



Understanding and Correcting Inaccurate Calorie Estimations on Amazon Mechanical Turk

Lillio Mok

Brenna Li

Stephen Gou

Computer Science, University of Toronto

Toronto, ON, Canada

{lillio,brli,gouzhen}@cs.toronto.edu

Joseph Jay Williams

Computer Science, University of Toronto

Toronto, ON, Canada

williams@cs.toronto.edu

ABSTRACT

Current research on technology for fitness is often focused on tracking and encouraging healthy lifestyles. In contrast, we adopt an approach based on improving consumer knowledge of food energy. An interactive survey was distributed on Amazon Mechanical Turk to assess how well crowdworkers can judge the calories in a series of foods. Our subjects yielded results comparable to traditional participants, exhibiting well-known phenomena such as underestimating the energy contained in foods perceived to be healthy. Several techniques from the online education literature, such as prompts for reflection, were also investigated for their efficacy at increasing estimation accuracy. Although calories were more accurately judged after applying these methods on aggregate, the effects of individual techniques on our participants were inconclusive. A more thorough investigation is thus needed into effective educational methods for correcting calorie estimations on the Web.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5971-9/19/05.

<https://doi.org/10.1145/3290607.3312764>

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; Collaborative and social computing; • **Applied computing** → *Consumer health*.

KEYWORDS

Nutrition; Mechanical Turk; online health education; learning at scale.

INTRODUCTION

The reduction of obesity is a research topic of increasing contemporary importance. Global prevalence of obesity has been on the rise, and, in the USA alone, almost 30% of adults are obese [11]. Consequences could be dire for the obese, with multiple physiological diseases linked to excess weight and mortality rates from cancer significantly elevated in those whose BMIs exceed 40 [2, 11].

Promoting healthier lifestyles and diets through technology has hence been the focus of several avenues of research in the human-computer interaction literature. These include, for example, systems for crowdsourcing nutritional information from food photos [6], mobile applications for personalising fitness and food suggestions [7], and multiple recommender systems for nutritional advice [8, 10]. Indeed, nutrition and activity trackers are now commonly used by the general public¹.

We take a different approach founded on health education. In particular, we attempt to help consumers pick healthier food choices by improving how well they assess the nutrition in common food items. This follows considerable evidence demonstrating that consumer estimates of food calories (kcal) are systematically inaccurate [1, 3–5, 9]. We thus distributed an interactive learning exercise on Amazon Mechanical Turk (MTurk), during which participants were tasked with guessing and learning the energy content of a selection of common foods. We aim to understand the following questions:


- **RQ1:** To what extent are erroneous calorie estimations made by online crowdworkers similar to those of the general populace?
- **RQ2:** Are there online education techniques that can help consumers estimate the energy in food items more accurately?

RELATED WORK

Existing work on computer-mediated solutions to the fitness problem are varied. Noronha et al., for example, designed an application through which photos of meals are passed through an MTurk-based workflow [6]. Information such as tagged ingredients, approximate serving sizes, and nutritional and energy content is crowdsourced from MTurk workers. The resulting system was demonstrated to be comparable in accuracy to human experts. Wayman and Madhvanath analysed the grocery receipts of 15 users to recommend food choices that filled nutritional gaps in their eating habits [10]. The

¹See e.g. MyFitnessPal, which can scan food barcodes to record nutritional content (www.myfitnesspal.com).

Consider the following serving of icecream:



The serving size is 3.9oz (110g), which is typical for a small soft-serve cone.

How many calories do you think this food item contains?

On the following scale, indicate how healthy you think this serving of icecream is.

| | | | | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------------------|
| | Not healthy | | | | | Very healthy | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Rate item healthiness | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> |

Figure 1: A screen-shot of the interface presenting test items.

²www.qualtrics.com

You were wrong. Your estimate was -52.4% away from the actual calories (positive means you've overestimated; negative means you've underestimated).

Ice cream is delicious but often not recommended by dieticians because of its high calories and fat content -- 100g of a vanilla ice cream can have over 135 calories, 11% of which come from fat.

Please reflect on why your calorie estimation is different from the actual calorie content.

Figure 2: A screen-shot of the intervening information presented after each test item; in this instance, the participant was assigned a harsh tone, percentage inaccuracy presentation, a specific nutritional fact, no human fact, and a reflection prompt.

authors observed that the items suggested to their participants were highly relevant, e.g. vegetables for those lacking fibre. Rabbi et al. used a Multi-Armed Bandit algorithm to personalise food and physical activities through a mobile interface [7]. The system employed a Pareto frontier algorithm to balance between user preferences and calculated health requirements.

None of these examples, however, enable their participants to make healthier decisions without the aid of their computing devices. We attempt to facilitate better *independent* food decisions by targeting inaccuracies in calorie estimation. In particular, there is an abundance of research suggesting that humans are systematically unable to accurately judge their caloric intake due to multiple biasing effects, which can have a significant negative impact on their wellbeing [1, 3–5, 9]. To correct these inaccuracies, we draw inspiration from literature on online education such as learning at scale. Techniques employed in this area include asking learners to explain their answers to questions [12], prompting them to reflect on misconceptions [13], and concurrently, sometimes adaptively, applying interventions deemed most conducive to learning [12]. Our study hence aims to investigate the calorie estimation inaccuracies of online participant pools (RQ1), and whether these inaccuracies can be reduced by some of the aforementioned educational techniques (RQ2).

STUDY DESIGN

We designed an interactive study using the Qualtrics platform for distributing surveys². This consisted of three main sections. The first, a *pre-intervention test* (“pre-test”), asked participants to estimate the calories in and perceived healthiness (1-7 Likert scale) of five food items: a burger, an icecream cone, a beer, a salad, and a muffin. These were chosen from fast-food chains with standardised meals, and were presented with their images and serving sizes obtained from the chains’ websites (Figure 1). Additional information about our participants was also collected, such as dieting status, age, and BMI.

The second, a *learning phase*, presented participants with the same five questions interleaved with “interventions”. After participants evaluated each food item, they were shown the difference between their estimation and the actual energy content (Figure 2). This was accompanied with intervening information and prompts varied in a 2x2x2x2 factorial experiment described in Table 1, the levels of which were determined at the beginning of the survey for all five items. For example, their estimation difference was either shown in calories or as a percentage of the correct answer. Additionally, a fact about each food item was always presented, but may be either specific to the item or general to its food group. The last factor displayed a reflection prompt asking the participant to enter thoughts about their estimation and why it was inaccurate [13].

Finally, participants were asked to undergo a *post-intervention test* (“post-test”) with five highly-similar food items. For example, beers of the same serving size were chosen from two different companies in the pre- and post-test items. To minimise confounds, the ordering of the items was randomised and serving sizes chosen to not be systematically larger or smaller than the pre-test [9].

Table 1: Factors and levels for intervening information

| Factor | Description | First level | Second level |
|-----------------------|-----------------------------|--|--|
| Tone | Tone of intervention | <i>Encouraging:</i> “Good attempt!” | <i>Harsh:</i> “You were wrong.” |
| Inaccuracy | Presented energy difference | <i>Calories:</i> “You were 300 cal away” | <i>Percentages:</i> “You were 30% away” |
| Nutrition Fact | Fact about test item | <i>Specific:</i> “Calories in burgers...” | <i>General:</i> “Fast food meals...” |
| Human Fact | Fact about estimations | <i>On:</i> “People often underestimate calories in “healthy” foods” | <i>Off</i> |
| Reflection | Prompt for reflection | <i>On:</i> “Reflect on how you arrived at your answer” | <i>Off</i> |

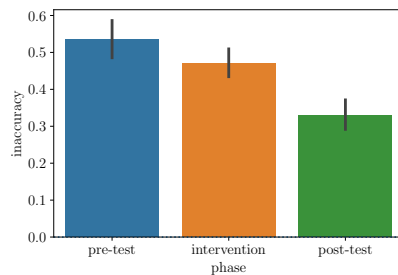


Figure 3: Barplot of average, absolute calorie estimation inaccuracies per phase (e.g. an inaccuracy of 0.5 implies an average 50% deviation from the actual calories of the items in a phase). Error bars are of the mean’s 95% confidence interval.

RESULTS

127 North-American participants were recruited on MTurk to participate in the study; 5 incomplete responses were removed yielding $N = 122$ responses. Participants were compensated \$1.20 USD for their time. Inaccuracies for each food item i and response j are calculated as $\delta_{ij} = (\gamma_{ij}^e - \gamma_i^a) / \gamma_i^a$, where γ_i^a and γ_{ij}^e are respectively the actual and estimated calories for that item. In other words, an inaccuracy of 0.3 is equivalent to a 30% overestimation. The mean inaccuracy for the food items F in a phase, averaged over every response j and item i , is calculated as $\Delta_F = \frac{1}{N|F|} \sum_{j=1}^N \sum_{i \in F} |\delta_{ij}|$. Note that the absolute difference is used to prevent cancellation of over- and underestimations.

The average absolute inaccuracy in the pre-test, intervention, and post-test phases are depicted in Figure 3. The decrease between the pre-test and post-test inaccuracies was significant ($M = 0.54$ $SD = 0.28$ vs $M = 0.33$ $SD = 0.23$; paired $t(121) = 8.03$, $p < 0.0001$), hinting at some efficacy of the interventions we applied. There was noticeable, albeit statistically marginal, support that dieters outperformed non-dieters in the pre-test (0.48 $SD = 0.26$ vs 0.58 $SD = 0.28$; $t(120) = 1.88$, $p = 0.06$).

Average inaccuracies were also calculated for each item in each phase, presented in Figure 4. There were significant differences between the inaccuracies of the salad item (-0.28 $SD = 0.36$) and, for example, the beer item (0.55 $SD = 0.78$; paired $t(121) = 10.87$, $p < 0.0001$) and the icecream item (0.60 $SD = 0.74$; paired $t(121) = 12.35$, $p < 0.0001$). This was likely due to the salad’s perception as the healthiest of the items (5.05 $SD = 1.29$ on a 1-7 Likert scale) and thus underestimation, whilst icecream (1.98 $SD = 1.14$) and beer (2.23 $SD = 1.28$) were two of the unhealthiest and overestimated.

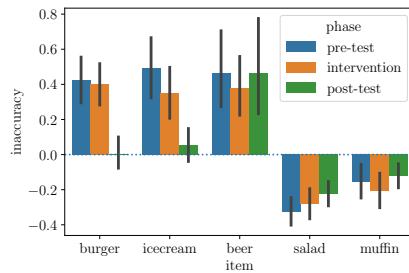


Figure 4: Barplot of average calorie estimation inaccuracies per item, per phase.

³Indeed, dieters appeared to outperform non-dieters in the pre-test but did not improve more than their counterparts. This could be explained if MTurk dieters operate similarly to dieters in the general populace, who tend to be more adept at guessing certain foods [3].

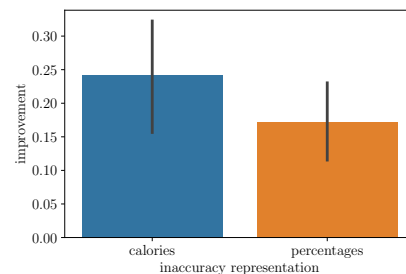


Figure 5: Barplot of pre-to-post accuracy improvements for the inaccuracy-representation experimental factor.

Pre-to-post *improvement* was then measured, defined as $\Delta_{pre} - \Delta_{post}$. For example, an improvement of 0.5 implies that each post-test item, on average, was guessed 50% more accurately than each item in the pre-test. Multiple hypothesis tests and regression analyses of improvement were run against the interventions. Displaying calorie differences appeared to have a slightly more positive impact on improvement than displaying percentage differences (0.24 $SD = 0.32$ vs 0.17 $SD = 0.24$; $t(120) = 1.38$, $p = 0.17$), shown in Figure 5, although this was statistically insignificant. Other factors we analysed did not demonstrate significant effects, even when second-order interaction terms were considered.

DISCUSSION

Perhaps the most compelling result from our study is a reproduction of earlier findings suggesting that the American populace tends to overestimate the calories in unhealthy foods, but underestimate healthy foods [3, 4]. Specifically, the healthiest test item in our experiment, salads, were clearly systematically underestimated. Contrastingly, burgers, icecreams, and beers were overestimated in our sample (see Figure 4). This suggests that online crowdworkers such as those on MTurk demonstrate a similar healthy-unhealthy bias as the general populace. With respect to RQ1, online participant pools hence show promise for future investigations into public nutritional knowledge³.

It is also encouraging that the intervening information and prompts we employed incurred an improvement in post-test accuracy (see Figure 3), particularly because human estimations of calories have notoriously high variance and implausibility [1]. Furthermore, our participants were not monetarily incentivised to provide improved estimates in the post-test. Nonetheless, the absence of significant first-order effects from well-established educational techniques like reflection prompts [13] is surprising. Thus, more work needs to be conducted to evaluate these techniques in order to answer RQ2, despite participant estimations improving as a whole in this study.

Our experiment suffers from several limitations. Firstly, it presents only a small number of items to participants, thus restricting generalisability of our results. It seems clear from Figure 4 that the burger and icecream items contributed significantly more to the pre-to-post improvement. A wider selection of test items will allow for deeper scrutiny into specific types of food. Additionally, our analyses could be expanded to account for contextual variables such as participants' BMIs. Interactions between these and our experimental factors may yield more significant effects for correcting estimation biases.

CONCLUSION

Online crowdworkers, such as those on MTurk, are a promising source of participants for studies on nutritional knowledge. Their behaviour when assessing food energy evidently mirrors those found in the general populace [3, 4], and their calorie estimations are improved by disclosing their inaccuracies with supplementary information. Although a more thorough examination of effective

learning strategies is needed, we are confident that online participant pools will allow for richer and larger-scale research into informing healthy food choices on the Web.

ACKNOWLEDGMENTS

We thank Dr. Cendri Hutcherson for her helpful comments on our experimental design and this document, and Samuel Maldonado for his help in constructing our Qualtrics study. We additionally thank members of the Multidisciplinary HCI course at the University of Toronto for piloting and commenting on this experiment. The authors gratefully acknowledge a grant from the Office of Naval Research (#N00014-18-1-2755).

REFERENCES

- [1] Edward Archer, Gregory A Hand, and Steven N Blair. 2013. Validity of US nutritional surveillance: National Health and Nutrition Examination Survey caloric energy intake data, 1971–2010. *PLoS one* 8, 10 (2013), e76632.
- [2] Eugenia E Calle, Carmen Rodriguez, Kimberly Walker-Thurmond, and Michael J Thun. 2003. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults. *New England Journal of Medicine* 348, 17 (2003), 1625–1638.
- [3] Robert A Carels, Krista Konrad, and Jessica Harper. 2007. Individual differences in food perceptions and calorie estimation: an examination of dieting status, weight, and gender. *Appetite* 49, 2 (2007), 450–458.
- [4] Pierre Chandon and Brian Wansink. 2007. The biasing health halos of fast-food restaurant health claims: lower calorie estimates and higher side-dish consumption intentions. *Journal of Consumer Research* 34, 3 (2007), 301–314.
- [5] Alexander Chernev. 2011. The dieter’s paradox. *Journal of Consumer Psychology* 21, 2 (2011), 178–183.
- [6] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z Gajos. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 1–12.
- [7] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 707–718.
- [8] Hanna Schäfer. 2016. Personalized Support for Healthy Nutrition Decisions. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 455–458.
- [9] Brian Wansink and Pierre Chandon. 2006. Meal size, not body size, explains errors in estimating the calorie content of meals. *Annals of internal medicine* 145, 5 (2006), 326–332.
- [10] Elizabeth Wayman and Sriganesh Madhvanath. 2015. Nudging Grocery Shoppers to Make Healthier Choices. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 289–292.
- [11] Ellen P Williams, Marie Mesidor, Karen Winters, Patricia M Dubbert, and Sharon B Wyatt. 2015. Overweight and obesity: prevalence, consequences, and causes of a growing public health problem. *Current obesity reports* 4, 3 (2015), 363–370.
- [12] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 379–388.
- [13] Joseph Jay Williams, Tania Lombrozo, Anne Hsu, Bernd Huber, and Juho Kim. 2016. Revising Learner Misconceptions Without Feedback: Prompting for Reflection on Anomalies. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 470–474.