Incentive Mechanism Design for Responsible Data Governance: A Large-Scale Field Experiment

Christina Timko*
Ruhr-Universität Bochum, Germany, christina.timko(at)rub.de

Malte Niederstadt
Ruhr-Universität Bochum, Germany, malte.niederstadt(at)rub.de

Naman Goel
University of Oxford, Oxford, UK, naman.goel(at)cs.ox.ac.uk

Boi Faltings
EPFL, Lausanne, Switzerland, boi.faltings(at)epfl.ch

Abstract

A crucial building block of responsible artificial intelligence is responsible data governance, including data collection. Its importance is also underlined in the latest EU regulations. The data should be of high quality, foremost correct and representative, and individuals providing the data should have autonomy over what data is collected. In this paper, we consider the setting of collecting personally measured fitness data (physical activity measurements), in which some individuals may not have an incentive to measure and report accurate data. This can significantly degrade the quality of the collected data. On the other hand, high-quality collective data of this nature could be used for reliable scientific insights or to build trustworthy artificial intelligence applications. We conduct a framed field experiment (N = 691) to examine the effect of offering fixed and quality-dependent monetary incentives, on the quality of the collected data. We use a peer-based incentive-compatible mechanism for the quality-dependent incentives without spot-checking or surveilling individuals. We find that the incentive-compatible mechanism can elicit good quality data while providing a good user experience and compensating fairly, although, in the specific study context, the data quality does not necessarily differ under the two incentive schemes. We contribute new design insights from the experiment and discuss directions that future field experiments and applications on explainable and transparent data collection may focus on.

CCS CONCEPTS

• Human-centered computing ~ Human computer interaction (HCI) • Computing methodologies ~ Artificial intelligence ~ Knowledge representation and reasoning • Information systems ~ Data management systems ~ Information integration

* First and corresponding author.

Additional Keywords and Phrases

Data Quality Assessment, Experimental, Quantitative Research, Incentive Mechanism Design, Responsible AI

ACM Reference Format:

Christina Timko, Malte Niederstadt, Naman Goel, and Boi Faltings. 2023. Incentive Mechanism Design for Responsible Data Governance: A Large-Scale Field Experiment. In ACM Journal of Data and Information Quality (ACM JDIQ). ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Technologies such as big data, data mining, and artificial intelligence are promising forces for improving data-driven decision-making. But there are also several societal concerns. For example, collecting user data through surveillance, without fair compensation and autonomy, has been a common practice in the tech industry [Zuboff 2019]. Similarly, the issue of data quality and its negative impacts often do not receive sufficient attention [Sambasivan 2021]. However, with several recently proposed and enacted regulations in the European Union (EU), such as the GDPR (General Data Protection Regulation), the AI Act, the DGA (Data Governance Act), and the DSA (Digital Services Act), these practices may be about to change. The regulations stress on social responsibility, trustworthiness, consumer protection as well as users' privacy, data autonomy and informational self-determination. Data minimization, fairness, and shareability become features of future-proof good practices [European Commission 2022]. Since these legal and ethical norms cannot be incorporated into data-driven systems in hindsight, recent research suggests viewing it as a system requirement calling for responsibility by design [Abiteboul 2019]. Motivated by this, we address the problem of designing for trustworthy data collection in this paper.

Specifically, we study incentive mechanism design for sharing high-quality personally measured fitness data. Instead of automatically collecting data from a tracked device, we preserve the users' autonomy to measure and report data voluntarily. While some data-providers may be intrinsically motivated to measure and report correct data, others may not be willing to do so without incentives. This may lead to low quality data that may be difficult or impossible to improve at a later stage [Mohan 2021]. Incentives can be manipulated by strategic misreporting, further degrading data quality. Explainable incentive-compatible mechanisms with game-theoretic guarantees [Faltings 2017] are a promising way to address these problems. Using fixed incentives and an incentive-compatible peer-consistency mechanism designed to elicit subjective and unverifiable data by incentivizing truthful data reporting [Goel 2019a], we conducted an explorative field experiment. The research questions that we aimed to explore in this experiment are (1) what is the effect of varying incentives on the quality of reported data and (2) whether the way we implemented the incentive-compatible design leads to a good user experience.

The experimental design and analysis focus on measuring and assessing data quality [Heinrich 2018, Madnick 2009]. For the analysis, we used a simple quality difference measure, which enabled the comparison of the fixed incentives and the incentive-compatible mechanism groups. The quality difference measure required a third group as a reference. This third group delivered proxy ground truth observations. We find in the specific study context that the quality of the data between the two incentive groups does not differ, while data quality improves if extreme outliers are excluded from both the groups for analysis. Further, we find that the incentive-compatible mechanism provides a good user-experience and compensates fairly. Based on our insights, we discuss specific directions that future studies may focus on, such as design improvements when applying the incentive-compatible mechanism.

Data collection is one of the earlier but important stages of data governance. If an overall responsible-by-design data governance is implemented, individuals are likely to act based on the applied incentives and provide trustworthy data. Our research contributes to real-world applications of data-based technologies, while informing future research and complementing regulatory requirements.

2 Related work

The design of robust incentive-compatible mechanisms is a central theme in economics and computation. The mechanisms of interest in the context of this paper are the mechanisms for information elicitation without verification. The pioneering work in this field is due to Prelec [2004] on the Bayesian Truth Serum and Miller et al. [2005] on the peer-prediction method. A lot of progress [Liu 2017, Radanovic 2016, Shnayder 2016] has since been made to make the mechanisms suitable for practical use in a variety of scenarios like opinion feedback elicitation (e.g. product reviews on e-commerce websites), participatory sensing (e.g. pollution measurements), human computation (e.g. microwork), etc. Faltings and Radanovic [2017] provide a comprehensive overview. Much of the work in this field is devoted to theoretical analysis and guarantees, but there have also been a few empirical studies in this area [Faltings 2014, Gao 2014]. These mechanisms work for discrete signals about phenomena that can be observed by multiple agents.

One exception is the Personalized Peer Truth Serum (PPTS) of Goel and Faltings [2019a]. PPTS is a game-theoretic incentive mechanism to elicit multi-attribute personal measurements from rational agents. Personal measurements may be about a phenomenon that can be observed only personally. To the best of our knowledge, this is the only incentive mechanism in the literature that meets the requirements of our experiment, in which we ask the participants to share their physical activity measurements. The mechanism is based on the logarithmic scoring rule [Gneiting 2007] and is a peer-consistency mechanism [Faltings 2017]. Peer-consistency mechanisms are incentive mechanisms to elicit correct information when there is no way to determine the correctness of the information, which is the case with subjective information related to physical activity.

The idea behind these mechanisms uses the fact that the information is often provided by multiple agents, the peers. A naive mechanism is the output agreement mechanism [Waggoner 2014]. In the output agreement, two agents may be asked to review the same product and if they both provide the same review, they both get $1. Otherwise, they both get $0. It is obvious that this naive mechanism works only under very strong assumptions. If an agent believes that the other agent is most likely to have the same opinion, then a truth-telling equilibrium prevails. This basic idea has been improved significantly in the literature and there are now several mechanisms that make truth-telling an equilibrium even under weaker belief assumptions. In case of personal measurements such as physical activity data, the peer-relationship between the agents is not clear because every agent measures and shares data about its own body (unlike products on an e-commerce website that can be used and experienced by many agents).

PPTS defines the peer-relationship by clustering agents based on similarity in correlated attributes. When agents share data about multiple correlated attributes (say X_1, X_2, X_3), the agents can be clustered based on similarity in the shared data in other attributes (say X_2, X_3). Now, the rewards of each of the agents for the remaining attribute (X_1) can be calculated. The score of an agent for an attribute is the ratio of the likelihood of the shared value of the attribute in the cluster of the agent and the likelihood of the shared value of the attribute in the overall population. Informally:

$$Score\ for\ X_1 = log \frac{Likelihood\ of\ X_1\ in\ the\ agent's\ cluster}{Likelihood\ of\ X_1\ in\ the\ overall\ population}$$

The higher this ratio, the higher is the score of the agent. This mechanism is incentive compatible, i.e. truth-telling equilibrium prevails and other (non-truthful) equilibria are not more profitable in this mechanism [Goel 2019a]. The score outcomes can then be scaled appropriately (as per budget constraints and fairness requirements) to calculate the rewards in euros for each agent. Depending on the structure of the collected correlated data, clustering may require large samples, which is usually the case in crowdsourcing and big data technologies.

3 Experimental design and hypotheses

3.1 Framing and task

We designed a web-based framed field experiment [Harrison 2004], which had a potential real-world context. We used oTree [Chen 2016] to implement the experiment. Participants were recruited between 11/2020 – 06/2021 through the Clickworker platform in Germany. They were presented with a crowdsourcing task aiming to collect fitness data suitable for training artificial intelligence algorithms. The task asked to report fitness data generated through a 15-20 minutes light or moderate outdoor activity. Light outdoor activity was defined by walking, and moderate outdoor activity as an activity that noticeably accelerates the heart rate, like Nordic walking, brisk walking, or light running. Fitness data to be reported contained seven measures correlated by individual physiology: *Walk or Run Time, Distance, Average Pace, Fastest Pace, Ascent, Descent,* and *Energy Burn*. The design did not require any personal identifiers or characteristics such as step length or body mass index, and there was no need to automatically track participants' fitness wearables. Thus, the design respected data regulations.

The study consisted of three easy steps: (1) agreeing to the informed consent, (2) downloading the Walkmeter app to the smartphone and collecting the fitness data, and (3) reporting the data with a chance to win 50 euros. The informed consent highlighted that we do not ask participants to do any outdoor physical activity that they would not usually do, and therefore we do not incentivize the execution of the physical activity. As a reference, crowdworking platforms usually require the minimum hourly wage, which then amounted to around 9.50 euros. By paying a fixed amount of 1.05 euros, we incentivized only data collection and data reporting, which took about 3-5 minutes. In a real-world application, additional factors such as value of data would also be considered. The freely available Walkmeter app ensured that all participants use the same means of data collection. Further, we explained to the participants that we decided for the Walkmeter app, as it is commonly used and the free version does not require registration with personally identifiable data such as an email address. Thus, we preserved anonymous data collection. For reporting the fitness data, participants had three days' time to return to the study website and submit their report. The chance to win one of ten lotteries worth 50 euros was available in each treatment group but varied by group. We did not offer the participants any personalized service, hence they did not benefit directly from measuring and reporting the data correctly and truthfully.

We also collected demographic and post-study feedback data. Demographic data includes information on weight, height, age, and gender. Post-study data explores feedback on the user experience related to the incentive design. For more details on the instructions, see Appendix B.

3.2 Treatments and hypotheses

The experiment had three treatment groups: the Proxy ground truth (P), the Fixed incentives (F), and the Quality incentives (Q) group. In the Quality incentives group, we rewarded participants based on the above described PPTS mechanism, so that increasing incentives were aligned to the increasing quality of the reported data. Depending on their PPTS-score for all the seven fitness data entries, participants had varying chances to win 50 euros. Using comprehension checks, we made sure that they understood the basic features of the PPTS mechanism (i.e. truthful and accurate data entries increase the quality of the automated peer grouping and entries score higher if they are more common in their own group than overall). Moreover, we made sure that they understood that they could influence their PPTS-scores, so that truthfulness and accuracy of their entries increases their PPTS-scores, and thus individual chances of winning. In the Proxy ground truth and Fixed incentives groups, the chance of winning the 50 euros was fixed (distributed equally among the participants) and independent from the content of the data entries. Comprehension checks made sure that participants understood this. In the Proxy ground truth group, we additionally required participants to submit a screenshot of the reported data, which served as a proof for data correctness. Thus, the Proxy ground truth group delivered a proxy of the ground truth data distribution to compare against. In the Fixed incentives group, we made it (by the instructions)

as salient as possible that data collection is uncontrolled. So, we expected to observe the most dishonesty and inaccuracy in this group.

In order to estimate the assumed fixed effect of quality difference (by lying or inaccuracy) in the Quality incentives and Fixed incentives groups, we normalize the data in these groups by the outcome of the Proxy ground truth group and pool the data as a panel. We then hypothesize that the quality difference in the Quality incentives group is below the quality difference in the Fixed incentives group. Since the PPTS mechanism requires clustering of agents, ideally, we targeted to recruit 500 participants equally in each treatment group. In this way, the experimental design, including the treatment groups, the analysis plan, and the hypothesis are consistently aligned with each other.

However, recruitment turned out to be very challenging and we will discuss in Section 5 on how to improve this in a future study or real-world application. We ended up recruiting the groups sequentially, and recruited only 691 participants altogether, out of which 501 were in (Q), 90 in (F), and 100 in (P). Each group was gender-balanced. In the Quality incentives and Proxy ground truth groups, the participation rate was around 60%, compared to the roughly 80% in the Fixed incentives group. We will discuss in Section 5 whether the reason for this difference in the participation rate might be due to a self-selection bias and how this could be addressed. For an overview of the treatment conditions, see Table 1. For details on the recruitment and participation rates see Section A.1 in Appendix A and for the balance table see Table A.2 also in Appendix A. For further details on how the 50 Euros lotteries were paid in the Quality incentives group at the end of the experiment, please see Section A.3 in Appendix A.

Table 1: Overview of the treatment conditions

Treatment	Number of Participants	Incentives (chance to win 50 euros)	Proof (take screenshot)
Proxy ground truth	100	Fixed chances	Yes
Fixed incentives	90	Fixed chances	No
Quality incentives	501	Quality-dependent chances (peer-based)	No

4 Results

4.1 Data quality

Descriptive statistics indicate that there are extreme outliers in the sample (see Table A.4 in Appendix A). For example, the maximum value for energy is 283,500.00 calories, which is clearly not possible for a human to burn in the course of a 20-minute-walk. Although one can argue that excluding extreme outliers may bias our results, not excluding them overlays our data with a huge variance and noise resulting from unsophisticated lying and inaccuracy that blurs the results.

To resolve this dilemma, we analyze all the three levels of data cleaning and will discuss the results for: (1) full sample with n=691, (2) sample excluding data points with invalid screenshots in the Proxy ground truth group resulting in n=668, and (3) sample additionally excluding extreme outliers in all the groups resulting in n=543. Invalid screenshots, such as spam pictures or screenshots from other apps than the Walkmeter app, were submitted in 23 cases in the Proxy ground truth group. Extreme outliers were identified using the interquartile range (IQR) method [NIST/SEMATECH 2013]. We calculate the IQR by subtracting the lower quantile (25%) from the upper quantile (75%). Extreme upper (lower) thresholds for outliers are defined by adding (subtracting) the triplicate of the IQR to (from) the upper (lower) quantile. Extreme outliers are identified for each fitness attribute. If a data row contains at least one extreme outlier value, the complete data row is excluded. Thus, we exclude

another 5% (4) data points in the (P) group, 20% (101) observations from the (Q) group, and 22% (20) from the (F) group.

The correlated nature of the outcome variables (the seven fitness attributes) suggests an analysis approach of fixed effects estimation of panel data with repeated measures. Assuming that there is a fixed effect of quality difference of the collected data over all the seven outcome variables in the Fixed incentives and Quality incentives groups respectively, compared to the Proxy ground truth group, we normalize the data. Normalization yields a measure of how many standard deviations of the Proxy ground truth group the effect is away from the mean in the Proxy ground truth group. We then run a regression with fixed effects, clustering at the individual level to account for the correlated standard errors in the outcomes. Using a linear hypothesis test, we test the null hypothesis that the difference between the quality differences in the Quality incentives and Fixed incentives groups is zero. As our main result, we fail to reject the null hypothesis, but only very tightly with a p-value of 0.054, if we exclude extreme outliers and keep only data with valid screenshots in the Proxy ground truth group. Table 2 reports the results in detail.

Table 2: Fixed effects analysis

		Without extreme outliers and invalid screenshots		**With extreme outliers**			
				Without invalid screenshots		**Full Sample**	
		n_Q=400, n_F=70, n_P=73		n_Q=501, n_F=90, n_P=77		n_Q=501, n_F=90, n_P=100	
		n=543, T=7, N=3801		n=668, T=7, N=4676		n=691, T=7, N=4837	
		(Q)	(F)	(Q)	(F)	(Q)	(F)
Quality difference	Estimate	0.118	0.192	5.605	4.748	4.801	3.964
	Standard error	0.038	0.050	3.597	4.561	3.170	4.205
		(0.047)	(0.067)	(1.594)	(1.426)	(1.682)	(1.537)
	t-value	3.138	3.879	1.558	1.041	1.515	0.943
		(2.499)	(2.847)	(3.516)	(3.329)	(2.855)	(2.579)
	p-value	0.002**	0.000***	0.119	0.298	0.130	0.346
		(0.013*)	(0.004**)	(0.000***)	(0.001***)	(0.004)	(0.010)
Residuals	Minimum	-1.180		-17.073		-17.107	
	Median	-0.221		-2.439		-2.241	
	Maximum	6.391		4812.525		4812.695	
R-squared		0.004		0.001		0.000	
Linear hypothesis test	Chi-squared	3.708		0.065		0.064	
	p-value	0.054		0.799		0.800	

[a] Fixed effects analysis with and without extreme outliers and with and without invalid proofing screenshots in the Proxy ground truth group (n=543, n=668, n=691). Wald Test with clustered standard errors reported in parentheses. Significance codes are: p***<0.001, p**<0.01, p*<0.05

Although the statistical significance of the main result is marginal (p=0.054), it is worth taking a more detailed look at the results [Jung 2017]. Looking at the uncertainty measures, such as standard errors and confidence intervals (CI), we see a smaller quality difference, larger group size, and smaller CI in the Quality incentives group (CI: 0.118 ± 0.07448 or 0.04352 to 0.1925) and a larger quality difference, smaller group size, and larger CI in the Fixed incentives group (CI: 0.192 ± 0.09839 or 0.09361 to 0.2904). Thus, further experiments with equal group sizes are necessary to better evaluate the statistical significance. In a pairwise power analysis, we find negligible to small Cohen's d's and small to medium power levels (comparing (Q) and (F) d = 0.092, power = 0.109; comparing (Q) and (P) d = 0.151, power = 0.219; comparing (F) and (P) d = 0.253, power = 0.323). In order to improve statistical significance, a future study would require around 1800 participants per treatment group

(comparing (Q) and (F) d = 0.092, significance level = 0.05, power = 0.8), which is not a big sample for real-world applications.

Next to statistical significance, we also look at the practical significance focusing on effect size. The closer the quality difference to zero, the higher is the data quality. If extreme outliers are included in the analysis, the mean quality difference does not differ in the Quality incentives and Fixed incentives groups. The quality difference is roughly four to five standard deviations of the Proxy ground truth group away from the Proxy ground truth group's mean. This might be interpreted as unsophisticated lying and inaccuracy being larger than without extreme outliers. If extreme outliers with large variances are excluded, we find that the data quality improves both in the Quality incentives and Fixed incentives groups. The values of the quality difference drop below 0.2 standard deviations of the Proxy ground truth group from the Proxy ground truth mean. Although there is no meaningful difference between the quality differences for the studied groups, the fact that their quality difference measure is close to zero means that they elicit good data. This might be interpreted as sophisticated lying and inaccuracy being reduced close to zero. Thus, our main result has the practical significance to provide evidence that the PPTS mechanism can elicit good data, especially if it further gets replicated in future experiments, where incentives to deliver incorrect data may be stronger. In the context of the present study, the varying incentives did not lead to a difference in the data quality.

Figure 1 visualizes the quality differences split into the seven fitness attributes. Similar to Table 2, the closer the quality difference to zero (the midpoint of the radar chart), the higher the data quality. There are some minor quality differences between the Quality incentives and Fixed incentives groups in case of *Average Pace*, *Ascent*, and *Descent*, such that the data quality improves in the Quality incentives group, but there are no differences for the other fitness attributes.

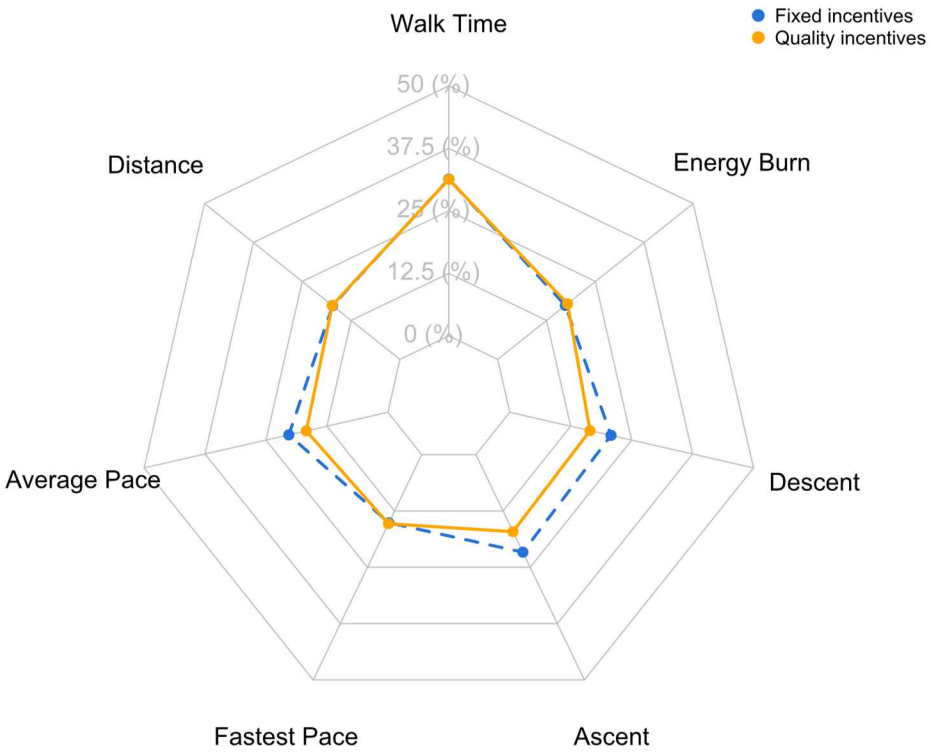

Figure 1: Radar Chart showing the mean values of the seven fitness attributes for the normalized quality differences in the Quality incentives (orange solid line) and Fixed incentives (blue dashed line) groups (n=543). Normalization is based on the Proxy ground truth group. The scaling is zoomed into the range 0-50%.

4.2 Feedback on user experience

Using post-study questions, we explore participants' feedback on their user experience with the implemented incentive design. Amongst others, we asked about *perceived effectiveness* of the applied method without emphasizing the implemented incentives: "Compared to automated data collection, what do you think about the [new/study] method?". Most participants in the Quality incentive group chose the answer option: "More people would allow having their data collected and data quality would be better.". The other two groups were less confident about the methods applied in their respective groups.

Regarding *fairness* related to the PPTS mechanism in the Quality incentive group, 70% of the participants expect to be scored fairly and 66% expect that other participants will also be scored fairly. While around a quarter of the participants are unsure about fairness.

Explainability scored around 5 percentage points lower in the Quality incentive group, compared to the other two groups. Still, 81% of the participants in the Quality incentive group felt that the new method was presented and explained properly, which is the vast majority. 13% felt confused and 5% felt overstrained and swamped, which could be improved. Interestingly, the Proxy ground truth group felt 2 percentage points less properly informed and instructed than the Fixed incentive group, which means that adding the screenshot-taking task

already increased complexity enough to have some participants more confused and overstrained and swamped. See Table A.5 in Appendix A for more details on the descriptive statistics of the post-study feedback.

5 Discussion

Studies that target the reduction of dishonest behavior [Frank 2017, Hussam 2017, John 2012, Rigol 2016], e.g. in reporting questionable research practices, show positive effectiveness of incentive mechanism designs based on peer-consistency methods. Despite their great potential, incentive designs are seldom validated nor adapted in the field and existing field studies rarely get replicated. Underlying algorithms are deemed to be too complex, so that e.g. in John et al. [2012] they are not explained in detail. Other barriers are the large sample size that is required by these big data algorithms and the lack of suitable platforms meeting the recruitment requirements. Conventional crowdsourcing platforms might have established cultures or recruitment difficulties, while online research platforms might have a strong culture to deliver reliable and high-quality data, which would require to design additional incentives to lie [Gneezy 2018] or defaults to nudge certain choices [Baillon 2022]. The present study is unprecedented in its design and was therefore risky to conduct, as the chances were high that it will be a learning-from-failure study. Although we find no significant quality differences between the studied incentive mechanisms, the applied quality-dependent incentives can elicit good data and yield a good user experience. Thus, we provide a first transparent proof-of-concept as a practical contribution, and identify dimensions that need improvement and might cause data quality problems [Madnick 2009]. These dimensions cover topics from sampling through design details to feedback loops and data cleaning.

The recruitment constraints during the data collection caused most limitations to our study. Thus, even though field experiments have a relatively high internal and external validity, their success is limited by the researchers' connections and recruitment possibilities [Roe 2009]. The conventional crowdsourcing platform that we chose could not deliver, under pandemic conditions, the required large sample. The resulting different sample sizes and room for inaccuracy by design underpower the results. For example, checking for the validity of the screenshots could have been done before calculating the payoffs. Moreover, extreme outliers could have been excluded by design, e.g. by implementing reasonable ranges and asking participants to cross-check their results for accuracy before submission. The problem of the extreme outliers seems to root in participants' negligence, inattention, and not taking the task seriously, rather than not having understood the task. What speaks for this are the descriptive statistics on explainability, which did not change in their composition after excluding extreme outliers.

Low incentives might be another problem. One way to increase incentives would be to implement the incentive design repeatedly and put individual reputation as a 'high-quality data provider' at stake in a realistic setting. Feedback on the reputation could be used as part of individual progress monitoring or in social comparison. Repeated interaction would also increase the chances for learning effects and reduce the costs of initial learning invested to understand the incentive design. Implementation in a real-world setting would address the challenging problem of recruiting a large number of participants too. Moreover, it would allow establishing a new and purposive crowdsourcing culture, aligned with the incentive design and its trustworthy and responsible framing. In contrast, conventional crowdsourcing platforms often already have a disadvantageous culture, such as inattention, self-misrepresentation, high attrition, social desirability bias, etc. that impacts data quality [Agley 2022, Aguinis 2021, Saravanos 2021].

If not caused by recruitment anomalies on the part of the crowdsourcing platform, the reported difference in the participation rates of 25 percentage points provides some evidence for a self-selection bias in the treatment groups. The self-selection might be driven by the varying complexity of the task. Any cognitive overload due to task complexity could be reduced by repeated interaction in a realistic setting, as suggested. Supporting this suggestion, Weaver & Prelec [2013] found that truthful behavior improved as participants learned the incentive mechanism through repeated interaction, even in the absence of explicit guidance. In another algorithmic context, Biermann et al. [2022] also showed that providing feedback to participants through repeated interaction is more

effective than explanation of the underlying mechanism. A future experiment may focus on, for example, the behavior of the participants in both groups as they learn through repeated interaction the incentive-compatibility of the peer-consistency mechanism and the absence of incentive-compatibility in case of fixed incentives.

It is worth noticing that even though we excluded outliers after and not during data collection, this procedure reduced the impact of unsophisticated cheating and inaccuracy. In order to have enough observations for the peer clustering amongst a group sample of 500 observations, we initially narrowed down walk time to a 15-20 minutes range. This way, all participants also got an anchor for applying any heuristic reporting strategy, rather than completing the task. Thus, in real-world applications, it may be a good idea to not only use individual reputation as an incentive, but also additional measures (e.g., including limited ground truth [Goel 2019b]) to make sophisticated cheating too complex and costly enough to raise the overall trustworthiness of the mechanism design.

In the context of fitness data, Zhou and Zhu [2022] recently showed that presenting calorie-equivalent exercise data is effectively nudging consumers to healthier food choices. Specifically, if food labels contain precise, rather than rounded exercise data, the intervention is more effective. "For instance, a chocolate bar with a calorie content of approximately 300 kcal may have a 5 km walk shown in its exercise data or, more precisely, a 4.87 km walk or a 5.13 km walk" [Zhou 2022]. Thus, data quality matters in our context for the most various applications, and its requirements (e.g. range, scale, interpretability, feasibility, acceptability, etc.) need to be defined in advance [Heinrich 2018].

Finally, the post-study survey could additionally ask whether participants were aware that they were in the respective uncontrolled group, as well as collect qualitative data via interviews or focus groups. Such feedback loops could be of key importance in repeated interaction settings in order to monitor how participants learn over time. Learning might also lead to abusing the mechanism, and thus further adjustment of the incentive design might be necessary. For monitoring continual and personalized adjustment in repeated and adaptive intervention settings, sequential randomized trials proved to be less limiting and more advantageous than traditional A/B-testing [NeCamp 2019]. Again, for this purpose, real settings can be more adequate than conventional crowdsourcing platforms. Another improvement might be to initially explain the motivation behind using the quality-dependent incentives. Motivations are, for example, the consumer protection requirements by the new EU regulations or simply the aims to increase data quality and trustworthiness in the data collection procedure. Moreover, it might be helpful to explain why the mechanism is complex. For example, PPTS does not require the participants to submit proof of correctness of their data points or execution of any spot-checking or surveillance on the participants to calculate incentives, which is compensated by comparing peers based on correlated data, making cheating complex and unnecessary. On top of that, PPTS can also be used for data cleaning, not only for incentives [Goel 2020]. Data cleaning by design may require telling the participants in advance about its procedures if, for example, this procedure may affect their payoffs or behavior.

6 Conclusion

Via a large-scale framed field experiment, we demonstrate and discuss the challenges and opportunities that lie in incentive mechanism design for high-quality data sharing or collection, respecting the human-centric European responsible data governance principles. We implemented and explained the incentive mechanism in a transparent and easily comprehensive way, informing about the payoff risks of dishonest behavior and rewards for honest behavior, and leaving participation anonymous, voluntary, and untracked. We observe that incentive design can be effective in eliciting high-quality data in the context of unverifiable and personally measured fitness (physical activity) data. Moreover, we discussed the pitfalls of a challenging large-scale experiment related to design, recruitment, and analysis. In the post-study survey, participants reported they had a good user experience. Most of them felt properly instructed with a relatively high perceived effectiveness and fairness of the incentive mechanism. An ideal future experiment would include a repeated interaction, individual reputation as an additional stake and a real-world setting by design, in addition to a larger number of participants. Our research

thus contributes to improving the design and processes of future experiments and real-world applications utilizing incentive-compatible peer-consistency mechanisms for improving the quality of data.

ACKNOWLEDGMENTS

We thank Gauthier Boeshertz from EPFL for his tireless assistance through many iterations and tests of the oTree code. Especially we thank Marc Kaufmann from the Central European University for his invaluable feedback during the analysis work. We also thank the participants of the ASFEE Conference 2022 in Lyon and colleagues for their valuable comments improving our work. Christina Timko was supported by a personal research grant of the Research Department Closed Carbon Cycle Economy at the Ruhr-Universität Bochum.

7 History dates

Received September 2022; revised February 2023; accepted March 2023

REFERENCES

<bib id="bib1" label="Abiteboul, 2019">Serge Abiteboul and Julia Stoyanovich. 2019. Transparency, Fairness, Data Protection, Neutrality: Data Management Challenges in the Face of New Regulation. Journal of Data and Information Quality 11, 3, Article 15.</bib>
<bib id="bib2" label="Agley, 2022">Jon Agley, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo. 2022. Quality control questions on Amazon's Mechanical Turk (Mturk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. Behavior Research Methods 54, 885–897.</bib>
<bib id="bib3" label="Aguinis, 2021">Herman Aguinis, Isabel Villamor, and Ravi S. Ramani. 2021. Mturk Research: Review and Recommendations. Journal of Management 47(4), 823–837.</bib>
<bib id="bib4" label="Baillon, 2022">Aurélien Baillon, Han Bleichrodt, and Georg D. Granic. 2022. Incentives in surveys. Journal of Economic Psychology 93, 102552.</bib>
<bib id="bib5" label="Biermann, 2022">Jan Biermann, John Horton, and Johannes Walter. 2022. Algorithmic Advice as a Credence Good. ZEW Discussion Paper. No. 22, 071.</bib>
<bib id="bib6" label="Chen, 2016">Daniel L. Chen, Martin Schonger, and Chris Wickens. 2016. oTree - An open source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9 (2016), 88–97.</bib>
<bib id="bib7" label="European Commission, 2022">European Commission. 2022. (Press Release) Data Act: Commission proposes measures for a fair and innovative data economy. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113. Accessed: 2023-01-18</bib>
<bib id="bib8" label="Faltings, 2014">Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. 2014. Incentives to Counter Bias in Human Computation. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 2(1), 59–66.</bib>
<bib id="bib9" label="Faltings , 2017">Boi Faltings and Goran Radanovic. 2017. Game theory for data science: Eliciting truthful information. Synthesis Lectures on Artificial Intelligence and Machine Learning 11, 2 (2017), 1–151.</bib>
<bib id="bib10" label="Frank, 2017">Morgan R. Frank, Manuel Cebrian, Galen Pickard, and Iyad Rahwan. 2017. Validating Bayesian truth serum in large-scale online human experiments. PLoS ONE 12(5), e0177385.</bib>
<bib id="bib11" label="Gao, 2014">Xi Alice Gao, Andrew Mao, Yiling Chen, and Ryan Prescott Adams. 2014. Trick or treat: putting peer prediction to the test. In Proceedings of the fifteenth ACM conference on Economics and computation. 507–524.</bib>
<bib id="bib12" label="Gneezy, 2018">Uri Gneezy, Agne Kajackaite, and Joel Sobel. 2018. Lying Aversion and the Size of the Lie. American Economic Review 108(2), 419–53.</bib>
<bib id="bib13" label="Gneiting , 2007">Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association 102, 477 (2007), 359–378.</bib>
<bib id="bib14" label="Goel , 2019a">Naman Goel and Boi Faltings. 2019a. Personalized peer truth serum for eliciting multi-attribute personal data. In Uncertainty in Artificial Intelligence. PMLR, 18–27.</bib>
<bib id="bib15" label="Goel , 2019b">Naman Goel and Boi Faltings. 2019b. Deep Bayesian trust: a dominant and fair incentive mechanism for crowd. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI'19). AAAI Press, Article 247, 1996–2003.</bib>
<bib id="bib16" label="Goel, 2020">Naman Goel. 2020. Truthful, Transparent and Fair Data Collection Mechanisms. PhD Thesis. EPFL.</bib>
<bib id="bib17" label="Harrison , 2004">Glenn W. Harrison and John A. List. 2004. Field experiments. Journal of Economic Literature 42 (4), 1009–1055.</bib>
<bib id="bib18" label="Heinrich, 2018">Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. 2018. Requirements for Data Quality Metrics. Journal of Data and Information Quality 9, 2, Article 12.</bib>
<bib id="bib19" label="Hussam, 2017">Reshmaan Hussam, Natalia Rigol, and Benjamin Roth. 2017. Targeting high ability entrepreneurs using community information: Mechanism design in the field. Unpublished Manuscript (2017).</bib>
<bib id="bib20" label="John, 2012">Leslie K John, George Loewenstein, and Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science 23, 5 (2012), 524–532.</bib>
<bib id="bib21" label="Jung, 2017">Inkyung Jung. 2017. Some Facts That You Might Be Unaware of About the P-Value. Arch Plast Surg. 44(2): 93–94.</bib>
<bib id="bib22" label="Liu , 2017">Yang Liu and Yiling Chen. 2017. Machine-learning aided peer prediction. In Proceedings of the 2017 ACM Conference on Economics and Computation. 63-80.</bib>
<bib id="bib23" label="Madnick, 2009">Stuart E. Madnick, Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. 2009. Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality 1, 1, Article 2.</bib>

<bib id="bib24" label="Miller, 2005">Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. Management Science 51, 9 (2005), 1359–1373.</bib>
<bib id="bib25" label="Mohan , 2021">Karthika Mohan and Judea Pearl. 2021. Graphical models for processing missing data. Journal of the American Statistical Association 116.534 (2021): 1023–1037.</bib>
<bib id="bib26" label="NeCamp, 2019">Timothy NeCamp, Josh Gardner, and Christopher Brooks. 2019. Beyond A/B Testing: Sequential Randomization for Developing Interventions in Scaled Digital Learning Environments. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19). Association for Computing Machinery, New York, NY, USA, 539–548.</bib>
<bib id="bib27" label="NIST/SEMATECH, 2013">NIST/SEMATECH. 2013. e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm Accessed: 2023-01-18</bib>
<bib id="bib28" label="Prelec, 2004">Dražen Prelec. 2004. A Bayesian truth serum for subjective data. science 306, 5695 (2004), 462–466.</bib>
<bib id="bib29" label="Radanovic, 2016">Goran Radanovic, Boi Faltings, and Radu Jurca. 2016. Incentives for effort in crowdsourcing using the peer truth serum. ACM Transactions on Intelligent Systems and Technology (TIST) 7, 4 (2016), 1–28.</bib>
<bib id="bib30" label="Rigol , 2016">Natalia Rigol and Benjamin Roth. 2016. Paying for the Truth: The Efficacy of a Peer Prediction Mechanism in the Field. Unpublished Manuscript (2016).</bib>
<bib id="bib31" label="Roe , 2009">Brian E. Roe and David R. Just. 2009. Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments, and Field Data. American Journal of Agricultural Economics 91, 5, 1266–71.</bib>
<bib id="bib32" label="Sambasivan, 2021">Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 39, 1–15.</bib>
<bib id="bib33" label="Saravanos, 2021">Antonios Saravanos, Stavros Zervoudakis, Dongnanzi Zheng, Neil Stott, Bohdan Hawryluk, Donatella Delfino. 2021. The Hidden Cost of Using Amazon Mechanical Turk for Research. In HCI International 2021 - Late Breaking Papers: Design and User Experience. HCII 2021. Lecture Notes in Computer Science 13094. Springer, Cham.</bib>
<bib id="bib34" label="Shnayder, 2016">Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. 2016. Informed truthfulness in multi-task peer prediction. In Proceedings of the 2016 ACM Conference on Economics and Computation. 179–196.</bib>
<bib id="bib35" label="Waggoner , 2014">Bo Waggoner and Yiling Chen. 2014. Output agreement mechanisms and common knowledge. In Second AAAI Conference on Human Computation and Crowdsourcing.</bib>
<bib id="bib36" label="Weaver , 2013">Ray Weaver and Drazen Prelec. 2013. Creating truth-telling incentives with the Bayesian truth serum. Journal of Marketing Research 50, 3 (2013), 289–302.</bib>
<bib id="bib37" label="Zhou , 2022">Li Zhou and Guowei Zhu. 2022. Mind the gap: How the numerical precision of exercise-data-based food labels can nudge healthier food choices. Journal of Business Research 139, 354–367.</bib>
<bib id="bib38" label="Zuboff, 2019">Shoshana Zuboff. 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs.</bib>

A APPENDICES

This section contains additional results.
Data and analysis codes in R are available at:
https://osf.io/kj95x/?view_only=f096326afa0e4506bc481d5b6b1152d1

A.1 Recruitment and participation rates

The data for the (Q) group was collected from 16/11/2020 to 12/03/2021, for the (F) group from 11/03/2021 to 03/05/2021, and for the (P) group from 28/04/2021 to 30/06/2021. Note that data was collected sequentially. The reason for this was limited performance met by the recruitment platform. They delivered almost 700 sequentially collected valid observations, despite the originally offered randomized data collection of a total of 1500 observations. The participants were not qualified to be part of more than one group and groups were gender-balanced. We asked participants whether local COVID-19 measures affected their usual fitness behavior during the study, but did not find it to be a confounding factor in our analysis.

In the (Q) group we had 784 persons clicking on the study link altogether. 24 persons dropped out right at the welcome screen before entering any information. 115 persons were not willing to download the Walkmeter app, another 19 stated that they usually are not physically active outdoors, and one person stated to be below age 18. Since these preconditions were necessary to enter the study, these persons were not allowed to participate. Six persons dropped out when reading or hearing the informed consent. 26 never returned after reading the instructions on how to use the Walkmeter app. 25 dropped out after reading the task description and detailed information about the earnings, out of which three persons clicked on the calculation formula. Based on the mistakes in the answers to the comprehension questions, we guess that these 25 persons had difficulties understanding the earnings calculation. At the reporting stage, 67 persons were excluded for entering invalid

entries, having missing data, or not filling out the report at all. Finally, we ended up with 501 observations, which is a participation rate of 63,90%.

In the (F) group we had 111 persons clicking on the study link altogether. 15 persons were not willing to download the Walkmeter app, and another two stated that they usually are not physically active outdoors. At the reporting stage, four persons were excluded for entering invalid entries, or not filling out the report at all. Finally, we ended up with 90 observations, which is a participation rate of 81,08%.

In the (P) group we had 170 persons clicking on the study link altogether. Five persons dropped out right at the welcome screen before entering any information. 26 persons were not willing to download the Walkmeter app, and another three stated that they usually are not physically active outdoors. Three persons dropped out when reading or hearing the informed consent. Five never returned after reading the instructions on how to use the Walkmeter app. One person dropped out after reading the task description and detailed information about the earnings. At the reporting stage, 27 persons were excluded for entering invalid entries, or not filling out the report at all. Finally, we ended up with 100 observations, which is a participation rate of 58,82%. On top of that, in parts of the analysis another 23 persons were excluded for uploading invalid screenshots. We used the valid screenshots to correct any typos and small mistakes by participants who otherwise filled out the report correctly.

A.2 Balance table for the full sample (n=691)

Table A.2: Balance table for the full sample (n=691) with tests for all group comparisons

		Share female	Share aged 18-27	Height	Weight
Proxy ground truth (P)	Mean	49%	55%	173.7	73.7
	St. Dev.			9.525	16.663
	N	100	100	100	100
Fixed incentives (F)	Mean	50%	67.8%	172.9	74.022
	St. Dev.			10.307	18.519
	N	90	90	90	90
Quality incentives (Q)	Mean	48.1%	53.9%	174.166	76.218
	St. Dev.			10.073	17.915
	N	501	501	501	501
Test	P=F	χ^2=0	χ^2=6.591	F=0.309	F=0.016
	P=Q	χ^2=0.003	χ^2=3.157	F=0.181	F=1.684
	F=Q	χ^2=0.047	χ^2=8.853	F=1.196	F=1.134

[a] None of the demographic variables shows statistically significant differences at conventional levels. Significance codes are: * p<0.1; ** p<0.05; *** p<0.01

A.3 Details on lotteries in the Quality incentives group

In our experiment, the participants did not interact with the incentive mechanism in a repeated manner. We explained the mechanism to the participants before they measured and reported their data, and they were paid after they had reported the data. Therefore, the actual payments were not directly related with the observed behavior of the participants in this experiment. However, for completeness, this section provides details on how the lotteries in the (Q) group were drawn and final payments were made according to the PPTS mechanism.

We computed the clusters for every participant and for each of the seven fitness attributes, using the values reported by the participant and the other participants on rest of the attributes. We treated the k (=50) nearest neighbors of any participant, according to Euclidean distance metric, as their cluster. Having defined the clusters, we computed the attribute score of any participant using the logarithm-based scoring formula given in [Goel 2020]. The same process was repeated for each of the attributes and for every participant. The final score for any participant was the average of all the attribute scores for that participant. To decide lottery winners from these

scores, we normalized the scores and drew lottery (averaged over several thousand draws), in which the participants were assigned probabilities of winning that were proportional to their respective final scores.

A.4 Descriptive statistics for the full sample (n=691)

Table A.4: Descriptive statistics for the full sample (n=691)

Statistic	N	Mean	St. Dev.	St. Err.	Min	Pctl(25)	Median	Pctl(75)	Max
Walk Time (min)									
(Q)	501	18.299	3.465	0.155	10.000	16.017	18.083	20.200	25.850
(P)	100	18.787	3.702	0.370	10.000	16.196	18.508	20.862	26.417
(F)	90	18.908	3.707	0.391	10.000	16.508	18.908	20.300	25.350
Distance (km)									
(P)	100	2.112	2.249	0.225	0.060	1.197	1.530	2.090	20.000
(F)	90	2.209	2.190	0.231	0.660	1.252	1.730	2.575	20.000
(Q)	501	4.385	50.323	2.248	0.000	1.200	1.561	2.160	1126.538
Average Pace (min/km)									
(P)	100	19.282	46.243	4.624	1.100	9.892	11.967	14.665	371.033
(Q)	501	52.876	139.355	6.226	0.000	9.567	12.317	16.200	779.300
(F)	90	53.428	129.012	13.599	0.167	8.496	11.291	16.363	727.050
Fastest Pace (min/km)									
(P)	100	15.959	71.502	7.150	1.100	5.900	7.875	9.790	720.000
(Q)	501	46.509	126.776	5.664	0.000	6.400	8.900	11.717	725.050
(F)	90	58.961	147.637	15.562	0.083	5.525	8.358	12.113	725.100
Ascent (m)									
(P)	100	52.697	99.761	9.976	0.000	0.000	12.000	59.750	609.000
(Q)	501	56.356	158.361	7.075	0.000	0.000	14.000	57.000	2000.000
(F)	90	63.997	129.703	13.672	0.000	0.088	19.000	93.750	1000.000
Descent (m)									
(P)	100	45.136	95.595	9.560	0.000	0.000	10.500	46.000	608.000
(Q)	501	51.930	142.001	6.344	0.000	0.000	12.000	51.000	2000.000
(F)	90	60.572	133.469	14.069	0.000	0.000	12.500	91.250	1000.000
Energy Burn (calories)									
(P)	100	131.360	103.618	10.362	38.000	74.000	89.000	133.800	591.000
(F)	90	182.647	340.424	35.884	20.000	75.000	102.000	169.500	3000.000
(Q)	501	700.738	12660.460	565.628	0.000	71.000	95.000	146.000	283500.000

A.5 Post-study feedback on user experience for the full sample (n=691)

Table A.5: Post-study feedback on user experience for the full sample (n=691)

Questions and answer options	**Counts and percentages**		
	n_Q=501	n_F=90	n_P=100
Do you think that truthful and accurate reporting was the best strategy in this study?			
Yes	477 (95.21%)	87 (96.67%)	97 (97%)
No	24 (4.79%)	3 (3.33%)	3 (3%)
Do you think that the (new) method used in this study is a useful method to collect truthful and accurate entries on personally observable data, such as fitness- and health-related data?			

Questions and answer options	Counts and percentages		
	n_Q=501	n_F=90	n_P=100
Yes	432 (86.23%)	82 (91.11%)	88 (88%)
No	69 (13.77%)	8 (8.89%)	12 (12%)
The (new) method used in this study allows the collection of personal data in high quality, while preserving anonymity and without the need to automatically track users via apps or wearable devices. As a user, what would you prefer?			
I prefer the new method.	269 (53.69%)	51 (56.67%)	61 (61%)
I prefer automated tracking and data collection.	117 (23.35%)	20 (22.22%)	20 (20%)
I don't care.	115 (22.95%)	19 (21.11%)	19 (19%)
Compared to automated data collection, what do you think about the [new/study] method?			
Less people would allow having their data collected and data quality would be worse.	91 (18.16%)	19 (21.11%)	33 (33%)
Less people would allow having their data collected and data quality would be better.	134 (26.75%)	26 (28.89%)	25 (25%)
More people would allow having their data collected and data quality would be worse.	52 (10.38%)	19 (21.11%)	13 (13%)
More people would allow having their data collected and data quality would be better.	224 (44.71%)	26 (28.89%)	29 (29%)
More and more smartphone apps offer personalized services that are based on automated data collection. Which data source would you trust more?			
I would trust more the data that was collected by using the [new/study] method.	289 (57.68%)	40 (44.44%)	48 (48%)
I would trust more the data that was collected automatically.	212 (42.32%)	50 (55.56%)	52 (52%)
How did you feel about the way the [new/study] method was presented and explained? Please mark the statement that comes closest to how you felt.			
I felt properly informed and instructed.	407 (81.24%)	79 (87.78%)	86 (86%)
I felt confused.	65 (12.97%)	8 (8.89%)	10 (10%)
I felt overstrained and swamped.	23 (4.59%)	2 (2.22%)	3 (3%)
I felt intimidated in some way.	6 (1.20%)	1 (1.11%)	1 (1%)
Do you expect the new method in this study to score you fairly?			
Yes	353 (70.46%)	***In the Quality incentives group only.***	
No	21 (4.19%)		
I don't know	127 (25.35%)		
Do you expect the new method in this study to score all the study participants fairly?			
Yes	333 (66.47%)	***In the Quality incentives group only.***	
No	36 (7.19%)		
I don't know	132 (26.35%)		
How high do you expect your individual chance of winning to be, compared to other study participants? Rating from: (1) = 'very low' to (7) = 'very high'			
Mean	3.93	***In the Quality incentives group only.***	
(Standard deviation in brackets)	(1.63)		

B APPENDICES

This section contains the design details of the experiment, including an extract from the instructions. The extract contains the explanation of the PPTS mechanism to the participants. For the oTree codes and the full-text of the instructions, both in English and the original German version, visit the online supplementary repository: https://osf.io/kj95x/?view_only=f096326afa0e4506bc481d5b6b1152d1

B.1 Scoring of single entries

We use a scientifically validated mechanism to assess the truthfulness of your **entries to every single fitness attribute** as listed above, such as walk or run time, distance, pace, energy burn, etc. Each entry gets scored.

Truthful and accurate entries get higher scores than untruthful and inaccurate entries. We calculate the individual chance of winning from the **sum of scores for every single fitness data attribute**.

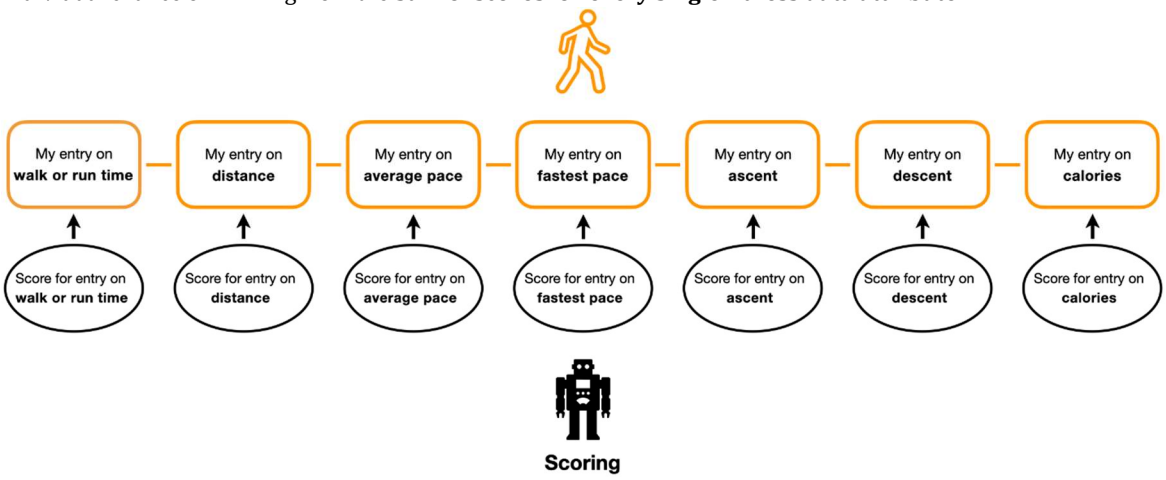

Figure B.1: Scoring of single entries explained graphically

B.2 Grouping

In order to assess the truthfulness of your single entries, for example your entry for distance, we compare your entry with the entries of 'peers'. 'Peers' are those study participants, who reported similar values than you to **all other entries**, other than distance in this example. Since all fitness data attributes are connected to each other through human physiology, 'peers' are very likely to be very similar in their physiology. Peer grouping is done by an algorithm. By this, we can still assess your entries in a personalized manner, yet anonymously. The more truthful and accurate your entries are, the better the quality of the peer grouping will be.

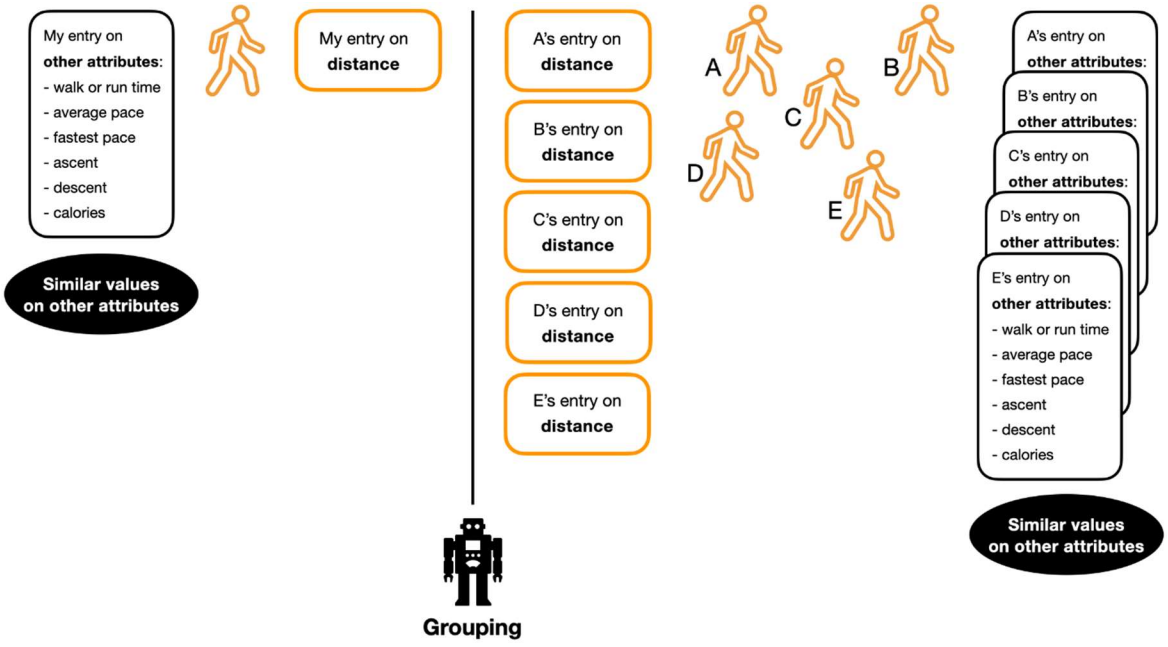

Figure B.2: Grouping explained graphically

B.3 Comparing

Your entries will score higher if they are **more common in your group than overall among all study participants**. Thus, to maximize your scores, make your entries as truthful and accurate as you can.

Click here, if you would like to see the **calculation formula**.

<For *distance* for example we apply:

$$\text{Score for } distance = log \frac{\text{Commonness of your } distance \text{ entry in your group}}{\text{Commonness of your } distance \text{ entry overall}}$$

where the logarithmic function (log) is an increasing function, and commonness is the likelihood of your entry in the distribution of the same entries.>

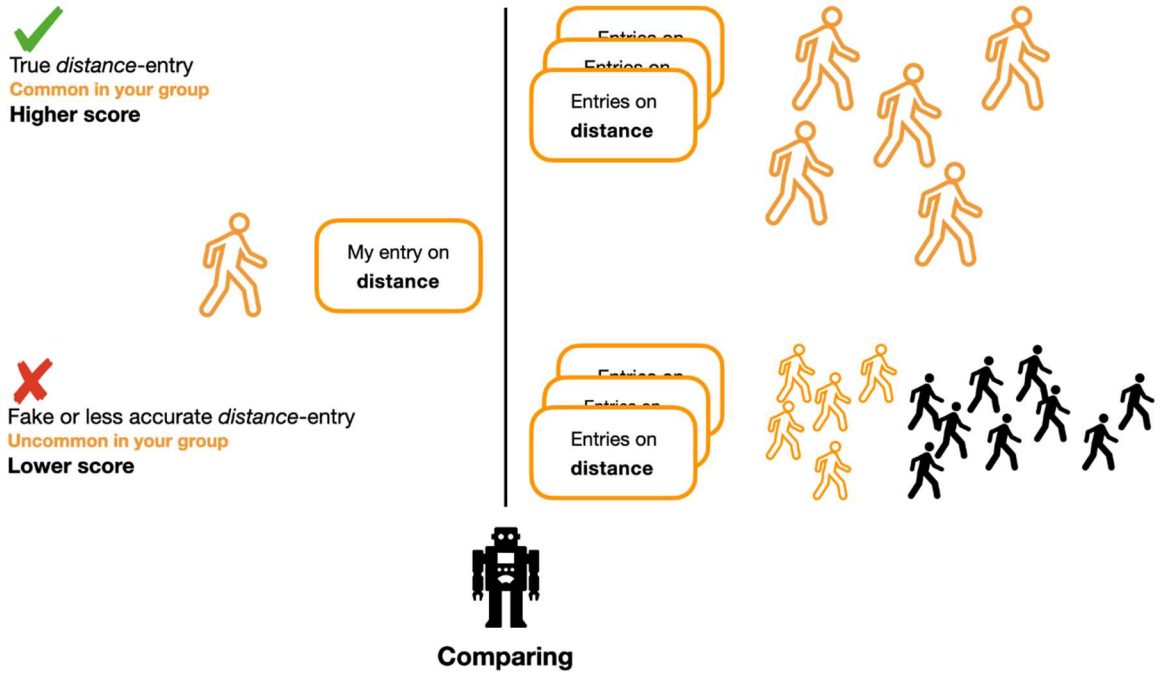

Figure B.3: Comparing explained graphically

Since it is individually the best strategy to report truthfully and accurately, you can assume that **also other study participants will report truthfully and accurately**.

Since the fitness data is not simultaneously collected from all study participants, it is not possible to provide you real-time feedback on your scores. The algorithms applied in this study may cause minor variation in scores. However, this does not affect the general tendency of **true and accurate entries getting higher scores than fake or less accurate entries**.