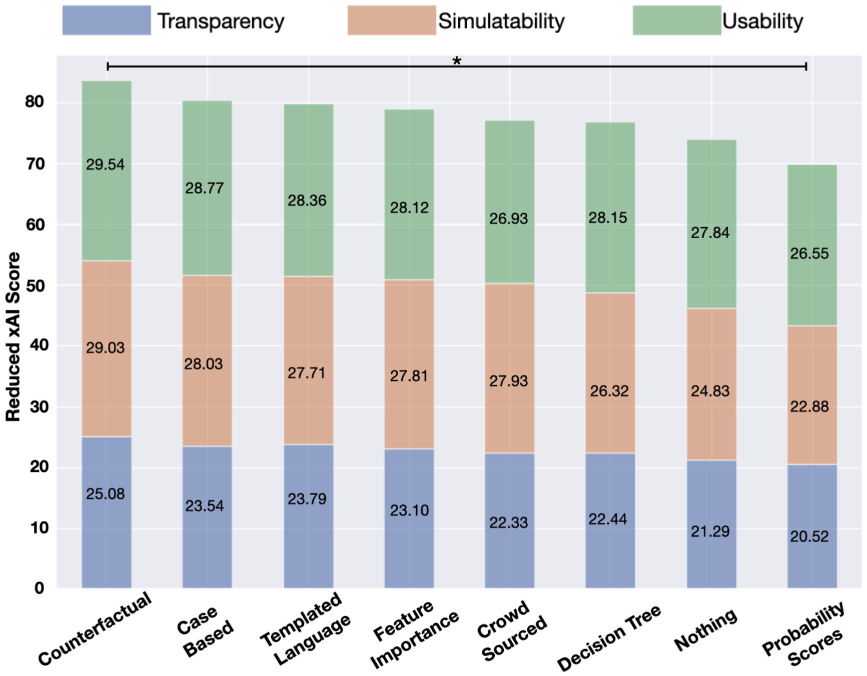


Figure A1.: Reduced xAI mean scores and sub-scale mean scores for all methods. The trends for the reduced scale match the full xAI scores, with the same ordering of conditions. While sub-scale scores for usability do not present much variance across conditions, the sub-scales of transparency and simulatability offer more variation across conditions. Statistical analysis of the aggregated reduced xAI scores reveal that counterfactual explanations score higher than probability scores ($p < 0.05$).

Appendix A. Reduced 14-Question xAI Survey

Echoing the results of our primary investigation with the full survey, here we present results according to our reduced xAI survey. An ANCOVA showed that certain conditions in our experiment were rated as significantly more explainable than others ($F(7, 277) = 3.14$, $p = 0.003$). Our independent variable is the explainability method and our dependent variable is the explainability score. We include as a covariate the participant's baseline explainability score. A Shapiro-Wilk test revealed that our data were not normally distributed, but we proceed with an ANCOVA due to a lack of non-parametric alternative and the robustness of the F-test (Cochran, 1947; Glass, Peckham, & Sanders, 1972; Hack, 1958; Pearson, 1931). A Tukey's HSD post-hoc analysis reveals that **Counterfactual** was rated as more explainable than **Probability Scores** ($p = 0.002$), as shown in Figure A1

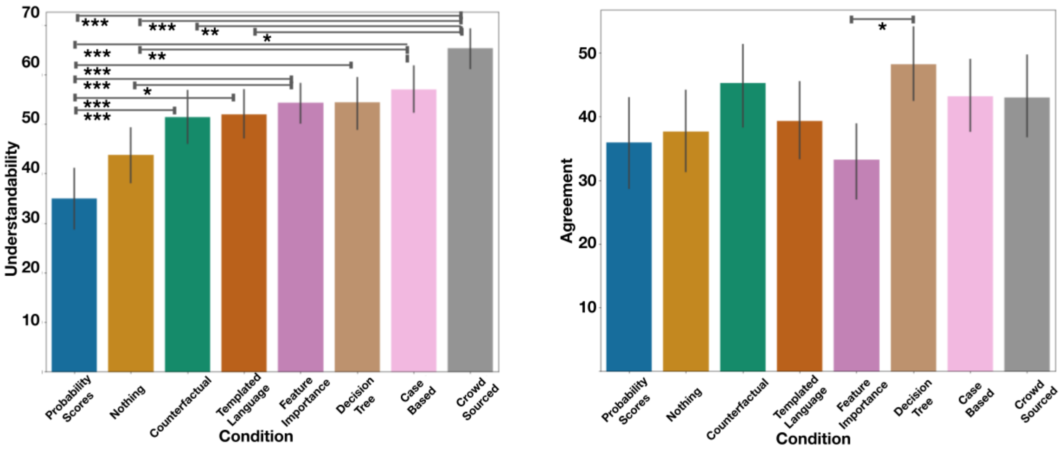
The reduced questionnaire, after a factor analysis and verification is given in Table A1.

Appendix B. Understanding and Agreement

Our scenarios included two Likert items for every question in the study: "I understand the reasoning behind the agent's suggestion" and "I agree with the agent's suggestion." Taking results from all participants on all questions, we have 3,672 responses for each

Table A1.: The reduced xAI Survey

Factor	**Question**
1	I had trouble using the explanations to answer the question.
1	I believe that the explanations would not help most people to answer the question.
1	Most people would not be able to apply the agent's explanations to the questions.
1	I would not understand how to apply the explanations to new questions.
1	The explanations were not relevant for the questions I was given.
2	The explanations were detailed enough for me to understand.
2	I understood the explanations within the context of the question.
2	The explanations provided enough information for me to understand.
2	The explanations were useful.
3	I am able to follow the agent's decision-making process step-by-step.
3	I would be able to repeat the steps that the agent took to reach its conclusion.
3	I understand why the agent used specific information in its explanation.
3	I could have applied the agent's reasoning to new problems, even if the agent didn't give me suggestions.
3	I believe that I could provide an explanation similar to the agent's explanation.



(a)

(b)

Figure B1.: (a) Condition has a significant effect on participant understanding of agent suggestions, revealing that all xAI techniques are superior to softmax confidence scores, and three techniques (feature importance, case-based reasoning, and crowd-sourced scores) are superior to the "no explanation" condition. (b) Condition has a significant effect on participant agreement with an agent, with decision tree explanations prompting significantly more agreement.

item. As each Likert item is not part of a full scale with additional context, prior work (Schrum, Johnson, Ghuy, & Gombolay, 2020) suggests that analysis on such data may lead to premature conclusions. However, owing to the added workload of a full Likert scale after each of the 20 scenarios in our study, we decided to reduce our data collection to only two items. Reducing the scale to two items drastically reduces the time and workload of our study, yet still presents interesting data for analysis. We acknowledge the limitations of statistical testing with single Likert items. As such, we present the following analyses as interesting case studies of single-item responses and as possible avenues for future work to explore further.

B.1. Understandability

An ANOVA showed that certain conditions in our experiment were rated as significantly more understandable for every question than others ($F(7, 3664) = 10.29$, $p < 0.001$). A Tukey's HSD post-hoc analysis revealed that the **Case Based** ($M = 56.98$, $SD = 50.85$), **Counterfactual** ($M = 51.41$, $SD = 59.79$), **Crowd Sourced** ($M = 65.37$, $SD = 38.43$), **Decision Tree** ($M = 54.45$, $SD = 58.39$), **Feature Importance** ($M = 54.35$, $SD = 48.24$), and **Templated Language** ($M = 52.02$, $SD = 58.12$) conditions were all rated as more understandable than the **Probability Scores** ($M = 34.98$, $SD = 60.82$) condition ($p < 0.001$). Similarly, the **Nothing** ($M = 43.84$, $SD = 61.54$) condition was rated as less understandable than the **Case Based** ($p = 0.006$), **Crowd Sourced** ($p < 0.001$), and **Feature Importance** ($p = 0.0497$) conditions. Finally, **Crowd Sourced** was rated as significantly more understandable than both the **Counterfactual** ($p = 0.007$) and **Templated Language** ($p = 0.011$) conditions. A comparison of all understandability ratings is shown in Figure B1a.

These results are very surprising. The **Crowd Sourced** condition presents the same information as the **Probability Scores** condition, the only difference is that the

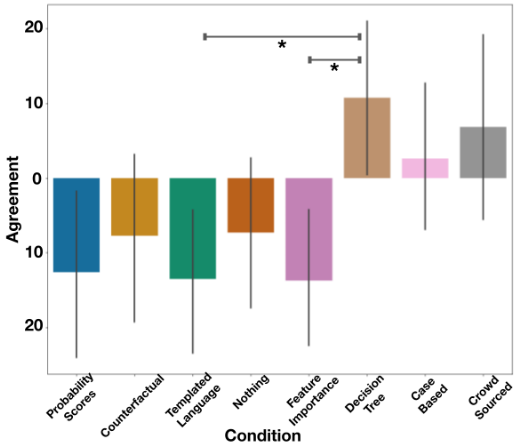


Figure B2.: Condition has a significant effect on participant agreement with an agent when the agent is offering incorrect suggestions, with decision tree explanations prompting significantly more agreement than feature importance scores or templated-language explanations.

top confidence score is placed into a natural-language sequence and the three unused confidence scores are removed. For example, instead of showing a table with 85%, 10%, 5%, and 0%, as in the **Probability Scores** condition, the **Crowd Sourced** condition presents the sentence "85% of experts agreed on this answer." Despite presenting the same probability for the suggested answer in slightly different ways, we observe a *significantly* higher tendency for users to rate the **Crowd Sourced** condition as more understandable.

One possible reason for this disparity is in the wording of the prompt: "I understand the reasoning behind the agent's suggestion." While a set of confidence scores do not offer insight into *why* the agent arrived at an answer, saying that "85% of experts agreed on this answer" provides participants with enough information to infer the agent's reasoning. It is reasonable to assume that the agent chose the answer because the largest portion of experts agreed upon the answer. Despite the quantitative information being identical, users have more to infer with the **Crowd Sourced** condition. This line of reasoning may also explain the relative superiority of the **Case Based** condition, as users may infer that the agent's decisions arise from past experience. It is possible that our participants interpreted the prompt to be "I understand how this agent was trained," and imagined possible training data involving expert opinions or prior cases.

Finally, and amusingly, we note that even the **Nothing** condition achieves a higher mean-understandability score than the **Probability Scores** condition. Our hypothesis is that confidence scores are not useful signals to untrained human users, offering little insight into the decision-making process or the imagined training process of an agent assistant. Even having no information at all may be less confusing to human users.

B.2. Agreement

An ANOVA showed that participant-rated agreement was also significantly affected by condition, albeit to a lesser degree ($F(7, 3664) = 2.63$, $p = 0.011$). A Tukey's HSD post-hoc analysis revealed that participants were more likely to agree with an agent

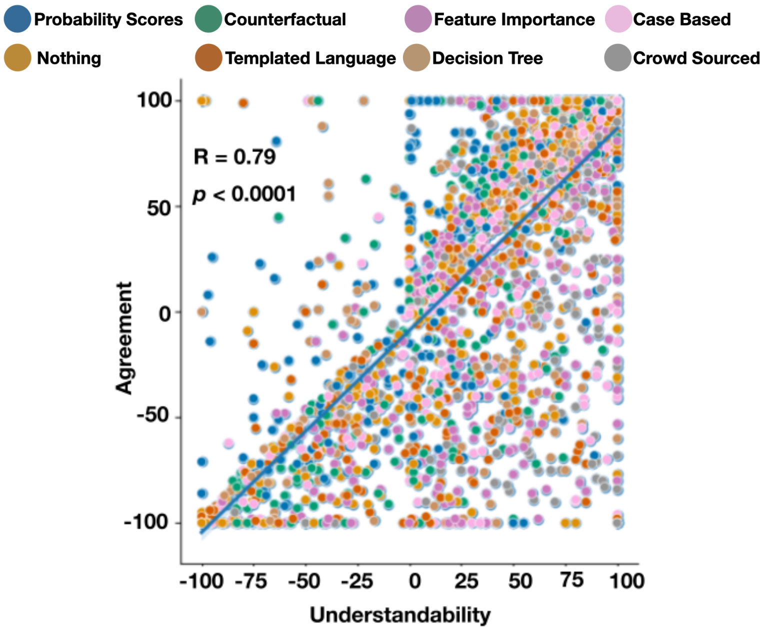


Figure B3.: Participant subjective agreement with agent suggestions is strongly correlated with participant understanding of agent suggestions ($R = 0.79$, $p < 0.0001$)

in the **Decision Tree** ($M = 48.23$, $SD = 63.13$) condition than in the **Feature Importance** ($M = 33.22$, $SD = 68.75$) condition ($p = 0.0136$). A comparison of all agreement ratings is shown in Figure B1b. We did not observe many statistically significant relationships between condition and participant-rated agreement with the virtual agent, which again corroborates our findings on xAI and compliance – namely, that compliance is unaffected by xAI condition.

When we specifically investigated agreement with *incorrect* suggestions, an interesting trend appeared. An ANOVA showed that participant-rated agreement was significantly affected by condition ($F(7, 1399) = 3.26$, $p = 0.0019$). A Tukey's HSD post-hoc revealed that, again, participants were more likely to agree with an agent in the **Decision Tree** ($M = 10.76$, $SD = 68.58$) condition than in the **Feature Importance** ($M = -13.71$, $SD = 66.20$) or **Templated Language** ($M = -13.51$, $SD = 71.35$) conditions at significance levels $p = 0.018$ and $p = 0.0203$, respectively. We also observe that, of all of our conditions, *only* **Case Based**, **Crowd Sourced**, and **Decision Trees** have *positive* average agreement scores. Results are shown in Figure B2.

Taking into context our results between xAI condition and the participants' compliance with the agent's suggestions, this result is surprising. Despite five of our eight conditions exhibiting *negative* average agreement with the virtual agent for incorrect suggestions, we do not observe significant differences between conditions for inappropriate compliance. In other words, our study suggests that untrained human users may accept an agent's suggestion *even if they disagree with the agent and the agent is wrong!* It is possible that, despite disagreeing with the agent, users were fooled by the agent's confidence in its suggestions (e.g., the agent never says "I'm not sure" or "Maybe the answer is..."), as there is abundant psychology research to suggest that humans tend to over-trust confidence (Elaad et al., 2015; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005; O'Mara, Kunz, Receveur, & Corbin, 2019; Rollwage et al., 2020; Schlenker & Leary, 1982; Schroeder, Tremblay, & Tremblay, 2021; Swann & Ely, 1984; Thomas & McFadyen, 1995). If someone got a previous question wrong, they might lose confidence in themselves and want to take the agent's suggestions (Hedlund, Johnson, & Gombolay, 2021). This result signals a need for xAI research to empower human users to actively challenge or interrogate their agent assistants, or for xAI agents to regularly remind users of their fallibility (Natarajan & Gombolay, 2020). At present, our results suggest that users may be feeling pressure to accept agent suggestions even if they do not agree with such suggestions.

Finally, Pearson's correlations revealed that agreement, understandability, accuracy, and compliance were all statistically significantly correlated ($p < 0.0001$). Of these correlations, understandability and agreement were the strongest ($R = 0.79$), followed by agreement and compliance ($R = 0.41$), understandability and compliance ($R = 0.32$), agreement and accuracy ($R = 0.16$), and understandability and accuracy ($R = 0.13$). A comparison of understandability and agreement is shown in Figure B3.

Appendix C. Scenarios

In this section we present all scenarios used in the study. Scenario 1 was presented alongside instructions with how to use the interface and work with the virtual robot, while the remaining scenarios did not include additional instructions or content (apart from associated explanations). Scenarios 2-6 were used as the priming task, and scenarios 7-20 were used as the main body of the study. For each question, the correct answer is highlighted in bold, and the robots incorrect suggestion (where applicable) is highlighted in bold and red.

C.0.0.1. 1. A soccer player arrives to the training facility early every day. After several months of rigorous training and practice, the player still hasn't managed to make it into the starting team, with too much competition for their preferred position. However, the player has significantly improved coordination, acceleration, and top-speed. Which position is the player most likely to play?

(1) **Attacker**
(2) Defender
(3) Midfielder
(4) Goalkeeper

C.0.0.2. 2. Mark has just started running, and is trying to train for a local marathon. The marathon is set to take place in a month, so Mark has been training very hard. Unfortunately, a week before the marathon, Mark suffered an injury. Where was Mark injured?

(1) Elbow
(2) Neck
(3) **Knee**
(4) Back

C.0.0.3. 3. John is preparing a garden behind his building. He dug up an old tree stump and cleared out weeds to prepare a vegetable box, and must now prepare the ground for seeds. How should John fill in the vegetable box before planting seeds?

(1) **Mixing soil and fertilizer**
(2) Mixing soil and clay
(3) Mixing clay and fertilizer
(4) Mixing clay and gravel

C.0.0.4. 4. Jane needs to attend a meeting on the other side of the country tomorrow. Her company will pay for her expenses, the top priority is for her to physically attend the meeting. What is the best way for Jane to get to the meeting on time?

(1) Take a train
(2) **Fly**
(3) Drive
(4) Take a bus

C.0.0.5. 5. Monica has been working from home for the past several months, and is constantly suffering from eye-strain and headaches from staring at her computer monitor all day. Which of the following is the **least likely** to help reduce Monica's headaches and eye-strain?

(1) Use a blue light filter on her computer
(2) **Do more work in the dark with the lights turned off**
(3) **Break up the day with walks outside**
(4) Regularly take breaks to stare at distant objects

C.0.0.6. 6. James stops at a lookout while driving across the country to rest. While there, he looks out across a herd grazing on a plain, composed of animals native to North America. Which animals is James looking at?

(1) Cattle
(2) Domestic Sheep
(3) Bobcats
(4) **Bison**

C.0.0.7. 7. Everyday at 8:00 AM and 6:00 PM, a person's pet needs to be fed a scoop of food. The pet's space in the house needs to be cleaned weekly and typically takes under an hour to clean. The pet needs to go to the vet every 6 months. What type of animal is the pet?

(1) Dog
(2) **Cat**
(3) Hamster
(4) Fish

C.0.0.8. 8. Jamie is an avid hiker. She loves to explore outdoors in cool weather with just a light jacket, without worrying about bugs or heavy rainstorms, and particularly enjoys venturing up into the mountains to walk along small streams. Unfortunately for Jamie, her allergies always flare up as flowers bloom. Which season is best for Jamie to go hiking

(1) Spring
(2) Summer
(3) **Fall**
(4) Winter

C.0.0.9. 9. Patrick has struggled to recreate a recipe he found online. He hasn't ever tried to cook this particular dish before, and he is finding it difficult to replicate precisely. Because of the ways that altitude can affect cook times in ovens, Patrick is finding that his finished product doesn't look like the example online. Which food is Patrick preparing?

(1) **Bread**
(2) Chicken
(3) Veggie Platter
(4) Homemade Chocolate

C.0.0.10. 10. Shelby loves to read. In the past year, she's finished several books by Dostoevsky and Tolstoy, and others set in a violent coup or in a gulag. Which genre does Shelby seem to prefer?

(1) Biographies
(2) Historical Fiction
(3) **Russian Literature**
(4) Gritty Fantasy

C.0.0.11. 11. Jean is busy training for the upcoming finals in her favorite sport. To train appropriately, Jean is dedicating an hour each day to stretching and warming up, and then alternating between 5 to 8 miles of distance training or an hour of speed training. Jean's teammates are also pushing themselves very hard, as they'll all be competing for first place. Which sport is Jean training for?

(1) Hurdles
(2) **Cross Country**
(3) **Swimming**
(4) Soccer

C.0.0.12. 12. Arlo spends his days on his feet. He is often talking to other people, though other people will only seldom have the opportunity to respond or to interject. Arlo's audiences often pay very close attention for an hour at a time, and then rotate out for a new audience. Which profession best matches Arlo?

(1) Doctor
(2) Stage Performer
(3) Lawyer
(4) **Teacher**

C.0.0.13. 13. Carl enjoys the same drink every day. After he arrives to work, stressed from the chaos of his commute, he always goes straight to the break room to catch up with co-workers. Carl usually takes this time to calm himself down and try to relax, not needing any more stimulation after his commute. Which drink does Carl prefer in the break room before beginning work?

(1) **Tea**
(2) **Coffee**
(3) Whisky
(4) Soda

C.0.0.14. 14. Charon's favorite music helps her get through tough workouts. When she is exhausted and worn-out, the predictable and energetic rhythms of her favorite songs will always help motivate her to finish. What type of music does Charon enjoy for her exercise?

(1) Cinematic Soundtracks
(2) **Jazz**
(3) **Rock**

(4) Classical

C.0.0.15. 15. Charlie lives 4 miles from his workplace in a city with heavy traffic. His workplace is near a subway station, but Charlie's house is 2 miles from a station and he doesn't like physical activity. Fortunately, his workplace offers bike racks in the parking deck. Which mode of transportation best fits Charlie's commute?

(1) Bike
(2) **Electric Scooter**
(3) Car
(4) Subway

C.0.0.16. 16. Jay is suffering from chronic headaches. He has been feeling bad for a few months, ever since a knee injury forced him to stop running every afternoon. With the added time, Jay has been much more active on social media, and he is excitedly considering a career as an influencer. Which of the following is likely the cause of Jay's headaches?

(1) **Increased screen time**
(2) Less running every afternoon
(3) A change in diet
(4) **Increased stress over his career choice**

C.0.0.17. 17. Taylor tried to bake bread for the first time last week. Unfortunately, he forgot to account for the mess created by kneading dough, causing him to coat his hands in sticky dough and his clothes in flour. He also misread the instructions, as his eyes were burning from chopping onions that he used in his dinner. Before he tries baking again tonight, which change to his outfit should Taylor make?

(1) Wear white clothes
(2) Put on gloves
(3) Wear goggles
(4) **Put on an apron**

C.0.0.18. 18. Persephone is trying to cut down a tree to make more space for her cars behind her house. After a few hours of exhausting work on a rainy summer day, she managed to get the tree down and out of her yard. However, she was left with a stump about four inches high and 10 inches across in the middle of her yard. How should Persephone deal with the stump?

(1) Use a sledgehammer to hit it down into the ground
(2) Use an axe to cut more of the trunk away
(3) **Use a shovel to dig it out**
(4) **Use a controlled fire to burn it away**

C.0.0.19. 19. George is trying to complete a tour of European capitals before he graduates. Next month, he will begin a study-abroad in Lyon, France where he will be able to visit new cities every month. Having never visited Europe before, where will George go first?

(1) **Paris**
(2) Madrid
(3) London
(4) Berlin

C.0.0.20. 20. When the total solar eclipse crossed the country on a Wednesday afternoon, thousands of tourists flocked to a narrow band of space where they would be able to see the total eclipse. There were not many restaurants or shops to visit in the path of the eclipse. How was traffic on the roads after the eclipse passed?

(1) No traffic
(2) **Heavy traffic**
(3) Light traffic
(4) Moderate traffic

References

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, *3*(1), 22–38.

Elaad, E., et al. (2015). The distrusted truth: Examination of challenged perceptions and expectations. *Psychology*, *6*(05), 560.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research*, *42*(3), 237–288.

Hack, H. (1958). An empirical investigation into the distribution of the f-ratio in samples from two non-normal populations. *Biometrika*, *45*(1/2), 260–265.

Hedlund, E., Johnson, M., & Gombolay, M. (2021). The Effects of a Robot's Performance on Human Teachers for Learning from Demonstration Tasks. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 207–215). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3434073.3444664`

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, *89*(6), 899.

Natarajan, M., & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (p. 33–42). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3319502.3374839`

O'Mara, E. M., Kunz, B. R., Receveur, A., & Corbin, S. (2019). Is self-promotion evaluated more positively if it is accurate? reexamining the role of accuracy and modesty on the perception of self-promotion. *Self and Identity*, *18*(4), 405-424. Retrieved from `https://doi.org/10.1080/15298868.2018.1465846`

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, 114–133.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature communications*, *11*(1), 1–11.

Schlenker, B. R., & Leary, M. R. (1982). Audiences' reactions to self-enhancing, self-denigrating, and accurate self-presentations. *Journal* *of Experimental Social Psychology*, *18*(1), 89-104. Retrieved from https://www.sciencedirect.com/science/article/pii/002210318290083X

Schroeder, E., Tremblay, C. H., & Tremblay, V. J. (2021). Confidence bias and advertising in imperfectly competitive markets. *Managerial and Decision Economics*, *42*(4), 885–897.

Schrum, M. L., Johnson, M., Ghuy, M., & Gombolay, M. C. (2020). Four years in review: Statistical practices of likert scales in human-robot interaction studies. In *Companion of the 2020 acm/ieee international conference on human-robot interaction* (p. 43–52). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3371382.3380739

Swann, W. B., & Ely, R. J. (1984). A battle of wills: self-verification versus behavioral confirmation. *Journal of personality and social psychology*, *46*(6), 1287.

Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, *16*(1), 97-113. Retrieved from https://www.sciencedirect.com/science/article/pii/0167487094000326