain and reconcile other-regarding action in BDM and its extensions with the self-interested acts of the same individuals in markets and other contexts—led to decades of experimental explorations. The puzzle also contributed to the discovery that *Sentiments* provided an independent means of interpreting and modelling action, wherein all individuals are strictly self-interested in preference, but follow rules that are other regrading. That development and its history in recounted in (Smith and Wilson, 2019, pp xiii-xx.)

## HUMANOMICS OF TRUST AND TRUSTWORTHINESS: WHY STRICTLY SELF-INTERESTED ACTORS CAN MAKE GOOD NEIGHBORS

The subtitle of *Sentiments* states succinctly its message, "An Essay towards an Analysis of the Principles by which Men Naturally Judge the Conduct and Character, First of their neighbors and then of themselves." (Smith, 1759; 1853, title page)

In *Sentiments* we learn an alternative to reciprocity and social preference explanations of cooperation in our lives and in our experimental trust game data. Cooperation stems from human sociability and is governed by our rule-following conduct; the very word "conduct" suggests a pattern of proper manners emanating from our judgement of each other. Our actions are other-regarding as well as own-regarding. Moreover, these actions are not direct consequences of our preferences, which are strictly self-interested, and not in any way conflictual with our actions either in markets or in our social world. To understand *Sentiments,* as economists who study behavior, we must distinguish between our self-interested preferences,

[12] Added to this procedure is the fact that "there is a professional tendency to view utility explanations as final—once a result is deemed due to utility the conversation stops, implying that there is nothing left to explain or test." Hoffman, McCabe, and Smith (2008, p 415).

and our actions, which need not take the form of acting in accordance with this principle; that is, action *need not have the form*: Action if and only if Max U(own).

> "Though it may be true, therefore, that every individual, in his own breast, naturally prefers himself to all mankind, yet he dares not look mankind in the face, and avow that he acts according to this principle. He feels that in this preference they can never go along with him, and that how natural soever it may be to him, it must always appear excessive and extravagant to them. When be views himself in the light in which he is conscious that others will view him he sees that to them be is but one of the multitude, in no respect better than any other in it. If he would act so as that the impartial spectator[13] may enter into the principles of his conduct, which is what of all things he has the greatest desire to do, he must upon this, as upon all other occasions, humble the arrogance of his self-love, and bring it down to something which other men can go along with." (Smith, 1759; 1853, p 120)

In fact, common knowledge that we are all self-interested is a necessary part of how we automatically know that a context-specific action is beneficial or hurtful to another and are thus able to implement the rules we follow in interacting with our neighbors. (Smith, 1759; 1853, p 112)

What are the circumstances of life that determine this rule-following means of disciplining our actions?

Smith observes that about the time we start to school and "mix with equals" we find that our play-fellows do not show the "indulgent partiality" of our parents in tolerating our expressions of anger; they use punishment to express their displeasure with our hurtful actions towards them, and find ways to reward our beneficent actions toward them. Thus, do we enter "the great school of self-command" in "which the practice of the longest life is very seldom sufficient to bring to complete perfection." (Smith, 1759; 1853, p 204, 206).[14]

---

[13] The "impartial spectator" in *Sentiments* is a metaphor for the means by which we learn to judge our own actions in the light or their impact on others—based on our sympathetic fellow-feeling toward others—and to choose in a manner that is properly other-regarding, and not only self-regarding. "We endeavour to examine our own conduct as we imagine any other fair and impartial spectator would examine it. If, upon placing ourselves in his situation, we thoroughly enter into all the passions and motives which influenced it, we approve of it, by sympathy with the approbation of this supposed equitable judge. If otherwise, we enter into his disapprobation, and condemn it." (Smith, 1759; 1853, p 162)

[14] For Adam Smith self-command (or self-government) is the omnipresent gatekeeper of virtue and many situations may allow the "voice of human weakness" to undermine this self-command. (Smith, 1759; 1853, p 29) Consequently, higher stakes may tempt a decrease in trust/trustworthiness. But meta-analysis finds this not to be the case; generally the results documented by BDM and OFB were further validated, indicating bedrock support for trust and trustworthiness as expressions of human beneficence even under conditions of anonymity. (Johnson and Mislin, 2011, p 874)

The first proposition implied by the analysis in *Sentiments*—that drives the experimental observations in BDM, OFB and MS—is the following: "Actions of a beneficent tendency, which proceed from proper motives, seem alone to require a reward; because such alone are the approved objects of gratitude, or excite the sympathetic gratitude of the spectator." (Smith, 1759; 1853, p 112)

By proper motives, Smith is referring to intentionality: I do something good for you because I wanted to do something good for you. And he is here asserting the strong directive that such action alone induces in us a compulsion to reward the action, and this is because of the spontaneous fellow-feeling of gratitude that we experience. Furthermore, this is not just what the individual who is the target of the action experiences and responds to, but it also commands the agreement of every indifferent spectator. That is, every third-party observer easily agrees, or fellow feels, with the target of the action.[15] In modern language we would call it a "social norm," and here Smith is articulating a theory with specific a priori predictions of its action consequences.

Regrettably, our abysmal ignorance of *Sentiments* prevented us, in the 1990s, from hypothesizing the implied behavior before we observed it.[16]

Later, Smith uses this proposition to derive "reciprocity" (logically, as an implication) although he does not use that word in that context. Rather he asks who above all we should be kind to. Those who have been kind to us. Thus "Kindness is the parent of kindness." (Smith, 1759; 1853, p 331) This is called the Principle of Beneficent Reciprocity by Smith and Wilson (2014, p 16). Hence, *Sentiments* provides the underlying explanation for reciprocity, and is not circular as in the original trust game literature.

Let us now think through the application of Smith's first beneficence proposition to BDM (OFB) and MS. The first mover is clearly under no obligation to send any

[15] Note the critical role for human fellow-feeling in *Sentiments* and the following corroborating results of trust game meta-analysis showing that it is important for senders to know that the receiver is a real person: "We also find that playing with a real person is associated with significantly more sent. While researchers sometimes employ simulated confederates to play the role of the receiver in the trust game, these studies rarely use a manipulation check to confirm that the experimenter's attempts to deceive the participants have been successful. Our findings suggest that participants in such trust game experiments may in fact not all believe, as the experimenters wish them to, that they are playing with real counterparts. This could be due to flaws in the experimental procedures employed or even early participants informing later participants of the deception." (Johnson and Misilin, p. 875). Recall that the BDM protocol made it credible that real people were matched with each other across two rooms.

[16] Harking back to those exciting discovery years, I should note that, even if we were given then our current understanding of *Sentiments*, there was reason aplenty for being skeptical that the proposition would be predictive—it might easily fail because of the deep cloak of secrecy implemented by BDM and OFB (but not MS). Hence, the BDM launch would have been no less path-breaking, by virtue of extending *Sentiments* significantly beyond its original presumed domain.

money, nor for the recipient to return money if any is sent. Moreover, the first mover is clearly at risk in getting nothing back. Knowing this, the recipient of any money sent can only infer that money was sent intentionally—an action that obviously and unambiguously benefits the recipient. Smith's proposition predicts that the recipient feels gratitude and is motivated to reward the action by returning some money. How much? Well, more in positive relation to the benefit and gratitude felt--"the greater exertions of that virtue appear to deserve the highest reward. By being productive of the greatest good, they are the natural and approved objects of the liveliest gratitude." (Smith, 1759; 1853, p 117) Smith's theory, culminating in this proposition, predicts the tendency expressed in the data of BDM and OFB, but found more prominently in MS.

Over and over, *Sentiments* stresses the importance of context, and across the BDM, OFC and MS treatments, context is being varied. (Also see Hoffman, McCabe and Smith, 2000) The data show clearly that context matters; the three studies all confirm *Sentiments* relative to utilitarian action in the self-interest, but actions in the MS game are most strongly consistent with Smith's first proposition on beneficence. In MS, Player 1 sends a strong and unambiguous signal of benefit, or none, and this dichotomy, we can conjecture, accounts for the proposition's greater consistency with MS.

It should be noticed how natural it is to think about how *Sentiments* applies to the actions of the subjects. If the model fails, we know where to look for the cause. Gratitude may not be felt as with sociopathic-like tendencies. Or, gratitude might be felt, but it is insufficient to overcome the temptation to defect, suggesting further experiments that vary payoffs. In contrast, if the traditional dominant-strategy, self-interested choice model fails, it says nothing about what to do next. We are left with no guidelines as to the next scientific step.

I will close by stating a second proposition in *Sentiments* on beneficence and apply it to a new trust game in the MS framework. "Beneficence is always free, it cannot be extorted by force, the mere want of it exposes to no punishment; because the mere want of beneficence tends to do no real positive evil. It may disappoint of the good which might reasonably have been expected, and upon that account it may justly excite dislike and disapprobation: it cannot, however, provoke any resentment which mankind will go along with." (Smith, 1759; 1853, p 112)

Two recent experimental studies are directly motivated by this important proposition in which the benefit-reward calculus governed by the first proposition is said to be voided if there is any threat of coercion, thus predictively bounding the domain of conditions over which the first applies.

The first study observes that the literature on the UG is replete with evidence that Responders feel and express much anger which in turn explains the pattern of rejections across UG treatments. Since UG participation is always involuntarily assigned by the experimenter, this suggests a treatment effect emanating from the influence of coercion—the implicit threat of veto by the Responder. In new UG experiments, the Responders move first, choosing to either exit the game along with their paired Proposer, each receiving $1, or voluntarily entering the UG stage by passing to the Proposer who choose between an equal split of $24, or the equilibrium outcome (Player 1 = $2. Player 2 = $22). Ninety-four percent of Responders signal willingness-to-play by passing play to the proposer. Remarkably, forty percent of the proposers offer the equilibrium option, and sixty-one percent accept—the highest known rate of equilibrium play, and of acceptance recorded in the extensive UG literature (Smith and Wilson, 2018; 2019, pp 135-141).[17]

The second study directly tests Adam Smith's second beneficent proposition in a new trust game, with the same extensive form structure as in MS, but different payoffs. Player 1's have two options: (1) pass to Player 2, who chooses the equilibrium ($12, $12), or ($10, $10); (2) pass to Player 2, who chooses either (Payoff 1, Payoff 2) = ($$18, $30), or ($6, $42). If Player 1 decides to not cooperate and choose the equilibrium option (1), and if Player 2 wishes to punish Player I at a cost, we have the outcome ($10, $10). No Player 2 so chooses: Of thirty-eight pairs in this game, twenty-three Player 1's select option (1), but none choose to punish the action.[18]

## CONCLUSIONS

That "Trust, Reciprocity and Social History" was a landmark paper in the history of experimental economics is indicated by several measures of academic impact: Citations, of course, but more precisely the results were unexpected, surprising and continue to inspire new trust game experiments; the results were replicable and robust; and they defined, along with the ultimatum game, the canonical structure and protocol for using trust games to examine human sociality.

However, with the discovery that the results are consistent with key propositions in *Sentiments*—the lesser known first book written by the founder of economics,

[17] Of course the Proposer may or may not also feel anger; allowing the Proposer to voluntarily choose whether or not to enter the game may further impact their joint outcome.

[18] Yet in sharp contrast, Justice Propositions in *Sentiments* predict that if Player 1 offers cooperation, and Player 2 defects, then—given a costly option to punish the defection—Player 1s will so choose, which indeed they do. (Smith and Wilson, 2019, pp 152-153) Hence, Adam Smith states general conditions that predict when people will use punishment strategies, and when they will not.

Adam Smith—I believe the paper became an important part of demonstrating the relevance of that monumental work for contemporary economics and the moral foundations of the human career.

## REFERENCES


Berg, Joyce, John Dickhaut, and Kevin A. McCabe, 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10, 122-142.

Camerer, Colin, 2003. *Behavioral Game Theory*. Princeton: Princeton University Press.

Chamberlin, Edward, 1948. An Experimental Imperfect Market. *Journal of Political Economy* 56(2), 95-108.

Fehr, Ernst, and Urs Fischbacher, 2002. Why Social Preferences Matter: The Impact of Non-selfish Motives on Competition, Cooperation and Incentives. *Economic Journal* 112, C1-C33.

Forsythe, Robert, Joel E. Horowitz, N.E. Savin, and Martin Sefton, 1994. Fairness in Simple Bargaining Experiments. *Games and Economic Behavior* 6 (3), 347-369

Guth, Werner, Rolf Schmittberger, and Bernd Schwarze, 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3(4), 367-388.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith, 1994. Preferences, Property Rights and Anonymity in Bargaining Games. *Games and Economic Behavior* 7(3), 346–380.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith, 2000. The Impact of Exchange Context on the Activation of Equity in Ultimatum Games. *Experimental Economics* 3(1), 5–9.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith, 2008. Reciprocity in Ultimatum and Dictator Games: An Introduction. *Handbook of Experimental Economics Results.* Edited by C. Plott and V. Smith. North Holland: Elsevier.

Holt, Charles A., 2019. *Markets, Games, and Strategic Behavior An Introduction to Experimental Economics.* Princeton: Princeton University Press.

Johnson, Noel D. and Alexandra A. Mislin, 2011. Trust games: A meta-analysis, *Journal of Economic Psychology* 32, 865–889. .doi:10.1016/j.joep.2011.05.007

Kagel, John H. and Alvin E. Roth,1995. *Handbook of Experimental Economics.* Princeton: Princeton University Press.

Mayo, Deborah, 1996. *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.

McCabe, Kevin A. and Vernon L. Smith, 2000. A Comparison of Naïve and Sophisticated Subject Behavior with Game Theoretic Predictions. *Proceedings of the National Academy of Sciences* 97 (7), 3777-3781.

Ortmann, Andreas, John Fitzgerald, and Carl Boeing, 2000. Trust, Reciprocity, and Social History: A Re-examination. *Experimental Economics* 3 (1), 81-100.

Polanti, Michael (1962) *Personal Knowledge*. Chicago: University of Chicago Press.

Smith, Adam (1759; 1853) *The Theory of Moral Sentiments; or, An Essay Towards an Analysis of the Principles by which Men Naturally Judge the Conduct and Character, First of their neighbors and then of Themselves. With a Biographical Critical Memoir of the Author, by Dugald Stewart.* London: Henry G. Bohm, Second Edition.

Smith, Vernon L., 1962. An Experimental Study of Competitive Market Behavior. *Journal of Political Economy* 70(2), 111-137.

Smith, Vernon L., 1982. Microeconomic Systems as an Experiential Science. *The American Economic Review* 72(5), 923-955.

Smith, Vernon L. (1994) "Economics in the Laboratory." *Journal of Economic Perspectives* 8 (1): pp 113–131.

Smith, Vernon L., 1998. The Two Faces of Adam Smith, Southern Economic Association Distinguished Guest Lecture. *Southern Economic Journal* 65 (1), 1-19.

Smith, Vernon L., 2008. *Rationality in Economics Constructionist and Ecological Forms*. Cambridge: Cambridge University Press.

Smith, Vernon L., 2017. Tribute to Sidney Siegel (1916-1961) A Founder of Experimental Economics. *Southern Economic Journal* 83 (3), 664 - 667

DOI: 10.1002/soej.12196

Smith, Vernon L. and Bart J. Wilson, 2014. Fair and Impartial Spectators in Experimental Economic Behavior. *Review of Behavioral Economics* 1 (1–2), 1-26.

http://dx.doi.org/10.1561/105.00000001

Smith, Vernon L. and Bart J. Wilson, 2018. Equilibrium play in voluntary ultimatum games: Beneficence cannot be extorted. *Games and Economic Behavior* 109(C), 452-464.

Smith, Vernon L. and Bart J. Wilson (2019) *HUMANOMICS Moral Sentiments and the Wealth of Nations for the Twenty-First Century.* Cambridge: Cambridge University Press.