

Survey Quality Audit

Aktywność fizyczna i samopoczucie studentów

2026-10-11 16:14

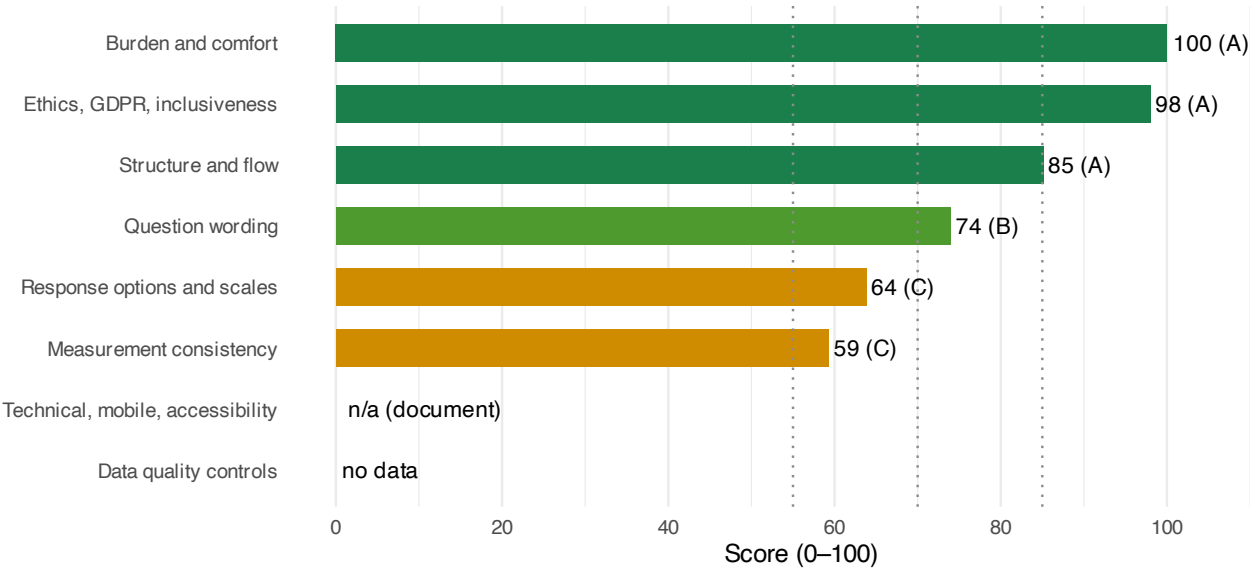
Contents

1 Summary	1
1.1 Most important fixes	3
2 How this audit was made	4
3 Survey map	5
4 Questions: problems and proposed fixes	6
5 Response options and scales	10
6 Data quality controls	11
7 Respondent burden and comfort	12
8 Measurement consistency and reliability	13
9 Structure and flow	13
10 Technical quality, mobile and accessibility	14
11 Ethics, GDPR and inclusiveness	14
12 Readiness for data analysis	15
13 Appendices	16
13.1 Full list of problems	16
13.2 Candidates rejected by the AI panel	17
13.3 Glossary of problem codes	17
13.4 Audit parameters	19

1 Summary

Survey quality index: 80/100 — grade B

good — minor fixes needed



Item	Value
Survey	Aktywność fizyczna i samopoczucie studentów
Document	Aktywność fizyczna i samopoczucie studentów.docx (DOCX)
Audit date	2026-10-11 16:14
Audit mode	quick
Data sources	The questionnaire was read from the document 'Aktywność fizyczna i samopoczucie studentów.docx' (DOCX): the model extracted 10 questions, and every question, option and grid row was checked against the text of the document.
Parts with questions	3
Questions	10
Items (incl. grid rows)	14
Estimated completion time	≈ 2.1 min
Required questions	not marked in the document
Problems found	29 (critical: 0, major: 13, minor: 16)
AI model	qwen3.6:35b-mlx

The questionnaire exhibits significant measurement validity threats, primarily driven by ambiguous wording and flawed response structures in Questions 1, 6, and 9. Question 1 imposes a specific World Health Organization threshold that may not align with the respondents’ personal definitions of physical activity, creating a filter bias that excludes valid participants who engage in shorter, frequent bouts of exercise. Question 6 suffers from a double-barreled structure that conflates fatigue with low energy, making it impossible to determine which construct drives the respondent’s answer, while also relying on the vague quantifier ‘little’ which lacks a common reference point. Question 9 fails to specify units for the birth year input, risking data entry errors or confusion regarding whether the respondent should enter the full year or an age-derived value. Additionally, Question 2 demands precise recall of specific days within the last week, a task known to induce satisficing and recall error, while Question 4 presents a heterogeneous battery of constructs that complicates the interpretation of any aggregated scores. These issues collectively threaten the reliability of the data and the validity of any subsequent correlations between physical activity and well-being.

Strengths

- The survey structure is logical and flows well, with a clear separation between behavioral questions, psychometric scales, and demographic data.
- Ethical considerations and GDPR compliance are robustly addressed, ensuring high inclusiveness and respect for participant privacy.
- The estimated completion time of two minutes is appropriate for the number of items, minimizing respondent burden and fatigue.
- The use of a grid format for Question 4 is efficient for collecting multiple related statements, provided the items are clarified.

1.1 Most important fixes

No.	Severity	Question	Problem	What to do
1	major	Q06	Double-barreled question	'zmęczony lub miałeś mało' — one answer covers two things; it is unclear which one it refers to. Fix: Split into two questions (statements) or keep one.
2	major	Q09	No unit or range	No unit — unclear whether to give hours, minutes, items, etc. Fix: State the unit and allowed range.
3	major	Q06	Vague terms or quantifiers	Vague terms: mało — they mean different things to different people. Fix: Replace with a concrete number, frequency or period.
4	major	Q02	Demanding recall or estimation	7. Requires precise memory and counting of days from the previous week, which is difficult and subject to recall error (Tourangeau). Respondents often guess or use heuristics. Fix: 8. Consider using a calendar or a list of days of the week to mark, to facilitate recall, or change the question to rating intervals (e.g., 0, 1-2, 3-4, 5-7 days).
5	major	Q04	Negation or double negative	11. The statement is negative and substantively reversed relative to the others (positive). In a Likert scale, the respondent must perform an additional cognitive operation (negation), which increases the risk of error and straightlining. Fix: 12. Change the statement to positive: 'Physical activity is valuable/important to me' or reverse the scale for this row (if the platform allows), but it is better to avoid negation in statements.
6	major	Q10	Gaps, non-exhaustive options	There is no option for non-full-time, evening, or those who have completed a bachelor's/engineering degree but continue studying in another form, which prevents them from providing a correct answer. Fix: Add an 'Other (please specify)' option or expand the list to include 'non-full-time/evening studies' and 'after completing a bachelor's/engineering degree'.
7	major	Q01	Vague terms or quantifiers	1. The phrasing 'at least 30 minutes in total per day' is misleading: it suggests that 30 minutes is the threshold required to classify activity as 'yes', whereas in epidemiological studies, total weekly time or days with activity >30 min are often counted. The respondent may not know whether to evaluate each day individually or the entire week. Fix: 2. Change the question to: 'Did you engage in physical activity lasting at least 30 minutes in at least one day during the last 7 days?' or clarify whether it refers to total time or days with activity.
8	major	Q04	Heterogeneous battery	15. The rows measure different constructs: pleasure (affect), well-being after exercise (physiological/psychological effect), perceived value of time (attitude), time availability (barrier), motivation (goal). Summing them into a single scale would be unjustified. Fix: 16. Treat these questions as separate indicators or subscales, not as a single 'attitude towards activity' scale. In analysis, do not create a total score without verifying unidimensionality.
9	major	Q01	Missing filter	3. The filtering question assumes that 'physical activity' means 30 min per day, which is a specific WHO definition, not a common one. Fix: 4. Add the option 'Don't know' or clarify the definition of activity to avoid incorrect 'No' responses.
10	major	Q01	Social desirability pressure	5. The question about physical activity is strongly subject to social desirability bias; the 'Yes' response is socially desirable, which can lead to underreporting of inactivity. Fix: 6. Apply social desirability neutralization, e.g., by adding information that 'many people have different habits' or using the 'randomized response' technique if possible, and in an online survey, ensure anonymity and lack of judgment.

Next steps

- Rewrite Question 6 to separate the concepts of 'feeling tired' and 'having little energy' into two distinct items to eliminate the double-barreled ambiguity.
- Revise Question 1 to remove the rigid 30-minute threshold from the filter question, allowing for a broader definition of physical activity, or clearly justify why this specific WHO standard is the only valid metric for this study.
- Add explicit units or instructions to Question 9, such as 'Enter the full year (e.g., 1995)', to prevent input errors.
- Simplify Question 2 by asking for total minutes per week or frequency categories rather than requiring precise daily recall, which is prone to estimation error.
- Pretest the heterogeneous battery in Question 4 with a small sample to verify that respondents interpret the items as distinct constructs rather than a single general attitude.
- Review Question 10 to ensure the options for 'year of study' are exhaustive and inclusive of non-traditional student statuses, such as part-time or continuing education.

Dimension	Weight	Score	Grade	Crit.	Major	Minor	Basis
Question wording	20	74	B	0	6	0	10 questions
Response options and scales	15	64	C	0	4	10	10 questions
Data quality controls	10	—	—	0	0	0	too few criteria

(cont.)

Dimension	Weight	Score	Grade	Crit.	Major	Minor	Basis
Burden and comfort	15	100	A	0	0	2	55/55 pts
Measurement consistency	10	59	C	0	2	1	40/60 pts
Structure and flow	10	85	A	0	1	2	92/100 pts
Technical, mobile, accessibility	10	—	—	0	0	0	n/a (document)
Ethics, GDPR, inclusiveness	10	98	A	0	0	1	100/100 pts

2 How this audit was made

The audit combines four independent sources of information; every finding has a severity (critical, major, minor), a confidence (0–1) and its sources.

- 1. Reading the questionnaire from the document.** The questionnaire was read from the document ‘Aktywność fizyczna i samopoczucie studentów.docx’ (DOCX): the model extracted 10 questions, and every question, option and grid row was checked against the text of the document. The text (paragraphs, automatic numbering, tables, answer boxes) was read straight from the file; the model identified the questions, their types, options, grid rows and columns, instructions, filters and required marks. Questions, options and sentences that are not in the document were dropped, and the introduction comes straight from the document.
- 2. Automatic rules.** Language patterns (negations, vague quantifiers, missing reference periods, leading wording, abbreviations, hypotheticals), answer-option analysis (overlapping/missing ranges, scale balance, ‘don’t know’/‘prefer not to say’, ‘other’), burden metrics (estimated time, grids, open questions, readability), control items, consent and GDPR elements.
- 3. AI expert panel** (model qwen3.6:35b-mlx): survey methodologist and psychometrician, respondent advocate (cognitive interviewing, plain language). Each role reviews all questions independently; findings of the same problem from several sources are merged and confidence grows with agreement ($1 - \prod(1 - p_i)$). Rule candidates the panel did not confirm are listed in the appendix. The panel also reviews the survey as a whole and proposes a new wording for every flawed question.
- 4. Technical tests.** An audit of a document does not cover accessibility (WCAG), the phone version, validation messages, page breaks or the progress bar — they depend on the platform the survey will be set up on. Once it is set up, audit the survey link as well.

Scoring. Question-level dimensions (wording, options): every question starts at 100% and loses 100% (critical), 60% (major) or 20% (minor) for its worst problem plus 10% for each further one (max. 30%), weighted by confidence (confidence ≥ 0.7 : $\times 1$; $0.45-0.7$: $\times 0.7$; < 0.45 : $\times 0.4$); the dimension score is the mean over questions minus survey-level problems (15/6/2 points). Checklist dimensions (ethics, flow, controls, measurement, burden): share of points for criteria met (partly met = half). Accessibility: 100 minus penalties for each axe-core rule (12/8/4/2 by impact) and mobile problems. One confident critical problem caps a score at 84 (no A), three or more at 69. The overall index is the weighted mean of the available dimensions.

Grades: A ≥ 85 (very good), B 70–84 (good, minor fixes), C 55–69 (fair, needs revision), D < 55 (poor, needs major rework).

Reading the document. The model may have missed a question or misread its type (e.g. one answer instead of several); check the survey map below. Required questions are not marked in the document, so forced answering is not assessed.

Limitations. AI and rule-based judgements are a preliminary expert review, not a replacement for a pilot with people: the model can miss problems or raise false alarms (hence the confidence and sources). Synthetic respondents answer more consistently than humans. axe-core detects roughly 30–40% of accessibility problems; a full assessment needs a screen-reader test. The estimated time comes from a reading-and-answering pace model (about 240 words per minute).

Dimension	Weight	Comment
Question wording	20	Whether questions are clear, neutral, unambiguous and ask about one thing at a time.
Response options and scales	15	Whether answer options are exclusive, exhaustive, balanced and fit the questions.
Data quality controls	10	Attention checks, reverse-keyed items, consistency pairs, seriousness check.
Burden and comfort	15	Length, grids, open questions, forced answers, readability, progress.
Measurement consistency	10	Multi-item, homogeneous scales and the expected consistency of answers.

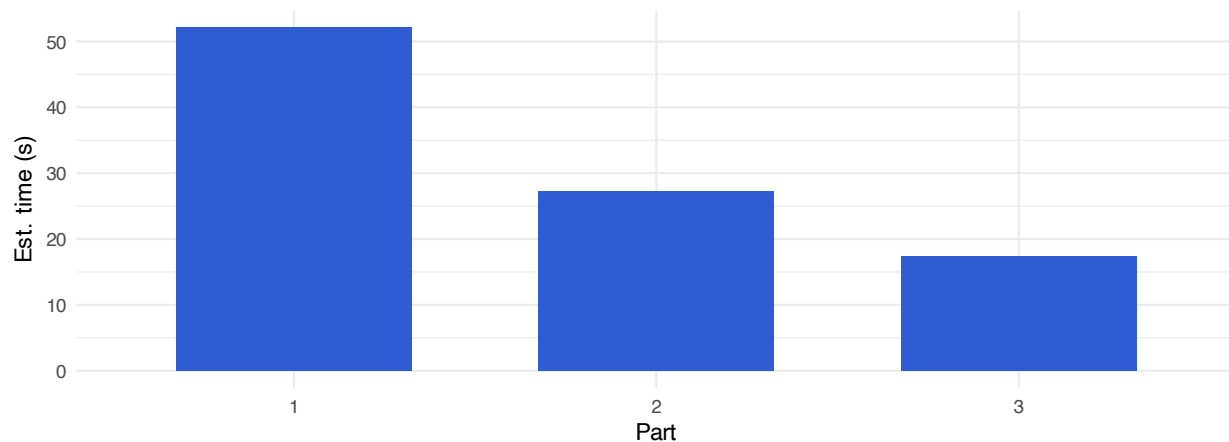
(cont.)

Dimension	Weight	Comment
Structure and flow	10	Introduction, question order, filters, terminology, closing.
Technical, mobile, accessibility	10	WCAG violations (axe-core), mobile usability, technical failures.
Ethics, GDPR, inclusiveness	10	Informed consent, data-protection notice, sensitive data, inclusive language.

Severity	Penalty (question)	Penalty (survey)	Comment
critical	100%	15	invalidates the data, makes answering impossible or breaches ethics
major	60%	6	likely to bias answers or confuse many respondents
minor	20%	2	style, small unclarity, convenience

3 Survey map

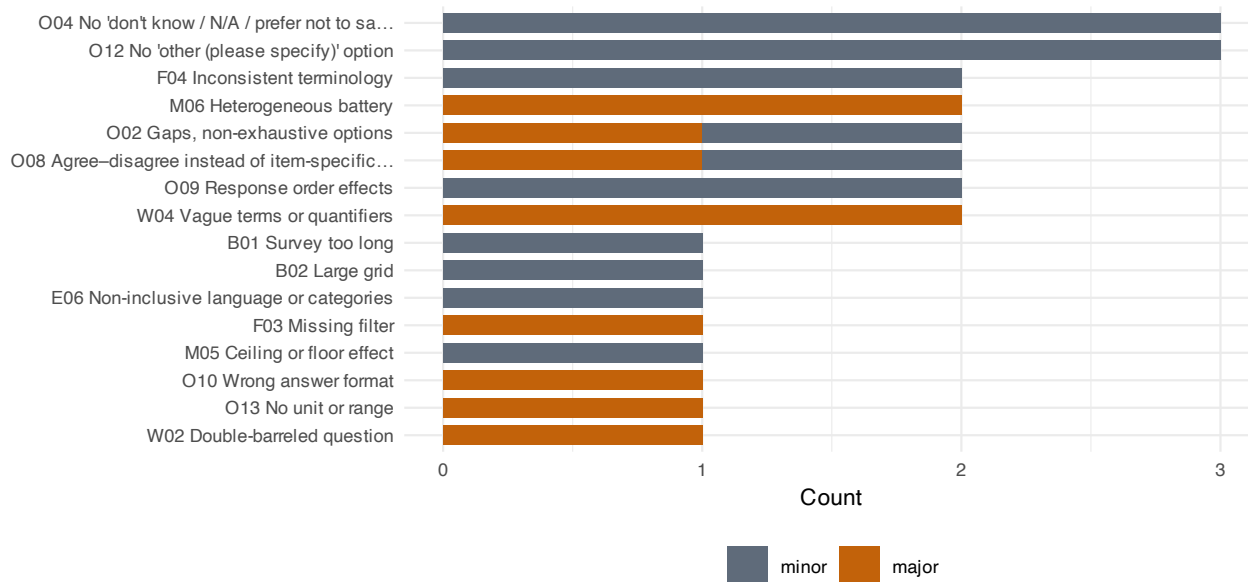
Part	Title in the document	Questions	Items	Est. time (s)
introduction		0	0	3
Part 1	Część A. Aktywność fizyczna	4	8	52
Part 2	Część B. Samopoczucie	3	3	27
Part 3	Metryczka	3	3	17
closing		0	0	3



ID (no.)	part	Type	Question	Opt.	Req.	quality
Q01 (A1)	1	single choice	Czy w ciągu ostatnich 7 dni podejmowałeś(-aś) aktywność fizyczną trwającą łącznie co najmniej ...	2		20
Q02 (A2)	1	number	W ile dniach w ciągu ostatnich 7 dni podejmowałeś(-aś) taką aktywność?	0		40
Q03 (A3)	1	multiple choice	Jakie formy aktywności podejmowałeś(-aś) w ciągu ostatnich 7 dni?	6		80
Q04 (A4)	1	grid	Na ile zgadzasz się z poniższymi stwierdzeniami?	5x5		10
Q05 (B1)	2	single choice	Jak oceniasz swoje ogólne samopoczucie w ciągu ostatnich 30 dni?	11		60
Q06 (B2)	2	single choice	Jak często w ciągu ostatnich 2 tygodni czułeś(-aś) się zmęczony(-a) lub miałeś(-aś) mało energ...	5		10
Q07 (B3)	2	ranking	Uszereguj poniższe powody podejmowania aktywności od najważniejszego (1) do najmniej ważnego (...)	4		20
Q08 (M1)	3	single choice	Płeć	4		80
Q09 (M2)	3	number	Rok urodzenia:	0		30
Q10 (M3)	3	single choice	Rok studiów	6		20

In brackets: the question number in the document (where it differs from the id). 'Req.' marks the questions the document labels as required.

4 Questions: problems and proposed fixes



Q01 (no. in the document A1)

Czy w ciągu ostatnich 7 dni podejmowałeś(-aś) aktywność fizyczną trwającą łącznie co najmniej 30 minut dziennie?

part 1 · single choice · quality 20/100

Filter / routing: Nie -> przejdź do pytania A4

Options: Tak / Nie

- **[major] Vague terms or quantifiers** (W04; confidence 0.94; AI: methodologist + AI: respondent advocate). 1. The phrasing 'at least 30 minutes in total per day' is misleading: it suggests that 30 minutes is the threshold required to classify activity as 'yes', whereas in epidemiological studies, total weekly time or days with activity >30 min are often counted. The respondent may not know whether to evaluate each day individually or the entire week. *Evidence:* trwającą łącznie co najmniej 30 minut dziennie *Fix:* 2. Change the question to: 'Did you engage in physical activity lasting at least 30 minutes in at least one day during the last 7 days?' or clarify whether it refers to total time or days with activity.
- **[major] Missing filter** (F03; confidence 0.75; AI: whole-survey review). 3. The filtering question assumes that 'physical activity' means 30 min per day, which is a specific WHO definition, not a common one. *Evidence:* Q01 ... trwającą łącznie co najmniej 30 minut dziennie? *Fix:* 4. Add the option 'Don't know' or clarify the definition of activity to avoid incorrect 'No' responses.
- **[major] Social desirability pressure** (W09; confidence 0.75; AI: methodologist). 5. The question about physical activity is strongly subject to social desirability bias; the 'Yes' response is socially desirable, which can lead to underreporting of inactivity. *Evidence:* Czy w ciągu ostatnich 7 dni podejmowałeś(-aś) aktywność fizyczną *Fix:* 6. Apply social desirability neutralization, e.g., by adding information that 'many people have different habits' or using the 'randomized response' technique if possible, and in an online survey, ensure anonymity and lack of judgment.

Proposal: Czy w ciągu ostatnich 7 dni wykonywałeś(-aś) aktywność fizyczną, która trwała łącznie co najmniej 30 minut w jednym dniu?

Rationale: The confusing term 'daily' referring to the entire week has been removed, replaced by a precise reference to 'one day', which facilitates assessment; the question remains closed to minimize social pressure through a clear 'yes/no' criterion based on a specific event.

Q02 (no. in the document A2)

W ile dniach w ciągu ostatnich 7 dni podejmowałeś(-aś) taką aktywność?

part 1 · number · quality 40/100

- **[major] Demanding recall or estimation** (W07; confidence 0.97; AI: methodologist + AI: respondent advocate). 7. Requires precise memory and counting of days from the previous week, which is difficult and subject to recall error (Tourangeau). Respondents often guess or use heuristics. *Evidence:* W ile dniach w ciągu ostatnich 7 dni podejmowałeś(-aś) taką aktywność? *Fix:* 8. Consider using a calendar or a list of days of the week to mark, to facilitate recall, or change the question to rating intervals (e.g., 0, 1-2, 3-4, 5-7 days).

Proposal: W ilu dniach w ciągu ostatnich 7 dni wykonywałeś(-aś) aktywność fizyczną trwającą co najmniej 30 minut?

Rationale: The word ‘such’ referring to the unclear ‘daily’ from the previous question has been removed, replaced by a repetition of the clear criterion ‘lasting at least 30 minutes’, which reduces memory burden and the risk of recall error by referring to a specific, easier-to-remember event.

Q03 (no. in the document A3)

Jakie formy aktywności podejmowałeś(-aś) w ciągu ostatnich 7 dni?

part 1 · multiple choice · quality 80/100

Options: Bieganie lub nordic walking / Ćwiczenia na siłowni / Pływanie / Jazda na rowerze / Zajęcia fitness lub taniec / Inne (jakie?)

- **[minor] No ‘other (please specify)’ option** (O12; confidence 0.96; AI: methodologist + AI: respondent advocate). 9. The ‘Other’ option exists, but the lack of a text field to enter answers prevents collecting data on unforeseen forms of activity, limiting data completeness. *Evidence:* Inne (jakie?) *Fix:* 10. Add a text field next to the ‘Other’ option so the respondent can enter the specific form of activity.

Proposal: Jakie formy aktywności fizycznej podejmowałeś(-aś) w ciągu ostatnich 7 dni? (Można zaznaczyć wiele odpowiedzi)

New options: Bieganie lub nordic walking / Ćwiczenia na siłowni / Pływanie / Jazda na rowerze / Zajęcia fitness lub taniec / Inne (proszę wpisać)

Rationale: A text field has been added to the ‘Other’ option to allow respondents to provide unforeseen forms of activity, which increases data completeness and reliability.

Q04 (no. in the document A4)

Na ile zgadzasz się z poniższymi stwierdzeniami?

part 1 · grid · quality 10/100

Options: Zdecydowanie się nie zgadzam / Raczej się nie zgadzam / Ani się zgadzam, ani nie zgadzam / Raczej się zgadzam / Zdecydowanie się zgadzam

Rows: (1) Ruch sprawia mi przyjemność.; (2) Po ćwiczeniach czuję się lepiej.; (3) Aktywność fizyczna to dla mnie strata czasu.; (4) Łatwo znajduję czas na ruch w ciągu tygodnia.; (5) Ćwiczę, bo chcę dbać o zdrowie.

- **[major] Negation or double negative** (W03, row 3; confidence 0.97; AI: methodologist + AI: respondent advocate). 11. The statement is negative and substantively reversed relative to the others (positive). In a Likert scale, the respondent must perform an additional cognitive operation (negation), which increases the risk of error and straightlining. *Evidence:* Aktywność fizyczna to dla mnie strata czasu. *Fix:* 12. Change the statement to positive: ‘Physical activity is valuable/important to me’ or reverse the scale for this row (if the platform allows), but it is better to avoid negation in statements.
- **[major] Agree–disagree instead of item-specific scale** (O08, row 3; confidence 0.75; AI: methodologist). 13. Using a Likert scale for a negative statement (‘waste of time’) is less precise than a ‘how true is this’ or ‘how much does this describe’ scale. The Likert scale is susceptible to acquiescence bias. *Evidence:* Zdecydowanie się nie zgadzam / Raczej się nie zgadzam... *Fix:* 14. Change the scale to ‘To what extent does this statement describe your feelings?’ or ‘How true is this?’ for all rows, to standardize the format and reduce methodological error.
- **[major] Heterogeneous battery** (M06; confidence 0.73; AI: whole-survey review + AI: methodologist). 15. The rows measure different constructs: pleasure (affect), well-being after exercise (physiological/psychological effect), perceived value of time (attitude), time availability (barrier), motivation (goal). Summing them into a single

scale would be unjustified. *Evidence:* Grid z 5 wierszami o różnych aspektach aktywności *Fix:* 16. Treat these questions as separate indicators or subscales, not as a single ‘attitude towards activity’ scale. In analysis, do not create a total score without verifying unidimensionality.

- **[minor] Large grid** (B02; confidence 0.56; AI: whole-survey review). 17. A short battery of 5 rows in grid form is acceptable, but on a phone, it may be difficult to use. *Evidence:* Q04 [grid (one answer per row)] ... 5 wierszy *Fix:* 18. On mobile devices, display questions as single sequences instead of a grid.

Proposal: Na ile poniższe stwierdzenia opisują Twoją sytuację?

New options: Wcale nie opisują / Raczej nie opisują / Trudno powiedzieć / Ani tak, ani nie / Raczej opisują / W pełni opisują

New rows: (1) Ruch sprawia mi przyjemność.; (2) Po ćwiczeniach czuję się lepiej.; (3) Aktywność fizyczna to dla mnie strata czasu.; (4) Łatwo znajduję czas na ruch w ciągu tygodnia.; (5) Ćwiczę, bo chcę dbać o zdrowie.

Rationale: The scale has been changed from ‘agree/disagree’ to ‘describe the situation’, which reduces the risk of acquiescence and facilitates cognitive processing, especially for negative statements that no longer require an additional negation operation in the respondent’s mind.

Q05 (no. in the document B1)

Jak oceniasz swoje ogólne samopoczucie w ciągu ostatnich 30 dni?

part 2 · single choice · quality 60/100

Options: 0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10

- **[minor] No ‘don’t know / N/A / prefer not to say’** (O04; confidence 0.97; AI: methodologist + AI: respondent advocate + rule). An opinion question without ‘don’t know / no opinion’. *Evidence:* Zaznacz jedną liczbę na skali od 0 do 10. *Fix:* Add ‘Don’t know’ as an off-scale option at the end.
- **[minor] Response order effects** (O09; confidence 0.93; AI: methodologist + AI: respondent advocate + rule). Long list (11 options) — the first ones will be chosen more often. *Evidence:* 0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10 *Fix:* Randomise the option order (keeping ‘Other’ last) or group the list.
- **[minor] Ceiling or floor effect** (M05; confidence 0.60; AI: whole-survey review). 19. A 0-10 scale may lead to a ceiling effect among students if most are satisfied. *Evidence:* Q05 Jak oceniasz swoje ogólne samopoczucie ... 0 – bardzo złe 10 – bardzo dobre *Fix:* 20. Ensure the scale is well-interpreted; alternatively, consider a 5-point scale for better distribution.

Proposal: Jak oceniasz swoje ogólne samopoczucie w ciągu ostatnich 30 dni?

New options: 0 - Bardzo złe / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10 - Bardzo dobre / Nie dotyczy / Wolę nie odpowiadać

Rationale: The ‘Not applicable / Prefer not to answer’ option has been added to avoid forcing respondents to randomly select values, which would introduce noise into the data, and ensured that scale labels are visible for all options.

Q06 (no. in the document B2)

Jak często w ciągu ostatnich 2 tygodni czuleś(-aś) się zmęczony(-a) lub miałeś(-aś) mało energii?

part 2 · single choice · quality 10/100

Options: Wcale / Kilka dni / Więcej niż połowę dni / Prawie codziennie / Wolę nie odpowiadać

- **[major] Double-barreled question** (W02; confidence 0.97; AI: methodologist + AI: respondent advocate + rule). ‘zmęczony lub miałeś mało’ — one answer covers two things; it is unclear which one it refers to. *Evidence:* czuleś(-aś) się zmęczony(-a) lub miałeś(-aś) mało energii? *Fix:* Split into two questions (statements) or keep one.
- **[major] Vague terms or quantifiers** (W04; confidence 0.97; AI: methodologist + AI: respondent advocate + rule). Vague terms: mało — they mean different things to different people. *Evidence:* mało energii *Fix:* Replace with a concrete number, frequency or period.
- **[minor] No ‘don’t know / N/A / prefer not to say’** (O04; confidence 0.60; AI: respondent advocate). 21. The presence of this option is good, but it is worth ensuring it is visible and understandable. *Evidence:* Wolę nie odpowiadać *Fix:* 22. Ensure this option is clearly marked and available to all respondents.

- **[minor] No ‘other (please specify)’ option** (O12; confidence 0.50; rule). The list has no ‘Other (please specify)’ option — it may not cover all cases. *Fix:* Add ‘Other (please specify)’ with a text field.

Proposal: Jak często w ciągu ostatnich 2 tygodni odczuwałeś(-aś) zmęczenie?

New options: Wcale / W kilka dni / W więcej niż połowie dni / Prawie codziennie / Wolę nie odpowiadać

Rationale: The two-element question has been split into ‘fatigue’, removing the imprecise ‘low energy’, which increases measurement precision and reduces answer ambiguity, while retaining the ‘Prefer not to answer’ option for the label.

Q07 (no. in the document B3)

Uszereguj poniższe powody podejmowania aktywności od najważniejszego (1) do najmniej ważnego (4).

part 2 · ranking · quality 20/100

Options: zdrowie / wygląd / kontakty towarzyskie / odstresowanie się

- **[major] Wrong answer format** (O10; confidence 0.75; AI: methodologist). 23. Ranking forces an arbitrary choice, even when reasons are equivalent. This cognitively burdens the respondent and may not reflect real motivations (e.g., two main reasons are equally important). *Evidence:* Uszereguj poniższe powody... od najważniejszego (1) do najmniej ważnego (4). *Fix:* 24. Change to ‘select up to 2 most important reasons’ or ask to rate each individually (e.g., 1-5 scale), which is less burdensome and more reliable.
- **[minor] Response order effects** (O09; confidence 0.75; AI: methodologist). 25. Lack of rotation of the order of reasons can lead to position effects (first options are more often chosen as ‘most important’ in rankings). *Evidence:* zdrowie / wygląd / kontakty towarzyskie / odstresowanie się *Fix:* Implement random ordering (randomization) of options in the ranking for different respondents.
- **[minor] Gaps, non-exhaustive options** (O02; confidence 0.60; AI: respondent advocate). The list may not include all important reasons, e.g., ‘relaxation’ or ‘mental health’. *Evidence:* zdrowie / wygląd / kontakty towarzyskie / odstresowanie się *Fix:* Consider adding an ‘Other’ option or ensure that the list is exhaustive.

Proposal: Proszę ocenić, jak bardzo każdy z poniższych powodów motywuje Cię do podejmowania aktywności fizycznej, wybierając jedną opcję dla każdego wiersza.

New options: Bardzo mało / Mało / Średnio / Dużo / Bardzo dużo

New rows: (1) Zdrowie; (2) Wygląd; (3) Kontakty towarzyskie; (4) Odstresowanie się

Rationale: The format has been changed from ranking to rating the intensity of each reason separately (Likert), which eliminates forced arbitrary choices and the cognitive burden associated with comparing reasons, allowing for an equivalent assessment of motivation.

Q08 (no. in the document M1)

Płeć

part 3 · single choice · quality 80/100

Options: Kobieta / Mężczyzna / Inna / Wolę nie odpowiadać

- **[minor] Non-inclusive language or categories** (E06; confidence 0.96; AI: whole-survey review + AI: methodologist + AI: respondent advocate). The ‘Other’ and ‘Prefer not to answer’ options improve inclusion, which is positive, but ‘Other’ is imprecise. *Evidence:* Q08 Płeć ... Inna / Wolę nie odpowiadać *Fix:* Consider adding an ‘Other (please specify)’ field for the ‘Other’ option if the number of responses is significant.

Proposal: Płeć

New options: Kobieta / Mężczyzna / Inna / Nie chcę podać

Rationale: Changed ‘Prefer not to answer’ to ‘Do not wish to provide’, which is a standard, neutral form in Polish demographic surveys, and retained ‘Other’ as an imprecise but necessary option for inclusion, without adding additional subcategories that could be misleading.

Q09 (no. in the document M2)

Rok urodzenia:

part 3 · number · quality 30/100

- **[major] No unit or range** (O13; confidence 0.97; AI: methodologist + AI: respondent advocate + rule). No unit — unclear whether to give hours, minutes, items, etc. *Evidence:* Rok urodzenia: *Fix:* State the unit and allowed range.
- **[minor] Inconsistent terminology** (F04; confidence 0.60; AI: whole-survey review). The question about year of birth is precise, but in the context of students, it may be less useful than 'Year of Study'. *Evidence:* Q09 Rok urodzenia: *Fix:* Ensure that the variable 'Year of Study' (Q10) adequately covers respondents' ages.

Proposal: W którym roku urodziłeś(-aś)?

Rationale: Added the word 'year' to the question to clearly indicate that the respondent should enter the year (e.g., 1990), not age, which prevents data coding errors.

Q10 (no. in the document M3)

Rok studiów

part 3 · single choice · quality 20/100

Options: I / II / III / IV / V / studia doktoranckie

- **[major] Gaps, non-exhaustive options** (O02; confidence 0.97; AI: methodologist + AI: respondent advocate). There is no option for non-full-time, evening, or those who have completed a bachelor's/engineering degree but continue studying in another form, which prevents them from providing a correct answer. *Evidence:* I / II / III / IV / V / studia doktoranckie *Fix:* Add an 'Other (please specify)' option or expand the list to include 'non-full-time/evening studies' and 'after completing a bachelor's/engineering degree'.
- **[minor] No 'don't know / N/A / prefer not to say'** (O04; confidence 0.96; AI: methodologist + AI: respondent advocate). Lack of 'Not applicable' or 'Prefer not to answer' options forces respondents to choose a mismatching category or skip the question, which distorts the data. *Evidence:* I / II / III / IV / V / studia doktoranckie *Fix:* Add a 'Not applicable' option (e.g., for non-students) or 'Prefer not to answer'.
- **[minor] No 'other (please specify)' option** (O12; confidence 0.50; rule). The list has no 'Other (please specify)' option — it may not cover all cases. *Fix:* Add 'Other (please specify)' with a text field.

Proposal: Jaki jest Twój aktualny rok studiów?

New options: I rok / II rok / III rok / IV rok / V rok / Studnia doktoranckie / Inne (proszę wskazać) / Nie dotyczy / Preferuję nie odpowiadać

Rationale: Expanded the list of options with missing categories (e.g., doctoral studies, other forms of study) and added 'Other', 'Not applicable', and 'Prefer not to answer' options to ensure completeness and avoid forcing respondents to choose a mismatching answer.

Remarks on all questions

- **[minor] Agree–disagree instead of item-specific scale** (O08). Agree–disagree format in 1 grid(s) (Q04). Item-specific scales are usually more reliable and less prone to acquiescence — unless the questionnaire is standardised.

5 Response options and scales

ID	Type	Pts	+0/–	Balanced	Off-scale	Direction	Options
Q01	nominal	2					Tak / Nie
Q03	nominal	6					Bieganie lub nordic walking / Ćwiczenia na siłowni / Pływanie / Jazda...
Q04	agreement	5	2/1/2	yes		neg-pos	Zdecydowanie się nie zgadzam / Raczej się nie zgadzam / Ani się zgadz...

(cont.)

ID	Type	Pts	+/-	Balanced	Off-scale	Direction	Options
Q05	numeric	11	0/0/0				0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10
Q06	frequency	4			Wolę nie odpowiadać		Wcale / Kilka dni / Więcej niż połowę dni / Prawie codziennie / Wolę ...
Q07	nominal	4					zdrowie / wygląd / kontakty towarzyskie / odstresowanie się
Q08	nominal	3			Wolę nie odpowiadać		Kobieta / Mężczyzna / Inna / Wolę nie odpowiadać
Q10	nominal	6					I / II / III / IV / V / studia doktoranckie

ID	Severity	Problem	Comment
Q03	minor	No 'other (please specify)' option	9. The 'Other' option exists, but the lack of a text field to enter answers prevents collecting data on unforeseen forms of activity, limiting data completeness.
Q04	major	Agree-disagree instead of item-specific scale	13. Using a Likert scale for a negative statement ('waste of time') is less precise than a 'how true is this' or 'how much does this describe' scale. The Likert scale is susceptible to acquiescence bias.
Q05	minor	No 'don't know / N/A / prefer not to say'	An opinion question without 'don't know / no opinion'.
Q05	minor	Response order effects	Long list (11 options) — the first ones will be chosen more often.
Q06	minor	No 'don't know / N/A / prefer not to say'	21. The presence of this option is good, but it is worth ensuring it is visible and understandable.
Q06	minor	No 'other (please specify)' option	The list has no 'Other (please specify)' option — it may not cover all cases.
Q07	major	Wrong answer format	23. Ranking forces an arbitrary choice, even when reasons are equivalent. This cognitively burdens the respondent and may not reflect real motivations (e.g., two main reasons are equally important).
Q07	minor	Response order effects	25. Lack of rotation of the order of reasons can lead to position effects (first options are more often chosen as 'most important' in rankings).
Q07	minor	Gaps, non-exhaustive options	The list may not include all important reasons, e.g., 'relaxation' or 'mental health'.
Q09	major	No unit or range	No unit — unclear whether to give hours, minutes, items, etc.
Q10	major	Gaps, non-exhaustive options	There is no option for non-full-time, evening, or those who have completed a bachelor's/engineering degree but continue studying in another form, which prevents them from providing a correct answer.
Q10	minor	No 'don't know / N/A / prefer not to say'	Lack of 'Not applicable' or 'Prefer not to answer' options forces respondents to choose a mismatching category or skip the question, which distorts the data.
Q10	minor	No 'other (please specify)' option	The list has no 'Other (please specify)' option — it may not cover all cases.
—	minor	Agree-disagree instead of item-specific scale	Agree-disagree format in 1 grid(s) (Q04). Item-specific scales are usually more reliable and less prone to acquiescence — unless the questionnaire is standardised.

6 Data quality controls

Criterion	Status	Pts	Comment
Attention check (longer surveys)	n/a	—	The short duration does not require attention checks. [ALL]
Reverse-keyed items in batteries	yes	25/25	In the Q04 battery, there is reverse positioning without labeling. [Q04: Q04_3]
Consistency pairs	n/a	—	The short duration does not require consistency pairs. [ALL]
Seriousness / honesty check	n/a	—	The short duration does not require severity checks. [ALL]

Control items found

None.

Suggested control items

Type	After question	Text	Expected answer	Rationale
instructed response	Q03	Proszę zaznaczyć tylko odpowiedź: Pływanie.	Pływanie	This element serves to identify respondents who do not read instructions or do not carefully mark answers in multiple-choice questions, which is a common source of noise in data.
consistency pair	Q02	Czy w ciągu ostatnich 7 dni nie podejmowałeś żadnej aktywności fizycznej trwającej co najmniej 30 minut?	Tak	This allows verification of consistency between the 'No' answer in Q01 and the number of days in Q02. If Q01 is 'No' but Q02 contains a number >0, there is a logical contradiction.
seriousness check	Q07	Czy uważasz, że regularna aktywność fizyczna jest korzystna dla zdrowia?	Tak	A simple question about conformity to social norms, which requires minimal cognitive effort, serves as an attention check after a difficult ranking task.

To detect careless responding in the main-study data, compute: the longest run of identical answers (*long-string*), intra-individual response variability (*IRV*), Mahalanobis distance, even-odd consistency and completion time (*speeding*, e.g. < 2 s per item). The code below computes them for Respondent AI data or a platform export (columns = scale items):

```
d <- read.csv("audit_data/items_values.csv", fileEncoding = "UTF-8") # or your platform export
items <- setdiff(names(d), c("persona_id", "style")) # scale items only
x <- as.matrix(d[items])
longstring <- apply(x, 1, function(r) max(rle(as.vector(r[!is.na(r)]))$lengths))
irv <- apply(x, 1, sd, na.rm = TRUE)
mahal <- if (nrow(x) > ncol(x) + 5) mahalanobis(x, colMeans(x, na.rm = TRUE), cov(x, use = "pairwise"))
flag <- longstring >= 0.8 * ncol(x) | irv < 0.3
data.frame(id = d[[1]], longstring, irv = round(irv, 2), mahal, flag)
```

7 Respondent burden and comfort

Criterion	Status	Pts	Comment
Estimated completion time	2.1 min	25/25	
Largest grid (rows)	5	15/15	
Required open questions	no data	—	The document does not mark required questions.
Share of required questions	no data	—	The document does not mark required questions.
Readability (FOG)	7.3	15/15	
Progress indicator	n/a	—	Not applicable to a document — depends on the survey platform.
Heaviest page (questions / items)	n/a	—	Not applicable to a document — pages are set up on the survey platform.

Readability of the questions: FOG = **7.3** (easy), 7.5 words per sentence, 11% hard words. Values up to about 9 are understood by most adults; above 12 the text requires higher education.

Linguistically hardest questions

ID	FOG	Question
Q09	20.8	Rok urodzenia:
Q07	20.0	Uszereguj poniższe powody podejmowania aktywności od najważniejszego (1) do najmniej ważnego (4).
Q03	12.5	Jakie formy aktywności podejmowałeś(-aś) w ciągu ostatnich 7 dni?
Q01	8.7	Czy w ciągu ostatnich 7 dni podejmowałeś(-aś) aktywność fizyczną trwającą łącznie co najmniej 30 minut dziennie?
Q02	8.0	W ile dniach w ciągu ostatnich 7 dni podejmowałeś(-aś) taką aktywność?
Q05	8.0	Jak oceniasz swoje ogólne samopoczucie w ciągu ostatnich 30 dni?

ID	Severity	Problem	Comment
Q04	minor	Large grid	17. A short battery of 5 rows in grid form is acceptable, but on a phone, it may be difficult to use.
—	minor	Survey too long	The discrepancy between the declared time (8 min) and the actual time (2 min) may discourage or mislead respondents.

8 Measurement consistency and reliability

Criterion	Status	Pts	Comment
Constructs measured with several items	yes	25/25	Key constructs (attitudes, well-being) are measured using multiple items. [Q04, Q06, Q07]
Homogeneous batteries	no	0/20	The Q04 battery combines pleasure with time availability, creating a heterogeneous construct. [Q04 Rows: (1) Ruch sprawia mi przyjemność ... (4) Łatwo znajduję czas]
Consistent scale format and direction	yes	15/15	
No keying problems (data)	no data	—	
Items discriminate (data)	no data	—	
Internal consistency of batteries (data)	no data	—	

Expected consistency of human answers. The greatest threat to data reliability is the Q04 battery, which combines different constructs (affective, cognitive, behavioral), potentially lowering internal consistency (Cronbach's alpha) and complicating the interpretation of a single indicator. Questions Q05 and Q06 are single measures of complex states (well-being, fatigue), making them susceptible to measurement error and subjective interpretation, although they are standard instruments in epidemiology. Answers in Q01–Q03 may be subject to recall bias due to the short reference period (7 days) and the lack of specific definitions of 'physical activity' (does household work count, walking?). The risk of fatigue (satisficing) is moderate, particularly in Q07 (ranking), which is cognitively demanding. The lack of control elements increases the risk of failing to detect respondents who do not follow instructions.

Expected α for human data from the Spearman–Brown formula: $\alpha = k \cdot \bar{r} / (1 + (k-1) \cdot \bar{r})$, where k is the number of items and $\bar{r}$ the typical inter-item correlation for the battery's homogeneity (high ≈ 0.45 ; medium ≈ 0.30 ; low ≈ 0.15).

Battery	Construct	k	Homogeneity	α humans (pred.)	α synth.	Reversed	Comment
Q04	Postawy i motywacja do aktywności fizycznej (postawy, odczucia po wysiłku, postrzegane bariery czasowe, instrumentalna motywacja zdrowotna)	5	medium	0.68 (0.56–0.77)	—	Q04_3	The scale mixes constructs of different psychometric nature: emotional (pleasure, well-being), cognitive (time loss), and behavioral/barrier (finding time). Items r1 and r2 are strongly correlated (a...

9 Structure and flow

Criterion	Status	Pts	Comment
Introduction and instructions	yes	20/20	The introduction includes a greeting, objective, instructions, and consent. [108 słów przed pierwszym pytaniem]
Closing and thank-you	yes	10/10	There is a closing page with thanks. [Dziękujemy za udział w badaniu! Jeśli chcesz poznać jego wyniki, napisz do nas na adres podany na początku ankiety.]
Logical order, easy opening, thematic blocks	yes	20/20	Questions are grouped thematically, with demographics at the end. [Część A. Aktywność fizyczna ... Część 3 ... Część 4]
Sensitive and demographic questions near the end	yes	20/20	Demographic and sensitive questions are placed at the end. [Q08 Płeć, Q09 Rok urodzenia, Q10 Rok studiów]

(cont.)

Criterion	Status	Pts	Comment
Consistent terminology	partly	8/15	The terminology is consistent, although Q04 uses the word 'Ruch' instead of 'physical activity'. [Q01: aktywność fizyczną ... Q03: formy aktywności ... Q04: Ruch]
No unnecessary repetition	yes	15/15	There are no repeated questions. [ALL]

ID	Severity	Problem	Comment
Q01	major	Missing filter	3. The filtering question assumes that 'physical activity' means 30 min per day, which is a specific WHO definition, not a common one. Fix: 4. Add the option 'Don't know' or clarify the definition of activity to avoid incorrect 'No' responses.
Q09	minor	Inconsistent terminology	The question about year of birth is precise, but in the context of students, it may be less useful than 'Year of Study'. Fix: Ensure that the variable 'Year of Study' (Q10) adequately covers respondents' ages.
—	minor	Inconsistent terminology	Criterion partly met: Consistent terminology.

10 Technical quality, mobile and accessibility

This audit concerns a document, so accessibility (WCAG), the phone version and validation messages were not tested — they depend on the platform the survey will be set up on. This dimension does not count towards the quality index. Once the survey is set up, audit its link: walkthroughs on a desktop and a phone with the axe-core test will show technical problems.

11 Ethics, GDPR and inclusiveness

Criterion	Status	Pts	Comment
Purpose of the study	yes	12/12	The study's purpose is clearly defined in the introductory text. [...gii Zdrowia Uniwersytetu Przykładowego. Celem badania jest poznanie związku między aktywnością fizyczną a samo...]
Who conducts the study	yes	8/8	The introduction includes the name of the research team and the institution. [...aństwo, zapraszamy do udziału w badaniu prowadzonym przez Zespół Psychologii Zdrowia Uniwersytetu Przykładowego...]
Researcher contact	yes	10/10	The email address for contacting the research team is provided. [...nie w celach naukowych, zgodnie z RODO. Kontakt z zespołem badawczym: badanie.aktywnosc@example.edu. Klikni...]
Voluntary participation	yes	10/10	It is indicated that participation is voluntary. [...ie około 8 minut. Udział w badaniu jest dobrowolny i anonimowy — nie pytamy o imię, nazwisko ani adres e-mail...]
Anonymity or confidentiality	yes	12/12	Anonymity is guaranteed, and there are no questions asking for identifying data. [...nut. Udział w badaniu jest dobrowolny i anonimowy — nie pytamy o imię, nazwisko ani adres e-mail. W każdej c...]
Data-protection (GDPR) notice	yes	14/14	There is a mention of compliance with GDPR and the provision of the data administrator. [...nianie ankiety bez podawania przyczyny. Administratorem danych jest Uniwersytet Przykładowy; dane będą przetwa...]
Right to withdraw at any time	yes	8/8	The participant has the right to interrupt the study at any time. [...tamy o imię, nazwisko ani adres e-mail. W każdej chwili możesz przerwać wypełnianie ankiety bez podawania przy...]
Explicit consent	yes	12/12	There is a clause expressing consent by clicking a button. [...(albo wypełnienie tej ankiety) oznacza zgodę na udział w badaniu. Badanie jest przeznaczone dla osób peł...]
Expected duration	yes	6/6	An estimated time is provided, but it is inconsistent with the model (8 vs. 2 min). [...entów. Wypełnienie ankiety zajmie około 8 minut. Udział w badaniu jest dobrowolny i anonimowy — nie pytam...]
Age requirement / target group	yes	8/8	The target group is defined as adults. [...niu. Badanie jest przeznaczone dla osób pełnoletnich. Część A. Aktywność fizyczna]

ID	Severity	Problem	Comment
Q08	minor	Non-inclusive language or categories	The 'Other' and 'Prefer not to answer' options improve inclusion, which is positive, but 'Other' is imprecise. Fix: Consider adding an 'Other (please specify)' field for the 'Other' option if the number of responses is significant.

12 Readiness for data analysis

Suggested codebook: variable names follow the question ids (grid rows as Qnn_r), values coded from 1 in the order of the options. Open questions need content coding; reverse-keyed items must be recoded before scoring.

Variable	Label	Type	Values
Q01	Czy w ciągu ostatnich 7 dni podejmowałeś(-aś) aktywność fizyczną trwającą łącznie co najm...	nominal	1 = Tak; 2 = Nie
Q02	W ile dniach w ciągu ostatnich 7 dni podejmowałeś(-aś) taką aktywność?	number	
Q03_1	Jakie formy aktywności podejmowałeś(-aś) w ciągu ...: Bieganie lub nordic walking	multiple (0/1)	0 = nie, 1 = tak
Q03_2	Jakie formy aktywności podejmowałeś(-aś) w ciągu ...: Ćwiczenia na siłowni	multiple (0/1)	0 = nie, 1 = tak
Q03_3	Jakie formy aktywności podejmowałeś(-aś) w ciągu ...: Pływanie	multiple (0/1)	0 = nie, 1 = tak
Q03_4	Jakie formy aktywności podejmowałeś(-aś) w ciągu ...: Jazda na rowerze	multiple (0/1)	0 = nie, 1 = tak
Q03_5	Jakie formy aktywności podejmowałeś(-aś) w ciągu ...: Zajęcia fitness lub taniec	multiple (0/1)	0 = nie, 1 = tak
Q03_6	Jakie formy aktywności podejmowałeś(-aś) w ciągu ...: Inne (jakie?)	multiple (0/1)	0 = nie, 1 = tak
Q04_1	Ruch sprawia mi przyjemność.	ordinal	1 = Zdecydowanie się nie zgadzam; 2 = Raczej się nie zgadzam; 3 = Ani się zgadzam, ani nie zgadzam; 4 = Raczej się zgad...
Q04_2	Po ćwiczeniach czuję się lepiej.	ordinal	1 = Zdecydowanie się nie zgadzam; 2 = Raczej się nie zgadzam; 3 = Ani się zgadzam, ani nie zgadzam; 4 = Raczej się zgad...
Q04_3	Aktywność fizyczna to dla mnie strata czasu.	ordinal	1 = Zdecydowanie się nie zgadzam; 2 = Raczej się nie zgadzam; 3 = Ani się zgadzam, ani nie zgadzam; 4 = Raczej się zgad...
Q04_4	Łatwo znajduję czas na ruch w ciągu tygodnia.	ordinal	1 = Zdecydowanie się nie zgadzam; 2 = Raczej się nie zgadzam; 3 = Ani się zgadzam, ani nie zgadzam; 4 = Raczej się zgad...
Q04_5	Ćwiczę, bo chcę dbać o zdrowie.	ordinal	1 = Zdecydowanie się nie zgadzam; 2 = Raczej się nie zgadzam; 3 = Ani się zgadzam, ani nie zgadzam; 4 = Raczej się zgad...
Q05	Jak oceniasz swoje ogólne samopoczucie w ciągu ostatnich 30 dni?	ordinal	1 = 0; 2 = 1; 3 = 2; 4 = 3; 5 = 4; 6 = 5; 7 = 6; 8 = 7; 9 = 8; 10 = 9; 11 = 10
Q06	Jak często w ciągu ostatnich 2 tygodni czuleś(-aś) się zmęczony(-a) lub miałeś(-aś) mało ...	nominal	1 = Wcale; 2 = Kilka dni; 3 = Więcej niż połowę dni; 4 = Prawie codziennie; 5 = Wolę nie odpowiadać
Q07	Uzereguj poniższe powody podejmowania aktywności od najważniejszego (1) do najmniej ważn...	nominal	1 = zdrowie; 2 = wygląd; 3 = kontakty towarzyskie; 4 = odstresowanie się
Q08	Płeć	nominal	1 = Kobieta; 2 = Mężczyzna; 3 = Inna; 4 = Wolę nie odpowiadać
Q09	Rok urodzenia:	number	
Q10	Rok studiów	nominal	1 = I; 2 = II; 3 = III; 4 = IV; 5 = V; 6 = studia doktoranckie

Draft R code for scoring the batteries and their reliability on the study data (psych::alpha with reverse keys):

```
library(psych)
d <- read.csv("dane.csv", fileEncoding = "UTF-8") # eksport z platformy / platform export
# Q04
b1_items <- c("Q04_1", "Q04_2", "Q04_3", "Q04_4", "Q04_5")
b1_keys <- ifelse(b1_items %in% c("Q04_3"), -1, 1)
print(alpha(d[b1_items], keys = b1_keys))
d$b1 <- rowMeans(reverse.code(b1_keys, d[b1_items]), na.rm = TRUE)
```

13 Appendices

13.1 Full list of problems

ID	Code	Problem	Severity	Conf.	Source	Comment
Q01	W04	Vague terms or quantifiers	major	0.94	AI: methodologist + AI: respondent advocate	1. The phrasing 'at least 30 minutes in total per day' is misleading: it suggests that 30 minutes is the threshold required to classify activity as 'yes', whereas in epidemiological studies, total we...
Q01	F03	Missing filter	major	0.75	AI: whole-survey review	3. The filtering question assumes that 'physical activity' means 30 min per day, which is a specific WHO definition, not a common one.
Q01	W09	Social desirability pressure	major	0.75	AI: methodologist	5. The question about physical activity is strongly subject to social desirability bias; the 'Yes' response is socially desirable, which can lead to underreporting of inactivity.
Q02	W07	Demanding recall or estimation	major	0.97	AI: methodologist + AI: respondent advocate	7. Requires precise memory and counting of days from the previous week, which is difficult and subject to recall error (Tourangeau). Respondents often guess or use heuristics.
Q03	O12	No 'other (please specify)' option	minor	0.96	AI: methodologist + AI: respondent advocate	9. The 'Other' option exists, but the lack of a text field to enter answers prevents collecting data on unforeseen forms of activity, limiting data completeness.
Q04/3	W03	Negation or double negative	major	0.97	AI: methodologist + AI: respondent advocate	11. The statement is negative and substantively reversed relative to the others (positive). In a Likert scale, the respondent must perform an additional cognitive operation (negation), which increase...
Q04/3	O08	Agree–disagree instead of item-specific scale	major	0.75	AI: methodologist	13. Using a Likert scale for a negative statement ('waste of time') is less precise than a 'how true is this' or 'how much does this describe' scale. The Likert scale is susceptible to acquiescence b...
Q04	M06	Heterogeneous battery	major	0.73	AI: whole-survey review + AI: methodologist	15. The rows measure different constructs: pleasure (affect), well-being after exercise (physiological/psychological effect), perceived value of time (attitude), time availability (barrier), motivati...
Q04	B02	Large grid	minor	0.56	AI: whole-survey review	17. A short battery of 5 rows in grid form is acceptable, but on a phone, it may be difficult to use.
Q05	O04	No 'don't know / N/A / prefer not to say'	minor	0.97	AI: methodologist + AI: respondent advocate + rule	An opinion question without 'don't know / no opinion'.
Q05	O09	Response order effects	minor	0.93	AI: methodologist + AI: respondent advocate + rule	Long list (11 options) — the first ones will be chosen more often.
Q05	M05	Ceiling or floor effect	minor	0.60	AI: whole-survey review	19. A 0-10 scale may lead to a ceiling effect among students if most are satisfied.
Q06	W02	Double-barreled question	major	0.97	AI: methodologist + AI: respondent advocate + rule	'zmęczony lub miałeś mało' — one answer covers two things; it is unclear which one it refers to.
Q06	W04	Vague terms or quantifiers	major	0.97	AI: methodologist + AI: respondent advocate + rule	Vague terms: mało — they mean different things to different people.
Q06	O04	No 'don't know / N/A / prefer not to say'	minor	0.60	AI: respondent advocate	21. The presence of this option is good, but it is worth ensuring it is visible and understandable.
Q06	O12	No 'other (please specify)' option	minor	0.50	rule	The list has no 'Other (please specify)' option — it may not cover all cases.
Q07	O10	Wrong answer format	major	0.75	AI: methodologist	23. Ranking forces an arbitrary choice, even when reasons are equivalent. This cognitively burdens the respondent and may not reflect real motivations (e.g., two main reasons are equally important).
Q07	O09	Response order effects	minor	0.75	AI: methodologist	25. Lack of rotation of the order of reasons can lead to position effects (first options are more often chosen as 'most important' in rankings).
Q07	O02	Gaps, non-exhaustive options	minor	0.60	AI: respondent advocate	The list may not include all important reasons, e.g., 'relaxation' or 'mental health'.
Q08	E06	Non-inclusive language or categories	minor	0.96	AI: whole-survey review + AI: methodologist + AI: respondent advocate	The 'Other' and 'Prefer not to answer' options improve inclusion, which is positive, but 'Other' is imprecise.
Q09	O13	No unit or range	major	0.97	AI: methodologist + AI: respondent advocate + rule	No unit — unclear whether to give hours, minutes, items, etc.
Q09	F04	Inconsistent terminology	minor	0.60	AI: whole-survey review	The question about year of birth is precise, but in the context of students, it may be less useful than 'Year of Study'.
Q10	O02	Gaps, non-exhaustive options	major	0.97	AI: methodologist + AI: respondent advocate	There is no option for non-full-time, evening, or those who have completed a bachelor's/engineering degree but continue studying in another form, which prevents them from providing a correct answer.

(cont.)

ID	Code	Problem	Severity	Conf.	Source	Comment
Q10	O04	No 'don't know / N/A / prefer not to say'	minor	0.96	AI: methodologist + AI: respondent advocate rule	Lack of 'Not applicable' or 'Prefer not to answer' options forces respondents to choose a mismatching category or skip the question, which distorts the data.
Q10	O12	No 'other (please specify)' option	minor	0.50		The list has no 'Other (please specify)' option — it may not cover all cases.
—	M06	Heterogeneous battery	major	0.75	AI: whole-survey review	Criterion not met: Homogeneous batteries.
—	B01	Survey too long	minor	0.90	AI: whole-survey review	The discrepancy between the declared time (8 min) and the actual time (2 min) may discourage or mislead respondents.
—	F04	Inconsistent terminology	minor	0.75	AI: whole-survey review	Criterion partly met: Consistent terminology.
—	O08	Agree–disagree instead of item-specific scale	minor	0.40	AI: whole-survey review rule	Agree–disagree format in 1 grid(s) (Q04). Item-specific scales are usually more reliable and less prone to acquiescence — unless the questionnaire is standardised.

13.2 Candidates rejected by the AI panel

Low-confidence rule findings that no AI expert confirmed. They do not count towards the scores but are worth a look.

None.

13.3 Glossary of problem codes

Code	Dimension	Problem	Comment
W01	Question wording	Leading question	Wording suggests the 'right' answer (e.g. 'Do you agree that...', loaded adjectives, appeals to authority or the majority).
W02	Question wording	Double-barreled question	The question or statement asks about two things at once; the answer cannot be attributed.
W03	Question wording	Negation or double negative	A negated statement rated on an agreement scale ('I disagree that I don't...') is hard to process.
W04	Question wording	Vague terms or quantifiers	Words such as 'often', 'regularly', 'recently', 'a lot' mean different things to different people.
W05	Question wording	No reference period	A behaviour or frequency question without a time frame (e.g. 'in the past 7 days').
W06	Question wording	Jargon, abbreviations, difficult words	Technical terms or abbreviations some respondents may not know.
W07	Question wording	Demanding recall or estimation	Requires recalling or counting something people rarely remember (long period, exact numbers).
W08	Question wording	Hidden presupposition	The question assumes something that may not be true for the respondent.
W09	Question wording	Social desirability pressure	One answer is clearly more socially acceptable, which invites over-reporting or under-reporting.
W10	Question wording	Sensitive question without softening	An intimate or distressing topic asked bluntly, without lead-in, normalisation or a way to skip.
W11	Question wording	Too long or complex	A long sentence with several clauses or conditions; the meaning gets lost before answering.
W12	Question wording	Ambiguous	The question can be read in more than one way (unclear word, subject or scope).
W13	Question wording	Hypothetical question	Asks about behaviour in an imagined situation; such answers predict real behaviour poorly.
W14	Question wording	Missing or unclear instruction	Unclear how many options to tick, what format to type, or what the scale refers to.
W15	Question wording	Language errors	Typos, grammar or punctuation errors that reduce credibility and clarity.
W16	Question wording	Assumes knowledge	Requires knowledge or information the respondent may not have.
O01	Response options and scales	Overlapping categories	The same value fits two options (e.g. '18–25' and '25–35').
O02	Response options and scales	Gaps, non-exhaustive options	Some respondents find no fitting option (a gap between ranges, missing categories).
O03	Response options and scales	Unbalanced scale	More positive than negative options (or vice versa) shifts the results.
O04	Response options and scales	No 'don't know / N/A / prefer not to say'	Respondents without an opinion, to whom it does not apply, or who prefer not to answer must guess.
O05	Response options and scales	Options do not match the question	Answer categories answer a different question (e.g. 'how often' with an agreement scale).
O06	Response options and scales	Inappropriate number of scale points	Too few points (poor discrimination) or too many (spurious precision).
O07	Response options and scales	Unclear scale labels	Labels are vague, unevenly spaced or only at the ends where full labelling is needed.

(cont.)

Code	Dimension	Problem	Comment
O08	Response options and scales	Agree–disagree instead of item-specific scale	Agree–disagree scales invite acquiescence; item-specific scales are usually more reliable (Saris et al., 2010), unless the instrument is standardised.
O09	Response options and scales	Response order effects	A long list without rotation — the first options are chosen more often (primacy effect).
O10	Response options and scales	Wrong answer format	E.g. single choice where several answers apply, or a free-text field instead of a list or number.
O11	Response options and scales	Inconsistent scale direction	Scales start from the negative end in some places and the positive end in others — errors become likely.
O12	Response options and scales	No 'other (please specify)' option	A nominal list does not cover all possibilities and does not allow adding one's own.
O13	Response options and scales	No unit or range	A numeric field without a unit (hours? minutes? currency?) or a sensible range.
C01	Data quality controls	No attention checks	A longer survey has no instructed-response item (e.g. 'select »somewhat disagree«').
C02	Data quality controls	Flawed control item	The control item is too obvious, ambiguous, offensive or badly placed.
C03	Data quality controls	No reverse-keyed items	All statements in a battery point in the same direction — acquiescence cannot be detected.
C04	Data quality controls	No consistency checks	No pairs of equivalent questions that would reveal random answering.
C05	Data quality controls	No seriousness check	No closing question on whether the respondent answered carefully and honestly.
B01	Burden and comfort	Survey too long	The expected completion time is long — break-off and satisficing increase.
B02	Burden and comfort	Large grid	Many rows on one scale invite straightlining and fatigue, especially on phones.
B03	Burden and comfort	Required open-ended questions	Required open questions add time and increase break-offs and filler answers.
B04	Burden and comfort	Forced responses	All (or almost all) questions are mandatory, including sensitive ones, with no way to decline.
B05	Burden and comfort	Low readability	High fog index — long sentences and hard words.
B06	Burden and comfort	No progress indicator	Respondents cannot see how much is left.
B07	Burden and comfort	Overloaded page	Too many questions on one page.
B08	Burden and comfort	Repetitiveness, fatigue	A run of very similar questions or scales tires respondents and lowers answer quality.
M01	Measurement consistency	Single-item measure of a construct	A complex construct (attitude, trait, state) measured with a single item — low reliability.
M02	Measurement consistency	Item does not discriminate	Almost everyone gives the same answer — the item carries no information.
M03	Measurement consistency	Keying problem (reverse item?)	The item correlates negatively with the rest of the scale — reverse it in the key or rephrase it.
M04	Measurement consistency	Weak fit with the rest of the scale	Low corrected item–total correlation — measures something else or is unclear.
M05	Measurement consistency	Ceiling or floor effect	Most answers fall in an extreme category — the scale does not discriminate there.
M06	Measurement consistency	Heterogeneous battery	Items in one scale tap different things — the total score will be hard to interpret.
M07	Measurement consistency	Unstable answers	The same respondent profiles answer differently each time.
M08	Measurement consistency	Divergent comprehension	In cognitive interviews respondents understand the question or its key terms differently.
F01	Structure and flow	Sensitive or demographic questions too early	Personal questions before engagement is built increase break-offs and refusals.
F02	Structure and flow	Context / order effect	An earlier question may prime the answer to a later one.
F03	Structure and flow	Missing filter	The question does not apply to everyone and there is no filter or 'not applicable' option.
F04	Structure and flow	Inconsistent terminology	The same concept named differently in different places (or different concepts named alike).
F05	Structure and flow	Missing or weak introduction	No welcome, instructions, expected duration or explanation of how to answer.
F06	Structure and flow	No closing page	No thank-you page confirming that the answers were saved.
F07	Structure and flow	Illogical order	Topic jumps, specific before general questions, no thematic blocks.
F08	Structure and flow	Duplicate question	The same question appears twice (almost verbatim) — unless it is an intended consistency pair.
F09	Structure and flow	Non-random assignment to conditions	Respondents choose the version or condition themselves (e.g. identical options 'xxx', 'just guess') — assignment is not random and groups may differ systematically; use the platform's randomisation.
A01	Technical, mobile, accessibility	WCAG violation (axe-core)	The automated axe-core accessibility test found a WCAG 2.x violation.
A02	Technical, mobile, accessibility	Horizontal scrolling on mobile	The page is wider than the phone screen.
A03	Technical, mobile, accessibility	Small touch targets	Targets smaller than 24×24 px (WCAG 2.2, 2.5.8) — easy to tap the wrong answer.
A04	Technical, mobile, accessibility	Small font	Text smaller than 12 px is hard to read.
A05	Technical, mobile, accessibility	Grid hard to use on mobile	A table with many columns does not fit a phone screen or has very narrow columns.
A06	Technical, mobile, accessibility	Unlabeled field	A form field has no label exposed to screen readers.

(cont.)

Code	Dimension	Problem	Comment
A07	Technical, mobile, accessibility	Technical failure	Synthetic respondents could not complete the survey (error, hang, validation that never passes).
A08	Technical, mobile, accessibility	Unclear validation	The form rejects an answer while the requirement (format, range) is not stated at the question.
A09	Technical, mobile, accessibility	Not mobile-optimised	No viewport meta tag — phones show a zoomed-out desktop page.
E01	Ethics, GDPR, inclusiveness	No information on the study purpose	Respondents are not told what the study is for and how answers will be used.
E02	Ethics, GDPR, inclusiveness	No informed consent	No information on voluntary participation and no explicit consent.
E03	Ethics, GDPR, inclusiveness	No data-protection (GDPR) notice	No data controller, purpose and legal basis, retention period and data-subject rights.
E04	Ethics, GDPR, inclusiveness	Special-category data	Questions about health, sexuality, religion, political views or ethnicity (GDPR art. 9) need a clear justification, explicit consent and safeguards.
E05	Ethics, GDPR, inclusiveness	Identifying data	Name, e-mail, phone or address remove anonymity — collect only when necessary.
E06	Ethics, GDPR, inclusiveness	Non-inclusive language or categories	E.g. only two gender options without 'prefer not to say', stereotypes, male-only wording.
E07	Ethics, GDPR, inclusiveness	No researcher contact	Respondents have no one to contact with questions or concerns.
E08	Ethics, GDPR, inclusiveness	No information on withdrawal	No statement that participation can be stopped at any time without consequences.
E09	Ethics, GDPR, inclusiveness	Distress risk without support	Questions about distressing experiences without a warning, a way to skip, and where to find help.
E10	Ethics, GDPR, inclusiveness	No anonymity / confidentiality information	Respondents are not told whether answers are anonymous or confidential.
E11	Ethics, GDPR, inclusiveness	Researcher not identified	No institution, team or person responsible for the study is named.

13.4 Audit parameters

Item	Value
Wersja Respondent AI / version	0.1.0
Model	qwen3.6:35b-mlx
Backend	ollama
Tryb / mode	quick
Języki / languages	pl, en
Role AI / AI roles	metodolog i psychometra, rzecznik badanego (wywiady poznawcze, prosty język)
Przebiegi na żywo / live	both
Urządzenie mobilne / device	Pixel 7
axe-core	4.12.1
Wywołania modelu / LLM calls	14
Nieudane wywołania / failed calls	0
Odrzuceni kandydaci / rejected candidates	0
Czas audytu (min) / duration	3.6
Źródła / sources	Aktywność fizyczna i samopoczucie studentów.docx