Supplementary Information: This Supplementary Information has not been peer reviewed.

Title: “It's a good reason to get up every morning!”: Perspectives on how feeding and observing birds affects people’s wellbeing

Appendix S1: Survey Questionnaires

As all of our surveys were extensive, we present only the questions here that are relevant to this manuscript. Questions appear here in the same order as in the original surveys, although numbers do not represent the question numbers in the original surveys (questions not relevant to this manuscript preceded them or were in between them).

Pre-season survey

Q1. Below are a few questions about your overall well-being and health. Remember that there are no right or wrong answers and you may skip any questions that you do not want to answer. We understand that these questions are personal and your answers will be confidential as they are elsewhere in this survey.

Please select the box for each statement that best describes your experience of each over the last 2 weeks.

The Short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS) © University of Warwick 2006, all rights reserved, was included here. To access the scale items and scoring guide, please register at https://warwick.ac.uk/services/innovations/wemwbs.

Below are a few questions to help us learn more about who participates in Project FeederWatch.

Q2. In what year were you born? *Please write in your birth year below.*

_____

Q3. What is your gender? *Please select one response or write in your answer below.*

- ○ Man (6)
- ○ Woman (2)
- ○ Non-binary (3)
- ○ I prefer not to answer. (4)

○ I prefer to self-describe. (5)

_____

Q4. What is your race and/or ethnicity? *Please select all that apply.*

☐ American Indian, Alaska Native, Native American, or Indigenous (1)

☐ Asian (2)

☐ Black or African American (3)

☐ Hispanic or Latino/Latina/Latine/Latinx (4)

☐ Middle Eastern or North African (5)

☐ Native Hawaiian or other Pacific Islander (6)

☐ White (8)

☐ Some other race or ethnicity (7)

☐ I prefer not to answer. (9)

☐ I prefer to self-describe. (10)

_____

Q5. Are you currently a guardian or primary caregiver for any of the following? *Please select all that apply and/or write in your answer below.*

- ☐ Minor child(ren) under the age of 18 (1)
- ☐ Adult child(ren) (2)
- ☐ Spouse or partner (3)
- ☐ secondary or other senior adults(s) (4)
- ☐ Other *(please specify)* (5)

_____

- ☐ None (6)
- ☐ I prefer not to answer. (7)

Q6. What is the highest level of education you have completed or the highest degree you have received? *Please select one response.*

- ○ High school diploma, equivalent, or less (1)
- ○ Some college (2)
- ○ Associate's or technical degree (3)
- ○ Bachelor's degree (4)

- ○ Professional, master's, or doctoral degree (5)

Q7. What is your annual household income? *Please select one response.*

- ○ Less than $24,999 (1)
- ○ $25,000 - $49,999 (2)
- ○ $50,000 - $99,999 (3)
- ○ $100,000 - $199,999 (4)
- ○ More than $200,000 (5)
- ○ I prefer not to answer. (6)

Q8. What is your sexual orientation? *Please select one response or write in your answer below.*

- ○ Heterosexual (1)
- ○ Gay/Lesbian (2)
- ○ Bisexual (3)
- ○ Pansexual (4)
- ○ Asexual or ace spectrum (5)
- ○ Sexually/romantically fluid (6)

○ I prefer not to answer. (7)

○ I prefer to self-describe. (8)

_____

Q9. Do you experience challenges associated with any of the following in your daily life?
Please select one response per category.

	Not at all (1)	Very little (2)	Somewhat (3)	Quite a bit (4)	A great deal (5)	I prefer not to answer. (6)
Vision (1)	○	○	○	○	○	○
Hearing and/or auditory processing (2)	○	○	○	○	○	○
Cognition and learning (e.g., autism, dyslexia, intellectual disability) (3)	○	○	○	○	○	○
Mobility/physical (e.g., spinal cord injury, difficulties with dexterity) (4)	○	○	○	○	○	○
Mental health (e.g., anxiety, depression, PTSD) (5)	○	○	○	○	○	○

Chronic illness or health condition (e.g., chronic pain, illness, autoimmune diseases: diabetes, celiac disease) (6)	○	○	○	○	○	○
Other *(please specify below)* (7)	○	○	○	○	○	○

Q10. Do you identify as disabled/are you disabled? *Please select one response.*

- ○ Yes (1)
- ○ No (2)
- ○ I don't know. (3)
- ○ I prefer not to answer. (4)

Q11. Do you identify as neurodiverse/are you neurodiverse? *"Neurodiverse" is a nonmedical term used to explain the unique ways people's brains work. Identifying as or being neurodiverse means having a brain that works differently from the average or "neurotypical" person. A medical condition or diagnosis is not a requirement for neurodiversity. Some of the conditions that are common among those who describe themselves as neurodiverse include autism, attention-deficit hyperactivity disorder (ADHD), dyslexia, and mental health conditions. Please select one response.*

- ○ Yes (1)

- ○ No (2)
- ○ I don't know. (3)
- ○ I prefer not to answer. (4)

Post-season survey

Q12. Below are a few questions about your mental well-being. Remember that there are no right or wrong answers and you may skip any questions that you do not want to answer. We understand that these questions are personal and your answers will be confidential as they are elsewhere in this survey.

Please select the box for each statement that best describes your experience of each over the last 2 weeks.

The Short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS) © University of Warwick 2006, all rights reserved, was included here. To access the scale items and scoring guide, please register at https://warwick.ac.uk/services/innovations/wemwbs.

Q13.To what extent has feeding and/or observing birds at your feeding areas affected your responses to the previous question?

_____

_____

_____

_____

_____

Q14. To what extent did feeding birds and observing your bird feeding area(s) have a positive or negative impact on your mental well-being between November 2024-April 2025? *Mental well-being enables people to feel happy, cope with the stresses of life, realize their abilities, learn well and work well, and contribute to their community. Please select one response.*

- ○ Extremely negative (2)
- ○ Very negative (3)
- ○ Somewhat negative (4)

- ○ Neither positive nor negative (9)
- ○ Somewhat positive (5)
- ○ Very positive (6)
- ○ Extremely positive (7)
- ○ I prefer not to answer. (8)

Q15. To what extent did other experiences in your life (not related to feeding birds or observing your bird feeding areas) have a positive or negative impact on your mental well-being between November 2024-April 2025? *Please select one response.*

- ○ Extremely negative (2)
- ○ Very negative (3)
- ○ Somewhat negative (4)
- ○ Neither positive nor negative (9)
- ○ Somewhat positive (5)
- ○ Very positive (6)
- ○ Extremely positive (7)
- ○ I prefer not to answer. (8)

Q16. As a result of feeding birds and observing my bird feeding areas between November 2024-April 2025, I have been feeling this way in my life…

	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)
Optimistic (1)	○	○	○	○	○
Useful (2)	○	○	○	○	○
Distressed (3)	○	○	○	○	○
Able to deal with problems well (4)	○	○	○	○	○
Anxious (5)	○	○	○	○	○
Close to other people (6)	○	○	○	○	○
Able to make up my mind about things (7)	○	○	○	○	○
Relaxed (8)	○	○	○	○	○
Able to think clearly (9)	○	○	○	○	○
Helpless (10)	○	○	○	○	○

Appendix S2: Supplementary Methods

Short Warwick Edinburgh Mental Wellbeing Scale

In the pre-season survey, we measured wellbeing using the Short Warwick-Edinburgh Mental Well-Being Scale (SWEMWBS, © University of Warwick 2006, all rights reserved), with seven five-point Likert-type items and possible scale scores ranging from 7 (all seven items answered with a "1") to 35 (all seven items answered with a "5") (Stewart-Brown et al. 2009; Appendix S1: Survey Questionnaires, Q1). This is an abbreviated version of the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS), which focuses on positive mental health, covers both hedonic (related to pleasure) and eudaimonic (related to doing good) aspects of wellbeing, and has been shown to be responsive to changes in wellbeing from interventions over longer periods of time (Maheswaran et al. 2012). Individual changes in wellbeing of at least 1 to 3 points on the SWEMWBS are considered meaningful (Shah et al. 2018). We measured SWEMWBS again in the post-season survey.

We calculated the difference between each respondent's pre- and post- season SWEMWBS score, and sorted respondents into three groups based on the upper threshold for meaningful change (Shah et al. 2018): those who experienced a negative change (≤ -3), a positive change (≥ 3), and no large change ($-3 < \text{change} < 3$). We then used a Kruskal-Wallis test and Dunn's post-hoc test to examine whether these three groups differed in how they rated the impact of feeding birds on their wellbeing. To ensure that our results were robust to different thresholds, we ran the same tests with respondents with the lower threshold for meaningful change in SWEMWBS (1 scale point, Shah et al. 2018).

This was done to detect whether there were any pronounced quantitative changes in wellbeing over the course of the PFW season, although participants in our study had likely been feeding birds before the season began, and many had participated in PFW in previous years. Likewise, change over such a long period of time, even if detected, would be difficult to relate directly to program participation, but for this reason we compared participant self-assessments with change over time. We did this to see whether there was any relationship between experiencing a change in wellbeing over 6 months and self-perceived impacts of bird feeding on wellbeing (that is, did those who believe bird feeding had a positive impact on their wellbeing experience change as measured by a psychometric scale).

Qualitative Coding

The lead author coded the responses deductively, utilizing a codebook based on Buijs & Jacobs' (2021) set of pathways for linking positive human–wildlife interactions to wellbeing. She adjusted categories to reflect participant responses and added tertiary codes (categories nested

within the “secondary” or overarching codes) that captured key themes repeated across participants. Coding was performed in three steps. First, the lead author coded responses into mutually-exclusive “primary” codes based on the type of impact respondents felt that feeding and observing birds had on their wellbeing: “Positive”, “Negative”, and “None”. She labeled responses “Non-directional” if they were too vague to interpret or did not fit into the coding scheme (for example, statements of magnitude without direction like “a lot” or “45%”). She then further sub-coded positive responses into the three secondary codes based on the pathways proposed by Buijs & Jacobs (2021): “Emotions”, “Engagement”, and “Meaning”. Finally, she performed sub-coding to tertiary codes within each of these three categories based on components from Buijs & Jacobs (2021) as well as emergent themes. For the secondary and tertiary coding, multiple codes could be applied to each response. All initial coding of the 440 responses was done by the lead author, with input solicited from co-authors on codes and example responses.

Coding with large language model

After coding a subset of responses (n=440) from our survey by hand, the lead author began attempting to code our responses with a large language model (LLM) with the plan to eventually use it to code our entire dataset. Due to suboptimal model performance at the tertiary coding level and consistent bias within some categories, we did not use an LLM to code our entire dataset. Instead, we performed analysis only on the subset of 440 human-coded responses, which had been coded to saturation within our framework and also is over the threshold of 244 respondents needed to generalize a population of 40,000 with skewed responses with 95% confidence and a 5% margin of error (almost 30,000 participants engage in PFW; Vaske 2019). We present the results of our interrater reliability testing in the manuscript as an intercoder reliability metric.

We performed coding in an open source model from OpenAI, gpt-oss-120b, on the [blinded university] computing cluster from August-October 2025 (see here for model documentation: https://huggingface.co/openai/gpt-oss-120b). This use of our data was allowed by the [blinded university] IRB after consultation, as the data are securely stored within university servers and are not accessible to OpenAI. The model was used in “instruct” mode with reasoning set to “high” and thinking enabled. We set aside half of our coded dataset (n = 220) for context learning and used the other half for accuracy testing. Context learning refers to when LLMs learn new tasks directly through examples provided in a prompt, rather than through training the model by adjusting parameters. To ensure proper context learning, given the small sample size in some of our categories, we split each coded category evenly instead of splitting the dataset completely randomly. In an interactive chat session, we provided the model with a written out form of the codebook, shared half of the data with codes attached to enable context learning, and then asked the model to code the remaining half of the data. We iterated on model prompts to improve model accuracy, both through giving more specific instructions on the task and creating more precise definitions in the codebook itself. Occasionally we provided the model the correct codes for the testing dataset and asked it to assess what it was mis-coding most

frequently as compared to the human coder, and then used those responses to remove ambiguities from the codebook and instructions.

We first asked the model to code our data using the first level of codes (the “primary” codes) from our codebook (“positive”, “negative”, “neutral”, and “NA” [called “Non-directional” in the manuscript]). Then, in a separate session, we asked the model to code all of the responses we had assigned to the “positive” category further into the two additional levels of codes (“secondary” and “tertiary” codes) specified in our codebook. We again split our dataset of positive codes in half, reserving approximately half (n = 154) for context learning and approximately half (n = 152) for accuracy testing, again splitting evenly within secondary and tertiary coding categories. We coded the secondary and tertiary codes together, as opposed to sequentially, because we found that improved the model accuracy.

Below is the full prompt used to perform the first round of coding. This is the final version for which we report accuracy in the main text.

> *You are a qualitative researcher analyzing survey data. You are looking at responses to how bird feeding and observing birds affects wellbeing. You are categorizing the dataset into the following themes:*
>
> *Positive: Participant explicitly states that bird feeding and observing birds has some positive impact on their wellbeing, potentially through increased positive emotions, decreased negative emotions, and reduced stress or increased calm. Even if participant mentions that they don't think feeding has an effect, the response should be coded as positive if they mention specific positive ways that it affected them.*
>
> *None: Participant explicitly states that bird feeding has no impact on their wellbeing. Does not include any other extents, for example "very little" is not counted in this category because that implies some effect.*
>
> *Negative: Participant explicitly states that bird feeding and observing birds has at least some negative impact on their wellbeing. Responses should not be included in this category if participants mention negative sentiment towards something other than feeding and observing birds.*
>
> *NA: Participant did not indicate whether bird feeding or observing birds had an impact on their wellbeing, or did not indicate what type of impact it had. Most of these will be the extent to which bird feeding affected their answer without indicating a direction. Responses also fall into this category if respondents state that feeding birds had no effect because they did not feed birds.*
>
> *Here are examples of responses and themes assigned to them (no need to do anything, these are just for you to learn):*

[enter example responses with codes set aside for context learning]

Please state which of these themes apply to each of the lines in the original dataset, attached. Do not skip or repeat any lines, and please provide responses in the same order as the original dataset. No need to explain why each theme applies, you can simply state the name of the theme or themes after each unique ID number associated with the response. More than one code can be applied to each response (for example, "positive & negative"), but please only apply the code if it matches the response. Remember that things can be coded as both positive and negative, if people mention both things that are positive about bird feeding and observing (feeling good, feeling relaxed) and negative (feeling stressed, feeling unhappy). If an extent is applied like "a little" or "a lot" but without mentioning the direction or positive or negative sentiment, the code should be "NA".

[enter responses for coding]

This produced one response that was "positive & none", which we did not allow - we then prompted the model to choose just one and use that final code for our accuracy testing.

Below is the full prompt used to perform the second round of coding. This is the final version for which we report accuracy in the main text. Similar levels of accuracy were achieved using prompts generated by the model based on prior coding attempts. In some cases, accuracy was higher for the secondary codes (Emotions/Engagement/Meaning) but lower for the tertiary codes.

You are a qualitative researcher analyzing survey data. You are looking at responses to how feeding and observing birds positively affects wellbeing. Note that themes should only be applied to parts of the response specifically about feeding and observing birds. You are categorizing the dataset into the following secondary and tertiary themes (tertiary themes are nested under secondary themes):

secondary theme 1 is "Emotions": Bird feeding and observing birds impacts wellbeing via changes in emotions, specifically increased positive emotions, decreased negative emotions, and reduced stress. This should include specific references to how feeding and observing birds makes people feel, including enjoyment, happiness, calm, peace, etc.

Emotions tertiary themes are "Positive emotion", "Negative emotion reduction", and "Stress reduction". The definitions are as follows.

Positive emotion: Bird feeding positively impacts wellbeing via increased positive emotions, like joy, excitement, happiness, gratitude, and hope. Participant must explicitly state that feeding and/or observing birds makes them feel a positive emotion, the mere

inclusion of a positive affect word is not sufficient. While participants speaking of enjoying, loving, or liking feeding/observing birds should be coded under this because that implies them feeling enjoyment during the activity. This includes statements of looking forward to the activity, and the activity being pleasant. This does not include statements of "feeling better", which would be coded in the negative emotion reduction category. Trigger words: “enjoy”, “joy”, “excited”, “happy”, “grateful”.

Negative emotion reduction: Bird feeding positively impacts wellbeing via decreased negative emotions. Participants may mention feeling less distressed, feeling better or uplifted, and being distracted from world events/politics/health crises/problems that are stated or implied to be distressing. Trigger words: “cheers me up”, “makes me feel better”, “lifts my spirits”, “helps me forget …”

Stress reduction: Bird feeding positively impacts wellbeing via decreased stress, worry, and anxiety. This includes mentions of the activity being calming (relaxed, peaceful) or inducing mindfulness (therapeutic, grounding), as well as mentions of decreasing stress or distracting from explicitly stressful thoughts or situations. Trigger words: “calm”, “relaxed”, “grounded”, “less anxious”, “less worried”.

secondary theme 2 is “Engagement”: Bird feeding and observing birds impacts wellbeing via feelings of engagement and a sense of "flow" or focus. This may be a result of awe or fascination, related to interest and curiosity, and the development of knowledge and skills. This should not include statements of how birding is part of routine.

tertiary themes for Engagement are “Interest”, “Focus”, “Wonder”, and “Growth”. The definitions are as follows.

Interest: Respondent feels that bird feeding has a positive effect on their wellbeing through engaging their curiosity or interest, including observing interesting behaviors. Respondent must state this explicitly, speaking about their curiosity and interest in the activity - simply stating they enjoy birding is not sufficient. If respondents mention the activity is interesting, it should be coded in this category. Trigger words: “interesting”, “curious”, “fascinating”, “intriguing”

Focus: Respondent feels that bird feeding has a positive effect on their wellbeing through allowing them to enter a state of focus and concentration. Some describe zoning out or entering a meditative state. This should only be applied when the participant explicitly discusses focusing on the activity. In line with Attention Restoration Theory. Trigger words: “zone-out”, “focus”, “meditative”

Wonder: Respondent feels that bird feeding has a positive effect on their wellbeing through fostering a sense of awe or wonder that captures their attention. The participant must explicitly describe the sense of wonder or awe. Trigger words: “wonder”, “awe”, “amazing”

Growth: Respondent feels that bird feeding has a positive effect on their wellbeing through helping them learn new things, stimulating them intellectually, and testing their skills. All responses about learning should be coded with growth. Trigger words: "learning", "testing", "skills"

secondary theme 3 is "Meaning": Bird feeding and observing birds impacts wellbeing via creating meaning in the life of the participant through feelings of connection with something larger than themselves (including both nature and larger communities of people) and doing good.
tertiary themes for Meaning are "Doing good", "Connection to nature", and "Human Connection". The definitions are as follows:

Doing good: Bird feeding positively impacts wellbeing via creating meaning through a sense participants are doing something good, whether helping birds, helping nature, or contributing to conservation. Feeling they have satisfied the birds, that the birds are relying on them, that they are making a difference, or that they are improving habitat for wildlife should be coded into this category. This is not about creating positive affect. Trigger words: "helping", "contributing", "providing", "making a difference", "impact"

Connection to nature: Bird feeding positively impacts wellbeing via creating meaning through connecting participants to nature and the larger world. Respondents must explicitly discuss connection to nature, just mentioning spending time in or appreciating nature is not sufficient. Trigger words: "connection to nature", "part of the world"

Human Connection: Bird feeding positively impacts wellbeing via creating meaning through allowing people to connect with other people. This may be friends or family that they do or discuss the activity with, or a sense of being part of a larger community. Trigger words: "together", "community", "connected to people"

secondary theme 4 is "General positive": General positive sentiments that cannot be sorted into any of the other secondary or tertiary categories, often because they lack specific information.

No need to respond, these are just for you to learn.

[let model process]

Here are a set of examples - no need to respond, this is just to help you learn. Please use these to refine and update your understanding of how to do the coding.

[provided half of responses with codes for context learning]

Please state which of these themes apply to each of the lines in the original dataset, attached. Do not skip or repeat any lines, and please provide responses in the same order as the original dataset. No need to explain why each theme applies, you can simply state the name of the theme or themes after each unique ID number associated with the response. More than one secondary theme and more than one tertiary theme can be applied to each response, but please only apply the theme if it matches the response–single themes are preferred. Only apply a theme if it is explicitly part of the response. If a secondary theme is applied, at least one of the tertiary themes nested under it must be applied as well. "General positive" is exclusive of all of the rest of the codes. If "Emotions", "Engagement", and "Meaning" do not apply, write "General positive". Base your coding off of the codebook and examples. The responses may be given in the following format:

ExternalReference: secondary code - tertiary code; secondary code - tertiary code

[provided remaining half of responses for coding]

It is important to note that coding with an LLM differs from a second human coding in a few key ways. First, the LLM had been provided with a set of already-coded examples from the data. While this could improve LLM accuracy and align them more with the first human coder than if a second human were coding, it is similar to the way that two human coders often align on a set of accepted codes for a subset of the data before diverging to each code an additional subset upon which subsequent accuracy testing is performed. Second, the LLM was provided with a written out version of the codebook that provided additional rules and context. However, this could be replicated in a two human coder situation, where these rules and context may be agreed upon more informally in conversations. Third, the prompts and codebook were iterated upon throughout the process of coding with the LLM; however, iteration and an evolving understanding of the codebook is a natural part of human-centered coding as well. Finally, LLMs occasionally ignore direct instructions communicated to them in the prompt. For example, we instructed the LLM to sort responses into specific categories when certain key words or phrases were mentioned, but there were multiple instances of it failing to do so. In this way, LLMs fall short of human coders who can understand direct instructions. This led to ongoing coding bias wherein the LLM would continue to assign certain types of responses to categories despite direct instructions not to. Therefore, we made the choice not to code our entire dataset with the LLM. While our average interrater reliability metrics were fairly good, bias within categories in the LLM-based coding was too large of a concern (see Appendix S3 below, where Cohen's kappa statistics show that some individual codes had worse performance).

The use of LLMs in qualitative research is still evolving, and there does not yet exist a common set of standards for best practices and in which cases it is appropriate to utilize LLMs instead of human researchers. In this case, we attempted to employ an LLM for a task that would take a human many hours (coding our entire dataset of many thousands of responses, and then coding several hundred just for interrater reliability), but still based the core of our research off of human-based decision-making: our codebook was put together by the lead author and agreed

upon by all co-authors, and based off of extant theory as well as the lead author's coding of the subset of data presented in this paper. There remain many concerns about the use of AI in qualitative research, including its environmental impacts, the responsibility to engage with data that a researcher owes research subjects, and what is lost when the inherently subjective process of interpreting qualitative data is no longer done by a human but instead by a model that has its own biases and which may be more prone to capturing surface level themes (Schroeder et al., 2025). Through just using the LLM to measure interrater reliability, and presenting only results actually generated by a human, we avoid some of these pitfalls, although we acknowledge that the LLM may still be more likely to capture broader themes rather than nuance, and perhaps more likely to "agree" with the human coder training the model than a second human would be. However, qualitative research is inherently subjective, and through presenting fully our codebook (based in theory and agreed upon by multiple human authors) and LLM prompts here, we hope that this work has been made as "repeatable" as qualitative coding can be.

References:

Maheswaran, H., Weich, S., Powell, J., & Stewart-Brown, S. (2012). Evaluating the responsiveness of the Warwick Edinburgh Mental Well-Being Scale (WEMWBS): Group and individual level analysis. Health and Quality of Life Outcomes, 10(1), 156. https://doi.org/10.1186/1477-7525-10-156

Schroeder, H., Le Quéré, M.A., Randazzo, C., Mimno, D. & Schoenebeck, S. (2025). "Large language models in qualitative research: uses, tensions, and intentions." In Proceedings of the 2025 chi conference on human factors in computing systems, pp. 1-17.

Shah, N., Cader, M., Andrews, W. P., Wijesekera, D., & Stewart-Brown, S. L. (2018). Responsiveness of the Short Warwick Edinburgh Mental Well-Being Scale (SWEMWBS): Evaluation a clinical sample. Health and Quality of Life Outcomes, 16(1), 239. https://doi.org/10.1186/s12955-018-1060-2

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., & Weich, S. (2009). Short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS) [Database record]. APA PsycTests.https://doi.org/10.1037/t80221-000

Vaske, J. J. (2019). Survey research and analysis (2nd ed). Sagamore-Venture Publishing.

Appendix S3: Interrater Reliability Metrics

Table S1. Interrater reliability metrics comparing the performance of LLM coding to human coding performed by the first author for the first level of coding. Metrics are based on the test set of n = 220, and are presented for each code. “Negative & Positive” is treated as its own code, as these were the only two codes that were allowed to be applied together. Cohen’s kappa is a chance-corrected reliability statistic, where .60-.79 is moderate agreement, .80-.90 is strong agreement, and >0.9 is almost perfect agreement (McHugh 2012). The number of times (n) and proportion of times the code was applied across the entire dataset is also shown for both the human and LLM coder, as well as raw agreement. Rarely-applied codes tend to have lower Cohen’s kappa values.

Code	Cohen’s kappa	*p*-value	Human n	LLM n	Human proportion	LLM proportion	Agreement
Positive	0.96	<0.001	149	151	0.68	0.69	0.98
None	0.98	<0.001	28	29	0.13	0.13	1
Negative & Positive	0.8	<0.001	3	2	0.01	0.01	1
Negative	0.89	<0.001	4	5	0.02	0.02	1
Non-directional	0.95	<0.001	36	33	0.16	0.15	0.99

Table S2. Interrater reliability metrics comparing the performance of LLM coding to human coding performed by the first author for the second level of coding. Metrics are based on the test set of n = 152, and are presented for each code. Cohen’s kappa is a chance-corrected reliability statistic, where .60-.79 is moderate agreement, .80-.90 is strong agreement, and >0.9 is almost perfect agreement (McHugh 2012). The number of times (n) and proportion of times the code was applied across the entire dataset is also shown for both the human and LLM coder, as well as raw agreement. Rarely-applied codes tend to have lower Cohen’s kappa values.

Code	Cohen’s kappa	*p*-value	Human n	LLM n	Human proportion	LLM proportion	Agreement
Emotions	0.74	<0.001	124	129	0.82	0.85	0.93
Meaning	0.83	<0.001	18	22	0.12	0.14	0.96
Engagement	0.6	<0.001	13	24	0.09	0.16	0.91
General positive	0.79	<0.001	21	16	0.14	0.11	0.95

Table S3. Interrater reliability metrics comparing the performance of LLM coding to human coding performed by the first author for the third level of coding. Metrics are based on the test set of n = 152, and are presented for each code. Cohen's kappa is a chance-corrected reliability statistic, where .60-.79 is moderate agreement, .80-.90 is strong agreement, and >0.9 is almost perfect agreement (McHugh 2012). The number of times (n) and proportion of times the code was applied across the entire dataset is also shown for both the human and LLM coder, as well as raw agreement. Rarely-applied codes tend to have lower Cohen's kappa values.

Code	Cohen's kappa	*p*-value	Human n	LLM n	Human proportion	LLM proportion	Agreement
Connection to nature	0.81	<0.001	7	10	0.05	0.07	0.98
Positive emotions	0.86	<0.001	68	77	0.45	0.51	0.93
Stress reduction	0.92	<0.001	59	63	0.39	0.41	0.96
Interest	0.53	<0.001	5	13	0.03	0.09	0.95
Negative emotion reduction	0.69	<0.001	21	15	0.14	0.1	0.93
Doing good	0.85	<0.001	10	11	0.07	0.07	0.98
Focus	0.66	<0.001	4	5	0.03	0.03	0.98
Wonder	1	<0.001	3	3	0.02	0.02	1
Growth	0.66	<0.001	2	4	0.01	0.03	0.99
Community	1	<0.001	1	1	0.01	0.01	1

References

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282.

Appendix S4: Demographics Tables

Table S4. Demographics of respondents included in the sample used for quantitative analysis (n = 3,658). The total number of respondents who answered each question is shown in the variable header, except for the question regarding accessibility challenges, for which each dimension of access was its own question which resulted in different sample sizes. The percentages are calculated out of the total number of respondents who answered each question. For accessibility challenges, the number and percentage of respondents who answered “somewhat”, “quite a bit”, or “a great deal” in response to how much the challenge affected them are shown. “White alone” and “BIPOC” are mutually-exclusive variables derived from ethnoracial identities that respondents selected, but do not add up to 100% because some respondents wrote in identities that could not be definitively placed in one category or another. Variables with an * were “select all that apply” questions, and the percentages do not add up to 100.

Demographic Variable	**Sample size (percent)**
Age	n = 3,573
18-24	2 (<0.1%)
25-34	57 (1.6%)
35-44	119 (3.3%)
45-54	228 (6.4%)
55-64	602 (17%)
65+	2,565 (72%)
Gender	n = 3,586
Man	949 (26%)
Woman	2,618 (73%)
Non-binary	12 (0.3%)
I prefer to self-describe.	7 (0.2%)
Education	n = 3,618
High school diploma, equivalent, or less	100 (2.8%)
Some college	248 (6.9%)
Associate’s or technical degree	269 (7.4%)
Bachelor’s degree	1,165 (32%)
Professional, master’s, or doctoral degree	1,836 (51%)
Sexuality	n = 3,093
Heterosexual	2,932 (95%)
Gay/Lesbian	79 (2.6%)
Bisexual	40 (1.3%)
Pansexual	15 (0.5%)
Sexually/romantically fluid	7 (0.2%)
Asexual or ace spectrum	5 (0.2%)
I prefer to self-describe.	15 (0.5%)

Income	n = 2,549
Less than $24,999	55 (2.2%)
$25,000 - $49,999	294 (12%)
$50,000 - $99,999	951 (37%)
$100,000 - $199,999	965 (38%)
More than $200,000	284 (11%)
Identifies as disabled	n = 3,448
No	3,265 (95%)
Yes	153 (4.4%)
I don't know.	30 (0.9%)
Identifies as neurodivergent	n = 3,422
No	3,084 (90%)
Yes	178 (5.2%)
I don't know.	160 (4.7%)
Caregiving*	n = 3,454
Minor child	150 (4.3%)
Adult child	94 (2.7%)
Partner	176 (5.1%)
Senior	112 (3.2%)
Other	13 (0.4%)
None	2,977 (86%)
Accessibility challenges	
Vision (n = 3,432)	998 (29%)
Hearing and/or auditory processing (n = 3,454)	843 (24%)
Cognition and learning (n = 3,474)	81 (2.3%)
Mobility/physical (n = 3,466)	485 (14%)
Mental health (n = 3,460)	427 (12%)
Chronic illness or health condition (n = 3,438)	706 (21%)
Other (n = 21)	18 (percent not calculated)
Ethnoracial identity*	n = 3,435
American Indian, Alaska Native, Native American, or Indigenous	21 (0.6%)
Asian	26 (0.8%)
Black or African American	7 (0.2%)
Hispanic or Latino/Latina/Latine/Latinx	38 (1.1%)
Middle Eastern or North African	0 (0%)
Native Hawaiian or other Pacific Islander	0 (0%)
White	3,316 (97%)
Other	19 (0.6%)

Self describe	55 (1.6%)
Ethnoracial identity summary	n = 3,435
BIPOC	110 (3.2%)
White alone	3,304 (96%)

Table S5. Demographics of respondents included in the sample used for qualitative analysis (n = 440). The total number of respondents who answered each question is shown in the variable header, except for accessibility challenges, for which each dimension of access was its own question which resulted in different sample sizes. The percentages are calculated out of the total number of respondents who answered each question. For accessibility challenges, the number and percentage of respondents who answered "somewhat", "quite a bit", or "a great deal" in response to how much the challenge affected them are shown. "White alone" and "BIPOC" are mutually-exclusive variables derived from ethnoracial identities that respondents selected, but do not add up to 100% because some respondents wrote in identities that could not be definitively placed in one category or another. Variables with an * were "select all that apply" questions, and the percentages do not add up to 100.

Demographic Variable	**Sample size (percent)**
Age	n = 393
18-24	0 (0%)
25-34	8 (2.0%)
35-44	15 (3.8%)
45-54	20 (5.1%)
55-64	70 (18%)
65+	280 (71%)
Gender	n = 396
Man	99 (25%)
Woman	292 (74%)
Non-binary	3 (0.8%)
I prefer to self-describe.	2 (0.5%)
Education	n = 398
High school diploma, equivalent, or less	10 (2.5%)
Some college	17 (4.3%)
Associate's or technical degree	17 (4.3%)
Bachelor's degree	131 (33%)
Professional, master's, or doctoral degree	223 (56%)
Sexuality	n = 303
Heterosexual	272 (90%)
Gay/Lesbian	13 (4.3%)
Bisexual	8 (2.6%)
Pansexual	2 (0.7%)

Sexually/romantically fluid	3 (1.0%)
Asexual or ace spectrum	2 (0.7%)
I prefer to self-describe.	3 (1.0%)
Income	n = 238
Less than $24,999	9 (3.8%)
$25,000 - $49,999	25 (11%)
$50,000 - $99,999	89 (37%)
$100,000 - $199,999	86 (36%)
More than $200,000	29 (12%)
Identifies as disabled	n = 347
No	329 (95%)
Yes	16 (4.6%)
I don't know.	2 (0.6%)
Identifies as neurodivergent	n = 343
No	308 (90%)
Yes	24 (7.0%)
I don't know.	11 (3.2%)
Caregiving*	n = 348
Minor child	13 (3.7%)
Adult child	11 (3.2%)
Partner	21 (6.0%)
Senior	8 (2.3%)
Other	0 (0%)
None	298 (86%)
Accessibility challenges	
Vision (n = 345)	90 (26%)
Hearing and/or auditory processing (n = 346)	85 (25%)
Cognition and learning (n = 347)	7 (2%)
Mobility/physical (n = 346)	51 (15%)
Mental health (n = 347)	53 (15%)
Chronic illness or health condition (n = 345)	73 (21%)
Other (n = 3)	3 (percent not calculated)
Ethnoracial identity*	n = 347
American Indian, Alaska Native, Native American, or Indigenous	3 (0.9%)
Asian	2 (0.6%)
Black or African American	0 (0%)
Hispanic or Latino/Latina/Latine/Latinx	2 (0.6%)
Middle Eastern or North African	0 (0%)

Native Hawaiian or other Pacific Islander	0 (0%)
White	335 (97%)
Other	1 (0.3%)
Self describe	8 (2.3%)
Ethnoracial identity summary	n = 347
BIPOC	8 (2.3%)
White alone	335 (97%)

Appendix S5: Supplementary Results

Non-conservative threshold for change in SWEMWBS

With a non-conservative threshold for meaningful change of one scale point, 1,401 respondents experienced a positive change in wellbeing, 1,310 little to no change, and 947 a negative change.

As with the three-point threshold, using a one-point threshold for meaningful change we found that respondents who experienced a positive change in SWEMWBS rated birds as having a more positive impact on their wellbeing in the retrospective post-survey than those who experienced no change (Kruskal-Wallis $X2$ = 10.0, p = 0.007; Dunn test Z = 3.1, p = 0.006).