Comparative Feedbacks under Incomplete Information
Nicolas Astier

To cite this version:
Nicolas Astier. Comparative Feedbacks under Incomplete Information. Resource and Energy Economics, 2018, 10.1016/j.reseneeco.2018.07.002. hal-01465189v4

HAL Id: hal-01465189
https://hal.science/hal-01465189v4
Submitted on 26 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Comparative feedbacks under incomplete information

Nicolas Astier*

August 26, 2018

Abstract

Comparative feedbacks, that is personalized messages describing how one’s behavior compares to that of relevant others, are currently widely used in order to change people’s behaviors. Such feedbacks may induce recipients to update their beliefs about both reachable material outcomes and perceived self/social esteem. Both channels are very hard to disentangle in the field, which notably makes welfare analysis a very challenging task. This paper uses an online experiment that makes it possible to focus, within the considered setting, on the role of pure information on material outcomes. Despite an absence of normative pressure, comparative feedbacks are found to trigger comparable or even greater changes in behaviors than other kinds of informative and more accurate feedbacks. A possible explanation may be that comparative feedbacks more effectively conveyed to participants the idea that it should not be too difficult for them to reach a better outcome.

JEL codes: D12, D83, Q50.

*Toulouse School of Economics (TSE), Université de Toulouse, Manufacture des Tabacs, 21 allée de Brienne, 31000 Toulouse, France. Email: nicolas.astier@m4x.org. I am grateful to Yves Le Yaouanq and Nicolas Treich for very insightful discussions that were at the origin of the project, as well as to the numerous friends and colleagues who spent some time playing with and providing feedbacks on different pilot versions of the experiment, and to seminar participants at TSE and CIRED. I also thank Thomas-Olivier Léautier for his guidance and his financial support for running the experiment, and Louis-Gaëtan Giraudet for very helpful comments on an earlier draft of the paper. All remaining errors are mine. Finally, I gratefully acknowledge financial support from the Corps des Ponts, des Eaux et des Forêts.
1 Introduction

Providing information about others’ choices or outcomes – often referred to as “comparative feedbacks” – has been demonstrated to significantly impact (or “nudge”) decision making (Thaler and Sunstein, 2008). Numerous utilities send such feedbacks to their customers in order to decrease their electricity, natural gas and/or water consumptions (Allcott and Mullainathan, 2010; Ferraro et al., 2014). Some of these programs have reached a very large scale.

In many situations where comparative feedbacks are used, consumers do not have all the information they would need to take optimal decisions. This is typically the case for residential energy or water consumers, who very often lack either detailed and frequent information about their consumption, or the time needed to process such information (Kempton and Layne, 1994). In such settings, there are at least two distinct reasons why people may change their behavior after receiving a comparative feedback:

- **An informative channel:** agents will update their beliefs about *the way their choices map into outcomes* (monthly bill, daily comfort, etc.), that is the net utility they would derive privately from choosing a particular action.\(^2\)

- **A normative channel:** agents will update their beliefs about *the way their choices map into self or social esteem*, that is how an external observer would assess their social “status” from their actions.

Both channels need not be orthogonal (Cialdini and Goldstein, 2004). For example, if esteem is derived from one’s ability at a given task previously unknown to the agent, a comparative feedback may be at the same time informative (inference about the ad-

\(^1\) For instance, the company Opower tailors comparative feedbacks for more than 50 million consumers around the world.

\(^2\) This channel supposes that agents can infer something from the average behavior in the population. For example some agents must be endowed with valuable private information.
ditional surplus I can hope to get if I increase effort) and normative (inference about how good I am at the task). Although acknowledged a while ago (Deutsch and Gerard, 1955), this duality remains poorly understood because of the difficulty to disentangle both channels in the field.

In this paper, we design an online experiment which allows to focus, within a specific controlled environment, on the informative channel. Our setting indeed let little scope for social stigma or esteem, which is corroborated by the observation that comparative feedbacks had no measurable effect on participants’ choices under complete information.

We study participants’ response to different feedbacks under incomplete information. In addition to peer comparisons, two additional types of feedbacks are tested. The first one tells participants what would have been their outcome if they had made optimal choices. The second one warns outliers that the outcome they have reached is very far from the optimal outcome. Both these types of feedbacks leverage our knowledge of optimal behaviors, a feature made possible by the fact that we do observe participants’ payoff function within the experiment.

We find that, although peer comparisons conveyed less accurate information, comparative feedbacks triggered comparable or even greater changes in behaviors than did the two other types of feedbacks. A possible explanation is that participants who received a comparative feedback exhibit an increased confidence in their ability to reach a better outcome. Indeed, in their answers to an incentivized question asked during the experiment, participants are slightly more prone to state that they could have reached a better outcome after having received a comparative feedback than after having received non-normative feedbacks. Different kinds of cognitive costs may explain this observation.

---

"We do not yet know of a clean design to separate observational learning from the conformity effect" (Cai et al., 2009, footnote 8).
notably the cost to internalize the feedbacks or participants’ beliefs about the cognitive cost they would have to incur to change their behavior.

The paper is organized as follows. Section 2 provides some background and motivates the experimental set-up. Section 3 describes the experiment, the results are presented in Section 4 and are then discussed in Section 5. Section 6 concludes.

2 Literature review and motivation

While the influence that comparisons with peers can have on beliefs and choices has been discussed in the psychology literature as early as the 1950s (Asch, 1952; Festinger, 1954), the idea to leverage this influence for large scale applications only emerged a few decades later (Berkowitz, 2004; Cialdini and Goldstein, 2004).

A vast literature has experimented the use of comparative feedbacks in the field within different contexts: electricity and natural gas consumption (Midden et al., 1983; Schultz et al., 2007; Nolan et al., 2008; Alcott and Mullainathan, 2010; Alcott and Rogers, 2014; Allcott and Kessler, 2015; Byrne et al., 2016), water consumption (Ferraro and Price, 2013), savings choices (Beshears et al., 2011), etc. These studies have most often found such feedbacks to be effective at meeting the desired goal, although they sometimes note that peer comparisons have several limits. For example, comparative feedbacks have been found to decrease the energy consumption of the average US household persistently by about 2% for electricity (Alcott and Mullainathan, 2010).

4 Consumers initially behaving as desired may revert to a less “virtuous” behavior (boomerang effect, as in Schultz et al. (2007); Byrne et al. (2016)). Consumers at the other end of the distribution may develop strategies allowing them to ignore the feedbacks (excuses, discouragement, as in Beshears et al. (2011)). Finally, for a variety of reasons (Bayesian inference, morality, etc.), people may dismiss either the relevance (e.g. the sender of the feedback may lack credibility, as in Craig and Mccann (1978); Roberts et al. (2004); Alcott (2015)) or the appropriateness (e.g. Ayres et al. (2009); Ferraro and Price (2013); Costa and Kahn (2013)) of comparative feedbacks.
Allcott and Rogers (2014), and 0.7% for natural gas (Allcott and Kessler 2015).

The reasons why consumers react to comparative feedbacks remain however very hard to precisely identify. Indeed, field experiments only observe easily measurable metrics (e.g. daily consumption) which only have a loose connection to the unknown utility functions of participants. As a consequence, field experiments usually lack the ability to make reliable welfare statements. To make things worse, comparative feedbacks may impose unobserved reputational costs and/or distort participants’ choices. To tackle these issues, Allcott and Kessler (2015) recently implemented a revealed preference approach, and notably found that a significant fraction of consumers are actually willing to pay positive amounts of money not to receive comparative feedbacks. While this work represents a big step forward, fully understanding the underlying mechanisms of participants’ response to comparative feedbacks remains an open question.

In that perspective, lab experiments, thanks to the controlled environment they create, represent a useful and complementary approach. In particular, suitable designs may allow to focus on either the normative channel (update of agents’ beliefs about the way their choices map into self or social esteem) or the informative channel (update of agents’ beliefs about the way their choices map into outcomes) of peer comparisons, which are otherwise very hard to disentangle in the field. For example, Gill et al. (2018) use a lab experiment to estimate people’s pure preference for their rank in the distribution of outcomes, hence providing insights on the normative channel.

This paper uses an online experiment to study the informative channel of compara-

5In the framework by Bénabou and Tirole (2011) for example, peer pressure is modeled as a zero-sum game which ends up distorting people’s choices.

6An alternative reason that may explain people’s response to comparative feedbacks within specific environments is anchoring: comparative feedbacks may act as a reference point to which consumers converge. This explanation is however of little relevance in environments where consumers’ payoffs arise from a complex combination of numerous choices such as residential energy consumption.
tive feedbacks. Indeed, an important reason why consumers may react to such feedbacks is *incomplete information*: consumers may not know very well how changes in their behavior may influence the outcome they reach (e.g. due to rational inattention, that is it may be optimal to remain incompletely informed when information acquisition is costly). In such situations, one may learn by observing others’ behavior. Starting with Banerjee (1992) and Bikhchandani et al. (1992), a vast literature on social learning has studied such situations from a theoretical standpoint, although this literature mainly focussed on identifying conditions under which crowds may (rationally) herd on a suboptimal action (Eyster and Rabin 2010, 2012; Eyster et al., 2015).

Our choice to focus on the informative channel stems from the observation that comparative feedbacks are widely used in the context of residential energy or water consumption. In such settings, consumers are typically very poorly informed about the relationship that exists between their behavior and the bills they receive. For example, although residential customers consume electricity services (lighting, heating, etc.), they are still often billed for *monthly electricity quantities*: their consumption is thus aggregated both over a long period of time and over many different appliances. Kempton and Layne (1994) made this point salient in a powerful comparison which motivated our experimental design: “consider groceries in a hypothetical store totally without price markings, billed via a monthly statement like ‘US$527 for 2362 food units in April’ ”. Consumers’ choices are thus made under highly incomplete information, since they do

---

Faruqui et al. (2010) drew a similar comparison: “Imagine a world in which Joe Smith drives up to the gasoline pump in his large SUV, fills up his truck, and drives away without paying a dime. The gasoline is not free, but Smith won’t know how much he purchased or how much he owes until a month later because he has a monthly account with the filling station. When his wife drives up to the pump in the family sedan, she goes through the same procedure; as does their high school senior, who drives up to the pump in her compact coupe. The Smiths get a combined bill a month later and don’t know how the charges accumulated. Was it Joe’s driving, his wife’s driving, or their daughter’s driving that accounted for the lion’s share of the bill? What makes life even more interesting for the Smiths is that none of their cars have a speedometer or a gasoline gauge. They get no feedback at all on how to manage their gasoline bill”. 

---
not know the individual energy consumption\textsuperscript{8}— and thus the price— of the services they consume. Such a complex environment may deter any attempt to learn at all\textsuperscript{9} and it is thus not very surprising that households have been found to hold biased beliefs about the electricity consumption of individual appliances \cite{Wood2003,Attari2010}.

In the absence of user-friendly information about per-service costs, most consumers are likely to choose to remain imperfectly informed, and thus to make choices based on heuristics and routines. Such choices would create a wide dispersion in individual consumption patterns, as observed by \cite{Iyer1998,Zimmermann2009,Kwac2014} for daily electricity consumption. Consistently \cite{Armel2013} note that “anecdotal reports from plug monitoring companies, disaggregation developers, and researchers overseeing feedback studies suggest that the largest savings often come from surprising places, such as an extra Tivo, a pool pump, a pottery wheel, or an electric towel or floor warmer, inadvertently left on”; and \cite{Leighty2011} report that sustainable savings were realized following an energy crisis by the simple fact that “many “forgotten” or “spare” devices like clocks and televisions in guest rooms and spare refrigerators or freezers remained unplugged”. As a consequence, within incomplete information environments, a comparative feedback may act as a warning device, allowing customers to realize there is something wrong with their consumption pattern\textsuperscript{10} in the spirit of the learning by noticing literature \cite{Hanna2014}.

\textsuperscript{8}To make things worse, energy consumption is measured in unfamiliar units such as “kilowatt hours” \cite{Kempton1982,Roberts2004} or “cubic feet/meters”. Besides, in many countries, it is still common that bills are based on actual consumption only once a year.

\textsuperscript{9}For example \cite{Brounen2013} found after surveying 1721 Dutch households that about half of participants answered “I have no idea” to the question “How much do you pay for your monthly electricity bill?”.

\textsuperscript{10}Consistently high baseline consumers typically realize higher savings \cite{Dolan2013,Ferraro2013,Byrne2016}.
incomplete information, we set up a web-based experiment that replicates some stylized features of the information structure faced by residential energy or water consumers. Our experimental design is indeed inspired by the metaphor by Kempton and Layne (1994) of a grocery shop without price tags that would charge monthly bills to customers. In the context of energy consumption, the commodities without price tags are energy services, for which the needed quantity of energy is most often unknown, creating uncertainty regarding the price of these services. By analogy, within our experiment, participants choose whether to feed a virtual pet cat with milk (resp. fish) instead of water (resp. dry cat food): while they know by how much this upgrade will make the cat happier, they do not know the corresponding price of the service. They hence face a similar environment as do residential consumers when they consider decreasing the AC thermostat by 1 degree during a hot summer day: although consumers usually have a rough idea of how much better they would feel if they decrease the temperature in the room, the vast majority of them do not know how many kilowatt-hours are needed to reach this outcome and thus the price of the service.

A key advantage of our experimental set up compared to field experiments is to enable us to perfectly observe the payoff function of participants within the game, and thus to know the optimal consumption of services. We now describe the experimental design.

3 Experimental design

3.1 Description of the experiment

We recruited participants on Amazon’s Mechanical Turk platform and asked them to play an online game. Conditional on passing a tutorial (see section 3.2), participants

---

\(^{11}\)The framing of the experiment was a deliberate attempt to harness participants’ intrinsic motivation to play seriously. Indeed participants registered on online platforms very often perform numerous boring
received a fixed payment as well as a variable bonus which was increasing in their score at the game. They were instructed to take care of a virtual pet called “Noney”, choosing every day which food items to give to her.\footnote{An additional non-instrumental task was performed in order to divert players’ attention - as do various real-life tasks - and to encourage players to adopt simplified heuristics (see Appendix A for more details).}

The game lasted two virtual weeks. First-week instructions were identical for all participants within a given information environment (see below). It allowed participants to practice the game and made it possible to send non-deceptive personalized feedbacks based on first-week outcomes. At the end of this first practice week, participants received different types of randomly assigned feedbacks, and then went on to play a second virtual week. Our experiment thus implements a random control trial in which the outcomes of interest are the choices made during the second week.

Participants played the game only once and were informed that all parameters remained the same during both weeks. Their score was measured in “smiley units”, which served as a utility measure. Each week, participants had to feed Noney everyday using a weekly budget of $B = 15$. Smiley units were collected (i) from food consumption and (ii) from the leftover money at the end of the week which was used to buy Noney a gift. The more money was left, the bigger the gift and the happier Noney: every remaining dollar earned the participant an additional 10 smiley units.

On a given day, participants had to feed Noney with one drink and one meal:

- They could choose either water or milk for the drink. Water cost $0.1$ and earned 1 smiley unit. Milk cost $0.6$ and earned more smiley units but at a decreasing
rate (see Figure 1): the more participants previously gave Noney milk during the week, the lower the incremental utility.

- They could choose either dry cat food or fish for the meal. Dry cat food cost $0.4 and earned 4 smiley units. Fish cost $1.5 and earned more smiley units but at a decreasing rate (see Figure 1).

Incremental utility (in smiley units)

\[
\begin{array}{c|c|c}
\text{Number of times (within a week)} & \text{Incremental utility (in smiley units)} \\
\hline
1 & U_M(.) & U_F(.) \\
2 & 18 & 17 \\
3 & 13 & 11 \\
4 & 9 & 4 \\
5 & & \\
6 & & \\
7 & & \\
\end{array}
\]

Figure 1: Demand curves for milk and fish

Figure 1 displays the demand functions for milk and fish, which were identical for all players. If we denote \( U_M(n_M) \) (resp. \( U_F(n_F) \)) the incremental utility of consuming milk (resp. fish) for the \( n_M \)th (resp. \( n_F \)th) time of the week, a participant choosing to feed Noney with milk \( n_M \) times and with fish \( n_F \) times during a given week got a score:

\[
\text{Weekly score} \equiv \sum_{i=1}^{n_M} U_M(i) + (7 - n_M) + \sum_{j=1}^{n_F} U_F(j) + 4(7 - n_F) + 10(B - 0.6n_M - 1.5n_F)
\]

\[
= \sum_{i=1}^{n_M} U_M(i) + \sum_{j=1}^{n_F} U_F(j) + 10(B - 0.6n_M - 1.5n_F)
\]
Note that in a partial equilibrium framework this weekly score corresponds to consumers’ surplus, a feature we discuss in section 5.2. Participants’ weekly bill is:

\[
\text{Weekly bill} \equiv 0.6n_M + 1.5n_F + 0.1(7 - n_M) + 0.4(7 - n_F) = 0.5n_M + 1.1n_F + 3.5
\]

One can easily check that the optimal strategy was to feed Noney with milk 5 times, and with fish once. The corresponding weekly bill was $7.1.

Players were randomized between two environments: one environment where the prices of milk and fish were known – the Complete Information (CI) environment; and one environment where the prices of milk and fish were unknown – the Incomplete Information (II) environment. The assignment to a given information environment was performed explicitly during the game tutorial. Players who ended up in the II environment were thus aware that some other players ended up in a CI environment.

At the end of the first week, players received their weekly bill. Control groups received no further feedback. However, one treatment in the CI environment and three treatments in the II environment were implemented:\[13]\[14\]

- **Comparative feedback:** players in this group received a feedback telling them how their first-week bill compared to the bills of the control groups.\[13]\[14]
- **Optimal bill:** players were told the amount of the score-maximizing bill.
- **Warning to outliers:** players having a first-week bill significantly higher than the score-maximizing one received a feedback warning them about their abnormally

\[13]\]The allocation across treatments was pseudo-random in the sense that data was first collected for both control groups and the optimal bill treatment, and then for the other treatments (comparative and warning feedbacks were then based on realized data, making the experiment non-deceptive). In addition, in order to get balanced group sizes, the allocation to treatments was random at first, and then forced towards treatments with the lowest number of observations. Given participants’ arrival on the platform is random, it is unlikely that this procedure created any bias.

\[14]\]The outcomes of both control groups were pooled to design the comparative feedback, so that the feedback was actually informative even in the incomplete information environment.
high bill (a threshold bill of $9.9 was chosen, corresponding to the eighth decile of the first-week bills observed in control groups).

Figures 4 to 6 in Appendix A show screenshots of the feedbacks received in the different treatments. Table 1 gives the number of participants in the different treatments.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Comparative Feedback</th>
<th>Optimal bill</th>
<th>Warning</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI</td>
<td>50</td>
<td>48</td>
<td>50</td>
<td>52</td>
</tr>
<tr>
<td>II</td>
<td>50</td>
<td>51</td>
<td>50</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 1: Number of participants per treatment

3.2 Strategies to increase data quality

Several screening devices were used to make sure participants paid attention to the instructions. First, both the descriptive summary of the experiment and the first page of instructions made it clear that any payment (including the fixed participation fee) was conditional on successfully answering a few comprehension questions. Second the first page of instructions was formal and lengthy. It explained that the variable bonus paid to participants was relatively steep with respect to the score obtained in the game (ranging from $0 to $3.5, the latter being more than three times the fixed participation fee), and that an average performance translated into a relatively low bonus. Third, the tutorial was quite long and reading it was necessary before answering the comprehension questions. Fourth, players failing to answer comprehension questions correctly at their first attempt were redirected to the beginning of the tutorial. Fifth, participants who failed at their second attempt were kicked out of the game without payment. A single log per participant was allowed so that kicked-out players could not log in again. Finally, the experiment was framed as a funny online game in an attempt to harness intrinsic motivation. These various strategies did prevent a significant number of would-be participants from enrolling into the experiment, as shown in Table 2.
Hence about 20% of players (73 out of 374) who logged in did not pass the tutorial. Since 86% of the players (258 out of 301) who took part in the experiment provided correct answers to the tutorial questions at their first attempt, it is reasonable to think that the main reason for failing to pass the tutorial questions was a lack of diligence. As such it is very likely that the screening strategies implemented did increase data quality.

### 3.3 Validity of randomization

#### 3.3.1 Complete Information environment

Table 3 provides summary statistics (means and standard deviations) for the first week of the game, that is when instructions were so far identical for both the control and the treatment groups under CI. The randomization procedure seems to have reasonably worked based on these statistics.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Control</th>
<th>Comp. feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-week bill (in $)</td>
<td>7.27 (1.34)</td>
<td>7.70 (1.60)</td>
</tr>
<tr>
<td>Time spent on tutorial (in min)</td>
<td>7.08 (2.84)</td>
<td>8.15 (4.00)</td>
</tr>
<tr>
<td>Time spent on week 1 (in min)</td>
<td>4.31 (1.49)</td>
<td>4.86 (1.36)</td>
</tr>
<tr>
<td>Tutorial passed at 1st attempt (%)</td>
<td>82</td>
<td>93.75</td>
</tr>
<tr>
<td>Sample Size</td>
<td>50</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 3: Summary statistics for week 1 (CI environment)
3.3.2 Incomplete Information environment

Table 4 provides summary statistics for the first week of the game under II. Not surprisingly, the distribution of first-week bills is more dispersed under II than it was under CI. Because of this higher dispersion in first-week outcomes, our preferred empirical strategy will include a control for the first-week bill.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Control</th>
<th>Comp. feedback</th>
<th>Opt. bill</th>
<th>Warning</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-week bill (in $)</td>
<td>8.94 (2.31)</td>
<td>7.90 (2.52)</td>
<td>8.09 (1.91)</td>
<td>8.47 (2.42)</td>
</tr>
<tr>
<td>Time spent on tutorial (in min)</td>
<td>8.06 (4.09)</td>
<td>7.29 (3.86)</td>
<td>7.27 (3.45)</td>
<td>7.46 (4.89)</td>
</tr>
<tr>
<td>Time spent on week 1 (in min)</td>
<td>4.90 (3.77)</td>
<td>4.26 (1.95)</td>
<td>4.51 (1.55)</td>
<td>3.93 (1.86)</td>
</tr>
<tr>
<td>Tutorial passed at 1st attempt (%)</td>
<td>90</td>
<td>84.3</td>
<td>88</td>
<td>76.9</td>
</tr>
<tr>
<td>Sample Size</td>
<td>50</td>
<td>51</td>
<td>50</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 4: Summary statistics for week 1 (II environment)

4 Results

4.1 Empirical strategy

As explained in the description of the experimental design, first-week instructions were identical for all participants within a given information environment. Our experiment thus seeks to implement a random control trial in which the outcome of interest is participants’ choices during the second week. This section will focus on participants’ second-week bills. For each treatment, we will use several specifications of the following regression:

\[ Y_i = \alpha + \tau R_i + \beta X_i + \epsilon_i \]

where \( Y_i \) is participant \( i \)'s second-week bill, \( R_i \) is an indicator variable equal to 1 if the participant received the considered feedback and 0 otherwise, and \( X_i \) is a vector of control variables. The coefficient of interest is \( \tau \) which represents the magnitude of the treatment effect (impact on the average second-week bill in $ controlling for variables \( X \)).
Specification (1) includes no control variable (simple comparison of means). Specification (2), which will be our preferred specification, includes the first-week bill as a control. Specification (3) adds the total time spent on week 1 as a control. Finally, specification (4) adds the total time spent on the tutorial as a control. Robust standard errors are given.

4.2 Complete Information environment

We start with the results obtained under complete information. Figure 2 shows the deviations from the optimal weekly bill for the control group (left panels), and for the treatment group who received a comparative feedback at the end of week 1 (right panels). Week 2 deviations are in red, while week 1 deviations are in blue. Graphically intuition suggests that participants did not change their behavior much after receiving the comparative feedback. Statistical analysis (see Appendix B.1) supports this intuition.

A possible explanation may be that the optimal bill ($7.1) happened to lie between the second decile bill ($6.7) and the average bill ($8.1) of the comparison group, so that most players got a first-week bill within the range described in the comparative feedback (see Figure 4 in Appendix A). However, the dispersion in second-week bills of participants who played suboptimally does not seem smaller in the treated group than in the control group. A more likely explanation is thus that since the bill does not relate to any intrinsic characteristic of the players (there is no horizontal differentiation in the game), participants had no reason to react to comparative feedbacks under CI.

Participants’ choices under complete information suggest that no social esteem or stigma was derived from the weekly bills received within the game. Hence, any change in behaviors under incomplete information was most likely due to a pure information
effect. As a consequence, our experiment allows to credibly focus on the extent to which the purely informative content of peer comparisons is able to induce changes in behaviors. We now describe the obtained results.

4.3 Incomplete Information environment

4.3.1 Participants’ response to comparative feedbacks

Figure 3 shows the deviations from the optimal weekly bill for the control group (left panels), and for the treatment group who received a comparative feedback (right panels) under incomplete information. Week 2 deviations are in red, while week 1 deviations are in blue. Graphical intuition now suggests that comparative feedbacks triggered a decrease in week 2 bills, which is supported by the statistical analysis.

This result illustrates the ability of comparative feedbacks to convey pure information, and thus to influence behaviors, even in the absence of strong self/social esteem
Figure 3: Deviations from optimal bill under II for week 1 (top) and 2 (bottom), with (right) and without (left) a comparative feedback at the end of week 1.

<table>
<thead>
<tr>
<th>Specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>8.63 (0.34)***</td>
<td>5.56 (0.98)***</td>
<td>6.09 (1.14)***</td>
<td>6.29 (1.29)***</td>
</tr>
<tr>
<td>Treatment</td>
<td>-1.72 (0.46)***</td>
<td>-1.36 (0.48)***</td>
<td>-1.43 (0.49)***</td>
<td>-1.45 (0.50)***</td>
</tr>
<tr>
<td>First-week bill</td>
<td>0.34 (0.10)***</td>
<td>-0.33 (0.10)***</td>
<td>0.33 (0.10)***</td>
<td>0.33 (0.10)***</td>
</tr>
<tr>
<td>Time first week</td>
<td>-0.09 (0.06)</td>
<td>-0.08 (0.06)</td>
<td>-0.02 (0.05)</td>
<td>-0.02 (0.05)</td>
</tr>
<tr>
<td>Time tutorial</td>
<td>0.123</td>
<td>0.236</td>
<td>0.247</td>
<td>0.249</td>
</tr>
</tbody>
</table>

Table 5: Comparative feedbacks under II (***: p < 0.01, **: p < 0.05, * : p < 0.1)

concerns. Within such an environment, it appears interesting to investigate how such feedbacks compare to different purely informative feedbacks, that is messages framed without referring to other participants. Two additional types of feedbacks were thus tested. First, one treatment group received a feedback indicating the bill they would have received if they had played optimally. Second, another treatment group received – when relevant – a feedback warning them that their first-week bill was abnormally high compared to the optimal bill. Note that both treatments rely on the knowledge of the actual optimal choices, a feature that is very hard to achieve in the context of field experiments. Further note that our purely informative feedbacks conveyed accurate
information, while comparative feedbacks only conveyed information about the realized bills of other participants (see Appendix A).

### 4.3.2 Participants’ response to purely informative feedbacks

<table>
<thead>
<tr>
<th>Optimal bill feedback</th>
<th>Specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>8.63 (0.34)***</td>
<td>4.61 (0.83)***</td>
<td>5.08 (0.97)***</td>
<td>4.47 (1.13)***</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>-1.02 (0.44)**</td>
<td>-0.63 (0.41)</td>
<td>-0.67 (0.42)</td>
<td>-0.61 (0.43)</td>
<td></td>
</tr>
<tr>
<td>First-week bill</td>
<td>0.45 (0.08)***</td>
<td>0.44 (0.09)***</td>
<td>0.45 (0.09)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time first week</td>
<td>-0.07 (0.05)</td>
<td>-0.09 (0.06)</td>
<td>0.07 (0.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tutorial</td>
<td>0.051</td>
<td>0.229</td>
<td>0.237</td>
<td>0.249</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Warning feedback</th>
<th>Specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>8.63 (0.34)***</td>
<td>6.32 (0.99)***</td>
<td>6.79 (1.10)***</td>
<td>7.27 (1.12)***</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.57 (0.48)</td>
<td>-0.45 (0.48)</td>
<td>-0.54 (0.49)</td>
<td>-0.55 (0.49)</td>
<td></td>
</tr>
<tr>
<td>First-week bill</td>
<td>0.26 (0.10)**</td>
<td>0.25 (0.10)**</td>
<td>0.27 (0.10)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time first week</td>
<td>-0.08 (0.05)*</td>
<td>-0.05 (0.04)</td>
<td>-0.10 (0.04)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tutorial</td>
<td>0.014</td>
<td>0.076</td>
<td>0.087</td>
<td>0.116</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Warning feedback (treatment on the treated)</th>
<th>Specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>9.44 (0.45)***</td>
<td>7.90 (2.74)***</td>
<td>7.52 (2.95)***</td>
<td>7.76 (3.20)***</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>-2.30 (0.65)***</td>
<td>-2.30 (0.66)***</td>
<td>-2.31 (0.66)***</td>
<td>-2.26 (0.67)***</td>
<td></td>
</tr>
<tr>
<td>First-week bill</td>
<td>0.14 (0.23)</td>
<td>0.14 (0.24)</td>
<td>0.13 (0.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time first week</td>
<td>0.08 (0.16)</td>
<td>0.08 (0.16)</td>
<td>-0.02 (0.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tutorial</td>
<td>0.256</td>
<td>0.265</td>
<td>0.268</td>
<td>0.269</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Purely informative feedbacks (***: p < 0.01, **: p < 0.05, *: p < 0.1)

Table 6 shows the results obtained for the two additional feedbacks implemented under incomplete information. The corresponding graphs are provided in Appendix B.2. Interestingly, although treatment effects have the expected sign, both their magnitude and their significance are weaker compared to the results obtained for comparative...
feedbacks. Only the warning feedbacks exhibit a significant treatment effect when restricting attention to participants with a first-week bill higher than $9.9 (17 players in the treatment group and 21 players in the control group), that is when focusing on participants who received or would have received the warning.

4.4 Summary

<table>
<thead>
<tr>
<th>Week 1 Information</th>
<th>Week 1 Average bill</th>
<th>Week 2 Information</th>
<th>Week 2 Average bill</th>
<th>Bill week 2(T) - Bill week 2(C) (with a control on week 1 bill)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prices</td>
<td>7.27</td>
<td>Prices</td>
<td>7.11</td>
<td></td>
</tr>
<tr>
<td>Prices</td>
<td>7.70</td>
<td>Prices and a comp. feedback</td>
<td>7.36</td>
<td>0.01 (0.28)</td>
</tr>
<tr>
<td>8.94</td>
<td>8.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.90</td>
<td>Comp. feedback</td>
<td>6.91</td>
<td>-1.36 (0.48)**</td>
<td></td>
</tr>
<tr>
<td>8.09</td>
<td>Optimal bill</td>
<td>7.62</td>
<td>-0.63 (0.41)</td>
<td></td>
</tr>
<tr>
<td>8.47</td>
<td>Warning</td>
<td>8.06</td>
<td>-0.45 (0.49)</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Summary of obtained results (** : \(p < 0.05\), * : \(p < 0.1\))

Table 7 summarizes our results from which several conclusions emerge. First, under incomplete information and within our specific environment, participants appear to have slightly overreacted to the comparative feedback received: the average second-week bill of the group who received a comparative feedback under incomplete information ($6.9) happens to be slightly lower than the optimal bill ($7.1). Yet, this statement – made possible by our knowledge of the optimal bill – lacks statistical significance.\(^{16}\) Second, comparative feedbacks seem to have triggered greater responses than purely informative feedbacks. It indeed appears that the treatment effect of comparative feedbacks is significantly higher\(^{17}\) than the treatment effect of both purely informative feedbacks pooled together, and close to significantly higher than the treatment effect of optimal

\(^{15}\)For example, \(p\)-values range from 0.11 to 0.16 for the optimal bill treatment for specifications (2) to (4).

\(^{16}\)A more thorough discussion of the impact of the feedbacks on the score obtained by participants is provided in section 5.2.

\(^{17}\)Treatment effect of -0.82 with a \(p\)-value of 0.03.
bill feedbacks. Appendix B.3 provides a graphical illustration of this result, which appears quite surprising given the fact that purely informative feedbacks conveyed accurate information while comparative feedbacks only conveyed information about the realized bills of other participants. Section 5 discusses possible explanations, using the detailed data collected during the experiment.

In addition, our experimental design allows us to track to what extent the outcome reached by participants is close to the optimal one, which we happen to have knowledge of. To do so, we ran regressions using the absolute value of the deviation from the optimal bill as the dependent variable. Table 8 reports the results obtained using this approach. Again, it appears that, despite being more accurate in terms of informational content, the two purely informative feedbacks do not exhibit higher treatment effects than comparative feedbacks.

<table>
<thead>
<tr>
<th>Information</th>
<th>Average deviation</th>
<th>Information</th>
<th>Average deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prices</td>
<td>0.88</td>
<td>Prices</td>
<td>1.12</td>
</tr>
<tr>
<td>Prices</td>
<td>0.98</td>
<td>Prices and a comp. feedback</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>2.34</td>
<td></td>
<td>2.46</td>
</tr>
<tr>
<td></td>
<td>2.11</td>
<td>Comp. feedback</td>
<td>1.71</td>
</tr>
<tr>
<td></td>
<td>1.74</td>
<td>Optimal bill</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>2.17</td>
<td>Warning</td>
<td>1.99</td>
</tr>
</tbody>
</table>

Table 8: Deviations from the optimal bill (*** : p < 0.01, ** : p < 0.05, * : p < 0.1)

18Running specification (2) for the comparative feedbacks group using the optimal bill feedbacks group as the control group yields a treatment effect of -0.63 with a p-value of 0.11.
5 Discussion

5.1 Heuristics and cognitive costs

5.1.1 Choice patterns

Since the incremental utility from consuming milk and fish is decreasing, an optimal choice pattern should consist in choosing milk and fish first, and then switching to the numeraire good after a few days.\textsuperscript{19} The percentage of players who exhibit such a “single-switching-point” pattern appears to be relatively low (Table 9). Hence it seems most participants relied on suboptimal heuristics (mostly diversifying choices), especially so under incomplete information. Cognitive costs thus played a significant role.

<table>
<thead>
<tr>
<th></th>
<th>CI</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>27.6%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Week 2</td>
<td>45.9%</td>
<td>22.7%</td>
</tr>
</tbody>
</table>

Table 9: Percentage of participants who exhibit a “single-switching-point” pattern

5.1.2 Detecting suboptimal behavior from first-week bill and feedback

At the end of week 1, players were asked the following question regarding their first-week choices: “What do you think are the chances (in %) that you could have increased the number of smiley units you have collected so far by making different food choices?”

Beliefs were elicited in an incentive-compatible manner. More precisely, we slightly modified the classic BDM method (Becker et al., 1964) in order to reward sure guesses symmetrically. The method used is described in Appendix C.\textsuperscript{19}

\textsuperscript{19}The incremental utility after having consumed a given amount of a good was indeed unknown until this amount was reached, so that there was an option value to consume milk and fish first.
II control group

Under incomplete information, the first-week bill should allow participants to learn about unknown prices. Indeed, if a participant chose milk (resp. fish) $n_M \in [0, 7]$ times (resp. $n_F$ times) during the week and received a bill $b$, then she should be able to infer that consistent subjective beliefs $\hat{p}_M$ (resp. $\hat{p}_F$) for the price of milk (resp. fish) must satisfy:

$$n_M\hat{p}_M + n_F\hat{p}_F + 0.1(7 - n_M) + 0.4(7 - n_F) = b$$

The fact that choices are discrete (the quantity of food items are integers) creates an additional opportunity to learn. More precisely, it can be shown that only choices $(n_M, n_F) \in \{(0, 0), (0, 1), (1, 4), (2, 4), (3, 4), (4, 4), (5, 0), (5, 1), (5, 2), (5, 3), (5, 4)\}$ can be rationalized given the received first-week bill. We can thus test whether the participants in the II control group who made first-week choices that cannot be rationalized were more likely to state they first-week choices were suboptimal than other participants. We find no evidence that this is the case.\(^{20}\)

CI groups

Answers by participants under complete information to the question whether they believed their first-week choices were suboptimal or not provide further evidence that players dedicated heterogenous levels of attention to the experiment. Indeed these participants had all the information they needed to play optimally. As such, if they made suboptimal choices, they should be aware of it and should state that they could have made better choices. To test this hypothesis, we run the regression $Y_i = \alpha + \beta Opt_i + \epsilon_i$ where $Y_i$ is participant $i$ stated belief that she could have made better choices during the first week, and $Opt_i$ is an indicator variable taking the value 1 if participant $i$ first-week choices were score-maximizing. This regression strongly suggests that participants who

\(^{20}\)Participants in the II control group even reported being more confident that they could have performed better when their first-week choices were rationalizable than when they were not (average stated probability of 72% vs 61%).
played suboptimally under CI were aware of it, highlighting again the role played by cognitive costs.

<table>
<thead>
<tr>
<th>Specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>69.10 (2.69)***</td>
<td>42.91 (10.08)***</td>
<td>27.60 (13.81)***</td>
<td>27.88 (13.80)***</td>
</tr>
<tr>
<td>Optimal</td>
<td>−38.72 (5.43)***</td>
<td>−36.62 (5.51)***</td>
<td>−33.84 (5.65)***</td>
<td>−33.85 (5.64)***</td>
</tr>
<tr>
<td>First-week bill</td>
<td>3.39 (1.21)***</td>
<td>3.27 (1.25)***</td>
<td>3.04 (1.31)***</td>
<td>2.54 (2.22)</td>
</tr>
<tr>
<td>Time first week</td>
<td>3.32 (1.94)*</td>
<td>2.54 (2.22)</td>
<td>0.65 (0.89)</td>
<td></td>
</tr>
<tr>
<td>Time tutorial</td>
<td>0.381</td>
<td>0.406</td>
<td>0.429</td>
<td>0.433</td>
</tr>
</tbody>
</table>

Table 10: Awareness of suboptimal choices under CI (***/p < 0.01, **/p < 0.05, */p < 0.1)

**II treatment groups**

The different feedbacks gave participants under II clear evidence on whether their first-week choices were score-maximizing or not. It is thus natural to wonder to what extent participants internalized the feedbacks provided. To answer this question, we regress participants’ stated belief about their ability to make better choices on a dummy variable for the different treatments under II using specification (2).

<table>
<thead>
<tr>
<th>Treatment</th>
<th>CF</th>
<th>Info-Only</th>
<th>Warning (all)</th>
<th>Warning (treated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>43.71 (10.10)***</td>
<td>40.02 (10.96)***</td>
<td>41.22 (9.79)***</td>
<td>95.00 (22.20)***</td>
</tr>
<tr>
<td>Treat. Dummy</td>
<td>13.40 (4.06)***</td>
<td>8.11 (4.49)*</td>
<td>4.18 (4.47)</td>
<td>11.94 (6.70)*</td>
</tr>
<tr>
<td>First-week bill</td>
<td>2.31 (1.04)***</td>
<td>2.72 (1.15)**</td>
<td>2.59 (1.01)**</td>
<td>−2.43 (2.02)</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.132</td>
<td>0.081</td>
<td>0.073</td>
<td>0.121</td>
</tr>
</tbody>
</table>

Table 11: Feedbacks’ effectiveness in raising awareness (***/p < 0.01, **/p < 0.05, */p < 0.1)

Surprisingly, participants who received a comparative feedback were slightly more likely to state that could have increased their score by playing differently than partici-
pants who received a purely informative feedback. This result appears counter-intuitive given the fact that the latter provided accurate information: both the knowledge of the score-maximizing bill and the reception of a warning message provided clear cut evidence on whether first-week choices had been optimal or not. By contrast, comparative feedbacks only provided information on the outcome reached by other participants, with no guarantee that their choices had been optimal.

As a consequence, participants’ elicited beliefs appear to be consistent with the hypothesis that comparative feedbacks could be more effective than purely informative feedbacks in making people aware of their suboptimal behaviors. Different kinds of cognitive costs could support this hypothesis. First, a comparative feedback may be easier to process, given its graphical presentation. However Table 12 suggests that participants seem to have spent enough time on the feedback page to read its content carefully.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Mean (Std dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control group CI</td>
<td>9.7 (7.2)</td>
</tr>
<tr>
<td>CF treatment CI</td>
<td>35.5 (23.1)</td>
</tr>
<tr>
<td>Control group II</td>
<td>12.2 (8.4)</td>
</tr>
<tr>
<td>CF treatment II</td>
<td>34.8 (16.1)</td>
</tr>
<tr>
<td>Optimal bill treatment</td>
<td>33.8 (24.9)</td>
</tr>
<tr>
<td>Warning treatment (all)</td>
<td>19.6 (17.2)</td>
</tr>
<tr>
<td>Warning treatment (treated)</td>
<td>25.1 (17.2)</td>
</tr>
</tbody>
</table>

Table 12: Time spent on the feedback page (in seconds)

Second, comparative feedbacks may involve learning about cognitive transaction costs. Indeed the fact that many other players have different bills does not only suggest that one’s current behavior is not optimal, it also provides convincing evidence that adopting a behavior closer to the norm is not too difficult. As a consequence, the effec-

---

21 By “purely informative” feedbacks we mean feedbacks framed without referring to other participants.
tiveness of comparative feedbacks in our experiment may be partly explained by factors beyond the mere knowledge of the realized average bill they provide.

5.1.3 End of game price guesses

At the end of the game, participants under II were asked to guess the prices of milk and fish. These guesses were strongly incentivized: close enough bets would typically increase participants’ payoff by $2-$2.5, an amount roughly equivalent to the total average payment received by players who provided wrong price guesses for their participation in the whole experiment. In other words, participants were offered a chance to earn in a few seconds an amount of money they had just spent (on average) 16 min 43 sec to earn. In addition, since participants had received two weekly bills by the time they were asked to guess prices, they were in a position where they had, in theory, all the needed data to retrieve actual prices.

None of the 203 participants under II provided accurate guesses for both prices. This result illustrates that bills aggregating consumption both across time and services make it very difficult for consumers to learn.

5.2 Obtained scores

So far, we measured the effectiveness of the different feedbacks through their impact on participants’ choices, aggregated in their second-week bill. However, one may wonder whether receiving feedbacks actually made participants better off. Because all players had the same known “preferences”, we can easily test the impact of feedbacks on participants’ second-week scores by regressing participants’ score in week 2 on a treatment dummy. Results are reported in Table 13.

---

22 Unless they made identical aggregate choices in both weeks, which was the case only for 30 participants (14.8%) under II, out of which 6 received an optimal bill feedback, and 9 received a comparative feedback.
Interestingly none of the feedbacks had a significant impact on the obtained scores. Worse, the impact is likely to have been negative for participants who received warnings, as they seem to have overreacted to the feedback. Although the weekly score may be interpreted as consumer’s surplus, we are of course in no position to infer something about the welfare impact of such feedbacks in real-life applications in the absence of external validity. However, our results illustrate why feedback programs should be evaluated, to the extent that it is feasible, in terms of their impact on social welfare – as attempted in Allcott and Kessler (2015) – and not solely in terms of their impact on easily measurable metrics. Indeed, although comparative feedbacks did have a significant impact on players’ second-week bill in our experiment, they had no significant impact on weekly scores.

### 6 Conclusion

This paper uses an online experiment to study people’s response to peer comparisons, within a controlled environment that replicates a few stylized features of the information structure faced by residential energy and water consumers. The chosen framing offered little scope for social esteem or stigma. Consistently, incomplete information is shown to be a necessary condition for comparative feedbacks to influence participants’ choices. Normative pressure was thus unlikely to play a major role within our specific setting, allowing us to credibly focus on participants’ response to the purely informational con-
tent of comparative feedbacks.

Our approach complements the findings of the vast literature on peer comparisons in field experiments on several dimensions. First, this literature rarely emphasizes the informative channel of comparative feedbacks. Second, our controlled environment provides us knowledge of the optimal level of participants’ expenditure, which cannot be observed in the field. This knowledge notably allowed us to tailor two accurate purely informative feedbacks, that is messages framed without referring to other participants. The first one told players the amount of the score-maximizing bill, while the second one warned outliers that their first-week bill was abnormally high.

Comparative feedbacks are found to have triggered comparable or even greater changes in behaviors than other purely informative feedbacks. This represents a relatively surprising result in the absence of any normative pressure. Even more surprisingly, although purely informative feedbacks provided clear cut evidence on whether first-week choices had been optimal or not – contrary to comparative feedbacks that only provided information on the outcome reached by other participants – the former were not more effective at convincing participants that they could have improved their score by making different choices. While the effectiveness of comparative feedbacks at changing behaviors is often attributed to a peer pressure effect, our experiment thus suggests that a possible additional driver may be that such feedbacks convey a useful signal about how difficult it may be to improve on one’s current outcome. Such a “proof by example” that a better outcome may easily be reached could indeed represent a complementary incentive to change behavior.

The knowledge of participants’ payoff function also made it possible to assess to what extent the received feedbacks helped participants to converge to optimal choices and to
get higher scores. Again, it appears that, within our experiment, purely informative feedbacks were not more effective than comparative feedbacks in that regard. In addition, all feedbacks are found – within our specific setting – to have had little impact on the obtained score.

Finally, ancillary results of our experiment illustrate that environments in which consumers receive bills aggregating consumption both over time and services make it very difficult for them to learn over time. Indeed most participants chose to rely on heuristics, and none of them was able to accurately retrieve the missing information despite high incentives and enough data to do so.
Appendices

A Screenshots of the experiment

A.1 Received feedbacks

The following screenshots show the feedbacks received by the different treatment groups. Note that for the optimal bill and warning feedbacks, the warning sign displayed was flashing, making it very difficult for participants not to notice the feedback.

![End of First Week](image)

Figure 4: Comparative feedback
Figure 5: Optimal bill feedback (the warning sign was flashing)

Figure 6: Warning feedback (the warning sign was flashing)
A.2 Distracting task used

After each choice of both a drink and a meal, participants were asked to play a non-instrumental distraction task in order to divert their attention – as do various real-life tasks – and to encourage players to rely on simplified heuristics. This task was a hangman game, as shown on the screenshot below. Although this approach is, to the best of our knowledge, not common in the literature, an additional motivation behind its use was to make the game less boring, and thus increase players’ level of attention.

Figure 7: Distracting task used

B Additional results

B.1 Statistical analysis under complete information

Table 14 shows the results of the statistical analysis of comparative feedbacks under complete information. It confirms the graphical intuition that comparative feedbacks had no impact on participants’ choices.
<table>
<thead>
<tr>
<th>Specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>7.11 (0.25)***</td>
<td>2.99 (0.86)***</td>
<td>2.83 (0.83)***</td>
<td>2.81 (0.80)***</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.25 (0.32)</td>
<td>0.01 (0.28)</td>
<td>−0.01 (0.28)</td>
<td>0.00 (0.28)</td>
</tr>
<tr>
<td>First-week bill</td>
<td>0.57 (0.12)***</td>
<td>0.56 (0.12)***</td>
<td>0.58 (0.12)***</td>
<td>0.00 (0.28)</td>
</tr>
<tr>
<td>Time first week</td>
<td>0.04 (0.11)</td>
<td>0.10 (0.12)</td>
<td>0.05 (0.04)</td>
<td></td>
</tr>
<tr>
<td>Time tutorial</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.006</td>
<td>0.285</td>
<td>0.286</td>
<td>0.294</td>
</tr>
</tbody>
</table>

Table 14: Comparative feedbacks under CI (*** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$)

B.2 Purely informative feedbacks - graphical results

Optimal bill feedback

Figure 8 shows the deviations from the optimal weekly bill obtained in the control group (left panels), and the treatment group who received a message displaying the amount of the score-maximizing bill (right panels).

Figure 8: Deviations from optimal bill under II for week 1 (top) and 2 (bottom) with (right) and without (left) an optimal bill feedback at the end of week 1.
Warning feedback

Figure 9 shows the deviations from the optimal weekly bill obtained in the control group (left panels), and in the treatment group who received warning feedbacks (right panels). Participants in the treatment group whose first-week bill was higher than $9.9 received a message warning them that their first-week consumption was abnormally high (in practice, 17 out of the 52 participants in this treatment received the warning).

Figure 9: Deviations from optimal bill under II for week 1 (top) and 2 (bottom) with (right) and without (left) a warning feedback at the end of week 1.

B.3 Changes in the distribution of weekly bills

Figure 10 represents the inverse cumulative distributions of weekly bills obtained in the four groups under incomplete information. The black horizontal line corresponds to the optimal bill. Bills are ranked by decreasing order for both weeks independently, so that two data points at a given x-value are generally observed on two different participants. These graphs further support the regression results suggesting that comparative feedbacks seem to have triggered greater responses than purely informative feedbacks.
Figure 10: Inverse cumulative distribution functions of the obtained bills under incomplete information.

C Belief elicitation method

The BDM method, in one of its most common applications, consists in asking a given participant to state his belief $\hat{g}_i$ (in %) regarding the realization of a binary outcome $s \in \{s_1, s_2\}$, say $\hat{g}_1 \equiv \Pr(s = s_1)$. A number $p$ between 0 and 100% is then randomly picked. If $\hat{g}_1 > p$, the participant receives a reward $R$ if the realized state is indeed $s_1$. If $\hat{g}_1 < p$, a lottery is run in which the participant has a probability $p$ of winning $R$ (and nothing otherwise).
While incentive-compatible, this method rewards accuracy asymmetrically. Indeed, assume the participant knows perfectly which state has realized. If the realized state is $s_1$, she will state $\hat{g}_1 = 100\%$ and earn $R$ for sure. However, if the realized state is $s_2$, she will state $\hat{g}_1 = 0\%$ and the lottery will be ran with an expected gain of $E[p]R$. Hence, the knowledge of the state guarantees a sure payment if the state is $s_1$, while it does not if the state is $s_2$.

In the context of an online experiment, in the absence of a direct interaction with the experimenter, this feature may have harmful consequences in terms of reputation. Indeed, participants enrolled in a platform like Mechanical Turk often interact on online forums (e.g. Mturk Forum, Turker Nation, etc.) in order to screen the most rewarding tasks and avoid tasks whose remuneration is set very low relative to the amount of work required. As a consequence, a participant being 100% sure of her answer and receiving no payment may infer that the experiment is tricked so as to decrease participants’ earnings and subsequently advise others away from it.

This paper thus uses a modified BDM method which allows to reward accuracy symmetrically at the cost of additional complexity. This method derives from a joint work with Yves Le Yaouanq (yves.leyaouanq@econ.lmu.de). We start by noticing that the asymmetry of the BDM method comes from the definition of $\hat{g}_1$, that is from the reference state chosen. Indeed, if we ask instead participants to guess $\hat{g}_2 = \text{Prob}[s = s_2]$, the knowledge of the state will guarantee a sure payment if the state is $s_2$, while it will not if the state is $s_1$.

The classic BDM method is itself often criticized for being hard to understand for participants. As such the incremental cost of this additional complexity is likely to be outweighed, in the context of an online experiment, by the gains from granting symmetric rewards. In practice, participants were advised that “The attribution of the reward [was] such that it [was] in [their] best interest to answer [their] true best guess”, while the instruction page of the experiment made it clear that an ethic code was followed and no deception was used.
This observation highlights that, for a given binary outcome \( s \in \{s_1, s_2\} \), two different BDM methods may be used: one using state \( s_1 \) as a reference, and one using \( s_2 \) instead.

The method we propose consists in mixing both methods in order to ensure that accuracy is rewarded. More precisely, the participant is asked to state her belief \( \hat{g}_1 \equiv \text{Prob}[s = s_1] \).

Then two situations are distinguished:

- if \( \hat{g}_1 \geq 50\% \), the BDM method using \( s_1 \) as a reference is used: if \( \hat{g}_1 \geq p \) she receives the reward if the state is \( s_1 \), otherwise we run a lottery were the chances to earn the reward are \( p \) percent.

- if \( \hat{g}_1 < 50\% \), the BDM method using \( s_2 \) as a reference is used instead: if \( \hat{g}_1 \leq p \) she receives the reward if the state is \( s_2 \), otherwise we run a lottery were the chances to earn the reward are \( 1 - p \) percent.

While more costly for the experimenter (the expected reward for participants is increased), the method remains incentive-compatible. Indeed, assume the actual belief of the participant is \( g_1 \) and that \( p \) is drawn from a uniform distribution. The expected gain (from the participant point of view) of stating having a belief \( \hat{g}_1 \) is:

\[
G(\hat{g}_1) = \begin{cases} 
\int_{p=0}^{\hat{g}_1} \mathbb{E}[1_{s=s_1}] dp + \int_{p=\hat{g}_1}^{1} p dp & \text{if } \hat{g}_1 \geq 0.5 \\
\int_{p=0}^{\hat{g}_1} (1 - p) dp + \int_{p=\hat{g}_1}^{1} \mathbb{E}[1_{s=s_2}] dp & \text{if } \hat{g}_1 < 0.5 
\end{cases}
\]

Given participant’s actual belief \( g_1 \) the expected gain may be rewritten:

\[
G(\hat{g}_1) = \begin{cases} 
\hat{g}_1 g_1 + \frac{1}{2} (1 - \hat{g}_1^2) & \text{if } \hat{g}_1 \geq 0.5 \\
(1 - \hat{g}_1)(1 - g_1) + \frac{1}{2} (1 - (1 - \hat{g}_1)^2) & \text{if } \hat{g}_1 < 0.5 
\end{cases}
\]

that is:

\[
G(\hat{g}_1) = \begin{cases} 
-\frac{1}{2} (g_1 - \hat{g}_1)^2 + \frac{1 + g_1^2}{2} & \text{if } \hat{g}_1 \geq 0.5 \\
-\frac{1}{2} (g_1 - \hat{g}_1)^2 + \frac{1 + (1 - g_1)^2}{2} & \text{if } \hat{g}_1 < 0.5 
\end{cases}
\]

As a consequence, \( G(\hat{g}_1) \) is maximized at \( \hat{g}_1 = g_1 \). The elicitation method is thus incentive-compatible.
References


Solomon E Asch. Group forces in the modification and distortion of judgments. 1952.


Erik Eyster, Matthew Rabin, and Georg Weizsäcker. An Experiment on Social Mislearning. 2015.


