We analyse pricing, effort and tipping decisions at the online service ‘Google Answers’. Users set a price for the answer to their question ex ante, and they can additionally tip the researcher who provided the answer ex post. A positive wage-effort relationship is frequently found in laboratory gift-exchange games, yet field evidence for reciprocity in two-stage settings (wage, effort) is mixed. Our field data confirms lab findings for the three-stage design (wage, effort, bonus). Tipping 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 sufficient tip history is available.

In addition, we analyse how tipping is adopted when the behavioural default is not to tip and suggest estimates for reciprocal and selfish (strategic and myopic) user types.

© 2014 Elsevier Inc. All rights reserved.
answer to the user who can optionally give a tip to the GAR. This data set allows us to analyse pricing, effort and tipping decisions in a real life labour market. In particular, we focus on the underlying drivers for voluntary payments and the effects of such a design on effort levels and efficiency. In a simple model we consider two possible motivations for the tipping of users and test empirically to what extent they drive the behaviour at Google Answers. Social preferences can motivate users to leave a tip. We specifically consider concerns for reciprocity following the theory of sequential reciprocity of Dufwenberg and Kirchsteiger (2004). Moreover, reputation concerns can motivate frequent users to leave a tip. Self-interested users may decide to imitate reciprocal ones in order to attract high effort answers. We model this strategic tipping out of reputation concerns as a Bayesian updating process in the style of the literature pioneered by Kreps et al. (1982).

The Google Answers environment resembles a gift-exchange game, specifically a three-stage design: 1) principal sets wage, 2) agent chooses effort level, 3) principal decides on bonus/tip. Hence, our study enables us to check the external validity of the results of related lab experiments. Fehr et al. (1997) and Fehr et al. (2007) employ such a three-stage design in order to study labour relations between firms and workers in the lab. Several other studies test a gift-exchange labour market context in the field, for instance List (2006), Gneezy and List (2006), Kube et al. (2006), Maréchal and Thöni (2007), Bellemare and Shearer (2009) and Hennig-Schmidt et al. (2010). They all focus on a two-stage design (i.e., principal sets wage, agent chooses effort level) and results with respect to a positive wage-effort relationship are mixed. Our study is also related to Gächter and Falk (2002) who study the interaction effects of reciprocity and repeated game incentives.

Our real life findings are distinctly in favour of reciprocity and reputation. In line with the existing experimental evidence, the tendency to tip is positively correlated with effort and a user’s total number of questions posted. Furthermore, our rich data set allows for an in-depth analysis of the underlying mechanisms behind the interplay of reciprocity and reputation as well as its consequences for contract design. We test whether GARs take the past tipping behaviour of users into account and put more effort into the answer, if the user has frequently tipped before. We also analyse overall efficiency, that is, whether the repeated bonus design that allows for tips results in higher effort and adequate GAR compensation for the extra effort via the tip. In addition, we shed some light on the adoption process of tipping and estimate the fractions of genuine reciprocators and imitators in the sample population. Finally, we discuss implications of our findings for the online world.

Other studies using Google Answers data exist, but they focus on GARs. Rafaeli et al. (2007) and Raban (2008) focus on the social incentives for GARs to work on an answer. Harper et al. (2008) investigate predictors of answer quality. Chen et al. (2010) find that GARs with a higher reputation provide significantly better answers. Edelman (2012) analyses labour market aspects like GARs’ experience, on-the-job training and specialisation. Instead, we analyse the data from both GAR 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 analyse 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 develops a simple model of the user-GAR interaction and derives testable hypotheses. Section 4 describes our data set, while Section 5 analyses it. Section 6 concludes.

2. The online service Google Answers

The web-based service Google Answers (http://answers.google.com/) complements Google’s well-known standard search tool. It offers the assistance of expert online searchers to users who are willing to pay for this. 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. Google Answers can be seen as a fee-based expert service in contrast to community-based services like Yahoo! Answers (a free answer service where users both ask and answer questions). According to Google GARs are screened to ensure they are expert searchers with excellent communication skills. The focus is on quality provided by paid, freelancing experts.

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 $.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 a GAR and no other GAR can answer it in that time. The GAR will try to

---

1. Fehr et al. (1997) 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 last (strong reciprocity) treatment and a contract that gives the opportunity for mutual reciprocity was found to improve efficiency. 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.

2. Google Answers closed in 2006. It is a natural question – although not central to our analysis – why the service was stopped and why Yahoo! Answers had been more popular in terms of question volume. There is no official explanation by Google, but the topic has been discussed in length online, probably best accessible at http://usecue.com/xqj-2452. Important for the purpose of this study is that the closing of Google Answers cannot be regarded as a failure of the service. Google Answers and Yahoo! Answers are similar at surface but hardly comparable (users receive researched facts at Google Answers, while they basically get opinions by peers at Yahoo! Answers). See also Harper et al. (2008).
obtain the requested information and will post the 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 re-posted or apply for a refund.\(^5\) 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 within the service. Finally, users can tip the answer of the GAR. This 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.

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. Model

We model the interaction between users and researchers (GARs) as a repeated game with incomplete information about the user type.\(^6\) The stage game is characterised by a Principal-Agent relationship with moral hazard. A GAR answers the question of a user. 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 are long-lived players and, depending on their type, may consider the repeated nature of the game. While reciprocal users always reciprocate (high effort with a tip, low effort with no tip) and myopic selfish ones never tip, strategic selfish types may imitate reciprocal types in order to maintain a good reputation. GARs know that users of each type exist, but they do not know which type they face.

First, we analyse equilibrium behaviour in the stage game under (i) self-interest and (ii) when concerns for reciprocity are considered. Then, we analyse the repeated game.

3.1. Stage game

The interaction between a user and a GAR consists of three stages. First, the user posts her question and sets the price \(p\). Then, the GAR who accepted the question chooses his effort level. We simplify and allow only a discrete choice of effort \(e \in \{e^0, e^l, e^h\}\). No effort at all means the GAR shirks \((e^0 = 0)\), a low effort level is denoted by \(e^l\) and a high effort level by \(e^h = 1\). The cost of effort is simply expressed by the effort level itself. We assume that there is no randomness, that is, the value of an answer depends solely on the effort level with the following functional properties: \(v = v(e), v' > 0, v'' < 0, v(e^h) - e^h > v(e^l) - e^l > v(0)\). The last condition guarantees that a high effort level is efficient. Finally, the user has the option to reward the GAR with a tip. The binary variable \(\tau \in \{0, 1\}\) denotes, whether a tip has been given or not. The user can also punish the GAR by rejecting the answer which can be seen as a fine \(f\), because such an incident has a negative effect on the GAR’s standing within Google Answers.

Assumption 1. The GAR’s effort level can be zero, low, or high: \(0 = e^0 < e^l < e^h = 1\).

Assumption 2. Either no tip is given or the tip amounts to \(\tau = 1\).

Denote \(\pi^u(e, p, \tau)\) or \(\pi^{\text{GAR}}(p, e, \tau)\) the monetary payoff of the user and of the GAR, respectively. Payoff functions are defined as follows (recall that GARs receive only 75% of the actual price of a question):

\[
\pi^u(e, p, \tau) = v(e) - p - \tau
\]

\[
\pi^{\text{GAR}}(p, e, \tau) = \frac{3}{4} p + \tau - e
\]

The following assumptions ensure that users benefit from high effort answers more than from low effort answers even if they give a tip, and that exerting high effort is beneficial for the GAR if a tip is given (keeping the payoffs of our game in line with the sequential prisoner’s dilemma game).

\(^5\) However, this is very rare. Only in .03% of all answers a refund was granted and the price was returned.

\(^6\) This strand of the literature on reputation started with the seminal work of Kreps et al. (1982), Kreps and Wilson (1982), and Milgrom and Roberts (1982). See Mailath and Samuelson (2006) for a survey of the literature on repeated games and reputation.
Assumption 3. When low effort is put in and no tip is given the user’s payoff equals zero: \( v(e^l) - p = 0 \).

Assumption 4. With high effort the payoff of the user is positive even when a tip is given: \( v(e^h) - p - \tau > 0 \).

Assumption 5. Putting in high effort is profitable for the GAR in case he receives a tip: \( \tau - (e^h - e^l) > 0 \).

Denote \( \Pi_{GAR}^G(e) \) the expected payoff of the GAR from effort level \( e \). When a GAR exerts zero effort, he incurs the fine \( f = 1 \) with probability \( 0 < q < 1 \). Recall that the fine is an imperfect instrument, that is, it cannot implement the efficient effort level. Rational GARs will put in enough effort to avoid getting fined. Hence, the exogenous parameters \( f \) and \( q \) enforce the low effort level:

\[
\Pi_{GAR}^G(e^l) = \frac{3}{4}p - e^l \geq \frac{3}{4}p - q - e^0 = \Pi_{GAR}^G(e^0)
\]

This incentive compatibility constraint determines the base effort level \( e^l \). Whether GARs exert effort beyond the base level depends on their belief about getting a tip as reward for high effort (we assume that a tip is never given following low effort by the GAR). Let \( \rho_0 \) be a GAR’s prior belief that the user will reciprocate high effort with a tip. Generally, the GAR’s effort choice can be expressed as follows. He chooses \( e^h \), if and only if his expected payoff after high effort, \( \Pi_{GAR}^G(e^h) \), is greater than the one after low effort, \( \Pi_{GAR}^G(e^l) \):

\[
\Pi_{GAR}^G(e^h) = \frac{3}{4}p - e^h + \rho_0 \cdot \tau > \frac{3}{4}p - e^l = \Pi_{GAR}^G(e^l)
\]

This implies that he only exerts high effort when his prior belief \( \rho_0 \) is high enough:

\[
\rho_0 > \frac{e^h - e^l}{\tau} = \rho^* > 0
\]

However, regardless of the effort of the GAR, under pure self-interest a user’s dominant strategy in the stage game is not to give a tip. If GARs expect users to be selfish and correctly anticipate that a self-interested user will never tip, then \( \rho_0 \) equals 0 and GARs will always put in low effort. Note that this is an inefficient equilibrium. Effort beyond the base level would be beneficial for the user, but she fails to incentivise the GAR to exert more costly effort, because she cannot credibly commit to high effort or enforce it by monitoring/punishment technology.

We now allow for social preferences and analyse the effects on the stage game’s equilibrium. As explained earlier we focus on concerns for reciprocity to model pro-social behaviour. Generally, people are considered to be reciprocal if they reward kind actions and punish unkind ones. We follow the sequential reciprocity model in Dufwenberg and Kirchsteiger (2004). Utility is expressed by the material payoff and an additional reciprocity term. ‘Reciprocity’ utility is created when the signs of kindness and perceived kindness match. Hence, reciprocation corresponds to responding to positive perceived kindness of someone with positive kindness of oneself, and to negative perceived kindness with negative kindness.

By applying sequential reciprocity theory – under certain conditions – users give a tip in the stage game (see Appendix A for an outline of how the sequential reciprocity equilibrium is determined). 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, that is, their sensitivity to reciprocity \( \alpha \) has to be large enough. Moreover, the GAR has to believe that the user’s \( \alpha \) is large enough in order to provide high effort in the first place. This result can be seen as a justification of the reputation model’s assumption that GARs know that users of each type, also reciprocal ones, exist.

### 3.2. Repeated game

We now analyse the repeated game. In each round one user, the long-lived player, and one short-lived GAR meet via a random matching protocol. The total amount of a user’s questions is finite and denoted by \( T < \infty \). The current question number of a user is marked with the index \( t \). Perfect monitoring is possible via a public history. GARs can observe the previous stage game tipping behaviour of users and they can also assess, whether the effort of other GARs has been high or low in order to tell if giving no tip by the user was justified or not.

GARs know that the following three types \( \theta \in \Theta \) exist among users: Reciprocal (\( R \)), strategic self-interested (\( S \)), and myopic self-interested types (\( M \)). Users of type \( R \) will always answer high effort with a tip: \( \Pr(\tau = 1 | e^h, \theta^R) = 1 \). Users of type \( S \) maximise their profit over the repeated game. They invest in reputation by tipping high effort answers as long as their future benefits from this reputation warrant the investment. That is, they are characterised by switching from tipping to not tipping, when a break even point (threshold time \( 1 < t^* < T \) has been reached and reputation investments are no longer profitable. The threshold time \( t^* \) is not common knowledge and we denote the GAR’s belief about whether the strategic user will continue to tip given high effort by \( \lambda = \Pr(\tau = 1 | e^h, \theta^S) \). Finally, users of type \( M \) always maximise their stage game profits, that is, they will never tip: \( \Pr(\tau = 1 | e^h, \theta^M) = 0 \).

The type of a user is not common knowledge and GARs’ beliefs about a user’s type are identified by \( \mu(\theta^R) + \mu(\theta^S) + \mu(\theta^M) = 1 \). The GAR’s prior belief that the user is, for instance, a reciprocal type, is denoted by \( \mu_0(\theta^R) > 0 \) and the posterior belief at question number \( t \) given the user has tipped by \( \mu_t(\theta^R | \tau = 1) \) and by \( \mu_t(\theta^R | \tau = 0) \) if she has not.
3.2.1. Belief updating of GARs

When GARs face a user who has already received an answer, they can update their belief $\mu_t$ about her type based on her decision to tip or not. Consider a GAR faces a user at $t = 2$.

If the user did not leave a tip for the (high effort) answer to her first question, then $\mu_1(\theta^R | \tau = 0) = \mu_1(\theta^S | \tau = 0) = 0$. In contrast, if the user has tipped the answer to her first question, then the GAR updates his type belief in the following way:

$$
\mu_1(\theta^R | \tau = 1) = \frac{\mu_0(\theta^R)}{\mu_0(\theta^R) + \mu_0(\theta^S) \cdot \lambda}.
$$

Now consider a GAR meets a user (who tipped her first question) at $t > 2$. If the user stopped tipping high effort answers before $t$, the GAR identifies her as a strategic type who stopped tipping and updates his beliefs accordingly ($\mu_t(\theta^S | \tau = 0) = 1$, $\lambda = 0$). If the user has tipped high effort answers up to $t$, then the GAR – via repeated Bayesian updating of his belief – becomes increasingly convinced that the user is a reciprocal type. That is, $\mu_{t+1}(\theta^R | \tau = 1) = \mu_t(\theta^R | \tau = 1)$. The GAR’s belief of whether he is facing a reciprocal user increases with the amount of observed tipped high effort answers. See Fudenberg and Levine (1989) for a more formal analysis of this aspect.

As GARs take the past tipping behaviour of their user into account their decision rule for the GAR is the following. He chooses $e^h$ if and only if his belief about getting a tip is high enough:

$$
\rho_t = \mu_t(\theta^R) + \mu_t(\theta^S) \cdot \lambda > e^h - e^l = \rho^* > 0
$$

Specifically, this means a GAR does not believe the user will tip a high effort answer ($\rho_t = 0$), if the user did not tip the previous answer. As a consequence the GAR chooses low effort. If the GAR observed that the user always tipped in the past and his perceived chance of receiving a tip is greater than the critical value $\rho^*$, then the GAR decides to put in high effort.7

3.2.2. Reputation building of users

Users could have an interest in investing in a reputation of appreciating high effort and acknowledging it with a tip, especially if they use the service frequently. This way they may attract GARs who take them as reciprocal types and as a consequence will deliver high effort answers in anticipation of a tip. Such a motivation may be of particular relevance in online environments, since transaction partners do not see each other online (Resnick et al., 2000). Essentially, strategic self-interested users face a trade-off between foregoing profits in the short run in order to maintain a good reputation (by giving tips) and the future benefits of this reputation (continuous high effort answers).

A Perfect Bayesian Nash equilibrium of the game is characterised as follows. GARs only exert high effort, if their belief about receiving a tip is high enough (a frequent user’s immaculate tip history is a necessary condition for that). Otherwise, GARs put in low effort. Strategic self-interested users respond to high effort by giving a tip, if and only if their future gains outweigh the cost of tipping in the current period.

A user’s profit in future periods depends on whether she tips the high effort answer in the current period or not. When she tips in $t$ (and continues to do so), then she gets $\Pi^{\tau = 1}_m = v(e^h) - p - 1$ as payoff in each period $m > t$. When the previous answer has not been tipped, GARs update their beliefs, do not deliver high effort answers anymore and future profits drop to the low effort, no tip payoff of $\Pi^{\tau = 0}_m = 0$. Future profits are discounted by the factor $0 < \delta < 1$. The user decides to tip as long as her profits from tipping ($\Pi^{\tau = 1}_m$) exceed the profits, if she does not tip ($\Pi^{\tau = 0}_m$):

$$
\Pi^{\tau = 1}_m = \left( v(e^h) - p - 1 \right) + \sum_{m=t+1}^{T} \delta^{m-t} \cdot \left( v(e^h) - p - 1 \right) > \left( v(e^l) - p \right)
$$

$$
\Pi^{\tau = 0}_m = \sum_{m=t+1}^{T} \delta^{m-t} \cdot \left( v(e^l) - p - 1 \right) > 1
$$

It follows that in equilibrium a strategic user tips a high effort answer, if and only if the increase in future payoffs exceeds the cost of the one-off tip. Tipping requires that the number of remaining questions and the surplus from high effort answers are high enough, relative to the discount factor.

3.3. Hypotheses

Based on these theoretical implications of reciprocity and reputation concerns we derive the following null hypotheses that guide our empirical analysis:

7 Seinen and Schram (2005) find that observed records of cooperativeness of a player induce others to cooperate with him. Similarly, GARs would take the kindness of ‘their’ user towards other GARs into account, if they are also motivated by such indirect reciprocity.
**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 a repeated bonus design – providing mutual opportunities to reciprocate – encourages voluntary payments (tips) by single users and whether these tips are motivated by reciprocity.

**Hypothesis 2 (Reputation).** The total amount of a user’s questions has no effect on the users’ tendency to tip.

Turning to the repeated nature of the interaction, tipping out of strategic considerations hinges on the frequency of use.

**Hypothesis 3 (End game).** The tip rate in a ‘last period’-like situation is, ceteris paribus, higher than the tip rate of single users.

By looking at the ‘last period’ we distinguish between reciprocal and strategic selfish users who tip. The latter imitate reciprocal types until there is no more reputational benefit to gain, i.e., they approach their final question.

**Hypothesis 4 (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 their tipping behaviour or whether they tend to be either self-interested non-tippers or tippers (reciprocal or strategic). Both would tend to stick to their strategy or preference, respectively. In order to verify this classification, users who (do not) tip must have had a tendency (not) to tip in the past.

**Hypothesis 5 (Sorting).** The tip history of a user has no effect on the effort level of the GAR.

When different tipping patterns can be distinguished, GARs may inform themselves about a user’s tip history and update their belief about the user’s type. 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 should be answered with more effort, questions of users with a reputation for not tipping should be answered with less effort.

**Hypothesis 6 (Efficiency).** Effort levels do not increase significantly compared to when tipping was not possible (phase 1) or when it was not employed (frequent myopic users in phase 2).

Finally, we test, whether a repeated bonus design has a positive effect on efficiency (for both users and GARs) in Google Answers.

4. Description of the data set

All questions posted at Google Answers are archived and accessible online. A Perl script extracted this information. It produced sequential URLs to download the page of every possible question number in the range Google Answers used (1 to 787,274) and collected data for all existing questions.

The 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 .03%) were rejected by the user. Thus, actual transactions amount to 57,651.

The observations of our data set are generated by 31,120 different users. The majority of them just asked a single question, while the highest number of questions posted by the same user is 599. Users who asked more than one question stayed, on average, active for about eight months. Overall, there are 571 GARs and the most active answered 3591 questions. The average tenure of GARs is about 1.5 years. See Figs. 1 and 2 for more details.

We collected the following data for each answer: The user ID of the person who posted the question, the price set, the tip possibly given, 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, see also Fig. 3), the word count (the sum of answer and clarification) and the total number of questions posted (answered or not) by each user.

An essential aspect 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\(^8\) for the effort the GAR put into the answer. Users motivated by reciprocity will base their decision to tip on

---

\(^8\) It is a noisy signal as some chance is involved as well that determines the value of an answer to the user. Nevertheless, the user’s perception of the GAR’s effort will be based on the answer’s value.
their perception of the GAR’s effort. 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 an answer’s value and, in turn, perceived effort: (i) its information content (the more information, the better, as long as it is applicable) and (ii) its timeliness (the faster an answer is returned, the better). Of course, we cannot assess the quality of an answer’s content, but we have a precise quantity measure (the word count). We also know the time between posting of the question and posting of the answer.

Word count is the amount of words of an answer and its clarification. While more words do not necessarily mean higher effort or value of an answer, evidence in related studies points to such a general relationship.9 We still have to consider that some questions will be more complex than others, so they will demand more words. No one seems better suited than the user to rate a question’s difficulty via the price she attaches to a question. Therefore, we take the users’ perspective and use price as a proxy for the question’s difficulty.10 As more is expected for a more demanding and thus higher-priced

9 Harper et al. (2008) find that answer length (as well as the number of hyperlinks) is positively correlated with quality (judged by a panel of six college students) and statistically significant at the 1%-level. Edelman (2012) finds a positive and significant correlation between answer length and the users’ rating of the answer. Using Yahoo! Answers data Adamic et al. (2008) find that answer length alone achieves 62% prediction accuracy of whether an answer will be selected by the user as the best answer.

10 The findings of Chen et al. (2010) substantiate this approach. They find a correlation between price and average question difficulty of .46 (p < .01) in their sample of 200 Google Answers question–answer pairs. Question difficulty was rated by a panel of Library and Information Science students.
question, we normalise the word count with respect to the price of the question (a correlation coefficient of .32 confirms this relationship). Hence, word count-based perceived effort equals word count divided by price. Still, the more words GARs have included in answers of equally priced questions, the higher their effort has been. Additionally, we have a way to compare differently priced questions. An alternative indication of an answer’s value is the rating given by the user. When given, rating and perceived effort are positively correlated (Spearman, 1%-level).

We can compute a time-based perceived 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 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. Hence, time-based effort equals price divided by the time difference. However, the variable has to be taken with some caution. Our measure for time is the difference between posting of the question and posting of the answer and we do not know the time when 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. This issue complicates the use of time-based effort. The absolute value of the time difference between posting question and answer should nevertheless be useful to test for the effect of very late answers on the tendency to tip.

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

An intriguing feature of the data set is the late introduction of the option to leave a tip (in October 2002). The 6206 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, Table 1) and phase 2 (when tipping was available, Table 2).

The price range is pre-determined by Google Answers. The lowest price users can set is $2, the highest price possible is $200. The average price conditional on the question being answered (57,833 observations) is $22.84, while the average price of the 88,823 questions that expired without an answer is only $20.19, significantly less at the 5%-level based on a Mann–Whitney test. Controlling for period and categories it appears as if the price plays an important role in the GARs’ decisions to answer a question or leave it in the pool.

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

---

11 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 time a GAR can lock a question, which reduces the sample by 50%). However, we do not know if otherwise equal questions that are on average answered within 1 hour are sometimes found, locked and answered right away (total time 60 min) or sometimes found only after 3 h (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.

12 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. Over time more questions are left unanswered. In 2006 28% of all questions were answered in contrast to 53% in 2003. According to Rafaeli et al. (2007) comments increase the chance of a question being answered.
Table 1
Descriptive statistics of phase 1 (no tips possible).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Median</th>
<th>St. dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>6206</td>
<td>14.9</td>
<td>8</td>
<td>24.05</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>Rating</td>
<td>3581</td>
<td>4.39</td>
<td>5</td>
<td>.96</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Time difference [min]</td>
<td>6206</td>
<td>2306.64</td>
<td>156</td>
<td>7311.59</td>
<td>1</td>
<td>129,799</td>
</tr>
<tr>
<td>Word count</td>
<td>6206</td>
<td>479.76</td>
<td>330</td>
<td>589.11</td>
<td>3</td>
<td>17,047</td>
</tr>
<tr>
<td>Answer clarification</td>
<td>6206</td>
<td>.3437</td>
<td>0</td>
<td>.475</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Perceived effort</td>
<td>6206</td>
<td>55.96</td>
<td>35.64</td>
<td>78.39</td>
<td>.2</td>
<td>3409.4</td>
</tr>
</tbody>
</table>

Note: St. dev. = standard deviation; before July 2002 questions expired after one year.

Table 2
Descriptive statistics of phase 2 (tipping possible).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Median</th>
<th>St. dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>51,445</td>
<td>23.79</td>
<td>10</td>
<td>37.31</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>Tip</td>
<td>12,109</td>
<td>9.13</td>
<td>5</td>
<td>14.79</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Rating</td>
<td>32,429</td>
<td>4.66</td>
<td>5</td>
<td>.679</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Time difference [min]</td>
<td>51,445</td>
<td>2616.35</td>
<td>241</td>
<td>6915.5</td>
<td>1</td>
<td>43,198</td>
</tr>
<tr>
<td>Word count</td>
<td>51,445</td>
<td>619.90</td>
<td>349</td>
<td>1152.99</td>
<td>1</td>
<td>81,851</td>
</tr>
<tr>
<td>Answer clarification</td>
<td>51,445</td>
<td>.2976</td>
<td>0</td>
<td>.4572</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Perceived effort</td>
<td>51,445</td>
<td>51.75</td>
<td>30</td>
<td>83.04</td>
<td>.005</td>
<td>7792</td>
</tr>
</tbody>
</table>

Note: St. dev. = standard deviation.

are 5 and the average rating is 4.66. Due to the web site structure there is a high correlation between a rating being given and a tip being left. If a user decides to give 'feedback', he first has to enter a rating (1 to 5) and then decides on a possible tip (zero (or simply no entry) to 100). Hence, users can leave a rating without having to enter a tip, but they cannot leave a tip without rating an answer.

5. Analysis of the data

We now analyse phase 2 data starting with users’ tipping behaviour. We then turn to the GARs’ perspective and analyse the relationship between belief updating, effort decision and efficiency. Finally, we study how tipping was adopted when it became available six months after the start of Google Answers.

5.1. Estimations

Our theoretical model proposes two motivations as determinants of tipping behaviour. Reputation may matter. Frequent users of the service have an incentive to maintain a good reputation and may regard tipping as a strategic device. Social preferences could also motivate people to tip. Users who are socially-minded should leave a tip as long as there is a reason to reciprocate positively.

Reputation concerns are proxied by the frequency with which a user asked questions. 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. Therefore, the more questions posted the more generous users should be with the tip – simply out of strategic considerations. However, one additional question asked should have a diminishing impact on reputation concerns with the total number of questions increasing. Hence, we use the logarithmic value of the total number of questions posted by a user in order to take the decreasing importance of an increase of the total number of questions into account.

The following proxies take account of behaviour that indicates a reason for the user to positively reciprocate: the user’s perception of the GAR’s effort, whether the GAR provided a timely answer, and whether an answer clarification has been provided.

A reciprocal user will base the decision to tip on the perceived effort of the GAR, that is, how much value the answer created for the user. Perceived effort is metered in terms of word count relative to the price (in order 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. Again, it seems reasonable that one additional unit has a diminishing impact on perceived effort with the effort increasing. We take this into account in additional regression specifications that use the logarithmic value of perceived effort.

In order to express whether the GAR provided a timely answer, we consider the absolute value of the time difference between posting question. Ten days is used as the threshold level and the dummy equals one if the answer took less than 10 days.

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. This is already captured in the word count, though. 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.

Since tips are only possible between 0 and 100$ a censored regression model appears suitable to ensure unbiased and consistent estimates. The standard Tobit model assumes a single distribution function for the dependent variable (Amemiya, 1984). However, there is 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 separate ones. Different distributions could be underlying and a two-step model of Cragg (1971) will take this into account. 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 (all users, single users, frequent users) and confirms our approach. This Probit-truncated regression approach is model I. Table 3 shows the results for decision 1 (whether to tip) and Table 4 contains the results for decision 2 (size of the tip). A random effects unbalanced panel model accounts for the individual heterogeneity among users.13

The rating may play an important role for the binary choice of whether a user tips or not as both decisions are intertwined.14 Only rated answers can be tipped, yet there does not seem to be a selection bias in the relationship between rating and tipping in the sense of Heckman (1979). If a user wanted to tip an answer, nothing prevents that except having to rate the answer (a mouse click) which is likely negligible. Hence, we also estimate a bivariate probit model (II) for the binary decisions whether to rate and tip (Table 3).

Regression models I and II are based on maximum likelihood and they assume a normal distribution of the error term. A Bera–Jarque test rejected the normality assumption. Interpretation of interaction effects is also not as straightforward as in linear models (Ai and Norton, 2003). Therefore, we report results of a linear probability model (III) with robust standard errors as well. The obvious drawback of such a model is that it may predict probabilities beyond the range of 0 to 1 (Wooldridge, 2002). However, only two observations of our sample were mistreated in this way. The underlying function may not necessarily be linear, but at least the LPM does not provide theoretically wrong predictions. Its estimates can be regarded as more robust and interaction effects can be interpreted without complication.

All three models are fairly similar in their results. The data 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 perceived effort is as well positive (1% significance level) for the decision to tip. It also clearly matters whether an answer clarification has been given. The coefficients are positive and highly significant in all models. The coefficients of the timely answer dummy are positive and significant at the 1%-level.15 Based on model III there appears to be a significant

13 A focus on users is straightforward, but what about potential heterogeneity among GARs? This could also explain part of the tipping. However, Fig. 4 shows that the distribution of GARs with respect to their received tip ratio is fairly normal. Regression results are robust for a GAR panel model and also for a mixed approach that clusters by equal user-GAR combinations.

14 If a user wanted to leave a tip, she will have to give a rating, too (due to the sequential design). When she wants to rate a question, she does not have to leave a tip. Out of 51,445 phase 2 answers 32,429 have been rated. 12,109 (rated) answers have been tipped.

15 When do users regard an answer to be still timely? Further analysis shows that there is no significant difference in the tendency to get tipped between the 12,080 answers that are posted after one day and the ones posted within a day. However, there is a noticeable (and statistically significant) difference between the 8289 answers that are posted after two days (tip rate .22) and the ones posted within two days (.238). This gap widens to tip rates of .24 and .161 when the threshold is 10 days. When tested against each other in a stepwise regression approach, the dummy for an answer that was posted more than 10 days after the posting of the question is the only significant one. Dummies for smaller threshold values (5 days, 3 days or 2 days) are not significant. Hence, we decided to use 10 days as the threshold level for the timely answer dummy.
interaction effect between reputation and reciprocity. According to the year dummies there is an increase of the tip rate compared to 2002. Finally, 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 in models II and III.

We proceed with several robustness checks. The relationship between perceived effort and the tendency to tip may be restricted to particularly short answers, as users may be insensitive to length once the answer is sufficiently long. We have tested this by excluding answers with less than 50/100/200 words, but the significance of perceived effort remains. Are results driven by high-volume users? There are 37 users with more than 50 answered questions and they combine for 3268 questions. Excluding these superusers does not change regression results in a meaningful way (likewise if only users with less than 25 or 100 answered questions are considered). The correlation between tip and answer clarification suggests that the increased social interaction due to requesting and receiving an answer clarification has a positive effect on giving a tip. Alternatively, the correlation might also be due to tipping users being inherently more inclined to request a clarification. We look at superusers since their tipping behaviour is most established in order to test whether the answer clarification dummy is correlated with users’ overall tip rate. The relationship is positive but not significant. However, answer clarification and overall tip rate are correlated if lower limits (30 or less) are applied to the number of answered questions. Hence, we cannot exclude this possibility. The behaviour of frequent users might not be motivated by concerns about their public reputation but due to a close relationship with specific GARs. Even though a user’s question can be locked and answered by any GAR, users may be insensitive to length once the answer is sufficiently long. We have tested this by excluding answers with less than 50/100/200 words, but the significance of perceived effort remains. Are results driven by high-volume users? There are 37 users with more than 50 answered questions and they combine for 3268 questions. Excluding these superusers does not change regression results in a meaningful way (likewise if only users with less than 25 or 100 answered questions are considered). The correlation between tip and answer clarification suggests that the increased social interaction due to requesting and receiving an answer clarification has a positive effect on giving a tip. Alternatively, the correlation might also be due to tipping users being inherently more inclined to request a clarification. We look at superusers since their tipping behaviour is most established in order to test whether the answer clarification dummy is correlated with users’ overall tip rate. The relationship is positive but not significant. However, answer clarification and overall tip rate are correlated if lower limits (30 or less) are applied to the number of answered questions. Hence, we cannot exclude this possibility. The behaviour of frequent users might not be motivated by concerns about their public reputation but due to a close relationship with specific GARs. Even though a user’s question can be locked and answered by any GAR,

![Table 3](image)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Model</th>
<th>I: Probit</th>
<th>II: Bivariate probit</th>
<th>III: LPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tip</td>
<td>Tip</td>
<td>Rating</td>
</tr>
<tr>
<td>Price</td>
<td></td>
<td>−.0001</td>
<td>.0001</td>
<td>−.0012***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0002)</td>
<td>(.0001)</td>
<td>(.0001)</td>
</tr>
<tr>
<td>Frequency of use</td>
<td></td>
<td>.2886***</td>
<td>.2139***</td>
<td>.3225***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0143)</td>
<td>(.0035)</td>
<td>(.0036)</td>
</tr>
<tr>
<td>Perceived effort*</td>
<td>FreqOfUse</td>
<td>.0004***</td>
<td>.00005***</td>
<td>−.0002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0001)</td>
<td>(.00002)</td>
<td>(.00002)</td>
</tr>
<tr>
<td>Perceived effort</td>
<td></td>
<td>.0004***</td>
<td>.0009***</td>
<td>.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0001)</td>
<td>(.00006)</td>
<td>(.00006)</td>
</tr>
<tr>
<td>Timely answer</td>
<td></td>
<td>.3682***</td>
<td>.3243***</td>
<td>−.5896**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0423)</td>
<td>(.024)</td>
<td>(.0241)</td>
</tr>
<tr>
<td>Answer clarification</td>
<td></td>
<td>.2987***</td>
<td>.2678***</td>
<td>.3273***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0192)</td>
<td>(.0114)</td>
<td>(.0114)</td>
</tr>
<tr>
<td>Arts/entertainment</td>
<td></td>
<td>.2577***</td>
<td>.2516***</td>
<td>.3163***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0342)</td>
<td>(.0189)</td>
<td>(.0189)</td>
</tr>
<tr>
<td>Business/money</td>
<td></td>
<td>−.0459</td>
<td>−.0437***</td>
<td>.0405**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0325)</td>
<td>(.0163)</td>
<td>(.0162)</td>
</tr>
<tr>
<td>Computers</td>
<td></td>
<td>.0288</td>
<td>.0405**</td>
<td>.1345***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0323)</td>
<td>(.0176)</td>
<td>(.0176)</td>
</tr>
<tr>
<td>Family/home</td>
<td></td>
<td>.0266</td>
<td>.0398</td>
<td>.0632</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0505)</td>
<td>(.0297)</td>
<td>(.0296)</td>
</tr>
<tr>
<td>Health</td>
<td></td>
<td>.0254</td>
<td>.0345</td>
<td>.0599***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0401)</td>
<td>(.0223)</td>
<td>(.0222)</td>
</tr>
<tr>
<td>Reference/education/news</td>
<td></td>
<td>.1177***</td>
<td>.1112***</td>
<td>.1542***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0347)</td>
<td>(.0187)</td>
<td>(.0187)</td>
</tr>
<tr>
<td>Relationships/society</td>
<td></td>
<td>.1788***</td>
<td>.1799***</td>
<td>.1807***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0459)</td>
<td>(.0279)</td>
<td>(.0278)</td>
</tr>
<tr>
<td>Science</td>
<td></td>
<td>.0472</td>
<td>.0379</td>
<td>.1146***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0421)</td>
<td>(.0223)</td>
<td>(.0222)</td>
</tr>
<tr>
<td>Sports/recreation</td>
<td></td>
<td>.0276</td>
<td>.0388</td>
<td>.1669**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0542)</td>
<td>(.031)</td>
<td>(.0309)</td>
</tr>
<tr>
<td>2002</td>
<td></td>
<td>−.3754</td>
<td>−.3142***</td>
<td>.0075</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0347)</td>
<td>(.0165)</td>
<td>(.0164)</td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td>.1339***</td>
<td>.1287***</td>
<td>.1541***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0255)</td>
<td>(.0135)</td>
<td>(.0134)</td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td>.1911***</td>
<td>.1731***</td>
<td>.1958***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0284)</td>
<td>(.0142)</td>
<td>(.0142)</td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td>.0485</td>
<td>.0621***</td>
<td>−.0435**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0319)</td>
<td>(.0156)</td>
<td>(.0156)</td>
</tr>
<tr>
<td>Constant</td>
<td></td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Log likelihood</td>
<td></td>
<td>−24,085.46</td>
<td>−48,760.08</td>
<td>rho = .99***</td>
</tr>
</tbody>
</table>

*N = 51,445; random effects unbalanced panel models, 31,120 users; marginal effects reported, standard errors in parentheses; statistical significance: * = 10%, ** = 5%, *** = 1%.

Author's personal copy
it might still be possible that user-GAR partnerships form that interact repeatedly. Among the 37 superusers it is rare that one GAR answered more than 20% of a users' questions. The few cases of a close relationship between a user and a GAR do not seem to be unusual with respect to the tipping behaviour.

Finally, how substantial are the estimated effect sizes in an economic sense? The entire data set contains many transactions of users who had no intention to tip, independently of the perceived effort, and is therefore not particularly informative in order to assess the economic significance of the effect of the perceived effort on the tip. Instead, we look at superusers known for tipping (reciprocal and strategic users; 2520 observations) and use a regression specification with the logarithmic value of perceived effort in order to facilitate interpretation. The coefficient of perceived effort is .06 and significant at the 1%-level. A variation of perceived effort in the size of one standard deviation (1.15) is estimated to result in a change of the likelihood to receive a tip of approximately 7%. Running individual regressions the coefficients range from .09 to .49 among superusers (given significance of at least 5%) and, hence, the predicted effect of a one standard deviation change is up to 60%.

16 For comparison, the corresponding coefficient of perceived effort for the entire sample is .02 (significant at the 1%-level). The coefficient for superusers categorised as myopic is .01 (748 observations; not significantly different from zero).
Next, we analyse what determines the size of the tip (Table 4). For the maximum likelihood-based models I and II we run a truncated regression given a user has tipped.\(^{17}\) Continuing the linear approach of model III we run a simple GLS panel regression when a tip was given. The truncated regression confirms the importance of the price for the size of the tip (1% significance level). While the main effects of perceived effort and frequency of use are not significant in model I/II, their interaction is significant. In model III the main effects are significant but the interaction is not. The answer clarification dummy is again highly significant.

Users decide to tip an answer based on reciprocity and reputation concerns, but it seems less clear whether they use these factors as guidelines for the size of the tip. Instead, the question’s price appears to play a significant role. A possible explanation could be that ‘price orientation’ is the simplest heuristic available for determining the size of the tip as it is common in many service professions.\(^{18}\) While the perception of effort is a sufficiently precise signal to evaluate whether a tip should be given or not, it may not be a clear enough signal to determine the exact size of the tip that should be given. A well-established and easy-to-use alternative procedure seems to be preferred. The answer clarification, i.e., a higher degree of interaction with the GAR, can also be regarded as a clear signal.

5.2. Reciprocity

In order to focus on reciprocity we control for reputation concerns and analyse the behaviour of single users. 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.\(^{19}\) Regressions only with single users for models I to III deliver similar results as the main regressions. Perceived effort is statistically significant at the 2%- (I), 6%- (II) and 2%- (III) level and the answer clarification dummy is always 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 cannot 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 9650 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.\(^\text{20}\) 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 positively correlated with effort. This rejects Hypothesis 1.

**Result 1.** A substantial amount of single users tips and the tendency to tip is positively correlated with effort.

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

\(^{17}\) Clustering standard errors by user is employed to account for arbitrary forms of serial correlation and heteroscedasticity among observations for each user. This procedure ignores potential serial correlation across observations for the same GAR but different users. However, results are robust for clustering by GARs and for a mixed approach that clusters by equal user-GAR combinations.

\(^{18}\) Conlin et al. (2003), 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).

\(^{19}\) See Table 5 in the next subsection for the data about single users in comparison to occasional and frequent users.

\(^{20}\) While perceived effort has apparently no effect on the drop out rate of first time users, it appears the time to get an answer plays a role. Fig. 5 compares single users’ answers (21,512) to the first answer all other users received (9650). The spike at the very end of the time scale indicates questions that have been answered just before the expiry deadline of 30 days. Users who quit after the first question (single users) experienced a very late reply more often it seems.
at a significant level are also observed in another field study where reputation effects can be ruled out (Regner and Barria, 2009).

5.3. Reputation concerns

In order to gain more insights about 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. This should result in a better proxy of how often a user intends to use the service than the number of answers he actually received. Table 5 shows the pricing and tipping behaviour of users clustered by the amount of questions they posted. 14.87% of all single users gave a tip. However, with increasing number of questions posted we observe an almost 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 around 35% for frequent users (10 or more questions posted). Table 5 also shows the respective significance levels of tip rate comparisons between one row and the row below. The tip rate for single users is different from the rate when two questions were posted (1%-level). 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 eight questions.

These findings confirm the regression results and lead us to conclude that occasional users already take reputation concerns into account. For frequent users reputation concerns seem to matter even more. In fact, the tip rate is positively correlated with the frequency of use, which rejects Hypothesis 2.

Result 2. The tendency to tip increases with the frequency of use.

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 is expected to break down. For this purpose, we look at the individual end games of users in order to test this. We distinguish between the last question of each user and all previous ones. The tip rate of occasional users’ last questions (.24) is significantly lower than their questions before (.28, ranksum test, p < .01). To a smaller extent this is also the case for frequent users (tip rate of .34 for the last question compared to .38, p < .08). However, also the effort level of the last questions’ answers is significantly lower (ranksum tests, p < .01) both for occasional and frequent users. The lower tip rate for the last question may be an indication that frequent strategic users stop to tip, because they have no more reputational gains to reap. But the behavioural pattern is also in line with users stopping to use the service, because they are disappointed about the effort of the last received answer.

An alternative way to analyse the possible fading of reputation concerns is the actual end of Google Answers itself. On November 28th, 2006, Google officially announced that the service would stop accepting new questions at the end of November (answers could be given until the end of 2006). In the remaining three days 158 questions have been asked (by 127 different users), an average of 52.67. Given that for the previous 27 days of November 2006 25.4 questions per day were posted, it seems that users were aware of the upcoming end and that it would make no sense anymore to invest in a good reputation by tipping answers. Until the end of 2006 a total of 316 questions have actually been answered (by 34 different GARs), many of them throughout December. Thus, it seems that during the end game GARs turned to questions that lingered in the pool already for some time, simply because no new questions were coming in. We know from previous

![Fig. 5. Histogram of the time to get answer of first question (time difference > 1000 minutes): single users (top) and all others (bottom).](image-url)
analysis that very late answered questions are treated differently by users. Hence, in order to have a meaningful comparison between the end game and the main body of data, we only look at answers that have been given within 10 days. Applying this restriction to the end game reduces observations to 40% or 129 (for comparison, 95% of phase 2 answers were given within 10 days). In this sample 15% of the 61 single users tipped, while multiple users received 68 answers and tipped 44% of them.21 The rate is 32% if we consider only the last questions of the multiple users (40 of 68). The tip rate of frequent users does not fall back to the level of single users, not even when it is the very last question. There does not seem to be a 'last period'-like effect in this natural setting. There is no difference in average effort for users' final, post-announcement answers compared to the previous answers they got.

The evidence about a drop of the tip rate in the individual end games of users is inconclusive, and in the overall end game of Google Answers tipping appears to be robust. Hypothesis 3 cannot be rejected as in a 'last period'-like situation the tip rate does not drop to the level of single users.

Result 3. The tip rate of frequent users in their last question is higher than the tip rate of single users.

5.4. Behavioural types

Subtracting the baseline of single user tipping (15%) from the level at which frequent users tip (35%) provides us with an estimate for the fraction of strategically motivated tippers. Around 20% would then imitate genuine reciprocal types out of reputation concerns in order to receive high effort answers. These transactions-based estimates for reciprocal (15%) and self-interested strategic types (20%) must be regarded as minimal values.22 Looking at the tip history of frequent users gives us some insight about the prevalence of self-interested myopic types. At 10 previously answered questions 73 of 324 users never tipped before. At 20 previously answered questions 29 of 128 users tipped not more than once. It appears that 25% is a reasonable estimate for self-interested types who are not aware of the benefits of employing tipping as a strategic tool.

The analysis of the 37 superusers provides us with an alternative way to estimate the distribution of types. With at least 50 observations available superusers’ tipping behaviour is well established. Following the model’s types we categorise them as myopic selfish users, if their overall tip rate is negligible (smaller than .1). Moreover, we distinguish reciprocal types from strategic selfish ones, if they have tipped the answer to their last question.23 This results in 10 myopic selfish (27%), 13 strategic selfish (35%) and 14 reciprocal (38%) superusers. In contrast to the transactions-based estimates that provided lower bounds these user-based estimates can be regarded as a point estimate.

Since superusers have an established pattern of behaviour, they are ideal for additional in-depth analysis of the relationship between the tendency to tip and perceived effort. We study their behaviour across types and at the individual level. In Table 6 we present results of LPM regressions of the superuser data for each of the three types. Regressors are those already introduced except frequency of use. As expected, in the subsample of myopic users (column I) the coefficient of perceived effort is not significantly different from zero. Column II contains results of the strategic user subsample and column III of reciprocal users. Perceived effort is significant for strategic (5%-level) as well as for reciprocal users (1%-level). The coefficient size for reciprocal users is approximately three times bigger, though. It appears that the relationship between the tendency to tip and perceived effort is much more pronounced among reciprocal users.

A possible explanation for this gap is that strategic users may have a more pragmatic approach towards tipping than reciprocal users.24 Fig. 6 provides scatterplots for all superusers to illustrate the relationship between the tendency to tip and perceived effort on an individual level. The diagrams are grouped by user type. Most users categorised as reciprocal seem to respond to perceived effort when they decide whether to tip. The situation is less clear for strategic users. While

---

21 The numbers are simply lower, if the entire end game sample is considered. The tip rate of single users is 7% and the one of multiple users is 27%.
22 The estimates are conditional on high effort being exerted, since genuine reciprocators would not tip when they receive a low effort answer. There are at least two more reasons why the level of 15% should be considered a lower bound measure for the genuine reciprocator type. Single users experienced a high amount of very late answers (see footnote 21). Our analysis suggests that such delays in receiving an answer have a detrimental effect on the tendency to tip. Hence, the level of genuine reciprocators in the population is probably higher than the 15% single user tip rate. Moreover, eventual frequent users could self-select into using the service more often, because they like aspects of Google Answers (beyond their first question’s answer quality and timeliness) more than single users. Relatedly, eventual frequent users may have lower expectations of answer quality/timeliness than single users and, hence, their expectations were met more often. If this is the case, the level of genuine reciprocators would also be underestimated.
23 Reciprocal users might not have tipped, because perceived effort was low. We check whether effort of the last question’s answer is within or below the 5% confidence interval for effort predicted by previously untipped answers and not within the one predicted by previously tipped answers. This applies to five users. They are categorised as strategic, because they did not tip their last question’s answer. The effort received would not have warranted a tip, though.
24 Strategic users tip in order to appear as reciprocal and, in turn, attract high quality answers. Since they are not driven by genuine reciprocal concerns, they may be pragmatic about an answer’s quality as long as tipping the answer conveys to GARs that they could be a reciprocator. Tipping not only high quality answers could make sense, since one could potentially signal faster that one could be a reciprocal user (without having to wait for a high quality answer). If past answers’ quality is perfectly observable, GARs would be able to detect that a user who tipped low quality answers engaged in strategic tipping and this approach would not work. However, it seems unlikely that GARs are able to verify the corresponding answer quality of a user’s tipping behaviour. Hence, it seems plausible that (some) strategic users imitate reciprocal users on the behavioural level (tipping), yet are less strict about adhering to the underlying motivation (being reciprocal, that is, tipping high effort answers), because deviating is virtually impossible to detect and may be advantageous.
Table 6
Decision 1 (whether to tip) of superusers.

<table>
<thead>
<tr>
<th>User type</th>
<th>I: Myopic</th>
<th>II: Strategic</th>
<th>III: Reciprocal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>−.0005</td>
<td>−.00001</td>
<td>−.0029**</td>
</tr>
<tr>
<td></td>
<td>(.0003)</td>
<td>(.0004)</td>
<td>(.0004)</td>
</tr>
<tr>
<td>Perceived effort</td>
<td>.044</td>
<td>.1325**</td>
<td>.0008***</td>
</tr>
<tr>
<td></td>
<td>(.0009)</td>
<td>(.0002)</td>
<td>(.0001)</td>
</tr>
<tr>
<td>Timely answer</td>
<td>.0578</td>
<td>.0636</td>
<td>.0226</td>
</tr>
<tr>
<td></td>
<td>(.0003)</td>
<td>(.0004)</td>
<td>(.0004)</td>
</tr>
<tr>
<td>Answer clarification</td>
<td>.0591***</td>
<td>.0823***</td>
<td>.1005***</td>
</tr>
<tr>
<td></td>
<td>(.019)</td>
<td>(.0279)</td>
<td>(.0263)</td>
</tr>
<tr>
<td>Constant</td>
<td>.306***</td>
<td>.3574**</td>
<td>.8369***</td>
</tr>
<tr>
<td></td>
<td>(.0208)</td>
<td>(.0384)</td>
<td>(.0382)</td>
</tr>
<tr>
<td>(R^2)</td>
<td>.03</td>
<td>.09</td>
<td>.19</td>
</tr>
<tr>
<td>Observations</td>
<td>748</td>
<td>1191</td>
<td>1329</td>
</tr>
</tbody>
</table>

Random effects unbalanced panel models, regressions include category and year dummies; statistical significance: * = 10%, ** = 5%, *** = 1%.

Fig. 6. Strategic (left) and reciprocal (right) superusers’ scatter plots of the decision to tip and perceived effort (logarithmic scale) with fitted values and 95% confidence values. Dots are jittered to avoid overlay.

for some a positive relationship between giving a tip and perceived effort is evident, for others the relationship appears to be modest if it is positive at all.

5.5. Updating, effort decision and efficiency

This subsection tries to shed more light on the decision making of GARs. They may update their beliefs about the likelihood 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 answers she received at that point. Recall that this information is not very straightforward to obtain for the GARs.25 Table 7 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.

25 It is not shown next to the user name as the past average like, for instance, the rating of GARs is 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).
When we consider the respective tip history of each user at each question number we see in Table 8 that there is a large spread between tipped and untipped questions. Naturally, no tip history exists 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. Hypothesis 4 can be rejected. 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. It seems that users have preferences or a strategy to tip (high effort answers) or not and they stick to it.

Result 4. Tipping users had a tendency to tip in the past.

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 effect of the tip history to be moderated by the number of past questions. A GLS panel regression (Table 9) confirms this.

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 logarithmic value of the question index.

Table 10 shows the effort levels of GARs (time-adjusted with 2002 as the baseline to control for the price increase over time and allow comparisons with phase 1). 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 higher for the tipped answers just as we should expect it since we know that effort is positively correlated with the tip. The split between questions with and without tip widens in the intermediate range and even more in the high range (Mann–Whitney tests, both 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 10
Question Nr. and perceived effort (time-adjusted).

<table>
<thead>
<tr>
<th>Question Nr.</th>
<th>Pre OCT 2002</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6098</td>
<td>55.77</td>
<td>55.77</td>
</tr>
<tr>
<td>Obs</td>
<td>Average perceived effort</td>
<td>Without tip</td>
<td>With tip</td>
</tr>
<tr>
<td>first</td>
<td>31120</td>
<td>73.39</td>
<td>71.52</td>
</tr>
<tr>
<td>2nd to 9th</td>
<td>14,256</td>
<td>73.8</td>
<td>68.58</td>
</tr>
<tr>
<td>10+</td>
<td>6069</td>
<td>83.72</td>
<td>75.33</td>
</tr>
</tbody>
</table>

Table 11
Question Nr. and pay (price + tip).

<table>
<thead>
<tr>
<th>Question Nr.</th>
<th>Without tip</th>
<th>With tip</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>23.93</td>
<td>23.94 (+8.28)</td>
</tr>
<tr>
<td>2nd to 9th</td>
<td>24.80</td>
<td>25.57 (+10.21)</td>
</tr>
<tr>
<td>10+</td>
<td>20.98</td>
<td>19.04 (+9.28)</td>
</tr>
</tbody>
</table>

Result 5. GARS’ effort choices depend on past behaviour of the user.

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. The fifth hypothesis can be rejected.

Table 10 also shows the effort level during phase 1 when tipping was not possible. There is a general increase of the effort level after the introduction of the tipping option. The repeated bonus design with its mutual opportunities to reciprocate can lead to a significantly higher effort level compared to the conventional design without mutual 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). Do high effort answers make users better off? It seems reasonable to assume that this is the case since users would not voluntarily give away as a tip more than they actually benefited from the answer. 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. Hence, does it pay off for GARS to put in high effort, when they work on questions of users who are known for tipping?

Table 11 provides mean prices for questions with and without tip as well as mean tips if applicable. There is no indication that questions with and without tip are priced differently (ranksum test for question number greater than nine, \( p < .03 \)). Moreover, the mean tip ranges between roughly a third and one half of the question’s price. It seems that, on average, GARS are rewarded substantially for providing higher effort.

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.

Result 6. The repeated bonus design increases effort level and efficiency.

It seems that the repeated bonus design encourages socially-minded users to reciprocate (tipping high effort answers) and that it makes self-interested users consider maintaining 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 6.

5.6. 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 (question price $10, tip $1), the second on the 9th ($40, $30) and the third tip on the 10th ($5, $3). Three more tips followed before the 20th. Only in the last third of October users slowly warmed up and started to tip more often as can be seen in Fig. 7. In total, 71% of the 1942 answers in October 2002 were tipped. However, tipping gained momentum rather quickly. In November 2002 18.63% of 2459 answers were tipped. In the following months tipping already reached the level known from the overall numbers: 19.26%, 20.76%, 24.38%, 23.90%, 21.31%, 25.80% and 23.40% (from December 2002 to June 2003).

Who are the users that ‘kickstarted’ tipping? The very first user who tipped an answer became a superuser (89 total questions posted) with an overall tip rate of .49. Two others turned out to be single users. The remaining three users who
tipped an answer during the first 20 days ended up posting a total of 4, 8 and 9 questions, respectively, with overall tip rates of .25, .25 and .78. What motivated the behaviour of users and GARs in this adoption phase? We look at users’ second question with an answer, that is, the first question of a user with past tipping behaviour available. In October 2002 there are 167 such questions and for 10 of them the user gave a tip in her first question. Tip rates are .03 (no tip given previously) and .3 (tip given). In November 2002 out of 309 ‘second’ questions of a user, 57 were preceded by a tipped question. Tip rates are .15 (no tip given previously) and .53 (tip given). It seems that from the very beginning users stick to their preference/strategy.

GARs, however, are slow to respond to the tip history signal. They exert, on average, more effort if a user’s first question was tipped (in October 61, in November 77) than if she did not tip (56, 72), but the difference is not statistically significant. A regression on the data from October 2002 with the same specification as in the main model shows that the answer clarification dummy is significant at the 1%-level but no significant effect of effort nor the frequency of use. Price explains the size of the tip at the 1%-level. In a regression for November 2002 the familiar significance of perceived effort and the frequency of use appear in addition to the described significance of the answer clarification and price. It seems that once some users reciprocated – not the least motivated by social interaction with the GAR, that is, an answer clarification – imitators react quickly and adopt tipping as an instrument. Tipping users stick to their preference/strategy even though the data indicates that GARs hesitate initially to increase their effort given the user tipped the first question.

6. Conclusions

We investigate the real-life pricing, effort and tipping decisions using all available data from the online service Google Answers (57,651 transactions). This rich data set puts us in a position to test the interplay of reciprocity and reputation in a real life environment. In particular, we are interested in the underlying motivations for the occurring voluntary payments and the efficiency of such a repeated bonus design. We relate our findings to the theory of sequential reciprocity of Dufwenberg and Kirchsteiger (2004). Applied to the Google Answers context, intentions-based reciprocity predicts that tipping takes place even among single users, if they are sufficiently sensitive to reciprocity. Considering the repeated nature of the interaction between users and researchers (GARs) the existence of reciprocal types facilitates tipping out of reputation concerns among strategic self-interested users in the style of reputation models based on Bayesian updating pioneered by Kreps et al. (1982).

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 tendency to tip is correlated with reciprocity proxies (‘Effort of the GAR as perceived by the user’, ‘Timeliness of the answer’, ‘Has an answer clarification been provided?’) and reputation proxies (frequency of use). These results confirm the model’s predictions and complement results observed in the lab.

The positive relationship between effort and tip is in line with the experimental evidence of related three-stage gift-exchange games (Fehr et al., 1997, 2007). Their bonus games feature a third stage where a bonus can be given, in addition to the wage and effort stage of standard gift-exchange games. Lab evidence of a positive wage-effort relationship in the gift-exchange game abounds, yet evidence in favour of reciprocity in the field is harder to find.26 Given these mixed results

---

for the two-stage design and our distinctly positive results for the three-stage design, it seems to be important to provide mutual opportunities to reciprocate, i.e., a three-stage design, in order to reap the benefits of reciprocity. The higher tip rates of frequent users are in line with the experimental evidence in Gächter and Falk (2002) in favour of reputation concerns.

In addition, we find that GAR effort depends on a user’s past tipping behaviour supporting the model’s prediction of belief updating. The effect of reciprocal behaviour and repeated game incentives appear to be complementary. Our findings shed some light on the interplay between reciprocity and reputation and why it is beneficial. Users tend to either stick to their preferences/strategy to tip high effort answers or they generally do not tip. GARs seem 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. It seems that the repeated bonus design with its mutual opportunities to reciprocate functions like a virtuous circle that alleviates moral hazard and increases efficiency.

Principal-agent theory traditionally propagates explicit monetary incentives in order to motivate agents, while only recently non-monetary instruments emerged as an alternative. In the repeated bonus design the voluntary component of the tip forms the basis as it allows principal as well as agent to reciprocate. This leads to an efficient result, but only if both principal and agent are sufficiently motivated by reciprocity. The dynamic aspect of the possibility of reputation building extends the appeal of a voluntary payment in order to motivate agents beyond reciprocal types. This combination of ‘behavioural’ and standard instruments may prove to be a particularly attractive alternative to alleviate moral hazard problems in principal-agent situations.

It seems that two conditions are essential to reap the benefits of a repeated bonus design in a moral hazard context. First of all, the existence of genuine reciprocators is crucial. Without them strategic types have no one to imitate and the positive feedback loops of mutual opportunities to reciprocate would not even start. Moreover, agents need to be able to update beliefs about principal types. Only then the strategy of imitators pays off and they attract high effort.

What are the implications of our findings? When principals and agents meet online, the high degree of anonymity complicates their relationship as traditional cues to foster trust are not available anymore. At first glance, it seems moral hazard should be more of a concern in the online world. Yet, online environments are better suited to implement transparent reputation systems that provide easy access to a comprehensive history of past transactions and, in turn, allow belief updating. Hence, voluntary payments, although so far a rather unusual practice in the online world, could be a suitable instrument in moral hazard contexts, especially if they are systematically reinforced by an environment that allows for the signalling of reputation. Markets for online labour (see Horton, 2010) are a natural candidate for connecting principals and agents in a repeated bonus design, but the scope of combining voluntary components and reputation mechanisms in order to combat moral hazard goes beyond.

What about the proportions of reciprocal, strategic and myopic types? Do genuine reciprocators exist and what is the potential for strategic imitation? Our transactions-based analysis suggests minimum levels for reciprocators and imitators based on the tip rate of single users (around 15%) and the difference between the tip rate of frequent (around 35%) and single users. Our categorisation of 37 superusers provides point estimates. With no reason to believe that the Google Answers sample population is not representative, we suggest estimates for the genuine reciprocator type of 38%, of 35% for the strategic self-interested type and of 27% for the myopic selfish type.

Finally, 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 gives us the means to analyse the adoption process. During the first 6 months of the service no tipping was possible and the behavioural default has been not leaving a tip. After a slow start (6 out of the first 1,000 answers were tipped) in October (only the answer clarification dummy is significant), reciprocity and reputation proxies explain the tip in November 2002. It appears tipping is adopted slowly by some users motivated solely by reciprocity and is then recognised quickly as a strategy motivated by reputational concerns.

Appendix A

The utility function of reciprocal types 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

---

27 It may also matter that we analysed creative tasks in contrast to the other field experiments who studied less creative tasks like data entry.
28 In order to reduce agency costs caused by moral hazard, ‘behavioural’ principal-agent theory suggests alternative, often non-monetary, routes how a second-best can be reached like reciprocity, status incentives or symbolic awards, see Charness and Kuhn (2011) for an overview.
29 The online consumption of artistic or journalistic products is another market that is marred by moral hazard. These products have negligible marginal costs of reproduction, but it can be very difficult to exclude non-payers from consuming them. The combination of voluntary payments and reputation mechanisms to signal one’s type may be a promising approach to reduce agency costs. In this context, agents would signal their generosity and consumption choices to their peers. Users of the online service Flattr (http://www.flattr.com) can make micro-payments in order to support creators of content, for instance, music, software or blogs. A connection between accounts at Flattr and social network platforms would allow Flattr users to convey their activity to their community. Users with substantial social-image concerns would have increased incentives to make voluntary payments.
30 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.
psychological games framework of Geanakoplos et al. (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 $\tilde{b}_{ij}$ 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 $\Pi$ to the equitable or fair payoff of a player, $\Pi^e$.

The equitable payoff of an individual is the average of his best and worst outcome based on the choices of the other individual.\footnote{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.}

For agent $j$ it is given by:

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

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

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

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

$$\tilde{\kappa}_{ij}(b_{ij}, \tilde{b}_{ij}) = \Pi_i(b_{ij}, \tilde{b}_{ij}) - \Pi^e_i(\tilde{b}_{ij}) \tag{3}$$

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 $\alpha$ ($\alpha = 0$ is the special case of pure self-interest).

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

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: $\alpha_u > \alpha_u$. The possibility of $\alpha_u < \alpha_u$ corresponds to the negative reciprocity 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 when $\alpha_u < \alpha_u$ and therefore he will never give high effort.

The GAR also knows that the user will act reciprocally once her sensitivity to reciprocity $\alpha_u$ 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. See Regner (2005) for more details.

References


