Social Preferences? Google Answers![1]

Tobias Regner
Max Planck Institute of Economics, Jena
regner@econ.mpg.de

February 2009
JEL classifications: C24, C70, C93, D82, L86
keywords: social preferences, reciprocity, moral hazard, reputation, Internet, psychological game theory

Abstract:

We analyse pricing, effort and tipping decisions in the online service 'Google Answers'. While users set a price for the answer to their question ex ante, they can additionally give a tip to the researcher ex post. The obtained data set is analysed and compared to the results of similar laboratory experiments, namely Fehr, Gächter and Kirchsteiger (1997) and Gächter and Falk (2002). Reciprocal theories of social preferences pioneered by Rabin (1993) and extended by Dufwenberg and Kirchsteiger (2004) are useful to explain the observed pattern of behaviour.

In line with the related experimental literature we conclude that an open contracts design encourages people to tip. We find evidence that this is motivated by reciprocity, but also by reputation concerns among frequent users. Moreover, researchers seem to adjust their effort based on the user's previous tipping behaviour. An efficient sorting takes place when a sufficient tip history is available. Users known for tipping in the past receive higher effort answers, while users with an established reputation for non-tipping tend to get low effort answers.

[1] I am grateful to Maija Halonen-Akatwijuka and Sebastien Mitraille for valuable discussions and to seminar participants at the University of Bristol, the Royal Economic Society Annual Congress 2005, the World Congress of the Econometric Society 2005, the Max-Planck Institute Jena summer school 2005 and the Verein für Socialpolitik Congress 2007 - in particular to David Winter, Jürgen Bracht, Osiris Parcero, Klaus Schmidt and Matthias Wibral - for their comments.

1 Introduction

While other-regarding behaviour of individuals has been found in numerous lab experiments, it is not too clear yet what the precise drivers of socially-minded behaviour are and whether they also pertain in real-life environments.

The experimental evidence of individuals who consistently make voluntary payments has been explained by theories that take the psychological underpinnings of economic behaviour better into account, namely social preferences.[2] However, the external validity of the lab results is far less studied and merits more attention. Can we observe the behaviour found in the lab as well in real-life contexts and what are the underlying motivations of the occurring voluntary payments?

We collected field data about the pricing and tipping behaviour of "Google Answers" users in order to shed more light on these aspects. In this online service (a sub-service of Google) users can post questions and set a fixed price for the answer. They can also give a tip to the researcher who answered the question. Our data set covers all questions asked at Google Answers. The service started in April 2002 and ended in December 2006. The data set contains 146,656 questions, 57,654 have been answered. The average price for an answer is more than $20. Google Answers researchers (later GARs) may best be described as freelancers.

The paper's goal is to analyse the pricing, effort and tipping decisions in this non-laboratory test-bed in order to validate the results of related lab experiments. In particular, we focus on the underlying cause for the voluntary payments and the effects of such a design on effort levels and efficiency. We discuss three possible motivations for the tipping of users and test empirically to what extent they drive the behaviour of Google Answers users. Tipping could be to conform to a social norm as it is the case in restaurants, for instance. Users may decide to tip out of strategic considerations in order to build up a good reputation. Finally, social preferences could motivate users to leave a tip.

Social dilemmas have been analysed in numerous lab experiments. The Google Answers environment resembles a gift-exchange game in a labour market setting. It is particularly similar to Fehr, Gächter and Kirchsteiger (1997) who study labour relations between firms and workers. When mutual opportunities to reciprocate are given (firms can reward or punish the worker ex post), higher effort levels than under stricter contract options are reached. They also find a significant positive correlation between workers' effort and the firms' reaction (reward or punishment). Based on Rabin (1993) they explain the observed behaviour with reciprocity concerns. We follow this approach also taking into account the theory of sequential reciprocity of Dufwenberg and Kirchsteiger (2004). Besides reciprocity frequent users may also be motivated by reputation to leave a tip in Google Answers. Gächter and Falk (2002) conducted experiments about interaction effects between reciprocity and reputation and we refer to them in our analysis.

[2] See Camerer (2003) and Fehr and Schmidt (2003).

Our real-life findings confirm the experimental results: i) about 23% of all answers have been tipped, ii) even single users tip (almost 15% of 21,512 single transactions), iii) reputation matters as the more questions users ask over time the more likely they are to tip and iv) tipping seems to pay off. Our data confirm that GARs take the past tipping behaviour of users into account and put more effort into the answer, if the user has frequently tipped before. The higher effort increases the benefit of the user and the researcher gets adequately compensated for the extra effort via the tip. In addition, we gain insight about the adoption process of tipping and quantify who in the sample population made use of the option to tip.

Other studies of Google Answers exist, but both focus on researchers. Edelman (2004) analyses labour market aspects like researchers' experience, on-the-job training and specialisation. Rafaeli et al. (2007) focus on the social incentives for researchers to work on an answer. Instead, we analyse the data from both researcher and user perspective. In addition, we use all data from Google Answers in contrast to previous studies. Two features make the complete data set particularly compelling. First, the service started without the possibility of leaving a tip. This option was only introduced six months after the start or roughly 10% into the data. It provides an opportunity to analyse the adoption process of tipping. Second, Google Answers closed in 2006. This was announced briefly before no more new questions were allowed and we study the effect of this news on tipping behaviour.

In the following section we describe the pitch of our field study - the online service Google Answers. Section 3 presents the related experimental and theoretical literature. Section 4 describes our data set, while section 5 analyses it. Section 6 concludes.

2 The Online Service Google Answers

Google is the most popular search engine and an essential tool to find information online. In addition to its standard search tool there is "Google Answers" as sometimes even experienced Internet users need help finding exactly the answer they want to a question.[3] The service Google Answers (www.answers.google.com) offers assistance from researchers with expertise in online searching.

Google Answers users ask questions and Google Answers researchers (GARs henceforth) try to answer them in return for a fixed price and a possible tip. After registering with the service users can post a question to Google Answers and specify how much they are willing to pay for an answer. Users can price their question anywhere between \$2 and \$200. In addition a non-refundable listing fee of \$0.50 applies for each question. There is a pool of roughly 500 GARs who have the possibility to answer. Once one of them decides to search for an answer, a question will get 'locked' (for 4 hours if the price is below \$100, for 8 hours if above). This means a question is actively worked on by

[3] Users might also have low Internet skills or simply no time to look for a thorough answer themselves.

a GAR and no other GAR can answer it in that time. The GAR will try to obtain the requested information and will post his answer back to the service. Users are only charged for their question when an answer is given. If the answer received is not satisfying, the user can first ask for additional research through an "answer clarification" request. If still unsatisfied, users can request to have the question reposted for a new answer or apply for a refund.[4] When the answer is completed, they can also rate the quality of the answer. The average rating of a GAR is easily accessible and has an effect on the standing of the GAR towards users and their employer Google. Finally, users can give a tip to the GAR who answered. The tip goes fully to the GAR in contrast to the price of a question where Google takes a 25% cut. If answering the question is not attractive to any GAR out of the pool, it will expire after 30 days.

According to Google all GARs are tested to make sure that they are expert searchers with excellent communication skills. Some of them also have expertise in a particular field. Additionally, answers are edited by Google to ensure quality. GARs are independent contractors and for only a few of them Google Answers is the main job.

Any question that can be answered with words or numbers can get posted. Many users are looking for a specific piece of information like "How much tea was sold in China last year?", "In which San Francisco club did I see the Chemical Brothers play in 1995/96?" or "Race results from Belmont Park 5/24/1990. Who won the 8th & 9th race? And the daily double?". If the answer to the request is online, chances are pretty good that it will be found by the GARs. Moreover, complex questions are posted where background information is demanded and further links are expected. Examples are "How to get information about life in London during the late 1970's: films, television, plays, home decor, music, restaurants, political events, etc." or "Mutual perceptions of Europe and Asia via portraits". Also a number of questions are about marketing or business strategies. Questions are grouped into several categories as explained later.

Naturally, detailed questions regarding financial, medical or legal advice are excluded from Google Answers as is anything related to illegal activities.

3 Related Literature

A great number of experiments studies behaviour in social dilemma games. We particularly refer to Fehr, Gächter and Kirchsteiger (1997), henceforth FGK, and Gächter and Falk (2002).

FGK analyse a simple labour market with firms, workers and excess supply of workers. Three different contracts are simulated in experiments. While contract terms were exogenously enforced in the first treatment, workers were able to reciprocate in the second and both firms and workers were able to reciprocate in the third treatment. Effort levels of workers were significantly higher in the

[4] However, this is very rare. Only in 0.03% of all answers a refund was granted and the price was returned.

last (strong reciprocity) treatment and a contract that gives the opportunity for mutual reciprocity was found to improve efficiency.

Gächter and Falk (2002) study the interaction effects of reciprocity and repeated game incentives. A gift-exchange game between firms and workers was played in a one-shot and a repeated game treatment. Correlation between wage and effort in both treatments confirms reciprocal motivations. Higher effort levels in the repeated game treatment confirm the positive impact of reciprocal concerns.

3.1 Reciprocity

The set up in FGK consists of two stages - a third one is added in their strong reciprocity treatment. First, firms announce the details of their contract (wage, desired effort, the possible fine for shirking). Then, workers choose an offer they like and their effort level. Shirking, e.g. low effort levels, is verifiable only by chance. Firms' profits depend on the effort. In the final stage firms can reward or punish their workers. Equilibrium effort levels are determined by the offered wage and the amount and likelihood of the fine. If firms and workers are purely selfish, the third stage will not have any impact on equilibrium behaviour as it is costly for firms to reward or punish. Still, FGK found that firms often reciprocated. There was also a significant correlation between workers' effort and the firms' reaction (reward or punishment). Effort levels and profits for workers and firms were higher when firms had the opportunity to reward or punish.

The strategic structure of the Google Answers environment is very similar. Users post a question and set a price. GARs "compete" for the right to answer. One GAR answers the question and posts it back. The value of the answer depends on the effort of the GAR, which is not verifiable. The user's value of the labour relation depends on the GAR's effort and is therefore subject to moral hazard. Users can reject answers based on their quality. A rejection and a subsequent refund can be seen as a fine for the GAR, because such an incident affects the GAR's standing within Google Answers.

FGK explain the observed behaviour in their experiments by taking reciprocity motives into account. They relate to the seminal work of Rabin (1993). Concerns for reciprocity seem to play a significant role for the relationship between users and GARs in the context of Google Answers and we adopt this approach. In addition we consider Dufwenberg and Kirchsteiger (2004) as their theory of sequential reciprocity is better suited for the sequential character of Google Answers. It is important to stress that this approach does not relax the assumption that individuals maximise their utility. It merely allows their utility to reflect social concerns, too. Besides their own payoff it matters to them as well what the payoffs and intentions of other individuals are. Appendix A outlines how the sequential reciprocity equilibrium is determined.

3.2 Repeated Interaction

Google Answers users have a unique ID which makes them recognisable to GARs. The previous tipping behaviour of users can be observed by GARs and they may also be able to evaluate whether the effort of the respective GAR justified giving a tip. The relationship between reciprocity and reputation concerns in such a repeated games environment has been experimentally analysed by Gächter and Falk (2002). They aim to separate between non-strategic (reciprocity) and strategic (reputation) motives in their set up of a gift-exchange game. In a one-shot treatment firms and workers were anonymously matched for 10 periods knowing that they couldn't face the same partner twice, in the repeated game treatment 10 periods were played with a known partner. While the authors do observe reciprocal behaviour in both treatments, the wage-effort relationship is steeper in the repeated game treatment and effort levels are significantly higher in the repeated game treatment (until the last period) than in the one-shot treatment. Moreover, they identify reciprocal, selfish and imitating types among workers.

A possible explanation for the multiple equilibria in repeated games is described by the folk theorem. Alternatively, repeated interaction can be interpreted as a reputation mechanism where an updating process about a players' "type" takes place. When the decision to cooperate depends on the type of a player, e.g. good or bad, Kreps, Milgrom, Roberts and Wilson (1982) for instance show that cooperative equilibria can be reached. This kind of reputation model is based on Bayesian updating of beliefs.

In the Google Answers context GARs would update their beliefs about the tipping behaviour of the user they face. We can distinguish two different preferences types, reciprocal users who tip high effort answers and selfish users who would never tip. The Bayesian updating of users' past tipping behaviour reduces the uncertainty the GARs face. The more they are able to inform themselves about the user's past behaviour, the better they are able to identify the user's type. They will have a better idea whether or not to expect a tip and will put in high effort when it is likely to be rewarded. Selfish frequent users may take the GARs' updating into account and they might decide to imitate the reciprocal type. By tipping high effort answers they build up a good reputation and encourage high effort answers in the future.

Social preferences among GARs would reinforce these strategic considerations. GARs are able to observe the previous tipping behaviour of users and they may also be able to evaluate, whether a tip was not given due to low effort. As explained before that means GARs will update their beliefs about the tipping behaviour of the user they face. They would take the kindness of 'their' user towards other GARs into account, if they are also motivated by indirect reciprocity.[5] Then the GAR's belief about the kindness of the user is updated based on the user's previous actions and the GAR will put in high effort, if the user has a good enough track record of tipping and rewarding high effort

[5] See Seinen and Schram (2005) for an experimental study of indirect reciprocity where observed records of cooperativeness of a player induce others to cooperate with him.

answers.

3.3 Summary

The section presented the results of two experimental studies and stressed the similaritites of their designs and the Google Answers environment. In line with FGK we relate our analysis to sequential reciprocity theory and study whether reciprocity can explain the voluntary payments.

Since Google Answers users may ask questions repeatedly, frequent users may anticipate the benefits from establishing a good reputation by tipping and the resulting high effort answers in the future. Therefore, reputation concerns may motivate kind behaviour (i.e. tipping) besides reciprocity. Similar to Gächter and Falk (2002) we analyse the impact of such repeated interaction on the voluntary payments.

The following set of null hypotheses guides our empirical analysis:

Hypothesis 1 (Reciprocity): The tip rate of single users is not significantly higher than 0. Effort has no positive impact on the tip.

We test whether an open contracts design - providing mutual opportunities to reciprocate - encourages voluntary payments (tips) by single users and whether these tips are motivated by reciprocity.

Hypothesis 2.a (Reputation): The frequency of use has no effect on the users' tendency to tip.

Turning to repeated interaction, tipping out of strategic considerations hinges on the frequency of use and the belief updating of GARs.

Hypothesis 2.b (Reputation): In a "last period"-like situation imitating frequent users stop tipping, the tip rate drops to the level of single users.

We also try to distinguish between truly reciprocal and selfish frequent users who tip. The latter imitate reciprocal behaviour until there is no more reputational benefit to gain, i.e. they approach their final question.

Hypothesis 3 (Types): There is no individual heterogeneity among users with respect to their tendency to tip. No behavioural pattern can be detected.

We test whether users are homogeneous with respect to tipping or whether they tend bo either self-interested non-tippers or tippers (truly reciprocal or strategic). Both would tend to stick to their strategy or preference, respectively. In order to verify this classification, users who tip must have had a tendency to tip in the past, likewise users who do not tip must have had a tendency not to tip in the past.

Hypothesis 4 (Sorting): The tip history of a user has no effect on the effort level of the GAR.

When different tipping behaviour can be distinguished, GARs may inform themselves about a user's tip history and update their belief about the probability with which a user might tip. We test whether that has any effect on their effort decision. After sufficient observations to establish a reputation the questions of users with a high tip history are answered with more effort, questions of users with a reputation for not tipping are answered with less effort.

Hypothesis 5 (Efficiency): Effort levels do not increase significantly compared to phase 1 when tipping was not possible.

Finally, we test, whether an open contracts design has a similarly positive effect on efficiency (for both users and GARs) in Google Answers as in FGK.

4 Description of the Data Set

All questions posted at Google Answers are archived and accessible online. The entire thread of a question including the answer, answer clarification, any comments plus information about the question's price, tip, rating and category is therefore in the public domain. An automated Perl script extracted the information.

Our data set covers the entire life of Google Answers (April 2002 to December 2006). In total we collected 146,656 questions, 57,833 of them were answered. The rest expired 30 days after the question was posted. A very small fraction of answers (182 or 0.03%) were rejected by the user. Thus, actual transactions amount to 57,651.[6] The number of answered questions over time is very stable. Overall, 12,112 answers have been tipped, which is a ratio of 0.2354.

The observations of our data set are generated by 31,120 different users. The highest number of questions posted by the same user is 599. Still, the majority of users just asked a single question. The average number of questions per user is 1.42.

We collected the following data for each answer: The user ID of the person who posted the question, the price he set, the tip he possibly gave, the ID of the GAR who answered, date and time of posting the question, date and time of posting the answer, the rating of the GAR that was possibly left, the category of the question, the word count of the answer and the word count of the possible answer clarification.

Out of this data we computed additional variables. We calculated the time it took to answer a question (the difference in minutes between when the question was answered and when it was posted), the word count (the sum of answer and clarification) and the total number of questions posted (answered or not) by each user.

[6] Since the focus of our analysis is the tipping aspect we decided to deliberately truncate the data set considering only answered questions as observations. We are aware of the fact that a more general model would analyse all questions and why some are not answered. We only touch this issue in our paper.

An essential part of the analysis is finding a good way to measure the value an answer has for the user, since this is the user's signal[7] for the effort the GAR put into the answer. Users motivated by reciprocity or strategic concerns will base their decision to tip on the effort of the GAR. The more effort, the more likely they are to tip. Putting aside a question's difficulty for a moment two aspects should matter most to determine value/effort: Content and time. Better content means more value/effort, a faster response as well. Of course, we cannot assess the quality of an answer, but we have a precise measure for its quantity (the word count). We also know the time between posting of the question and posting of the answer.

Word count is the raw amount of words of an answer. We still have to consider that some questions will be more complex than others, so they will demand more effort hence more words. No one seems better suited than the user to rate a question's difficulty via the price they attach to a question. Therefore, we take the users' perspective and use price as a proxy for the question's difficulty. Hence, 'EffortWC' equals word count divided by price. Since naturally more is expected for a more demanding and thus higher-priced question, we normalise the word count with respect to the price of the question (a correlation coefficient of 0.32 confirms this relationship). The reasoning is that the more words GARs have included in answers of equally priced questions, the higher their effort has been.

We can compute a time-based effort variable in similar fashion. The faster an answer has been returned to the user, the higher should be the valuation of the answer and in turn the perceived effort of the GAR. Again, we have to normalise with respect to the price in order to take a question's difficulty into account. The quicker GARs have delivered answers of equally priced questions, the higher their effort has been. 'EffortTD' calculates then as the price divided by the time difference. The variable has to be taken with some caution, since our measure for time is the difference between posting of the question and posting of the answer and we do not know the time a user locked a question. Therefore the 'time difference' might not always be the time a GAR has worked on a question. It is exactly that, if the GAR started to work right after the question has been posted. However, questions might remain in the pool of unanswered questions for a while before a GAR decides to work on the answer. This can be up to 30 days after the posting of the question. The 'time difference' is then the time worked on the answer plus the time that passed until the GAR started working.[8]

This bias can be avoided when the sample is reduced to answers that have been returned within a rather short time (for instance 4 hours, the maximum

[7] It is a noisy signal as some chance is involved as well that determines the value of answer to the user. Nevertheless, the user's perception of the GAR's effort will be based on the value.

[8] However, from the perspective of a user this may not matter that much. Ceteris paribus, the user may care mostly about the time that passed to receive an answer to the question – the perceived effort – and not about the time the GAR really worked on the answer – the actual effort.

time a GAR can lock a question, which reduces the sample by 50%).[9] However, we do not know if otherwise equal questions that are on average answered within 1 hour (25% of the total sample) are sometimes found, locked and answered right away (total time 60 min) or sometimes found only after 3h (total time 240 min). This is avoided by setting the ceiling to 30 minutes or even less. But then the question is whether users consistently check in so frequently that such a fast answer is always recognised as a fast (i.e. high effort) answer. These issues confound the meaning of the time difference between posting question and answer and we do not use it in further analysis.

Finally, we created a dummy, if there was an answer clarification as well as various category dummies.

An intriguing feature of the data set is the late introduction of the option to leave a tip (in October 2002). The 6,206 answers during the first 6 months could not be tipped. This provides a great opportunity to study adoption behaviour, but it also requires adjustments in the data analysis. We distinguish between phase 1 (before the introduction) and phase 2 (when tipping was available).

Table 1: Descriptive Statistics of Phase 1 (No Tips Possible)

variable	obs	mean	median	st. dev.	min	max
price	6,206	14.9	8	24.05	2	200
rating	3,581	4.39	5	0.96	1	5
time difference [min]	6,206	2,445.72	156.5	135.26	1	449,689
word count	6,206	479.76	330	589.11	3	17,047
answer clarification	6,206	0.3437	0	0.475	0	1
effortWC	6,206	55.96	35.64	78.39	0.2	3,409.4
effortTD	6,206	405.5	19.9	1,808.41	0.11	32,449.75

where obs = number of observations, st. dev. = standard deviation

Table 2: Descriptive Statistics of Phase 2 (Tipping Possible)

variable	obs	mean	median	st. dev.	min	max
price	51,445	23.79	10	37.31	2	200
tip	12,109	9.13	5	14.79	1	100
rating	32,429	4.66	5	0.679	1	5
time difference [min]	51,445	2,616.35	241	6,915.5	1	43,198
word count	51,445	619.90	349	1152.99	1	81,851
answer clarification	51,445	0.2976	0	0.4572	0	1
effortWC	51,445	51.75	30	83.04	0.005	7,792
effortTD	51,445	318.8	23	1,431.96	0.075	21,583

where obs = number of observations, st. dev. = standard deviation

The price range is pre-determined by Google Answers. The lowest price users can set is $2, the highest price possible is $200. These are also minimum and maximum price of the sample. The average price conditional on the question being answered (57,833 observations) is $22.84, while the average price of the

[9] Edelman (2004) also addresses this bias imposing a ceiling for the time difference of the maximum lock period (4h) plus 1h.

88,823 questions that expired without an answer is only \$20.19, significantly less at the 5%-level based on a Mann-Whitney test.[10] With no more relevant information available it appears as if the price plays at least a partial role in the GARs' decision to answer a question or leave it in the pool.

Minimum and maximum values for the tip are also pre-set by the service. There is an upper limit of \$100 for the tip.

The time difference between question and answer is expressed in minutes. The quickest answer came after only two minutes, the slowest was given just before the 30 day expiration deadline. The median of the distribution is 241 minutes. That means half of all answers were posted within 4 hours.

The word count is the number of words of an answer. The shortest answer was a single word ("No" to be precise) and the longest contained 11,482 words (a \$190 question with \$65 tip).

A rating has been given for 32,429 answers, roughly two thirds of the total. The possible range is from 1 to 5, with 5 being the top rating. If users decided to give a rating, they did not mind giving the highest possible as median and mode are 5 and the average rating is 4.66. There is also a high correlation between a rating being given and a tip being left due to the web site structure. If a user decides to give "feedback", he first has to enter a rating (1 to 5) and then decides on a possible tip (0 or simply no entry to 100).

5 Analysis of the Data

We first present the results of a panel regression of phase 2 data, then we analyse some specific aspects in more detail. In contrast to the tipping in many service professions[11], there is a high variation in Google Answers tipping. That's why we disregard conforming to a social norm as a motivation for tipping in further analysis and focus on reciprocity and strategic considerations due to reputation concerns. We then turn to the GARs' perspective and analyse the relationship between updating, effort decision and efficiency in the data set. Finally, we study how tipping was adopted when the option to tip was introduced six months after the beginning of Google Answers.

5.1 Estimations

Three different motivations appear plausible to explain the tipping behaviour in the data. Firstly, reputation may matter. Frequent users of the service

[10] This is confirmed by a Probit regression in which also the categories Arts/Entertainment, Health, Reference/Education/News, and Relationships/Society have a significantly positive effect on the question being answered. The categories Business/Money, Computers, and Sports/Recreation have a significantly negative effect.

[11] Azar (2004) and Lynn (2005) survey tipping behaviour in common service situations like a restaurant visit, for instance. While originally (16th and 17th century in Europe) people tipped out of gratitude for extra service, out of compassion or to encourage better service, it soon became a social norm. In many occasions tipping is very institutionalised and a quite precise fraction of the bill ought to be tipped. In restaurants people would tip roundabout the same percentage of their respective bill. (Azar 2004)

have an incentive to build up a good reputation and may regard tipping as a strategic device. Secondly, social preferences would make people tip. Users who are socially-minded should leave a tip as long as there is a reason to reciprocate positively. Thirdly, the tip should simply be affected by the price of the question. Users may tend to tip proportionally to the price, giving a high tip for a highly priced question and vice versa.

Reputation concerns are proxied by the frequency with which a user asked questions. The more questions posted the more generous users should be with the tip – simply out of strategic considerations. A high frequency of using the service means the user should have much to gain from high effort answers in the future and this can be positively affected by tipping now. We use the logarithmic value of the total number of questions posted by a user in our regression, because the impact of reputation concerns on the tipping behaviour should decrease with the total number of questions increasing.

We use the following proxies to take account of behaviour that indicates a reason for the user to positively reciprocate (effort exerted by the GAR or whether an answer clarification has been provided) or a tendency to reciprocate of the user himself (the rating given by the user).

The effort involved in a given answer indicates how hard a GAR worked for the answer and how much value it created. Effort is metered in terms of word count (relative to the price to control for the difficulty of a question). Everything else equal, a very comprehensive answer with a lot more background information than expected will be perceived as a "high effort"-job and should have a higher value for the user. When a question has been answered with high effort, users sufficiently motivated by reciprocity would tend to return the perceived kind behaviour of the GAR and give a tip.

An answer clarification is given only on request, after the answer itself has been posted. It is likely that the clarification adds more value to the answer, which is captured in the word count. However, the clarification may also be perceived by the user as an extra effort of the GAR and this should trigger reciprocal behaviour of the user. It can also be regarded as increased social interaction between user and GAR. Hence, we use the answer clarification dummy as another proxy for reciprocity.

When a user leaves a rating, it seems reasonable to assume that he is not entirely self-interested. It only costs time and a positive impact on a user's reputation seems hard to imagine. It shows on the other hand that the user cares about the benefit of the GAR, since GARs' ratings are fairly important to them. There is no monetary sharing of course, but leaving a positive rating can be seen as a sign for a tendency to reciprocate greater than zero, a necessary condition for caring in a monetary sense, i.e. giving a tip. Leaving a high rating is of course an indication that the user is content with the answer, another pre-requisite for a monetary reciprocation.

We are aware that these variables can only be rather crude surrogates for what motivates voluntary payments, yet we believe that this quantification can nevertheless contribute to a better understanding of social preferences.

The rating plays an important role in the analysis of tipping as both decisions are intertwined.[12] If I wanted to leave a tip, I will have to give a rating, too (due to the sequential design). When I want to rate a question, I do not have to tip it. Only rated answers can be tipped, yet there does not seem to be a selection bias in the relationship between rating and tipping. If I wanted to tip, I am not prevented by anything except having to rate the answer which is probably negligible. Hence, we estimate a bivariate probit model for the binary decisions whether to rate and tip.

Since no negative tip or rating can be given the distributions are left-censored at zero. Therefore, a censored regression model appears appropriate. The Tobit model takes limits of the range of the dependent variable into account to ensure unbiased and consistent estimates. The standard Tobit model assumes a single distribution function for the dependent variable. However, there is good reason to believe that the decision on whether to tip or not and the decision how much to tip (given one has chosen to tip) are separated. The same applies to the rating decision. Different distributions could be underlying and a two-step model of Cragg (1971) will take this into account. (Amemiya 1984) A Probit model estimates the binary decision of whether to tip or not and a truncated regression is used to estimate the size of the tip. A likelihood ratio test of the restricted Tobit model against the unrestricted composite model of Probit and truncated regression rejects the null hypothesis clearly for all specifications and confirms our approach.[13] Table 3 lists the variables, their coefficients and respective standard errors for our estimations.

[12] Out of 51,445 phase 1 answers 32,429 have been rated. 12,109 (rated) answers have been tipped.

[13] Our censored regression models are based on maximum likelihood and they assume a normal distribution of the error term and homoscedasticity. A Bera-Jarque test rejected the normality assumption. Therefore, we used a model that bootstraps standard errors. The robust Huber–White sandwich estimator is employed to control for potential panel heteroscedasticity.

Table 3: Bivariate Probit Model (Tip and Rating):				
	Tip		Rating	
Explanatory variable	coeff.	st. error	coeff.	st. error
Price	.0001	.0002	-.0014 ***	.0002
Frequency of use (log(Total Questions Posted))	.1694 ***	.0170	.2568 ***	.0197
EffortWC	.0004 ***	.0001	.0002 ***	.0001
Answer Clarification	.2366 ***	.0172	.3046	.0198
Arts/Entertainment	.2338 ***	.0326	.2791 ***	.0306
Business/Money	-.0908 ***	.0334	.0077	.0400
Computers	-.0018	.0310	.1001 ***	.0292
Family/Home	.0431	.0447	.0556	.0413
Health	.0214	.0339	.0529 *	.0319
Reference/Education/News	.0926 ***	.0342	.1332 ***	.0299
Relationships/Society	.1545 ***	.0427	.1809 ***	.0351
Science	-.0028	.0500	.0617	.0384
Sports/Recreation	.0876 *	.0475	.1499 ***	.0403
2002	-.2815 ***	.0370	-.0027	.0289
2004	.1133 ***	.0255	.1412 ***	.0231
2005	.1267 ***	.0286	.1760 ***	.0247
2006	.0638 **	.0335	.0008	.0284
Constant	-1.143 ***	.0327	-.1929 ***	.0299
Sample size: 51,445; standard errors adjusted for 31,120 clusters				
Log pseudolikelihood: -52,178.93				
Statistical significance: *=10% / **=5% / ***=1%				

One argument for separating the tipping decision and the decision of how much to tip was that the price of the question might not affect the first, but even more the second decision. In fact, the regressions confirm that the price does not affect the decision, whether to tip. The data also confirms the significance of reputation concerns. The estimators for the coefficient of the frequency of use explain both tip and rating at a statistically significant level (1%-level). The effect of the word count-based effort is clearly positive as well (1%-level for both tip and rating). It also clearly matters whether an answer clarification has been given. The coefficients are positive and highly significant.

There is also a clear increase of the tip rate compared to 2002 captured by the year dummies. Behaviour appears to be different across the various categories. Answers in Arts/Entertainment, Reference/Education/News and Relationships/Society are more likely to be tipped/rated. Answers in Business/Money are less likely to be tipped.

A truncated regression confirms the importance of the price for the size of the tip (1% significance level).

5.2 Reciprocity

Reputation concerns may influence the tipping behaviour of users. We need to study the behaviour of single users in order to control for reputation and focus on reciprocity. During the entire life of Google Answers there are 21,512 users who posted only one question (that got answered). 14.87% of them did leave a tip.[14], significantly more than 0. A regression only with single users delivers equivalent results as the main regression. The word count-based effort is statistically significant at the 1% level. Also a non-parametric Wilcoxon rank-sum test confirms that the effort level is significantly higher when single users decided to tip (1%-level).

While it is a fact that these users asked just one question, we can not be certain that they had no intention to use the service again. Maybe they planned to use it often, but in retrospect they were disappointed by the answer quality and stopped using the service. In that case effort levels of the answers the single users received should be significantly lower than the effort levels of the 9,650 first answers that multiple users received. The effort levels are 50.81 and 50.77, a non-significant difference based on a Mann-Whitney test.

After controlling for the impact of reputation concerns we find that tips are still prevalent, albeit at a lower rate than among frequent users. Moreover, single users' tips are explained by effort. This rejects hypothesis 1. Our approach to control for repeated game incentives is naturally limited by the field data set and cannot be regarded as bullet proof. Nevertheless, the results are in line with comparable experimental and field studies. Voluntary payments at a significant level are also observed in another field study where reputation effects cannot play a role. (Regner and Barria, 2009)

5.3 Reputation Concerns

Frequent users could have an interest in building up a reputation of appreciating good value and acknowledging it with a tip. This way they may attract GARs who recognise them as generous and will deliver high effort answers in anticipation of a tip. This motivation may be of particular relevance in online environments, since transaction partners do not see each other online. (Resnick *et al.* 2000)

In order to test the impact of reputation concerns on the tipping behaviour we clustered the data by the amount of questions a user posted. Recall that this variable counts also questions that did not get answered. Thus, it should give a better proxy of how often a user intends to use the service than the number of answers he actually received. Still, some users may not have a clear idea of how often they are going to use the service when they start with the first question, but on average they should be aware of that. Therefore, we believe the frequency of use is a good indicator for the reputation concerns of users.

[14]See Table 4 in the next subsection for the data about single users in comparison to occasional and frequent users.

The following table shows the pricing and tipping behaviour of users clustered by the amount of questions they posted:

TABLE 4: SUBGROUPS BY FREQUENCY OF USE

posted questions	observations	tip rate	significance	avg. price
1 question	21,512	.1487	***	23.55
2 questions	5,909	.2288	***	26.68
3	3,146	.2540	0	26.89
4	2,163	.2621	0	25.83
5	1,478	.2848	0	28.94
6	1,351	.2850	0	23.95
7	1,100	.2736	0	26.17
8	937	.2636	***	25.32
9	778	.3509	0	27.41
10+ questions	13,071	.3492	–	20.68
all	51,445	.2354	–	23.79

14.87% of all single users gave a tip. However, with increasing number of questions posted we observe a steadily increasing tip rate. Already about a quarter of the transactions by users who asked three to four questions were tipped. The tip rate goes up to almost 35% for frequent users (10 or more questions posted). Statistically, the tip rate for single users is different from the rate when two questions were posted (1%-level). Table 4 also shows the respective significances of tip rate comparisons. No difference is found in the range of 3 to 8 questions posted. The tip rate of frequent users is again significantly different from the level of users who posted less than nine questions.

These results lead us to conclude that occasional users already take reputation concerns into account. For frequent users reputation concerns matter even more.

Strategic considerations are an explanation for tipping, but when the end of using the service is near – when there is no more reason to maintain a good reputation – tipping out of strategic considerations should break down. If we are able to observe a "last period"-like effect, we can further distinguish behaviour motivated by reputation from reciprocity and possibly quantify the difference.

The natural way to analyse this possible fading of reputation concerns is Google Answers' "end game". On November 28th, 2006, Google officially announced that the service will stop accepting new questions in a few days (answers could be given until the end of 2006). At this time users should have been aware that it makes no sense anymore to invest in a good reputation by tipping answers. 158 questions have been answered after November 28th and 36 (or 22.8%) had been tipped. But 108 of the 158 questions have actually been last questions of the respective user and 14 of those (13%) have received a tip. When facing the imminent end of the service, the fraction of users who still tip goes down significantly - into roughly the range of single users - and this level (circa 15%) can probably be seen as the level of intrinsic motivation or the fraction of genuine socially-minded individuals.

Subtracting this baseline from the level at which frequent users tip (34.92%)[15] should provide us with a good estimate for the fraction of strategically motivated tippers. Around 20% would then imitate genuine socially-minded individuals out of reputation concerns in order to receive high effort answers. The remaining 65% appear to be self-interested, not willing to (knowingly or not) employ tipping as a strategic tool. Of course, these numbers are conditional on high effort being exerted which is rather unlikely. The estimates for truly reciprocal (15%) and strategic imitators (20%) must be regarded as minimal values. The true values are most likely higher.

Frequent users tip consistently more often than single users, similar to the experimental findings of Gächter and Falk (2002). In fact, the tip rate increases with the frequency of use, which rejects hypothesis 2.a and b.

5.4 Updating, effort decision and efficiency

This section tries to shed more light on the decision making of GARs. They may update their beliefs about the likeliness the user they face will tip (if effort is high). In the data set we can specify the tip history of each user at each number of question she answered. It is the amount of answers she tipped divided by the total of answers she received at that point. Recall that this information is not very straightforward to obtain for the GARs.[16] Table 5 splits the sample into different sub groups with respect to the question number asked. Essentially we see that the tip rate increases for users who keep on asking questions which is not surprising as we know that frequent user tend to tip more often.

Table 5: Question Nr. and Tip Rate

question nr.	obs	tip rate	avg. price	avg. tip (if tipped)
first	31,120	0.18	23.93	8.28
2nd to 9th	14,256	0.28	25.02	10.21
10+	6,069	0.41	20.19	9.28

When we consider the respective tip history of each user at each question number we see in Table 6 that there is a large spread between tipped and untipped questions. Naturally, the tip history does not exist at question number 1. In the intermediate range of question numbers users who did not tip had an average tip history of just 18%, while users who left a tip had one of 56%. The spread is very similar in the high range of question numbers as shown in Table 6.

Users who tip an answer clearly had a tendency to do so in the past as well. On the other hand, users who did not give a tip have a rather low tip history.

[15] This value is transactions- and not user-based. Distinguishing between tipping (tip history $> 1/2$) and not-tipping user types (tip history $< 1/2$) delivers similar results. At 25 answered questions there are 38 tipping and 62 non-tipping types (a ratio of 0.38) and at 50 answered questions there are 13 tipping and 38 non-tipping types (a ratio of 0.34).

[16] It is not shown next to the user name as the past average like the rating of GARs is for instance or the seller's reputation on eBay. GARs have to enter the user's ID in a search mask and the user's previous questions are shown with price (and tip).

It seems that users have preferences or a strategy to tip (high effort answers) or not and they stick to it, which rejects hypothesis 3.

Table 6: Question Nr. and Tip History			
question nr.	avg. tip history	/ without tip	/ with tip
first	–	–	–
2nd to 9th	0.29	0.18	0.56
10+	0.38	0.22	0.59

If GARs do in fact update their beliefs about the chances to get a tip for high effort work, then they should anticipate that and make their effort decision based on this updated belief. They should put in "low effort" when they face a user with a low tip history who likely will not tip anyway, while they should exert "high effort" when they meet a user who has tipped in the past and might well do so again. But GARs can only reliably update their beliefs about the user's tendency to tip, when previous questions are available. The more past questions available, the better is the GARs' signal. Hence, we should expect the tip history to be be mediated by the number of past questions. An OLS panel regression confirms this:

Table 7: OLS Panel Regression: Effort

Explanatory variable	coeff.	st. error
Tip History	-5.5457 *	2.172
log(Question Index)	.5963	.4596
Tip History * log(Question Index)	4.388 ***	1.204
2002	.8891	1.255
2004	-8.169 ***	.9941
2005	-17.00 ***	1.058
2006	-22.325 ***	1.168
Arts/Entertainment	-.8922	1.073
Business/Money	18.22 ***	1.951
Computers	14.39 ***	1.414
Family/Home	4.180 ***	1.201
Health	12.866 ***	1.783
Reference/Education/News	12.99 ***	1.487
Relationships/Society	4.849 **	2.123
D_SCI	-1.741	1.442
Constant	55.555 ***	.8222
Sample size: 51,445	Number of groups: 430	
Statistical significance	*=10% / **=5% / ***=1%	

where D_* = dummy variable for #

ART = arts & entertainment, BIZ = business & money

COM = computers, FAM = family & home, HEA = health

REF = reference, education & news, REL = relationship & society

SCI = science

if all category dummies = 0, we have observation in 'miscellaneous'

Tip history alone does not explain the effort level, in fact it is a negative determinant. It is positive and significant (1%-level) only when it interacts with the log of question index.

Table 8 shows the effort levels of GARs. When a user asks the first question, no tip history exists and the effort decision cannot be based on the user's past. Effort is slightly higher for the tipped answers just as we should expect it since we know that effort explains the tip. While the split between questions with and without tip is similar in the intermediate range of question numbers (Mann-Whitney test between tipped and untipped samples, 5% significance level), the gap widens in the high range. With ten or more questions available to assess the user's tendency to tip the tipped questions have been worked on with significantly more effort than questions without tip (Mann-Whitney test, 1% significance level). The average effort is also higher compared to earlier questions that were tipped (Mann-Whitney test between tipped samples of occasional and frequent users, 1% significance level).

Table 8 also provides the effort level during phase 1 when tipping was not possible. Frequent users who tip have received answers with significantly higher effort than answers during phase 1 even though a positive time trend of the price (see regression in Table 8) reduces the effort level (word count / price) over time.

Table 8: Question Nr. and Effort

question nr.	obs	average effort1	/ without tip	/ with tip
pre OCT 2002	6,098	55.77	55.77	–
first	31,120	50.82	49.88	55.11
2nd to 9th	14,256	51.41	48.88	57.86
10+	6,069	57.26	51.42	65.82

It seems that indeed GARs update their beliefs based on the tip history and that they make their effort decision according to that belief. Moreover, users stick to their behaviour (due to preference or strategy) and they reward high effort, if they are sufficiently motivated by reciprocity or reputation. Effort and tip history are correlated in the frequent user sample (Spearman correlation coefficient, 5% significance level). This rejects the fourth hypothesis.

The open contracts design with its mutual opportunities to reciprocate can lead to a significantly higher effort level compared to the conventional design without opportunities to reciprocate (used in phase 1) and its counterpart in phase 2 (mutual opportunities to reciprocate are available, but an extensive history shows the user disregards them). High effort levels can be assumed to translate directly into more value for the user. They are made better off as they would not voluntarily give away as a tip more than they actually want. But are GARs compensated for the higher effort they put in? Or are they hunting for tips that at the end of the day do not pay them adequately? Maybe non-tipping frequent users move their incentives into the price and the tip given is fairly small. So, does it pay off for GARs to put in high effort, when they work on questions of users who are known for tipping?

Table 9 shows that there is very low variation of the price across the groups.

There is no indication that frequent non-tipping users price their questions differently from their counterparts who make use of the tip. Also, the size of tips is substantial across groups and it seems that it rewards higher effort adequately.

Table 9: Question Nr. and Pay (Price + Tip)

question nr.	/ without tip	/ with tip
first	23.93	23.94 (+ 8.28)
2nd to 9th	24.80	25.57 (+ 10.21)
10+	20.98	19.04 (+ 9.28)

Users known for tipping get higher effort answers than new users, but they also reciprocate and apparently let the GARs participate in the gain from a high value answer by returning some of the surplus and leaving a high tip.

The open contract design increases the effort level and the efficiency. It seems that it encourages socially-minded users to reciprocate (tipping high effort answers) and that it makes self-interested users consider building up a good reputation (in order to motivate future high effort answers). Through belief updating the GARs are able to match their effort decision better to the user types. Consistent high effort answers are possible in contrast to a more complete contract that does not allow a tip. Such a strict contract type is simulated, when users reveal that they are not going to tip (long enough low tip history). Then GARs update their beliefs accordingly and put in relatively low effort. Hence, we can reject hypothesis 5.

5.5 Adoption process

For the first six months of Google Answers no tips could be given. Only in October 2002 the option to tip an answer was introduced. It appears this feature was not welcomed with open arms, but rather greeted with some healthy reservation. With on average 63 answers per day at that time, the first tip ever was given on the 7th of October, the second on the 9th and the third tip on the 10th. Only in the second half of October users slowly warmed up and started to tip more often as can be seen in Figure 1. In total, 7.1% of the 1,942 answers in October 2002 were tipped. However, tipping gained momentum rather quickly. In November 2002 18.63% of 2,459 answers were tipped. In the following months tipping already reached a level known from the total numbers: 19.26%, 20.76%, 24.38%, 23.90%, 21.31%, 25.80% and 23.40% (from December 2002 to June 2003).

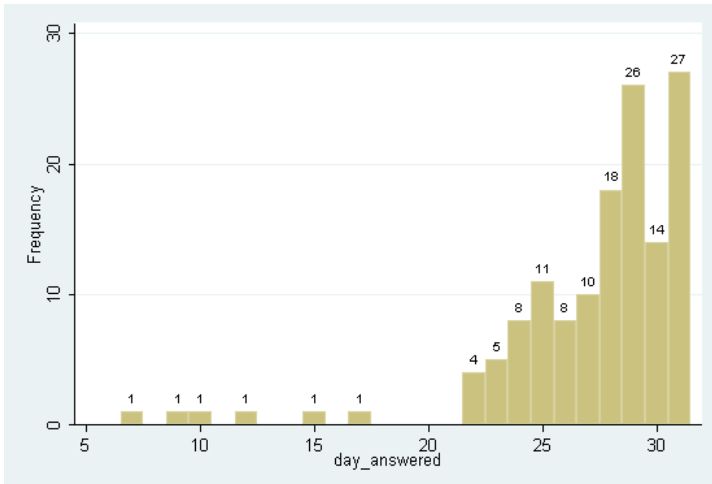

Figure 1: Frequency of tips in October 2002

What has been the motivation of those early tippers? A regression on the data from October 2002 with the same specification as in the main model shows no significant effect of effort nor the frequency of use. The answer clarification dummy and rating explain selection at 1%-level and price exlains the size of the tip also at the 1%-level. In a regression for November 2002 the familiar significance of the word count-based effort (5%- in selection, 1%-level in size) appears in addition to the described significance of the answer clarification, rating and price. We get the same results for all 2002 (6,050 transactions). Only in the first quarter of 2003 (4,565 transactions) frequency of use starts to become significant (at the 5%-level).

6 Conclusions

We investigate the real-life pricing, effort and tipping decisions using all available data from the online service Google Answers (57,654 transactions). This rich data set puts us in a position to test the relevance of social preferences in a real-life environment complementing behaviour observed in the lab. In particular, our interest is in the underlying motivations for the occurring voluntary payments and the efficiency of such an open contracts design. We relate our findings to the theory of sequential reciprocity of Dufwenberg and Kirchsteiger (2004). Applied to our context, an intentions-based reciprocity model predicts that tipping takes place even among single users, if they are sufficiently sensitive to reciprocity.

Almost 15% of all single users left a tip, occasional (circa 25%) or frequent (circa 35%) users tip even more often. Our regression analysis shows that the tip can be explained by reciprocity proxies ("Effort of the researcher (GAR)", "Rating given by the user" and "Has an answer clarification been provided") and reputation proxies (frequency of use). The higher tip rates of frequent users are in line with the experimental findings of Gächter and Falk (2002). The effect of reciprocal behaviour and repeated game incentives appear to be complementary.

The data from Google Answers also confirms the positive effects of an open contracts scheme on the effort level as found in Fehr, Gächter and Kirchsteiger (1997). Our data shows that users tend to either stick to their preferences/strategy to tip high effort answers or they do not tip. GARs will try to update their beliefs about the user's type when they make their effort decision. The uncertainty about whether a user will tip is reduced the more history of the user's decisions is available. GARs can then update their beliefs more reliably and are able to make an educated effort decision. When GARs face a frequent user (10 or more answers available), high effort is matched to rewarding users and low effort is matched to users who do not tip. The open contract design can be seen as a virtuous circle that increases efficiency.

It seems that two conditions are essential for the success of an open contracts design. GARs need to be able to update beliefs about user types. Only then the strategy of imitators pays off and they attract high effort answers. The existence of genuine tippers motivated by reciprocity is particularly crucial for the open contracts design. Without them strategic players have no one to imitate and the positive feedback loops of mutual opportunities to reciprocate would not even start.

How does tipping evolve, in particular how does it start when the default is not to give a tip? The late introduction of the option to tip (during phase 1 no tipping was possible) gives us the means to analyse the adoption process. For 6 months the behavioural default was not leaving a tip. After a slow start (6 out of 1,000 answers were tipped in the first half of October) reciprocity proxies explain the tip in the remainder of 2002, but no reputation proxies. It takes until the first quarter of 2003 for the frequency of use to become significant. It appears tipping is adopted slowly by some users motivated solely by reciprocity. After a while tipping is recognised as a strategy motivated by reputational concerns.

What are proportions of reciprocators, imitators and the remaining self-interested users? We can classify into these types based on the data from phase 2. Single users who tip (around 15%) may indicate the fraction of reciprocators. The difference between frequent users who tip (around 35%) and the single user baseline may be taken as the fraction of imitators (20%). Data from the end game confirms these relations. After Google Answers already announced to its users that it will be closed soon 13% tipped the final answer they received. With no reason to believe that the Google Answers sample population is not representative, we propose minimum levels[17] for the reciprocator type of 15% and of 20% for the imitator type.

[17] Users – no matter whether motivated by reciprocity or reputation – would only tip high effort answers. Perceived low effort answers would never be tipped. Hence, the fractions are potentially higher and the estimates are minimum levels.

7 Appendix

The utility function of socially-minded individuals increases not only in their material payoffs but also in the psychological payoffs which depend on the individuals' kindness to others and beliefs about that. The resulting games are solved using the psychological games framework of Geanakoplos, Pearce and Stacchetti (1989). While the action set a_i describes the choices of player i (e.g. the effort of the GAR or the chosen price and tip of the user), b_{ij} defines the belief of i about the choices of player j, whereas $\widetilde{b}_{iji}$ is i's belief about what j believes are i's choices. This framework of beliefs allows us to express the kindness and beliefs about the kindness of individuals towards another individual. This is done by comparing an actual payoff Π to the equitable or fair payoff of a player, Π^e.

The equitable payoff of an individual is the average of his best and worst outcome based on the choices of the other individual.[18] For agent j it is given by:

$$\Pi_j^e(b_{ij}) = \frac{1}{2}(\max\{\Pi_j(a_i, b_{ij})\} + \min\{\Pi_j(a_i, b_{ij})\}) \tag{1}$$

It can be seen as a reference point for how kind i is to j as this kindness κ_{ij} is expressed by relating the actual payoff j is given by i to the equitable payoff of j:

$$\kappa_{ij}(a_i, b_{ij}) = \Pi_j(a_i, b_{ij}) - \Pi_j^e(b_{ij}) \tag{2}$$

Similarly i's belief about the kindness of j to i is:

$$\widetilde{\kappa}_{iji}(b_{ij}, \widetilde{b}_{iji}) = \Pi_i(b_{ij}, \widetilde{b}_{iji}) - \Pi_i^e(\widetilde{b}_{iji}) \tag{3}$$

Incorporating kindness and the beliefs about it gives the following utility function with a material payoff as the first term and the reciprocity payoff in the second term that is weighted by the reciprocity sensitivity α ($\alpha = 0$ is the special case of pure self-interest).

$$U_i = \Pi_i(a_i, b_{ij}) + \alpha_i \cdot \kappa_{ij}(a_i, b_{ij}) \cdot \widetilde{\kappa}_{iji}(b_{ij}, \widetilde{b}_{iji}) \tag{4}$$

The condition to solve the game is that in equilibrium all beliefs and second order beliefs are correct. It is also important to mention that beliefs of players are updated over the course of the game. The individuals apply Bayesian updating.

A positive reciprocity equilibrium exists. The user will give a tip, if his sensitivity to reciprocity is large enough: $\alpha_u > \overline{\alpha}_u$. The possibility of $\alpha_u < \overline{\alpha}_u$ corresponds to the nasty equilibrium.

After establishing conditions for the user to give a tip once the GAR has put in high effort, it has to be analysed whether the GAR will ever work at a high effort level in the first place. He knows that the user will never give a tip

[18] The average is used here, because it is straightforward. Using another intermediate value is also possible and it does not affect the qualitative results. See also Dufwenberg and Kirchsteiger (2004) footnote 7.

when $\alpha_u < \overline{\alpha}_u$ and therefore he will never give high effort. This constitutes the sequential reciprocity equilibrium of (low effort, (no tip, no tip)).

The GAR also knows that the user will act reciprocally once her sensitivity to reciprocity α_u is large enough. That means he assumes the user will reward the choice of high effort with a tip and will reply to low effort by not giving a tip. It can be shown that the condition for the GAR to make the high effort decision is always fulfilled and the sequential reciprocity equilibrium of (high effort, (tip, no tip)) results.[19]

By applying sequential reciprocity theory we can explain when users give a tip. Social preferences are necessary which are incorporated into the utility function with a reciprocity payoff. Once reciprocity gains (from returning kind behaviour) outweigh the material loss of paying a tip, users will prefer to tip. However, users and GARs have to be sufficiently motivated by reciprocity, e.g. α – their sensitivity to reciprocity – has to be large enough. Moreover, the GAR has to believe that the user's α is large enough in order to provide high effort in the first place.

Table 2 lists the ten different categories in which users can post their questions. We created dummies for all of them except the last one: 'Miscellaneous'. Their popularity is quite different. While only 216 observations are in category 'Sports and Recreation', the most popular category after 'Miscellaneous' was 'Computers' with 1,209 entries. About 31% of all observations in 'Arts & Entertainment' or 'Sports and Recreation' have been tipped. Users in the 'Business & Money' category appear to be the least generous as only 21.68% of these questions have been tipped. The tip rate of the other categories is fairly close to the overall average of 25.46%. The 'Business & Money' category also features the highest average price ($34.32).

Table 2 : Question Categories

category name	answers	tip ratio	avg. price
Arts/Entertainment	5,674	0.285	15.32
Business/Money	8,572	.2012	37.45
Computers	7,840	.2330	21.55
Family/Home	1,923	.2350	18.55
Health	3,937	.2291	29.87
Reference/Education/News	5,834	.2478	21.52
Relationships/Society	2,345	.2912	22.95
Science	3,513	.2265	20.42
Sports/Recreation	1,604	.2475	19.51
Miscellaneous	10,203	.2218	20.70
all	51,445	0.2354	23.79

[19] See Regner (2005) for more details.

References

[1] AMEMIYA, T. (1984). "Tobit Models: A Survey", *Journal of Econometrics* 24: 3-61.

[2] AZAR, O. (2004). "The History of Tipping - from Sixteenth-century England to United States in the 1910s", *Journal of Socio-Economics*, 33: 745-764

[3] CAMERER, C. (2003): "Behavioral Game Theory: Experiments in Strategic Interaction", Princeton University Press, Princeton

[4] CRAGG, J. G. (1971): "Some Statistical Models for Limited Dependent Variables with Application to Demand for Durable Goods", *Econometrica*, 39, 829-844.

[5] DUFWENBERG, M., and G. KIRCHSTEIGER (2004): "A Theory of Sequential Reciprocity", *Games and Economic Behavior*, 47, 268-298.

[6] EDELMAN, B. (2004): "Earnings and Ratings at Google Answers", *unpublished manuscript*

[7] FEHR, E., S. GÄCHTER, and G. KIRCHSTEIGER (1997): "Reciprocity as a Contract Enforcement Device: Experimental Evidence", *Econometrica*, 65, 833-860.

[8] FEHR, E., and K. M. SCHMIDT (2003): "The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories", *Handbook of Reciprocity, Gift-Giving and Altruism*

[9] GÄCHTER, S. and A. FALK (2002): "Reputation and Reciprocity: Consequences for the Labour Relation", *Scandinavian Journal of Economics, 104(1), 1-26*

[10] GEANAKOPLOS, J., D. PEARCE, and E. STACCHETTI (1989): "Psychological Games and Sequential Rationality", *Games and Economic Behavior*, 1, 60-79

[11] KREPS, D., MILGROM, P., ROBERTS, J. and WILSON, R. (1982). "Rational Cooperation in the Finitely Repeated Prisoners Dilemma", *Journal of Economic Theory* 27, 245-252

[12] LYNN, M. (2005): "Tipping in Restaurants and Around the Globe: An Interdisciplinary Review" in: "Foundations and Extensions of Behavioural Economics: A Handbook", edited by M. Altman, M. E. Sharpe Publishers

[13] RABIN, M. (1993): "Incorporating Fairness into Game-Theory and Economics", *American Economic Review*, 83, 1281-1302

[14] RAFAELI, S., D. RABAN and G. RAVID (2007). "How social motivation enhances economic activity and incentives in the Google Answers knowledge sharing market", *International Journal of Knowledge and Learning,* Volume 3, Number 1 / 2007

[15] RESNICK, P., R. ZECKHAUSER, E. FRIEDMAN and K. KUWABARA (2000). "Reputation systems", *COMMUNICATIONS OF THE ACM,* Volume 43, Number 12, p. 45-48

[16] REGNER, T. (2005). "Why Voluntary Contributions? Google Answers", *University of Bristol working paper 05/115*

[17] REGNER, T. and J. BARRIA (2009). "Do Consumers Pay Voluntarily? The Case of Online Music", *Journal of Economic Behavior and Organization, forthcoming*

[18] SEINEN, I. and A. SCHRAM (2005). "Social status and group norms: Indirect reciprocity in a mutual aid experiment", *European Economic Review*