AI-Moderated Interviews vs. Static Online Surveys: A Comparative Study on Qualitative Response Quality

Aylin Idrizovski | Chair of Quantitative Marketing & Consumer Analytics | University of Mannheim

December 2025

Abstract

Traditional online surveys are widely used, yet they often produce short and superficial open-ended responses. In contrast, AI-moderated interviews are rising, offering a more natural and responsive format with conversational interaction and contextual follow-up questions. This study compares AI-moderated interviews with static online surveys in terms of qualitative response quality. A between-subjects design was used with two samples of 100 participants each, collected through an AI-moderated interview tool (Glaut's AIMI) and a static online survey platform (SoSci). Participants answered the same questionnaire on healthy lifestyle choices. The analysis includes linguistic, thematic, and response-quality metrics, as well as participant experience. AI-moderated interviews produced richer responses, including higher verbosity, more unique words, greater lexical diversity, and a broader range of themes. No differences appeared for content-word share or reading ease. AIMIs also showed a lower gibberish rate and higher participant experience ratings. Overall, the findings suggest that AI-moderated interviews generate richer and more reliable qualitative data than static online surveys, while maintaining participant comfort. This makes AIMI a powerful tool for organizations that depend on higher-quality insights for research and decision-making.

Keywords: AI-moderated interviews, static online surveys, response quality, linguistic metrics, qualitative research

Acknowledgment

I would like to thank Glaut Inc. for providing access to the AIMI interview platform and supporting the data-collection process by covering panelists' incentive costs. My sincere thanks also go to Dr. Veronica Valli, for her valuable guidance in my Master thesis development.

1. Introduction

Companies increasingly rely on fast and meaningful qualitative insights, yet traditional online surveys often produce short and superficial responses. AI-moderated interviews, as implemented via the Glaut AIMI platform, represent a promising alternative, but scientific evidence of their comparative advantage remains underexplored. A systematic comparison is therefore both practically important and scientifically relevant. This white paper examines whether AI-moderated interviews conducted through Glaut's AIMI generate higher-quality qualitative responses than traditional online surveys when both use the same questions and comparable samples. The scope is limited to a methodological comparison between AI-moderated interviews and traditional surveys, excluding an analysis of the substantive topic "healthy lifestyle choices" as well as human-moderated formats or different response modalities. In the following, the methodology is outlined, followed by the empirical results. The paper concludes with a discussion that includes a critical evaluation and managerial implications.

2. Methodology

This section provides the methodological basis of the study and ensures transparency in how the research question was examined. It begins with the assumption and design, continues with the detailed procedure, and concludes with the definitions of all measures used in the analysis.

2.1 Assumption and Design

The following outlines the conceptual assumptions and experimental design that form the basis of the study's comparative approach.

This study assumes that AI-moderated interviews generate richer and more qualitative responses than static online surveys. The expectation is based on the conversational nature of AIMI, which can ask adaptive follow-up questions, respond to participants' input, and maintain engagement capabilities shown to increase elaboration (Lee et al., 2020).

A between-subjects design was used to test this assumption. This design was chosen because exposing participants to both formats could create learning or carryover effects that might influence the quality of their responses. Participants were recruited via the panel provider *PureSpectrum* and selected based on predefined eligibility criteria. The sample was restricted to U.S. citizens aged 18–55 with an interest in health and fitness, and gender distribution was balanced across conditions. These criteria were chosen to ensure a relatively homogeneous sample in terms of language use, digital proficiency, and familiarity with the questionnaire topic, and to minimize systematic response biases. Individuals who met these requirements were then randomly assigned to either an AI-moderated interview conducted via Glaut's AIMI platform or a static online survey administered through SoSci Survey, resulting in two samples of n = 100.

All participants completed the same questionnaire on healthy lifestyle choices, which included six open-ended questions, each supported by follow-up prompts, and three structured items: a single-choice difficulty rating, a multiple-choice behavioral item, and a participant experience scale.

This design ensures that any differences observed in the results can be traced back to the data-collection format rather than to other external factors.

2.2 Procedure

This section describes how participants moved through the study, from recruitment to task completion, to ensure a clear understanding of the data-collection flow.

2.2.1 Similarities across Conditions

To provide a clear procedural foundation, the shared elements of both conditions are described first.

Recruitment. Participants were recruited through *PureSpectrum* and entered the study only after meeting predefined screening criteria. Eligible participants were then randomly assigned to either the AI-moderated interview (AIMI) or the static online survey (SoSci Survey). Participation was voluntary, compensated and anonymous.

Introduction. All participants viewed a standardized onboarding screen outlining the purpose of the study, confidentiality, approximate duration, and the absence of right or wrong answers.

Main Questions. After the introduction, all participants completed the same questionnaire on healthy lifestyle choices, with identical item order across conditions. The first three items addressed weekly habits, personal health priorities, and perceived difficulty. If participants selected "1 = not difficult," the subsequent open question investigating perceived barriers (Q4) was automatically skipped; otherwise, it appeared as intended. The remaining items asked about information sources related to food and exercise, personal motivations, recent purchases on health-related products and a final seven-item participant-experience scale. All items were mandatory. A full copy of the questionnaire is provided in **Appendix A.**

Closing and Data Handling. Participants completed the study at their own pace and in their natural environments. Both formats automatically flagged incomplete or abandoned sessions so that only fully completed responses were retained for analyses. In addition, AIMI's uncooperative-response detector identified gibberish or otherwise nonsensical input. This included illogical statements or repetitive patterns such as "yes, yes, yes." Respondents who triggered these low-quality indicators were excluded during data cleaning. Data collection continued until 100 valid datasets were obtained in each condition, and all procedures complied with GDPR and APA ethical standards.

Together, these elements ensured a unified procedural flow across conditions.

2.2.2 Differences between Conditions

Although the questionnaire content was identical, the two formats differed in interaction style, follow-up behavior, interface design, and the tone of the instructions and onboarding.

Follow-Up Logic. In the static survey, each open-ended question was followed by two predefined follow-up prompts (e.g., “Can you give me an example?” or “Can you tell me more about it?”), which participants could answer voluntarily. In the AIMI condition, follow-up prompts were generated dynamically based on the participant’s preceding response, creating a unique and customized conversational flow. Because of this dynamic nature, AIMI’s follow-ups could exceed two probes. For comparability, only responses from the first two follow-ups were analyzed.

Interface and Interaction Structure. The static survey used a traditional form-based layout with visible text boxes. AIMI presented the interaction sequentially in a chat-style interface. Even though Glaut’s platform allows respondents to choose between text and voice input, we deliberately restricted responses to text only to ensure comparability.

Instructions and Onboarding Tone. The introductory style also differed slightly. The static survey opened with a neutral instruction screen, whereas AIMI began with a brief conversational greeting, creating a more dialogic entry into the task in line with AIMI’s interaction guidelines.

These differences illustrate how the two formats shaped participants’ interaction with the task.

Overall, this procedure provides a consistent foundation for comparing the two response formats.

2.3 Measures

This section explains how each variable was operationalized and measured to enable a structured analysis of response quality.

All measures were derived from participants’ responses. Linguistic metrics, thematic variety, and the gibberish rate were computed from all open-ended answers, including follow-up responses. Participant experience was measured using the seven Likert-scale items at the end of the questionnaire.

Linguistic quality metrics. These metrics assessed the richness and structure of participants’ written responses. **Verbosity** captured the total number of words (e.g., “I walk daily” = 3 words vs. “I try to walk every morning to clear my mind” = 10 words). **Unique words** reflected the number of distinct word types (e.g., “I walk, walk, walk” contains 4 total words but only 2 unique words). **Lexical diversity (TRR)** was calculated as the ratio of unique to total words (e.g., “I walk to the park to walk my dog” contains 9 total words and 7 unique words, TRR = 7/9 ≈ .78). **Content-word share (CWS)** measured the proportion of nouns, verbs, adjectives, and adverbs (e.g., "I cook healthy meals at home" contains 4 content words out of 6, whereas "Well, you know, I mean…" contains none). **Reading ease (FRE)** was computed using the Flesch Reading Ease formula, which considers average sentence length and syllable density (Flesch, 1948). The score is calculated as:

$$\text{FRE} = 206.835 - 1.015\left(\frac{\text{words}}{\text{sentences}}\right) - 84.6\left(\frac{\text{syllables}}{\text{words}}\right)$$

Scores above 60 are commonly interpreted as easy to read (e.g., "I eat more vegetables now," FRE = 66.4), whereas scores below 30 indicate complex or dense writing (e.g., "I consistently incorporate high-fiber ingredients into my daily meals," FRE = 19.0).

Together, these metrics provide an overview of linguistic complexity and expressiveness.

Thematic variety. It was assessed through two complementary metrics. **Theme Count** captured the overall amount of thematic content by counting all theme mentions in a response, including repeated references within and across categories (e.g., exercise, exercise, nutrition = 3 themes in total). **Unique Themes** measured how many different idea categories participants mentioned in their responses (e.g., exercise, nutrition, sleep = 3 unique themes). A data-driven codebook (see **Appendix C**) was developed inductively from recurring themes across the dataset. Each open-ended answer was mapped to all applicable categories, and the unique-themes score reflected the number of distinct themes contained in the response. For instance, a statement like "I go for a walk to clear my mind" was assigned to the themes *physical activity* and *mental well-being*, resulting in 2 unique themes.

Response validity. This was evaluated using the **gibberish rate**. Responses were classified as gibberish when they contained random character strings, repeated characters, symbol-only input, non-English text, or extremely short and repetitive phrases without meaningful linguistic content. The gibberish rate represents the proportion of these invalid responses among all open-ended entries and serves as an indicator of data quality across conditions.

Participant experience with the study format. It was measured through several closed questions on a seven-point Likert scale. The items measured: "ease of expression", "feeling of comfort", "perceived repetitiveness", "perceived conversational quality", "data-handling trust", and "willingness to recommend the format." The **participant-experience score** was calculated as the mean across all seven items and summarizes participants' overall evaluation of the interview style.

These measures provide the analytical basis for identifying how the two formats differ in linguistic richness, thematic breadth, data validity, and participant perceptions.

Overall, this section provides the methodological basis for the study and sets the foundation for the empirical analysis.

3. Empirical Analysis and Results

This section summarises the main empirical findings of the study. The results follow a simple structure: first, the dataset is briefly introduced, and then each metric is analysed using descriptive and inferential statistics.

3.1 Dataset

To give orientation, this section outlines how the final dataset was prepared.

All incomplete responses were removed. In the AI-moderated interview condition, gibberish entries were excluded, resulting in $n = 100$ valid responses. In the static survey condition, gibberish responses were retained so that the sample size remained at $n = 100$. Screening criteria and random assignment ensured that both groups were comparable. Participants were between 18 and 55 years old, and the gender distribution was balanced in both conditions.

This balanced sample allows differences in the results to be attributed to the data-collection format rather than demographic factors.

3.2 Descriptive Statistics and Analysis

The following subsections summarise descriptive patterns and statistical comparisons for each metric. Detailed statistical tables are provided in **Appendix B.**

All statistical analyses were performed using parametric or non-parametric tests, selected based on the distributional characteristics of each measure. Normality was assessed using the Shapiro–Wilk test, and variance homogeneity using Levene's test. When variances were unequal, Welch's t-tests were applied. When distributions violated normality, Mann–Whitney U tests were used instead. Effect sizes (Cohen's d and rank-biserial r) were reported for all comparisons to ensure interpretable magnitude estimates. All linguistic metrics and statistical analyses were computed in R.

3.2.1 Linguistic Response Quality

This part evaluates participants' responses across five linguistic dimensions: response length (Verbosity), lexical richness (Unique Words), lexical diversity (TTR), content density (Content-Word Share), and readability (Flesch Reading Ease).

Verbosity. Figure 1 shows that responses in the AI-moderated interview were substantially longer than those in the static survey. On average, AIMI participants wrote about 131 words, whereas SoSci participants wrote about 94 words, meaning AIMI elicited **39% more words.** This difference was statistically significant ($p = .021$, $d = 0.331$), indicating that the conversational format encouraged participants to elaborate more on their thoughts.

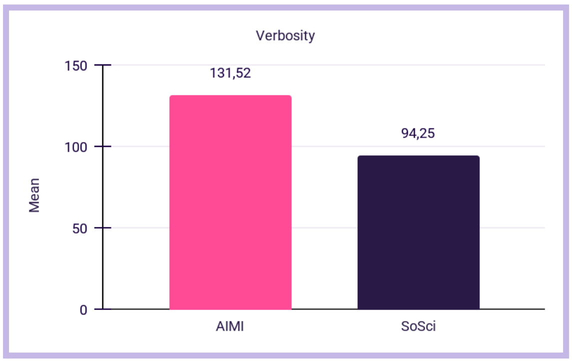

Figure 1

Unique Words. AIMI also produced far more unique words than SoSci (see figure 2). Participants in the AI condition used around 84 distinct words, compared to 55 in the survey condition, showing that the AI interview prompted **51% more unique words**. This difference was highly significant ($p < .001$, $d = 0.533$), suggesting richer vocabulary use and more varied expression in the AI-based format.

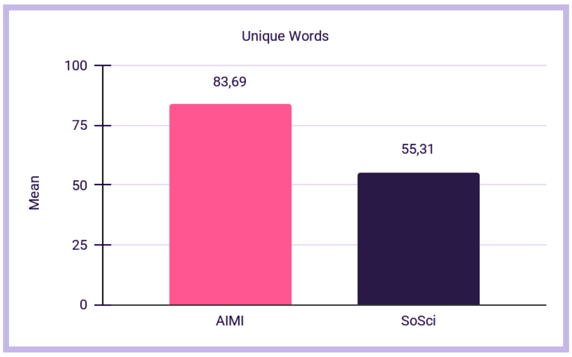

Figure 2

Lexical Diversity (TTR). Responses in the AI interview showed higher lexical diversity meaning participants used a more differentiated set of words rather than repeating the same ones. As shown in Figure 3, the AI condition reached a TTR of 0.704 on average, compared to 0.626 in SoSci, representing a **12% increase in lexical diversity**. This effect was highly significant ($p < .001$, $d = 0.562$), confirming that the AI interview stimulated more varied language production.

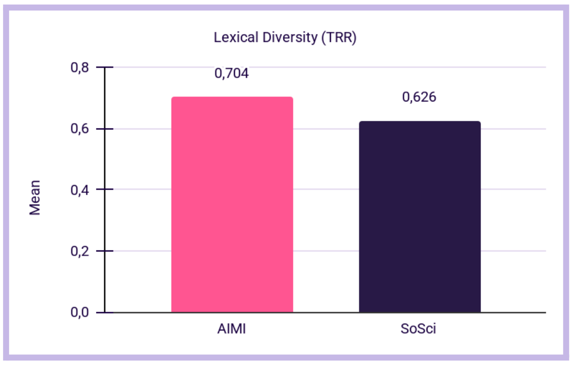

Figure 3

Content-Word Share (CWS) and **Flesch Reading Ease (FRE).** Both metrics were comparable across formats. Participants used similar proportions of content words in both conditions, with mean CWS scores of 0.62 in the AI interview and 0.64 in the static survey ($p = .376$, $r = -0.13$, ns). Readability also showed no meaningful differences, with FRE scores of 77.76 in the AI format and 79.60 in SoSci ($p = .434$, $r = -0.11$, ns). These results indicate that the AI interview produced longer and richer responses without reducing clarity or readability.

Overall, these metrics show that the AI-moderated interview encouraged participants to produce longer, richer, and more varied responses without reducing clarity or readability.

3.2.2 Thematic Variety

Thematic breadth was analysed to see how many different themes participants mentioned in their responses.

UNIVERSITÄT MANNHEIM

Theme Count. The total number of theme mentions was similar across conditions, with mean scores of 19.77 in the AI interview and 18.68 in the static survey ($p = .687$, $r = 0.03$, ns), meaning that both groups produced a comparable amount of content overall.

Unique Themes. Participants in the AI-moderated interview mentioned more distinct themes than those in the static survey (see Figure 4). AIMI responses included on average 8.76 themes, compared to 6.42 in SoSci, amounting to **36% more unique themes**. This difference was highly significant ($p < .001$, $r = 0.359$), indicating that the AI format stimulated broader and more diverse idea generation.

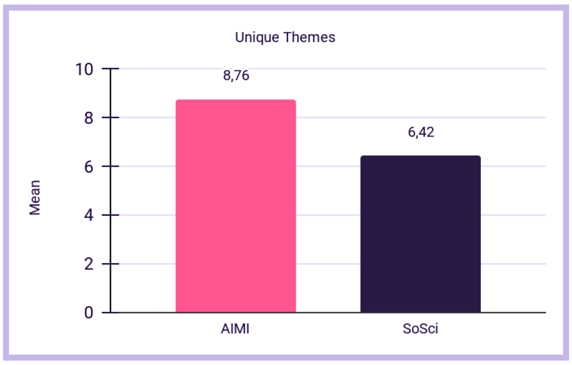

Figure 4

Together, these results show that the AI-moderated interview encouraged participants to express a wider range of ideas, resulting in more thematically diverse responses.

3.2.3 Response Validity

This subsection compares how often participants produced meaningless or nonsensical responses in each format, which serves as an indicator of data quality.

Gibberish responses. These occurred only in the static online survey. Table 1 indicates that the SoSci condition contained 10 invalid answers, corresponding to a 10% gibberish rate, while the AI interview produced **no gibberish entries at all.**

Group	n	Gibberish (n)	Rate (%)
AIMI	100	0	0.0%
SoSci	100	10	10.0%

Table 1

The findings indicate that the AI-moderated interview generated substantially more valid responses, whereas the static survey was more prone to meaningless input.

3.2.4 Participant Experience

This part examines participants' experience with each format, reported both as an aggregated overall score and as item-level evaluations of ease of expression, comfort, perceived repetitiveness, perceived conversational quality, data-handling trust, and willingness to recommend the format.

Aggregated Experience Score. Participants in the AI-moderated interview reported a more positive overall experience than those in the static online survey, as presented in Figure 5. Because the internal consistency of the seven items was acceptable (α ≥ .70), the items were aggregated into a single Participant Experience Score. On average, AIMI participants scored 4.22 versus 3.98 in SoSci, indicating a **6% better reported experience** in the AI format. This difference was statistically significant ($p = .02$, $d = 0.33$).

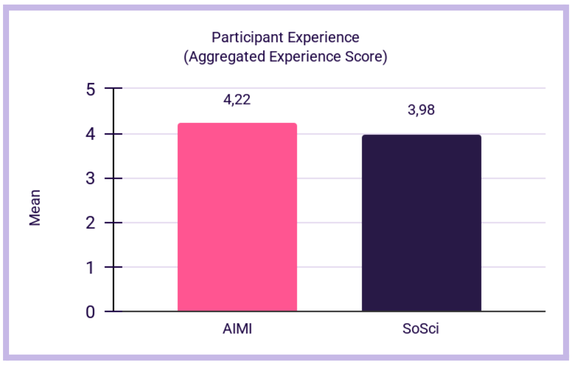

Figure 5

Item-Level Experience. Across the seven items, the AI moderated interview received more favorable evaluations on most dimensions. All values reported below refer to mean scores on the 5-point scale. **Ease of expression** and **comfort** were similar in both formats, with scores slightly above 4 in each group, indicating no meaningful differences (both ns). Clear differences emerged for the interaction-related items. Participants described the AI interview as more conversational, scoring 4.26 compared to 3.80, which corresponds to roughly **12% stronger conversational ratings** ($p = .001$, $d = 0.45$). They also felt more listened to and understood, with 4.38 versus 3.93, amounting to about **11% higher perceived understanding** in the AI condition ($p = .001$, $d = 0.47$). Trust in how data were handled was also evaluated more positively, with 4.45 compared to 3.95, which is around **13% better trust ratings** ($p < .001$, $d = 0.53$). Repetitiveness was significantly higher in the static survey, with mean scores of 3.97 compared to 3.30, meaning that the AI interview was perceived as about **17% less repetitive** ($p < .001$, $d = -0.52$). Participants were also more willing to recommend the AI interview, with scores of 4.41 compared to 3.88, a difference of about **14% in favor of the AI format** ($p < .001$, $d = 0.52$).

Taken together, the item-level results show a consistent pattern: the AI interview provided a more engaging, varied, and supportive experience across several key dimensions.

Overall, the results show that the AI-moderated interview consistently produced richer and more diverse responses, fewer invalid entries, and a more positive participant experience than the static online survey, indicating clear advantages of the conversational format across all major outcome dimensions.

4. Discussion

This section provides a critical evaluation of the study's findings and outlines the managerial implications for organizations.

4.1 Critical Evaluation

The following points reflect how the findings should be interpreted by considering alternative explanations and robustness.

Alternative Explanations. The effects favoring the AI interview are clear, but part of this advantage may stem from the interaction style rather than solely from participant motivation. While both formats included follow-up prompts, only the AI interviewer used adaptive, context-sensitive questions that can nudge participants toward elaboration or additional topics. Prior research shows that system behavior in AI-driven interactions can shape how people formulate responses and engage with a task (Benbasat & Wang, 2020). This suggests that the observed richness is driven by the AI's adaptive guidance rather than by differences in the participants themselves, indicating that the conversational dynamics of the AI actively encourage more elaborate responses.

Robustness of Findings. The effects remained stable across age and gender groups, which supports the internal validity of the results and suggests that the differences can be attributed to the data-collection format rather than demographic variation. In **Appendix D** you can find the full demographic tables.

Together, these points provide a broader perspective on how the results should be understood.

4.2 Managerial Implications

The findings offer several practical implications for organizations that rely on open-ended insights.

B2C and B2B applications. The AI-moderated interview generated longer responses, a higher number of unique words, and greater lexical diversity. These qualities benefit both consumer facing brands and B2B research contexts. For B2C, the added richness helps uncover deeper motivations, emotional drivers, and contextual factors behind customer attitudes. For B2B, the format supports projects with doctors, HR professionals, and employees, where more elaborate and reflective answers provide stronger insights than traditional surveys.

Market research agencies. The AI format revealed a broader range of themes, enabling agencies to capture multidimensional perspectives without human moderation. At the same time, the rate of meaningless or nonsensical responses was substantially lower, improving data quality and reducing the need for manual cleaning. This combination supports more efficient research operations and enhances the analytical value of open-ended data. While not replacing humans and qualitative research methodologies (e.g. focus groups), AI probing can improve scalability, agility and speed of qualitative studies and add depth to quantitative, long and complex surveys.

In-house insights and CX teams. Participants rated the AI interview as more engaging, less repetitive, and more trustworthy. A more positive experience can reduce survey fatigue and increase willingness to participate in recurring feedback cycles. Additionally, the absence of gibberish responses suggests that conversational formats yield more reliable and consistent qualitative input, which is critical for continuous listening programs in HR or customer experience settings.

Overall, the results demonstrate that AI-moderated interviews are a promising tool for organizations seeking deeper and more scalable qualitative insights.

5. Conclusion

This study examined whether AI-moderated interviews using Glaut's AIMI platform elicit richer and higher-quality responses than static online surveys. Using identical open questions and balanced samples, the AI format produced richer and more diverse responses, fewer invalid entries, and a more positive participant experience. These findings should be viewed in light of the study's scope: the investigation focused on one topic, one population, and one AI-based interviewing implementation, without a human-led comparison. Future research should examine additional topics, more diverse populations and human-moderated interviewers.

References

Benbasat, I., & Wang, W. (2020). Trust in artificial intelligence: The role of system behavior in shaping human responses. MIS Quarterly, 44(4), 1325–1359. https://doi.org/10.25300/MISQ/2020/15398

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233. https://doi.org/10.1037/h0057532

Lankton, N. K., McKnight, D. H., & Thatcher, J. B. (2015). The role of trust in human-robot and human-computer interactions. MIS Quarterly, 39(2), 463–486. https://doi.org/10.25300/MISQ/2015/39.2.09

Lee, J. J., Kim, Y., & Sundar, S. S. (2020). Why A.I.? Effects of anthropomorphic cues and conversational relevance on responses to chatbots. Computers in Human Behavior, 113, 106508. https://doi.org/10.1016/j.chb.2020.10650

Appendix A Questionnaire

Table A1 Full questionnaire

A.1 Onboarding & Instructions

Welcome!
Thank you for participating in this study.
This survey is about healthy lifestyle choices and will take approximately 5–7 minutes.
There are no right or wrong answers — we are only interested in your personal views.
All responses are anonymous.

A.2 Main Questionnaire Items

Q1. Habits

Question: "Which healthy habits do you personally keep up every week?"
Response format: Open-ended text
Follow-up prompts (predefined): "Can you give me an example?" and "Can you tell me more about it?"

Q2. Personal Priorities

Question: "When you choose a 'healthier' option, what do you prioritize first?"
Response format: Open-ended text
Follow-up prompts: "Can you share an example?" and "Can you explain it in more detail?"

Q3. Difficulty

Question: "How difficult is it for you to maintain your healthy routine right now?"
Response scale: 1 = Not difficult, 2 = Slightly difficult, 3 = Moderately difficult, 4 = Quite difficult, 5 = Very difficult
Skip logic: If participants selected 1 = Not difficult, Q4 was automatically skipped.

Q4. Reasons Barriers

(shown only if Q3 > 1)
Question: "What are the main obstacles that prevent you from keeping a healthy routine?"
Response format: Open-ended text
Follow-up prompts: "Can you give me an example?" and "Can you describe that a bit further?"

Q5. Information Sources — Food

Question: "When you need advice on healthy eating, where do you usually go first?"
Response format: Open-ended text
Follow-up prompts: "Can you share an example?" and "Can you elaborate a bit more?"

Q6. Information Sources — Exercise

Question: “When you need advice on exercise or plans, where do you usually go first?”
Response format: Open-ended text
Follow-up prompts: “Can you give an example?” and “Can you describe that a bit further?”

Q7. Motivation

Question: “What is the main result you aim to achieve by living a healthy lifestyle?”
Response format: Open-ended text
Follow-up prompts: “Can you share an example?” and “Can you tell me more about it?”

Q8. Health-related Purchases (Multiple Choice)

Question: “Which of the following have you spent money on in the past 30 days to support your health?”
Options (multiple selections allowed): Gym membership / fitness class, Fitness app subscription / nutrition app subscription, Wearable device / premium plan, Supplements / vitamins, Telehealth session / mental health session, Healthy meal kit / ready-meal plan, None of these

Q9. Participant Experience (7-item Likert Scale)

Question: “At the end, we would like to understand what you think about this interview format.”
Items (1 = strongly disagree, 5 = strongly agree): It was easy to express my thoughts,I was comfortable sharing personal thoughts, I encountered repetitive questions, This felt like a conversation, I felt listened to and understood during this process, I trust this organization to use my answers appropriately, I would recommend this format to others.

A.3. Notes on Format Differences

The follow-up prompts listed above were used only in the static survey.
In the AIMI condition, follow-up questions were generated dynamically.
Item order and skip logic were identical across both conditions.

Appendix B Statistical Tables

Table B1 Linguistic Response Quality

Metric	n (AIMI)	Mean (AIMI)	SD (AIMI)	n (SoSci)	Mean (SoSci)	SD (SoSci)	Test	t	df	p	Sig.	d
Verbosity	100	131.52	136.09	100	94.25	82.87	Welch t-test	2.339	163.55	.021	*	0.331
Unique Words	100	83.69	63.20	100	55.31	41.00	Welch t-test	3.767	169.80	< .001	***	0.533
Lexical Diversity (TTR)	100	0.704	0.102	100	0.626	0.166	Welch t-test	3.974	164.34	< .001	***	0.562
Content-Word Share (CWS)	100	0.618	0.097	100	0.634	0.153	Welch t-test	−0.887	167.93	.376	ns	−0.125
Flesch Reading Ease (FRE)	100	77.76	12.19	100	79.60	20.10	Welch t-test	−0.785	163.15	.434	ns	−0.111

Table B2 Thematic Variety

Metric	n (AIMI)	Mean (AIMI)	SD (AIMI)	n (SoSci)	Mean (SoSci)	SD (SoSci)	Test	U	p	Sig	r
Unique Themes	100	8.76	2.83	100	6.42	3.35	Mann-Whitney	7079	< .001	***	0.359
Theme Count	100	19.77	7.85	100	18.68	8.95	Mann-Whitney	5165	.687	ns	0.029

Table B3 Response Validity

Group	n	Gibberish (n)	Rate (%)
AIMI	100	0	0.0%
SoSci	100	10	10.0%

Note: $\chi^2(1, N = 200) = 10.53$, $p = .001$, sig. ***, $V = 0.23$

Table B4 Participant Experience

Item	n (AIMI)	Mean (AIMI)	SD (AIMI)	n (SoSci)	Mean (SoSci)	SD (SoSci)	Test	t	df	p	Sig.	d
Aggregated Experience Score	100	4.22	0.64	100	3.98	0.80	Welch t-test	2.20	180.0	.02	*	0.33
Ease of expression	100	4.36	0.93	100	4.17	1.05	Welch t-test	1.35	194.8	ns	ns.	0.19
Comfort	100	4.36	0.85	100	4.16	1.04	Welch t-test	1.48	190.0	ns	ns.	0.21
Repetitiveness	100	3.30	1.43	100	3.97	1.16	Welch t-test	-3.64	190.1	< .001	***	-0.52
Conversational feel	100	4.26	0.94	100	3.80	1.10	Welch t-test	3.18	193.2	.001	**	0.45
Feeling understood	100	4.38	0.87	100	3.93	1.05	Welch t-test	3.30	191.8	.001	**	0.47
Trust	100	4.45	0.82	100	3.95	1.05	Welch t-test	3.73	186.6	< .001	***	0.53
Recommendation	100	4.41	0.83	100	3.88	1.17	Welch t-test	3.69	178.1	< .001	***	0.52

Appendix C Thematic Variety Codebook

Table C1 Thematic Variety Codebook

Code	Description
physical_activity	Sport and physical activity (walking, running, gym, yoga, biking, stairs, etc.).
health_nutrition	Healthy eating and nutrient-related aspects (diet patterns, healthy food, balanced meals, fruits/vegetables, whole foods, calories, sugar, fat, protein, vitamins, macros, nutrition facts).
hygiene_selfcare	Hygiene and self-care habits (shower, bath, washing face, brushing teeth, skincare, laundry).
sleep_rest	Sleep and rest (sleep, going to bed, rest, naps).
medical_care	Medical care and health conditions (doctor, medicine, check-ups, cancer, allergies, blood pressure).
substance_use	Smoking, alcohol, and other substances.
taste_enjoyment	Taste and enjoyment (taste, delicious, yummy, flavor, favourite).
price_cost	Price and cost considerations (cheap, expensive, budget, money, cost).
convenience_time	Convenience and time (quick, easy, fast, convenient, ready-made, frozen, busy, no time).
quality_freshness	Freshness and quality of food (fresh, organic, quality).
dietary_restrictions	Dietary restrictions and conditions (allergy, gluten, lactose, diabetic, vegan, vegetarian, celiac).
health_goals	Health-related goals (be/stay healthy, improve or maintain health).
appearance_goals	Appearance and body-shape goals (look good, in shape, lose weight, abs, fit body).
time_stress_schedule	Time and stress barriers (busy, no time, stressful schedule, tired after work).
money_barrier	Financial barriers (too expensive, cannot afford, no money).
physical_limitations	Physical limitations and pain (injury, pain, sickness, disability).

cravings_habits	Cravings and habitual preferences (cravings, sweets, snacks, junk food, fast food).
motivation_barrier	Lack of motivation (lazy, no motivation, don’t feel like it, hard to maintain).
social_environment_barrier	Social environment barriers (friends, family, partner making it harder).
info_professional	Information from professionals (doctor, nutritionist, trainer, coach, therapist).
info_family_friends	Information from family and friends (relatives, partner, peers).
info_social_media	Information from social media and influencers (TikTok, Instagram, YouTube, podcasts).
info_online_search	Information from online search and websites (Google, internet, websites, blogs, apps).
info_traditional_media	Information from traditional media (TV, books, magazines, newspapers, radio, articles).
info_none_na	No information sources (none, no source, does not look it up).
intrinsic_wellbeing	Intrinsic wellbeing and feeling (feeling better, energy, less stress, mood, mental health, happiness).
longevity_ageing	Longevity and ageing motives (live longer, long life, live up forever, to live to old age).
family_responsibility	Responsibility for family (for my kids, children, family, to be there for them).
self_image_identity	Self-image and identity (better version of myself, pride, confidence, self-esteem).
other_meaningful	Meaningful content that does not clearly fit any of the specific categories above.
no_content	No meaningful content (empty, 'no', 'none', 'idk', thanks, etc.).

Appendix D Robustness Checks

Table D1 Demographic Balance

Variable	Test	df	p	Sig.
Gender	Chi-square	1	1	ns
Age	Chi-square	2	1	ns

Table D2 Group Sizes by Gender

Gender	AIMI	SoSci
Female	52	52
Male	48	48

Table D3 Group Sizes by Age

Age Group	AIMI	SoSci
18–29	34	34
30–44	33	33
45–55	23	23

Table D4 Robustness Model (Regression model: Metric ~ Group + Age + Gender)

Metric	**Group Effect (SoSci → AIMI)**	**p**	**Sig.**
Verbosity	-37.27	0.02	*
Unique Words	-28.38	0.0002	***
Lexical Diversity (TTR)	-0.077	0.00009	***
Content Word Share (CWS)	0.02	0.379	ns
Flesch Reading Ease (FRE)	1.84	0.437	ns

All effects remain directionally consistent when controlling for age and gender, indicating that the observed advantages of the AIMI format are robust across demographic subgroups.