Comparing informativeness of an NLG chatbot vs graphical app in diet-information domain

Simone Balloccu
University of Aberdeen, UK
simone.balloccu@abdn.ac.uk

Ehud Reiter
University of Aberdeen, UK
e.reiter@abdn.ac.uk

Abstract

Visual representation of data like charts and tables can be challenging to understand for readers. Previous work showed that combining visualisations with text can improve the communication of insights in static contexts, but little is known about interactive ones. In this work we present an NLG chatbot that processes natural language queries and provides insights through a combination of charts and text. We apply it to nutrition, a domain communication quality is critical. Through crowd-sourced evaluation we compare the informativeness of our chatbot against traditional, static diet-apps. We find that the conversational context significantly improved users understanding of dietary data in various tasks, and that users considered the chatbot as more useful and quick to use than traditional apps.

1 Introduction

Visual representations of data is commonly used to communicate insights to the reader. However, understanding the meaning of charts or other visualisations can be challenged by visual deficit, information context, or just the required cognitive effort. Previous research investigated on generating textual explanations of data and comparing them with visualisations (Gatt et al., 2009; Molina et al., 2011; Gkatzia et al., 2017). Approaches like these are particularly useful in healthcare, where lots of data get produced and communication plays a critical role (Zolnierek and DiMatteo, 2009; Brock et al., 2013). Most of these works showed that combining text and visuals improve users' understanding of data but they explored static contexts only, where information is presented in a fixed way and there is no active interaction with the reader. Little is known about the effects of text and charts combination in dynamic scenarios, such as conversational ones. Since chatbots are emerging as tools for healthcare (Zhang et al., 2020), it is important to assess if they can provide better communication than static tools (e.g. e-health apps).

In this work we develop and evaluate an NLG-chatbot that generates insights explanation by combining graphics and text. Using our chatbot, users do not need to explore or interpret data themselves, as they can directly ask what they're looking for and get it, along with explanation. We apply it to diet coaching, a domain where communication quality is critical (Van Dorsten and Lindley, 2008; Savolainen, 2010; Michie et al., 2011) and often overlooked by existing tools (Balloccu et al., 2021; Balloccu and Reiter, 2022). To assess the effectiveness of this approach, we run a human evaluation in which we compare our chatbot with traditional diet apps. Participants were assigned to either our chatbot or an app, and used it to take a 10-point quiz concerning the extraction of insights from a simulated food diary. At the end, participants expressed a feedback on the assigned tool. Results show that using our chatbot led to significantly higher scores compared to using traditional apps, both in general and with regards to particular subtopics. Feedback analysis also reveal that participants perceived our chatbot as more useful for finding diet problems and quicker to use than traditional diet apps.

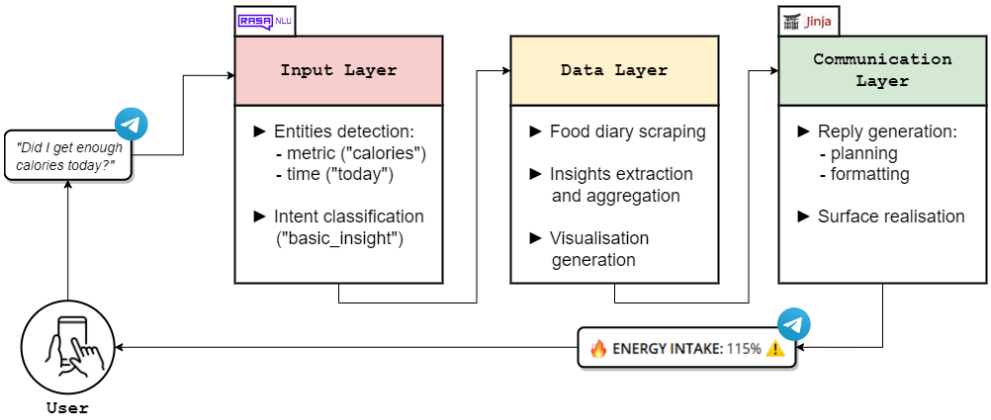


Figure 1: Chatbot architecture and interaction flow.

2 Related work

In this section we recap past research on charts and text combination for insights explanation. We first look at more general work, then move to healthcare and diet-coaching.

2.1 Text vs Graphics in NLG

Previous work investigated how NLG can enhance understanding of data by combining textual content and images. Work on weather data (Gkatzia et al., 2017), showed mixed text and pictures improving decision-making over images alone. Dashboards (Ramos-Soto et al., 2017) benefit from textual explanation of charts as well, as it helps assessing learning in students. Combining charts with explanation of sensors data (Molina et al., 2011) helps insights understanding for general users. Driving reports (Braun et al., 2015) are more helpful if presented as a mix of pictures and text. Healthcare data can also be explained through NLG (Pauws et al., 2019). Experiments in NICU (Law et al., 2005; van der Meulen et al., 2010) suggest that combining charts and text could be the preferred approach by clinicians.

2.2 Text vs Graphics in diet-coaching

Information quality and communication plays a big role in diet (Van Dorsten and Lindley, 2008; Savolainen, 2010; Michie et al., 2011). This applies to apps as well: comprehensibility showed to be a predictor of prolonged app use (Lee and Cho, 2017). Suboptimal communication can confuse and demotivate users, leading to early abandonment (Murnane et al., 2015; Mukhtar, 2016). Despite this, diet apps (like MyFitnessPal [1] or FatSecret[2]) typically come as calorie counters, where users log their meals to obtain insights. These tools adopt very limited textual communication and make extensive use of visualisations that must be interpreted by users themselves (Balloccu and Reiter, 2022). Considering the relationship between numeracy and nutrition literacy (Mulders et al., 2018), this poses a barrier between users and the delivered information. Our previous work (Balloccu et al., 2021) showed similar issues for conversational agents: chatbots adopt fixed educational material (Casas et al., 2018; Stephens et al., 2019; Davis et al., 2020), such as PDFs containing guidelines, and expose lack of reasoning over user queries (Maher et al., 2020). Similarly to apps, chatbots show plain reports, with little to no feedback on goals, progress or mistakes (Casas et al., 2018; Prasetyo et al., 2020).

3 NLG chatbot to improve communication quality

Our chatbot consists of an Input Layer for users' input understanding; a Data Layer that extracts insights and generates visualisations; a Communication Layer that per-

[1] www.myfitnesspal.com
[2] https://www.fatsecret.com/

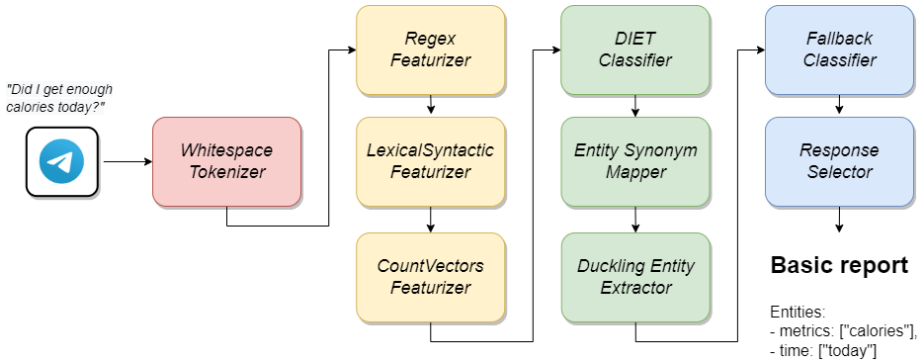


Figure 2: Overview of the NLU pipeline.

forms planning and surface realisation (Figure 1). We use RASA Open Source 2.0[3] as the main infrastructure for the entire system, and exploit its NLU component (Figure 2) for the Input layer; the Data Layer adopts a custom data analysis logic; the Communication Layer adopts rule-based NLG and variable templates (through Jinja 3.0 [4]).

We adopt an hybrid architecture: we use machine-learning for NLU but restrict text generation to rules. This is mainly for two reasons: 1) diet domain imposes strict accuracy requirements that cannot be met by current E2E NLG (Thomson and Reiter, 2020; van Miltenburg et al., 2021) and 2) to the best of our knowledge, there is no publicly available diet-coaching corpus which can be used to train or fine-tune generative models. On the other hand, machine-learning offers good generalisation for NLU with the only risk being unexpected inputs or failure in intent classification.

We model two main interactions into the chatbot: basic reports and comparisons (Figure 3). Basic reports show insights about a single time frame, either as brief information on energy and nutrients balance or combinations of charts and text. Comparisons extend basic reports to multiple time frames by informing users about progress (e.g. improved intake; changes in food choices etc..). For each request, users can specify metrics (calories and five nutrients: carbohydrates, protein, fat, sugar and sodium) and time (detected via Duckling Entity extractor[5]). This approach offers more flexibility than traditional apps, that typically aggregates all the metrics in a single section (e.g. a table) and present pre-defined comparisons (e.g. every month).

3.1 Explanation through text and charts

Users can access two typologies of insights: basic and advanced. Basic insights show energy and nutrients intake (see Figure 3) as brief textual messages. This is thought for users that need a quick glance at their data. Advanced insights deliver more information and are presented as a combination of text and charts. Users can obtain the following advanced insights (Figure 4):

1. **Intake analysis:** reasons and explains intakes with regards to user goals.
2. **Trend and consistency:** detects if trends match recommended changes in diet (e.g. getting less calories to fix an excess) and checks intake consistency (maintaining a stable intake across days).
3. **Food analysis:** reasons and explains intakes at food level, by showing which food has the biggest impact.

Advanced insights naturally extend to comparisons as well (Figure 4). To let both novice users (that need supervision) and advanced ones access advanced insights, they can be obtained in two ways (Figure 5):

[3] https://rasa.com/docs/rasa/
[4] https://jinja.palletsprojects.com/en/3.0.x/
[5] https://duckling.wit.ai/

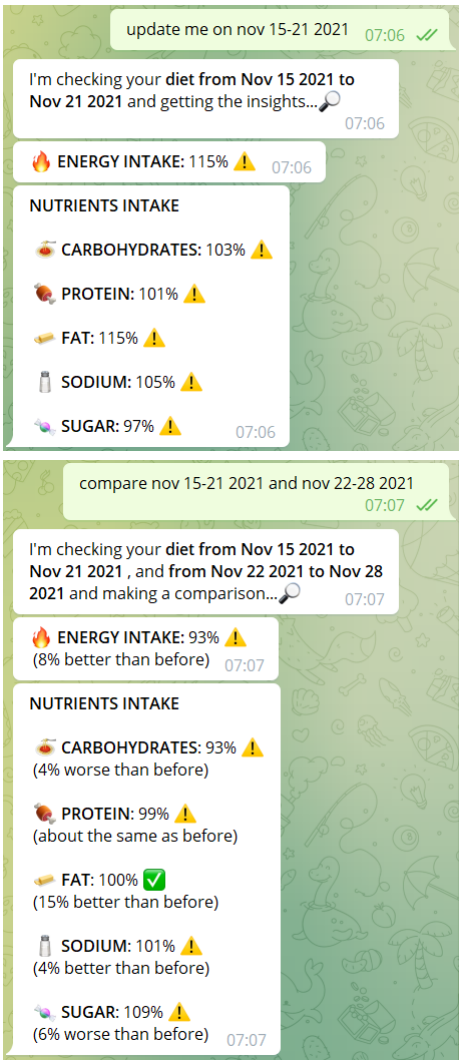


Figure 3: Basic report and comparison as presented by the chatbot.

1. **Guided navigation:** through generic queries (e.g. "tell me more about this" or "anything else?"). Following this trigger, the chatbot presents a button interface for each available advanced insight. Buttons can be checked and unchecked to obtain only those insights that are of interest.

2. **Natural language query:** by directly asking for specific insights and metrics. This can be done by specifying a particular insight (e.g. "food" or "intake") on a specific period.

For both interactions, users can specify one or more metrics.

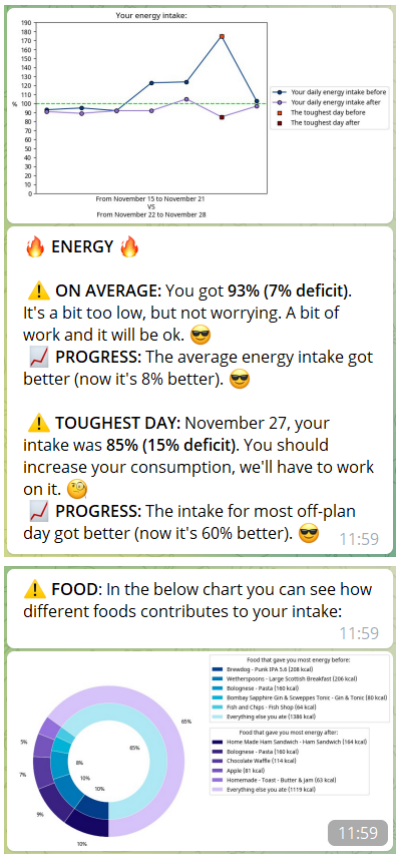


Figure 4: Example of advanced insights (intake and food analysis) for comparisons.

3.2 Other features

We implement a number of supplementary best practices (Ferman, 2018) to further improve usability and clarity. The chatbot actively provides feedback for each input (while informing users on the pending task); adopts emojis to make insights more understandable; splits the content in multiple messages and introduce a dynamic delay between them to avoid flooding.

4 Experiment setup

We deploy our chatbot on Telegram Bot API[6] and compare its informativeness with traditional diet apps. We gather our test population (**workers**) through crowd-sourcing on Amazon Mechanical Turk[7]. Details of recruitment, pay and sanity checks are available in the Appendix A. We choose to compare our

[6] https://core.telegram.org/bots/api
[7] https://www.mturk.com/worker/help

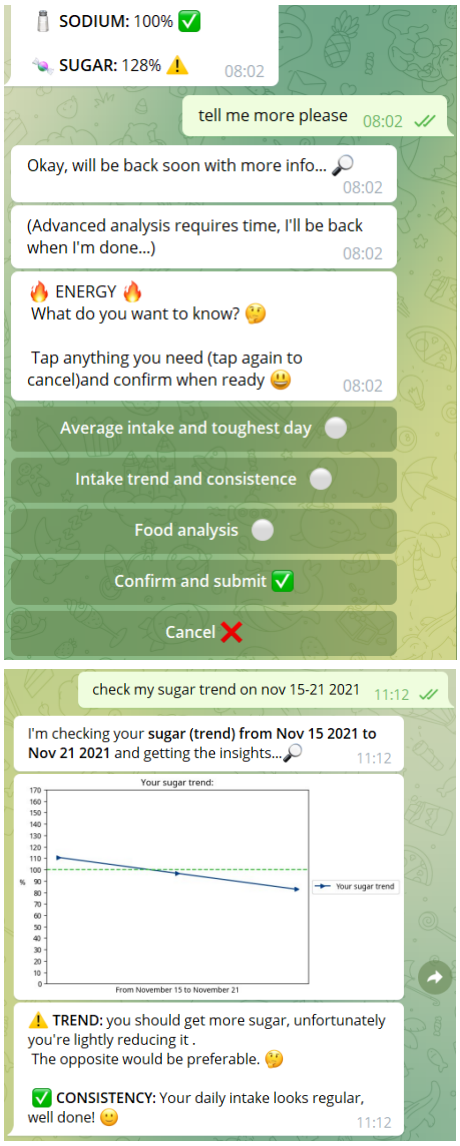


Figure 5: Obtaining advanced insights: Guided navigation with buttons (top) VS Natural language uery about trend and consistency (bottom).

chatbot with MyfitnessPal[8] (**MFP**) and FatSecret[9] (**FS**). An example of the two apps UI can be seen in Figure 6. We choose these two apps based on their popularity and downloads number on the Apple and Android app stores. We do not compare against any dieting chatbot as none of those present in literature is publicly available.

4.1 Measuring informativeness

Aiming at communication improvement, we need to find a measure to capture whether one specific tool performs better than others. From communication theory (Webster and Morris, 2019) we adopt the concept of "informativeness", defined as "how successfully a person is able to convey an intended message". We extend this definition to diet systems as "how successfully a tool is able to convey an intended message". To capture informativeness we create a ten questions quiz regarding diet analysis (a sample is provided in Appendix B). The quiz consists of 4 macro-tasks:

1. **Day analysis:** understanding if calories and carbohydrates are balanced on a single day (2pts).

2. **Food analysis:** understanding what food provided most calories and fat on a single day, along with quantities (4pts).

3. **Week analysis:** understanding if calories and carbohydrates are balanced across a week (2pts).

4. **Weeks comparison:** understanding if, by comparing two weeks, calories and carbohydrates improved or worsened (2pts).

Each question is worth 1 point, for a total of 10 points. We choose to develop a custom quiz because no available questionnaire can be used evaluate the informativeness of a diet-coaching tool. In creating it, we analyse existing apps and all the information that they deliver; we incorporate experts recommendations from previous surveys (Vasiloglou et al., 2020); we consider the theoretical constructs of self-regulation (Zahry et al., 2016), with a particular focus on the measure of informativeness. We avoid evaluating "trend and consistency" feature for fairness, as apps don't offer a way for the user to infer such information without long and tedious calculations.

Workers were randomly assigned to either our chatbot, MFP or FS, each of which was pre-filled with a simulated food diary (none of the data belonged to the users) consisting of 2 weeks of logged meals. We obtained

[8]https://www.myfitnesspal.com/
[9]https://www.fatsecret.com/

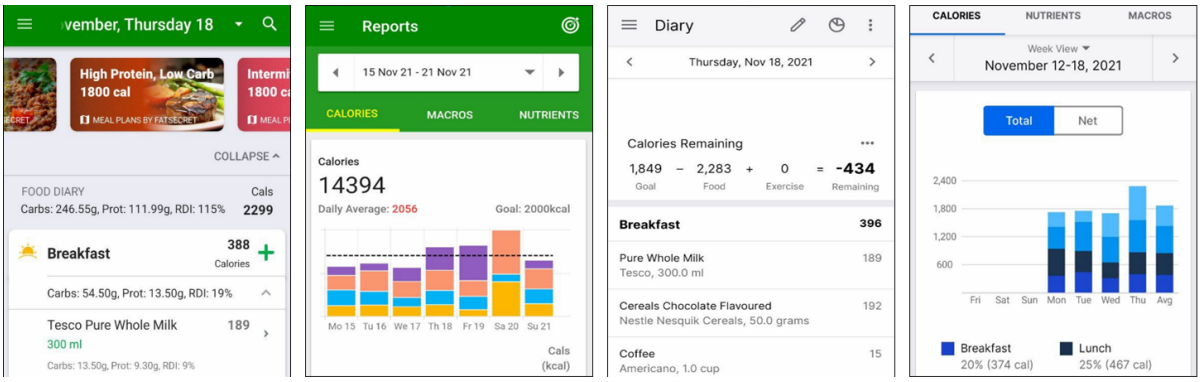


Figure 6: Food Diary and nutrition reports as showed to the user in FatSecret (left) and MyFitnessPal UI (right).

n=27 workers assigned to our chatbot; n=31 workers to MFP; n=29 workers to FS. Besides the tool itself, workers were provided with a PDF guide on how to use it and a glossary explaining the meaning of the terms used in the quiz. Each worker took the quiz and was asked to answer the questions to the best of their knowledge by using the tool. Through the quiz we test the following hypothesis:

Hypothesis 1 (H1): *Chatbot workers scored higher on informativeness quiz than MFP or FS workers.*

4.2 Measuring nutrition literacy

Previous research highlighted the importance of nutrition literacy in dieting (Michie et al., 2011), so we analyse its impact on our experiment. We also analyse if our chatbot communication can reduce the score gap between different literacy levels. Before taking the quiz, each worker completed Pfizer's Newest-Vital-Sign (**NVS**) (Weiss et al., 2005; Powers et al., 2010), consisting of 6 questions (each one worth 1 point) regarding an ice-cream label. NVS scores are grouped in ranges: 0-1 refers to "high likelihood of limited literacy", 2-3 refers to "possibility of limited literacy"; 4-6 refers to "adequate literacy". We compare NVS scores with quiz scores to test the following hypothesis:

Hypothesis 2 (H2): *There was a positive correlation between NVS score and quiz score in our experiment, but not for chatbot workers.*

4.3 Measuring perception of the tool and past experience

Finally, we inspect workers opinion on the tool they used. We ask each worker to rate the tool under different characteristics (see Figure 8) through Likert-5 scale. Through this approach we test the following hypothesis:

Hypothesis 3 (H3): *Our chatbot received higher ratings across the proposed questions.*

Finally, we ask workers to specify whether they had past experience with dieting tools (including the one they were assigned to) and to specify how often they used it (often; occasionally; rarely; never).

5 Results analysis

For variance analysis, we adopt One-Way ANOVA and Tukey's post-hoc test (replaced respectively by Kruskal-Wallis test and Dunn's post-hoc test if ANOVA's normality requirement is not met). To test variable dependence we adopt Chi-squared test and Bonferroni's post-hoc test. For correlation test we adopt Pearson correlation (substituted by Spearman correlation if Pearson's normality requirement is not met).

5.1 Preliminary checks

Before analysing results, we verify nutrition literacy uniformity across our population, to ensure that none of the groups contained mostly workers with high/low nutrition literacy. We discover that nutrition literacy

	Average score			Score differences		
Topic	CB	FS	MFP	CB-FS	CB-MFP	MFP-FS
Overall (10pt)	**6.65**	4.13	5.22	+2.52**	+1.43	+1.09
Day analysis (2pt)	1.15	0.76	**1.32**	+0.40	-0.16	+0.56
Food analysis (4pt)	**2.85**	2.14	0.91	+0.71	+1.94***	-1.23*
Week analysis (2pt)	**1.35**	0.66	1.05	+0.70**	+0.30	+0.39
Weeks comparison (2pt)	**1.31**	0.59	1.14	+0.72**	+0.17	+0.55**

Table 1: Results from informativeness quiz. On the left side: average scores, overall and for specific tasks. Highest score for each category are in bold. On the right side: score differences between tools. Green is for higher scores, red is for lower score. CB = Chatbot; MFP = MyFitnessPal; FS = FatSecret. Significance: * for p<0.05; ** for p<0.01; *** for p<0.001.

	Workers per class		
NVS class	CB	FS	MFP
LOW (0-1pt)	1	0	9
MID (2-3pt)	5	3	5
HIGH (4-6pt)	21	26	17

Table 2: Distribution of nutrition literacy for our population. CB = Chatbot; MFP = MyFitnessPal; FS = FatSecret.

distribution is unbalanced among apps, with the majority of workers with low nutrition literacy assigned to MFP sample, none to FS only one to our chatbot (see Table 2). We re-balance the samples by removing all the such workers workers. This limits our inspections on nutrition literacy but keeps the comparison fair. From now on, all results will refer to the re-balanced sample unless otherwise specified. We also check for meaningful difference in workers past experience with diet tools, but find none neither in general ($p = 0.47$) and by considering only those workers who had past experience and ($p = 0.27$).

5.2 Quiz scores

We first check total and per-task quiz scores (see Table 1). We find that, overall, the highest average score was reached by chatbot workers. The difference was statistically significant when compared to FS workers. Regardless of the group, average scores were low, not going much higher than 6/10. We consider this as a further confirmation of how hard understanding dietary insights is for the average user, especially in our context where data was simulated. By inspecting individual quiz tasks, we see that chatbot workers scored significantly higher in week analysis and comparison than FS workers, and in food analysis than MFP workers. We also find that MFP workers scored significantly higher than FS workers when comparing weeks, while the opposite happened for food analysis. Overall, chatbot workers always scored the highest score in every case, except for the day analysis, where MFP workers scores were slightly higher.

Next we look at the percentage of correct answers to check if any of the tools were associated with reaching specific scores (e.g. maximum points or 0 points). First, we find that our chatbot was positively associated ($p = 0.0001$) with an overall score of 9/10 points. This tells us that the chatbot made it easier to reach higher scores in general. We then proceed to analyse individual quiz tasks (Figure 7). Our chatbot was positively associated with maximum score in food analysis and week analysis. For chatbot workers it was easier understanding food details and insights based on aggregation in general. It was also negatively associated ($p = 0.001$) with 0 points in weeks comparison. In fact, every chatbot worker managed to answer at least one of the two questions about comparison right. Interestingly, we find the opposite for FS, that was positively associated with scoring 0 points in weeks comparison. This tells us that FS workers struggled considerably in this task. Lastly, using MFP was negatively associated with maximum score in food analysis: understanding food details was one of the hardest tasks with MFP.

5.3 Nutrition Literacy effect on scores

We check if nutrition literacy influenced quiz score. In here we discover a discrepancy between the balanced and unbalanced sample. MFP workers show a significant difference

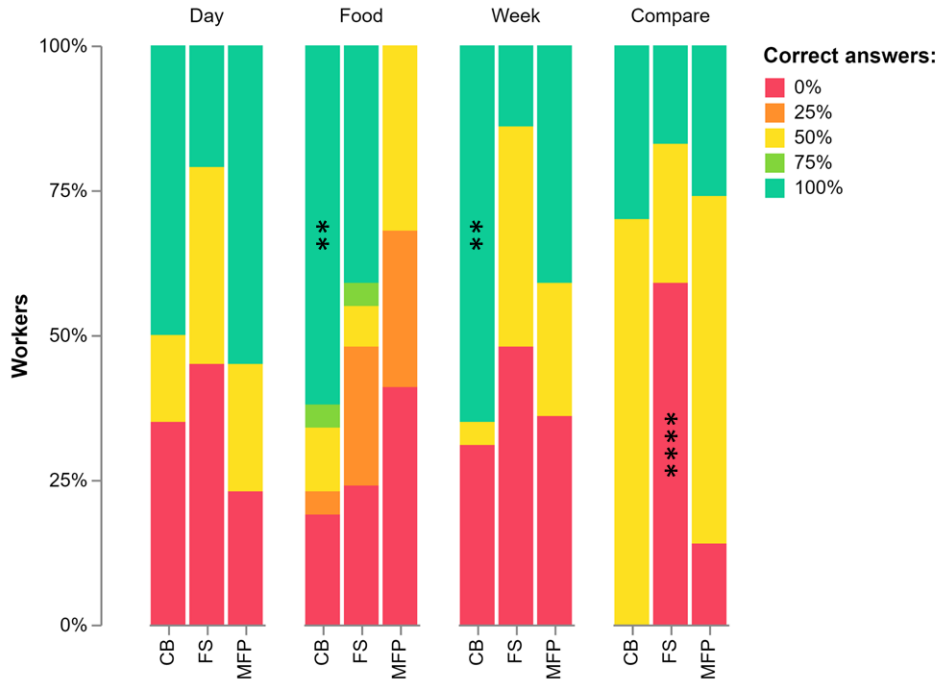


Figure 7: Percentage of correct answers by task, for each tool. CB = Chatbot; FS = FatSecret; MFP = MyFitnessPal. For day, week analysis and comparisons (2-points) we check no right answer (0%), 1 right answer out of 2 (50%) and all right answers (100%). For food analysis (4 points), we check quarters as well. Significance: * for p<0.05; ** for p<0.01; *** for p<0.001; **** for p < 0.0001.

($p = 0.03$) in scores between high and low nutrition literacy. By re-balancing the sample, we lose this significance. We also discover a moderate correlation ($\rho = 0.48, p = 0.02$) between nutrition literacy and quiz score for MFP workers, even after balancing the samples.

5.4 Users perception of the tool

Finally, we check workers feedback (see Figure 8). We notice a generally positive evaluation for every tool, with the chatbot getting an higher amount of "Agree" ratings across every question. By single-item analysis, our chatbot was positively associated with "Agree" in Q1 ($p = 0.01$), where it also shows a better mode value than the other tools. Chatbot workers felt it more useful for finding problems in the food diary. We also find a better mode than both apps in Q3, meaning that workers found it to be quicker to use. This result in particular is unexpected considering that there was no significant difference in the quiz execution time ($p = 0.22$). It could be that using natural language in our chatbot was felt as faster than navigating through different app sections. No app showed better mode than our chatbot in any question. Finally, it is interesting to notice that FS scored higher than MFP in Q5 despite being the tool with the lowest scores across every task except food analysis.

6 Discussion

From quiz results, chatbot workers scored the highest in informativeness across every scenario except for a slight advantage of MFP in day analysis. In multiple contexts, the difference with MFP and FS was statistically significant. We also found that using the chatbot was associated with higher completion rate in different tasks, and very high overall scores like 9/10. With these results we confirm H1. We could not inspect nutrition literacy properly, as the different samples were too unbalanced and introducing low-literate workers would have made the comparison between MFP and our chatbot unfair. We saw a relationship between lower nutrition literacy and quiz scores, but isolated to MFP workers, and could not verify

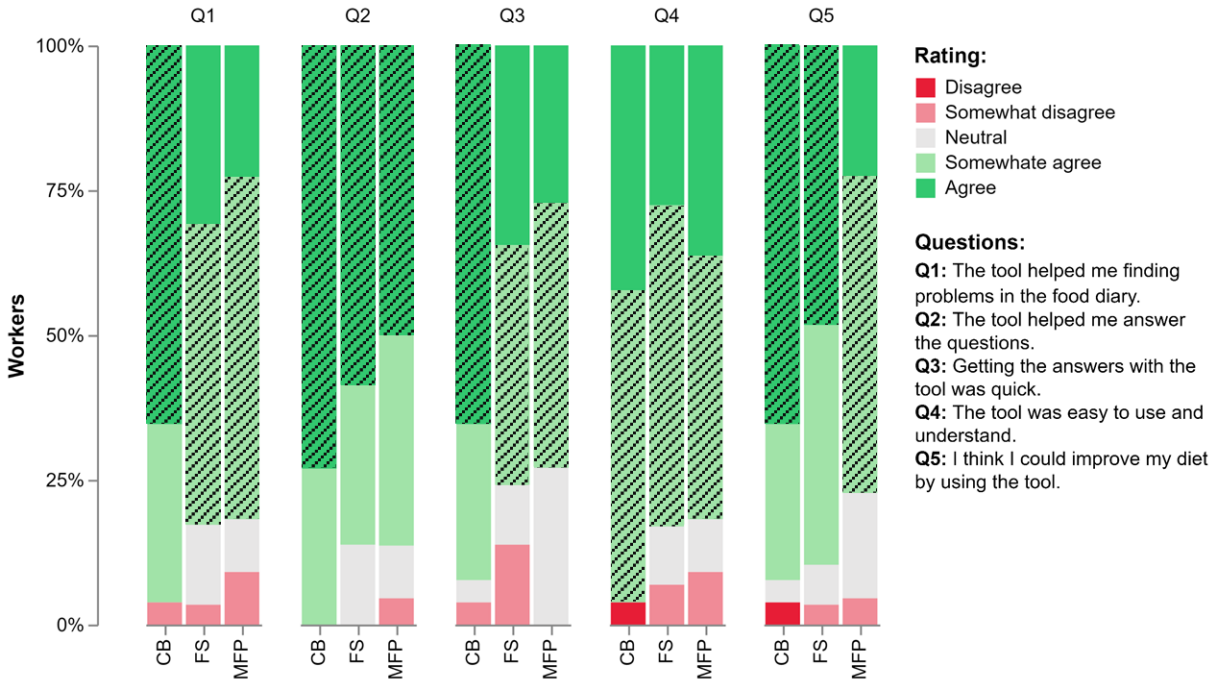


Figure 8: Feedback from users, based on used tool. CB = Chatbot; FS = FatSecret; MFP = MyFitnessPal. Lined bars indicate the mode for each question.

it across the whole population. With these results we neither confirm or reject H2 because of the lack of data. Looking at feedback, we found out that our chatbot received a higher amount of "Agree" ratings across every question. It was also the only tool that showed association with maximum usefulness in finding diet problems. By analysing the mode of each question, we discovered that our chatbot was evaluated as quicker to use than the other apps. We also see that, unlike MFP and FS, it never showed a lower mode than any other tool. With these results we confirm H3.

7 Conclusion and future developments

In this work we evaluated the combination of charts and textual explanation for diet coaching, in the conversational scenario. We implemented an NLG-chatbot that understands natural language input and returns dietary insights as a combination of textual explanations and visualisations. We compared the chatbot with traditional static diet apps by inspecting informativeness and user feedback. Results shows that the combination of visuals and text efficiently delivers information in diet-coaching, and makes it more understandable. Improved informativeness could play a critical role in diet outcome. Feedback was generally more positive for the chatbot, meaning that it can be a valid tool for diet-coaching, potentially substituting static apps.

For future work we plan to investigate if our approach can lead to actual learning from the user, for example through spaced repetition (Ausubel and Youssef, 1965; Tabibian et al., 2019) that can positively affect users' forgetting curve (Ebbinghaus, 2013). We also commit on addressing the limits of our setup, to properly inspect the relationship between nutrition literacy and informativeness. We also plan to inspect more personalised approaches to information tailoring, namely by considering users' stress and emotional state that showed to be promising research directions (Balloccu et al., 2020; Balloccu and Reiter, 2022). Lastly, we consider this result as a sign of the maturity of our approach and we plan to run a trial to measure its effect on diet-coaching (e.g. weight control).

References

David P Ausubel and Mohamed Youssef. 1965. The effect of spaced repetition on meaningful retention. *The Journal of General Psychology*, 73(1):147–150.

Simone Balloccu and Ehud Reiter. 2022. Beyond calories: evaluating how tailored communication reduces emotional load in diet-coaching. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 42–53, Dublin, Ireland. Association for Computational Linguistics.

Simone Balloccu, Ehud Reiter, Matteo G. Collu, Federico Sanna, Manuela Sanguinetti, and Maurizio Atzori. 2021. *Unaddressed Challenges in Persuasive Dieting Chatbots*, page 392–395. Association for Computing Machinery, New York, NY, USA.

Simone Balloccu, Ehud Reiter, Alexandra Johnstone, and Claire Fyfe. 2020. How are you? introducing stress-based text tailoring. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 62–70, Santiago de Compostela, Spain. Association for Computational Lingustics.

Daniel Braun, Ehud Reiter, and Advaith Siddharthan. 2015. Creating textual driver feedback from telemetric data. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 156–165, Brighton, UK. Association for Computational Linguistics.

Douglas Brock, Erin Abu-Rish, Chia-Ru Chiu, Dana Hammer, Sharon Wilson, Linda Vorvick, Katherine Blondon, Douglas Schaad, Debra Liner, and Brenda Zierler. 2013. Republished: Interprofessional education in team communication: working together to improve patient safety. *Postgraduate medical journal*, 89(1057):642–651.

Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2018. Food diary coaching chatbot. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, page 1676–1680, New York, NY, USA. Association for Computing Machinery.

Courtney R Davis, Karen J Murphy, Rachel G Curtis, and Carol A Maher. 2020. A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant. *International journal of environmental research and public health*, 17(23):9137.

Hermann Ebbinghaus. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4):155.

Maria Ferman. 2018. Towards best practices for chatbots.

Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using nlg technology for decision support and information management. *AI Commun.*, 22(3):153–186.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2017. Data-to-text generation improves decision-making under uncertainty. *IEEE Computational Intelligence Magazine*, 12(3):10–17.

Anna S Law, Yvonne Freer, Jim Hunter, Robert H Logie, Neil McIntosh, and John Quinn. 2005. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of clinical monitoring and computing*, 19(3):183–194.

H Erin Lee and Jaehee Cho. 2017. What motivates users to continue using diet and fitness apps? application of the uses and gratifications approach. *Health communication*, 32(12):1445–1453.

Carol Ann Maher, Courtney Rose Davis, Rachel Grace Curtis, Camille Elizabeth Short, and Karen Joy Murphy. 2020. A physical activity and diet program delivered by artificially intelligent virtual health coach: Proof-of-concept study. *JMIR mHealth and uHealth*, 8(7):e17558.

Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science*, 6(1):1–12.

Martin Molina, Amanda Stent, and Enrique Parodi. 2011. Generating automated news to explain the meaning of sensor data. In *Advances in Intelligent Data Analysis X*, pages 282–293, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hamid Mukhtar. 2016. Using persuasive recommendations in wellness applications based upon user activities. *International Journal of Advanced Computer Science and Applications*, 7(8).

Maria D.G.H. Mulders, O. Corneille, and O. Klein. 2018. Label reading, numeracy and food nutrition involvement. *Appetite*, 128:214–222.

Elizabeth L. Murnane, David Huffaker, and Gueorgi Kossinets. 2015. Mobile health apps: Adoption, adherence, and abandonment. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp/ISWC'15 Adjunct, page 261–264, New York, NY, USA. Association for Computing Machinery.

Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. 2019. *Making Effective Use of Healthcare Data Using Data-to-Text Technology*, pages 119–145. Springer International Publishing, Cham.

Benjamin J Powers, Jane V Trinh, and Hayden B Bosworth. 2010. Can this patient read and understand written health information? *Jama*, 304(1):76–84.

Philips Kokoh Prasetyo, Palakorn Achananuparp, and Ee-Peng Lim. 2020. *Foodbot: A Goal-Oriented Just-in-Time Healthy Eating Interventions Chatbot*, page 436–439. Association for Computing Machinery, New York, NY, USA.

Alejandro Ramos-Soto, Borja Vazquez-Barreiros, Alberto Bugarín, Adriana Gewerc, and Senen Barro. 2017. Evaluation of a data-to-text system for verbalizing a learning analytics dashboard. *International Journal of Intelligent Systems*, 32(2):177–193.

Reijo Savolainen. 2010. Dietary blogs as sites of informational and emotional support.

Taylor N Stephens, Angela Joerin, Michiel Rauws, and Lloyd N Werk. 2019. Feasibility of pediatric obesity and prediabetes treatment support through tess, the ai behavioral coaching chatbot. *Translational behavioral medicine*, 9(3):440–447.

Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2019. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10):3988–3993.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Marian van der Meulen, Robert H. Logie, Yvonne Freer, Cindy Sykes, Neil McIntosh, and Jim Hunter. 2010. When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, 24(1):77–89.

Brent Van Dorsten and Emily M Lindley. 2008. Cognitive and behavioral approaches in the treatment of obesity. *Endocrinology and metabolism clinics of North America*, 37(4):905–922.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Maria F. Vasiloglou, Stergios Christodoulidis, Emilie Reber, Thomai Stathopoulou, Ya Lu, Zeno Stanga, and Stavroula Mougiakakou. 2020. What healthcare professionals think of "nutrition amp; diet" apps: An international survey. *Nutrients*, 12(8).

Janet Webster and Julie Morris. 2019. Communicative informativeness in aphasia: Investigating the relationship between linguistic and perceptual measures. *American Journal of Speech-Language Pathology*, 28(3):1115–1126.

Barry D Weiss, Mary Z Mays, William Martz, Kelley Merriam Castro, Darren A DeWalt, Michael P Pignone, Joy Mockbee, and Frank A Hale. 2005. Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine*, 3(6):514–522.

Nagwan R Zahry, Ying Cheng, and Wei Peng. 2016. Content analysis of diet-related mobile apps: A self-regulation perspective. *Health Communication*, 31(10):1301–1310.

Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. 2020. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: Viewpoint. *J Med Internet Res*, 22(9):e22845.

Kelly B Haskard Zolnierek and M Robin DiMatteo. 2009. Physician communication and patient adherence to treatment: a meta-analysis. *Medical care*, 47(8):826.

A Ethics

This section sums up the procedure we adopted to ensure the ethical compliance of our experiment.

A.1 Preliminary review

Before starting the experiment, procedure and materials were carefully reviewed by our institution Ethics Board (omitted for the sake of double-blinded review). Our experiment proposal was accepted without major revisions.

A.2 Platforms

For the quiz, we adopted Microsoft Forms[10] because of its compliance with GDPR policy. For hiring, we used Amazon Mechanical Turk. No recruitment qualification was specified, beside custom ones to prevent the same worker from submitting multiple HITs. Participants were showed a consent form containing all the information regarding the experiment procedure. They were also informed about the requirements that had to be satisfied to obtain the remuneration. All worker had to confirm their acceptance of these conditions (through checkboxes) in order to proceed with the experiment. Workers were given an email contact in case of problems during the experiment.

A.3 Pay and workload

Before launching the experiment, we verified the average completion time with 10 test users. The average result for completing the whole experiment (reading information; downloading and setting up material; taking NVS; taking the quiz; expressing the feedback) was 20 minutes. We gave each worker 45 minutes, and paid 15USD for the HIT. Workers were informed that if they ran out of time on Mturk they could just finish the quiz (on Microsoft Forms web platform) and contact us through the provided email address to still get paid.

[10] https://forms.office.com/

A.4 HITs sanity checks

We received a total amount of 250 applications for our task. Most of these application were fraudulent, with random answers or unrealistic completion times. In order to recognise legit HITs we set up multiple sanity-checks, both in general and depending on the tool each worker was assigned.

A.4.1 Global sanity checks

To check on the attention of workers during Pfizer's NVS, a fake price was added to the ice-cream label. Consequently, we added a (non scored) question to the form, asking "what's the price of the ice-cream?". Moreover, each worker received a completion code that they had to submit on Mechanical Turk platform after completing all the tasks.

A.4.2 Sanity check for chatbot worker

The chatbot was programmed to accept some custom queries that led to specific answers. The workers were asked, at multiple times, to trigger one of these query. We manually checked the answers for HITs, in order to verify whether workers actually used the chatbot. In addition, conversations were logged and anonymised, and the provided WorkerID was used to track down specific workers and verify the sanity of interaction.

A.4.3 Sanity check for FS and MFP worker

To verify that workers actually used the diet apps they were asked to provide a description of the app logo, and to check which particular food (among three alternatives) could be seen in a specified day. As this tasks are subjective and could be failed by legit workers who struggled to use the app, each HIT was manually evaluated to avoid unfair treatment.

B Appendix A: Quiz sample

Evaluating the informativeness of various diet-coaching tools

* Required

Introduction

Please read the following instructions carefully before proceeding.

What is this experiment about?
This research aims at evaluating whether common diet-coaching apps are **informative** for users.
In other words, how easy it is for users to find the information they need and, most importantly, how comprehensible they are.

What will I have to do?
For this experiment, we ask you to do 3 main tasks.

1. **Preliminary form:** during this step, you'll be asked to answer a short form **(5-6 questions)** regarding nutrition. This will involve extracting information from a sample nutritional label and reasoning about them.

2. **Main form:** following the completion of the previous point, you'll be assigned to a tool (a diet-coaching app). You'll receive instructions on **how to download (through Play/Apple Store), install and use the app** on your phone. Each app has been pre-compiled with food diaries (imagine this as someone else record of what they ate).
   You will be asked to explore this data to answer **10 questions**.

3. **Final feedback:** finally, we will ask you to give us your opinion on the overall experiment **(7 questions)**, with a particular focus on the tool you used in step 2. You'll be asked for your **worker ID** and be given a **completion code**. Return it to us to process your HIT.

Total time for doing this experiment should be between **30-45 minutes.**

Additional details (1/2)

Some important things to keep in mind:

1. You'll need to **install and use** the assigned app on your phone (Android/IOS) to complete the experiment.
   Failure in complying with this requirement will cause **HIT invalidation**.

2. The experiment is monitored. Fraudulent behaviour such as **completing the form without reading the questions** or **giving random answers** will be detected and will result in the invalidation of your HIT.

3. Note that the previous points does not apply to the cases in which, **despite using the app, you're still not able to give an answer**. Regardless of the amount of correct answer you give, **you will still receive your remuneration.**

4. You will be assigned to only **one app** for this experiment. You won't have to repeat it multiple times.

5. Most of the apps don't require any registration: we'll give you login credentials (username and password).
   In only one case, the app will require your phone number for access. We won't be able to see or access this as it is a chat app **(Telegram)**.

6. You don't have to keep the app installed after the experiment. You can uninstall it immediately when done.

7. None of the apps have been developed by us and therefore **we won't receive any data except the form answers.**

8. As said before, your assigned app will show you some data regarding food and meals across different days.
   Please only read the data, **avoid changing, deleting or altering that data in any way**.
   Data alteration/augmentation will result in experiment invalidation **(and HIT invalidation for MTurkers)**

9. In any case, **no data (outside of form answers) will be gathered.**

10. Should you change your mind, you can withdraw from the experiment at any given stage and without giving a reason, until the point in which data analysis shall be done with your (completed) results

Additional details (2/2)

Data management and storage
No personal data about you shall be collected or stored beside the data which will be put in the forms. All your answers will be anonymously and safely stored in devices belonging to University of Aberdeen. None of these data shall be released to the public.

Confidentiality and anonymity
Raw data and the identity of participants will not be released to anyone outside the research team. The data you provide will be analysed and may be used in publications, dissertations, reports or presentations derived from the research project, but this will be done in such a way that your identity is not disclosed.

Consent
If you agree to take part in the research, you will be asked to indicate your consent by ticking the following checkboxes.

Risk
We foresee no risk for any participant involved.

Sponsor
This research is being funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812882.

1

Question *

☐ I confirm that the research project **"Evaluating the informativeness of various diet-coaching tools"** has been explained to me. I have had the opportunity to ask questions about the project and have had these answered satisfactorily.

2

Question *

☐ I consent to the material I contribute being used to generate insights for the research project **"Evaluating the informativeness of various diet-coaching tools"**.

3

Question *

- [ ] I understand that my participation in this research is voluntary and that I may withdraw from the project at any time (until the point of data analysis) without providing a reason. **I understand that (for MTurkers) withdrawal will invalidate my HIT.**

4

Question *

- [ ] I consent to allow the **fully anonymised** data to be used for future publications and other scholarly means of disseminating the findings from the research project.

5

Question *

- [ ] I understand that the information/data acquired will be securely stored by researchers, but that appropriately anonymised data may in future be made available to others for research purposes. I understand that the University may publish appropriately anonymised data in its research repository for verification purposes and to make it accessible to researchers and other research users.

Preliminary form (1/3)

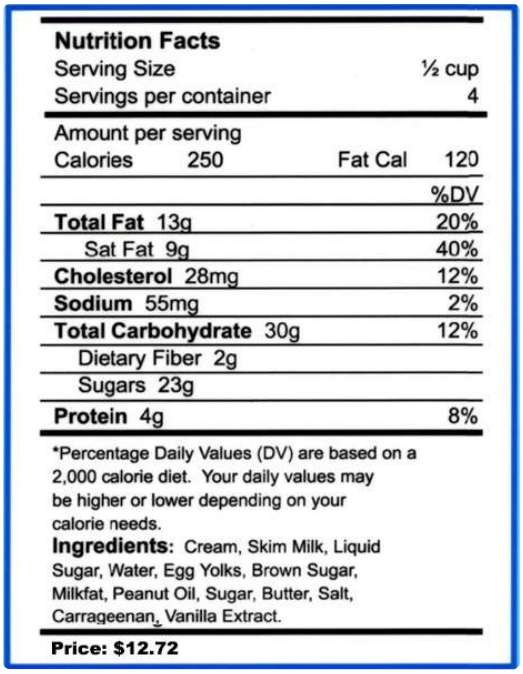
In this first form, we ask you to answer some questions related to **the nutritional label displayed below**. Answer to the best of your knowledge.

Please do not seek help from anyone else to complete this form. The aim is not to score maximum points at any cost. None of your answers will be shared with anyone and your identity will be kept anonymous.

Additional help:
- You are allowed to use a calculator if you would like to.
- You do not need any app for this part of the experiment.

6

*



Nutrition Facts

Serving Size	½ cup
Servings per container	4
Amount per serving	
Calories 250	Fat Cal 120
	%DV
Total Fat 13g	20%
Sat Fat 9g	40%
Cholesterol 28mg	12%
Sodium 55mg	2%
Total Carbohydrate 30g	12%
Dietary Fiber 2g	
Sugars 23g	
Protein 4g	8%

*Percentage Daily Values (DV) are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

Ingredients: Cream, Skim Milk, Liquid Sugar, Water, Egg Yolks, Brown Sugar, Milkfat, Peanut Oil, Sugar, Butter, Salt, Carrageenan, Vanilla Extract.

Price: $12.72

☐ **Check this box to proceed.**

7

If you eat the entire container, how many calories will you eat? *

(1 Point)

8

If you are allowed to eat 60 grams of carbohydrates as a snack, how much ice cream could you have? *

(1 Point)

9

What's the price of the ice-cream? *

10

Your doctor advises you to reduce the amount of saturated fat in your diet. You usually have 42 g of saturated fat each day, which includes one serving of ice cream. If you stop eating ice cream, how many grams of saturated fat would you be consuming each day? *

(1 Point)

11

**If you usually eat 2,500 calories in a day, what percentage (%) of your daily value
of calories will you be eating if you eat one serving?** *

(1 Point)

12

Pretend that you are allergic to the following substances: penicillin, peanuts, latex gloves, and bee stings. Is it safe for you to eat this ice cream? *

(1 Point)

- ◯ Yes
- ◯ No

13

If you replied "No" to the previous question, motivate your choice: *

(1 Point)

Main form (2/3)

To complete this form **you will need your assigned tool**.

Your assigned tool is: **$tool_name**

Please do not seek help from anyone else to complete this form.
The goal of this experiment is to assess your ability to use the tool, not to score maximum points at any cost. Your identity will be kept anonymous.

Additional help:

- You are allowed to use a calculator if you would like to.
- We suggest you to use the glossary to better understand the questions.

How to download, setup and use your tool:

Below you can find two download links:

1. **Glossary:** we made this file to make it clearer what certain terms means. You can use it to better understand what we're asking you.
2. **User guide:** this file shows you how to **download, install and setup** the app. It also guides you through all the features that you can use to answer the following questions.

Download links:

- Glossary: $glossary_link
- User guide: $guide_link

Credentials:

- Username: $user
- Password: $password
Please open the app and login now before proceeding.

Additional support:

If you have questions or something doesn't work, feel free to contact us at the following email:

14

Please read everything before proceeding, otherwise you could struggle while doing the experiment. *

☐ I read everything!

Food diary on November 28 2021

Following the **user guide**, you can access **a food diary**. That is, for two consecutive weeks **you can see every meal and some related information** (e.g.: nutrients and calories).

Through the app, check **November 28 2021** only and answer the questions to the

15

Which one of the following is true for November 28 2021? *

(1 Point)

- ◯ The calorie intake is **too high.**
- ◯ The calorie intake **is balanced.**
- ◯ The calorie intake **is too low.**
- ◯ I don't know.

16

Which one of the following is true for November 28 2021? *

(1 Point)

- ◯ The carbohydrates intake is **too high.**
- ◯ The carbohydrates intake **is balanced.**
- ◯ The carbohydrates intake **is too low.**
- ◯ I don't know.

17

Write the single food with most calories on November 28 2021:
(If you're not able to answer just type "unknown" and proceed) *

(1 Point)

18

How many calories does that food contain?
(If you're not able to answer just type 0 and proceed) *

(1 Point)

19

Write the single food with most fat on November 28 2021:
(If you're not able to answer just type "unknown" and proceed)
*

(1 Point)

20

How many grams of fat does that food contain?
(If you're not able to answer type 0 and proceed) *

(1 Point)

21

Describe $tool_name app logo in your own words: *

Food diary on November 22-28 2021

Following the **user guide**, you can access **a simulated food diary**. That is, for two consecutive weeks **you can see every meal and some related information** (e.g.: nutrients and calories).

Through the app, check the week **November 22-28 2021** only and answer the questions to the best of your knowledge

22

Which one of the following is true for November 22-28 2021? *

(1 Point)

- ◯ The calories intake is **too high.**
- ◯ The calories intake is **balanced.**
- ◯ The calories intake is **too low.**
- ◯ I don't know.

23

Which one of the following is true for November 22-28 2021? *

(1 Point)

- ◯ The carbohydrates intake is **too high.**
- ◯ The carbohydrates intake is **balanced.**
- ◯ The carbohydrates intake is **too low.**
- ◯ I don't know.

24

Go to the home section of $tool_name. At the top, you will see a recap of your profile, with a picture. What do you see as the profile picture? *

Food diary on **November 15-21 2021** and on **November 22-28 2021**

Following the **user guide**, you can access **a simulated food diary**. That is, for two consecutive weeks **you can see every meal and some related information** (e.g.: nutrients and calories).

Through the app, check both:
- the week **November 15-21 2021**

25

Which one of the following is true? *

(1 Point)

- ◯ The calorie intake is **better on November 22-28 2021**
- ◯ The calorie intake was **better on November 15-21 2021**
- ◯ The calories intake **is the same** for both weeks
- ◯ I don't know.

26

Which one of the following is true? *

(1 Point)

- ◯ The carbohydrates intake is **better on November 22-28 2021**
- ◯ The carbohydrates intake was **better on November 15-21 2021**
- ◯ The carbohydrates intake **is the same** for both weeks
- ◯ I don't know.

27

On November 19 2021, which one of these can you see in "Snacks/Other"? *

- ◯ Spaghetti bolognese
- ◯ Espresso
- ◯ Gin and Tonic

Final feedback (3/3)

Thank you again for your help. In this final form, we ask you to evaluate your overall experience by using your assigned tool.

Please do not give the most positive answer if you don't fully agree with the statement. The goal of this form is to see how good the tool was for you.

28

Please give a score to each statement, based on how much you agree with each one: *

	Disagree	Somewhat disagree	Neither agree or disagree	Somewhat agree	Agree
$tool_name helped me finding problems in the food diary.	◯	◯	◯	◯	◯
$tool_name helped me answer the questions.	◯	◯	◯	◯	◯
Getting the answers with **$tool_name** was quick.	◯	◯	◯	◯	◯
$tool_name was easy to use and understand.	◯	◯	◯	◯	◯
I think I could improve my diet using **$tool_name**.	◯	◯	◯	◯	◯

29

Did you use any diet-coaching tool (even $tool_name itself) before this experiment? *

- ◯ Yes
- ◯ No

30

If you chose yes, how often do you use the assigned or similar tool? *

- ◯ Often
- ◯ Occasionally
- ◯ Rarely
- ◯ Never