Data quality assessment of the *Youth Talks* dataset

This document exposes the content of the *Youth Talks* dataset. A minimal exploratory data analysis has been performed to insure data quality.

1. Data source

The first edition of the consultation was held from October 2022 to May 2023.

It was compound of 3 sets of questions:

- 8 Segmentation closed questions,
- 11 Open questions, enabling several answers per question,
- 16 Empathy questionnaire, with closed questions.

The "Participant name" field enables to identify the different participants. Its values are in the format "Participant-[number]" where the number is an integer. We create a "participant_id" field using this number.

1.1 Segmentation questions

The 8 segmentation questions are:

- Country or territory where I grew up
- Current situation
- Year of birth
- Gender
- Level of education (completed or in progress)
- Parents' highest level of education
- Income scale (1 = lowest ; 10 = highest)
- Which of the following do you identify with MOST?

1.2 Open questions

The 11 open questions are formulated and numbered as is:

1. When I think about the future, what I wish… …for myself:
2. When I think about the future, what I wish… …for the world (others, society, planet…):
3. When I think about the future, what worries me… …for myself:
4. When I think about the future, what worries me… …for the world (others, society, planet…):
5. What collective issues do we need to address to build the future I want?

6. To build this desired future, I would be ready to give up the following things:
7. On the contrary, I would not be willing to give up the following things:
8. Why?
9. To build this desired future, what we must all learn… …at school is:
10. To build this desired future, what we must all learn… …in life is:
11. What question would you like to ask young people around the world?

1.3 Empathy questionnaire

The 16 empathy questions are:

- I often have tender, concerned feelings for people less fortunate than me.
- I really get involved with the feelings of the characters in a novel.
- In emergency situations, I feel apprehensive and ill-at-ease.
- I try to look at everybody’s side of a disagreement before I make a decision.
- When I see someone being taken advantage of, I feel kind of protective toward them.
- I sometimes try to understand my friends better by imagining how things look from their perspective.
- After seeing a play or movie, I have felt as though I were one of the characters.
- Being in a tense emotional situation scares me.
- When I see someone being treated unfairly, I sometimes don't feel very much pity for them.
- I would describe myself as a pretty soft-hearted person.
- When I watch a good movie, I can very easily put myself in the place of a leading character.
- I tend to lose control during emergencies.
- When I’m upset at someone, I usually try to “put myself in his shoes” for a while.
- When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.
- When I see someone who badly needs help in an emergency, I go to pieces.
- Before criticizing somebody, I try to imagine how I would feel if I were in their place.

2. Data Processing

The dataset to be produced is compound of 3 data blocks:

- **Data block 1**: participant_id, Participant name, and the 8 Segmentation questions.
- **Data block 2**: participant_id, for each Open question, 3 contributions maximum per participant, including Answer text, Cluster title and Macro-cluster title
- **Data block 3**: participant_id, and the 16 Empathy questions.

2.1 Data block 1 – Segmentation questions

A first dataframe is calculated with 10 columns : "participant_id", "Participant name", and the 8 Segmentation questions.

The answer value "No answer" to the last question is replaced by a blank.

Columns are renamed as is:

- participant_id
- participant_name
- country
- situation
- yob
- gender
- education
- parents_education
- income_scale
- most_identification

2.2 Data block 2 – Open questions

The answers to the open question have been classified using clusters and macro-clusters. For methodological reasons, about 20% of data have not been clusterized. The "NC" symbol is used when macro-cluster or cluster are lacking.

For each participant and for each question, we keep the 3 first anwsers with their respective cluster and macro-cluster.

We get a second dataframe with 100 columns: "participant_id" and for each of the 11 questions, 3 contributions with the 3 columns "Answer text", "Cluster title", "Macro-cluster title". When a participant did not make any extra contributions, the data are left blank.

Columns are renamed as is by concatenating the question number (from 01 to 11), the contribution number (from *1* to *3*) and the 3 columns of interest (answer, cluster, macro_cluster):

- participant_id
- question_01_contrib1_answer
- question_01_contrib1_cluster
- question_01_contrib1_macro_cluster
- ...
- question_11_contrib3_answer
- question_11_contrib3_cluster
- question_11_contrib3_macro_cluster

2.3 Data block 3 – Empathy questionnaire

In this data block, answers are integers in the range *0* to *5*. The *0* integer denotes no answer and is replaced by a blank in the tabular datasets.

We added an indicator, "empathy_answers", with the number of answers given by each participant, therefore between 1 and 16.

Warning: For 1,363 participants who answered to the full empathy questionnaire, a technical problem set to *0* the answer to the sole question "When I see someone being treated unfairly, I sometimes don't feel very much pity for them" (EC-18). To keep a trace of this information, those answers have been switched to a *-1* and is replaced by a * character in the tabular datasets (CSV and Excel). In that case, the indicator "empathy_answers" is 15.

We get a third dataframe with 18 columns: "participant_id", the 16 Empathy questions, and the "empathy_answers" column.

The 16 columns are renamed using the *Interpersonal Reactivity Index* (*) as is:

- empathy_EC_2
- empathy_F_5
- empathy_PD_6
- empathy_PT_8
- empathy_EC_9
- empathy_PT_11
- empathy_F_16
- empathy_PD_17
- empathy_EC_18
- empathy_EC_22
- empathy_F_23
- empathy_D_24
- empathy_PT_25
- empathy_F_26
- empathy_PD_27
- empathy_PT_28

(*) Ingoglia, S., Lo Coco, A., & Albiero, P. (2016). Development of a brief form of the Interpersonal Reactivity Index (B– IRI). Journal of Personality Assessment, 98(5), 461–471. https://pubmed.ncbi.nlm.nih.gov/27050826/

3. Production of the datasets

We merge the 3 data blocks by performing a join on 'participant_id' and we filter empty rows without any usable data.

We added an indicator "total_answers" with the numbers of answer per participant. This indicator computes for each participant the sum of the number of answers to the Segmentation questions (8 max), the number of answers to the Open questions (11 max, counting 1 answer per question), and the number of answers to the Empathy questionnaire (1 min, 16 max). The range of this indicator is 1 to 35.

The dataset has been output in 3 different formats:

- **Tabular formats** including 127 columns and 41,753 records, limited to 3 answers to the open questions:
    - CSV: YT_dataset-v1.csv
    - Excel: YT_dataset-v1.xlsx
    - SQL: YT_dataset-v1.sql
- **Other format** including all answers to the open questions:
    - JSON: YT_dataset-v1.json

A supplementary Excel file, YT_notes-v1.xlsx provides extra information about the datasets:

- Tabular columns
- JSON schema

- Macro-clusters
- Clusters

4. Global picture

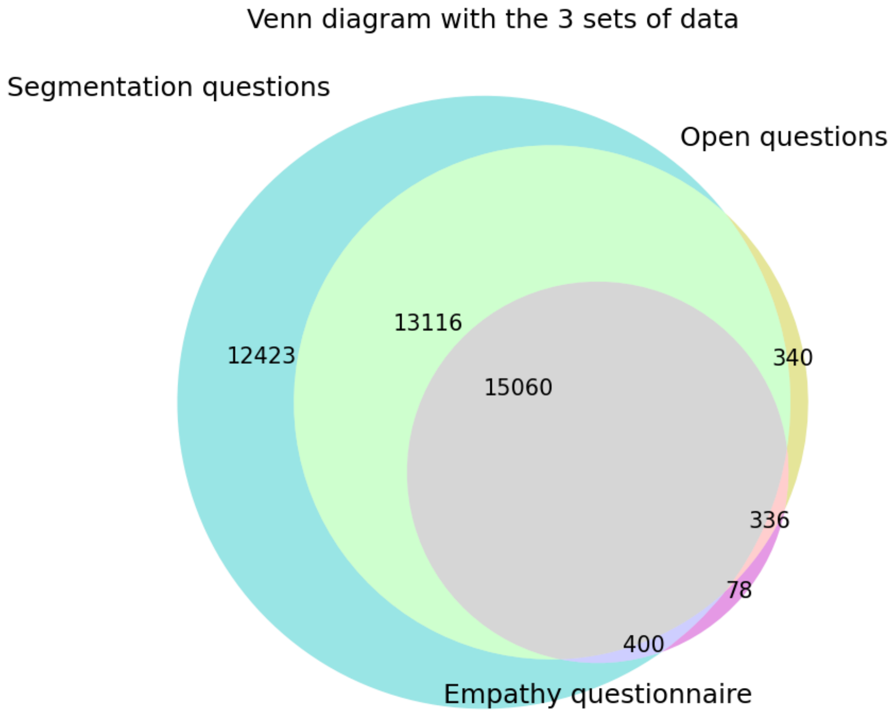
The final dataset counts 41,753 records for as many participants, and including:

- 40,999 participants with at least one answer to the Segmentation questions,
- 28,852 participants with at least one answer to the Open questions,
- 15,874 participants with at least one answer to the Empathy questionnaire.

The table below sums up the calculated numbers according to there is at least one answer, or none, to the different questions.

	Open questions ≥ 1		**Open questions = 0**	
	Empathy questions ≥ 1	**Empathy questions = 0**	**Empathy questions ≥ 1**	**Empathy questions = 0**
Segmentation questions ≥ 1	15,060	13,116	400	12,423
Segmentation questions = 0	336	340	78	–

We present below a Venn diagram with the 3 sets of data.



5. Data quality assessment

This section computes a few insights such as the completion rates of the different fields.

The percentages and other insights are computed upon the total number of participants collected in the dataset, i.e. 41,753.

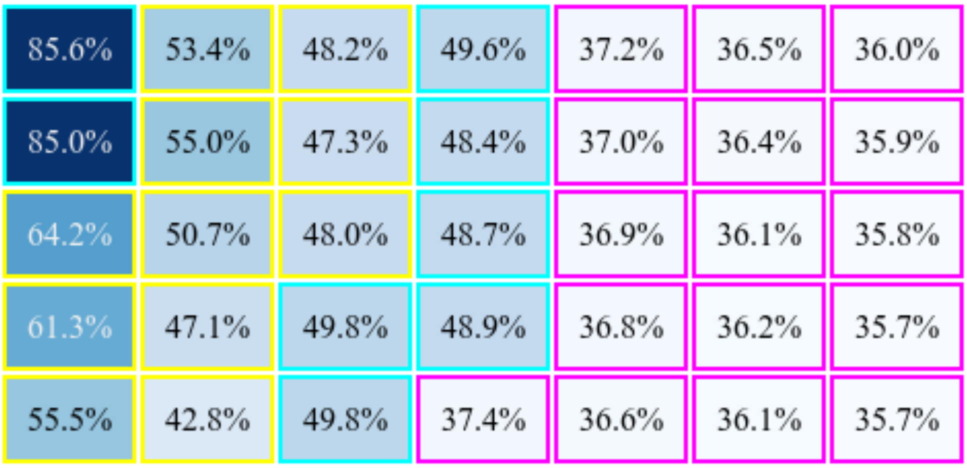
5.1 Global completion rates

We have computed the global completion rates of all questions upon the 41,753 participants and in the exact order of how the 35 questions where set in the *Youth Talks* consultation.

To ease the reading, the results have been reorganized in a 7 columns x 5 rows table, including:

- the 2 leading segmentation questions: "Country or territory where I grew up" and "Which of the following do you identify with MOST?",
- the 11 open questions,
- the 6 remaining segmentation questions,
- the 16 empathy questions.

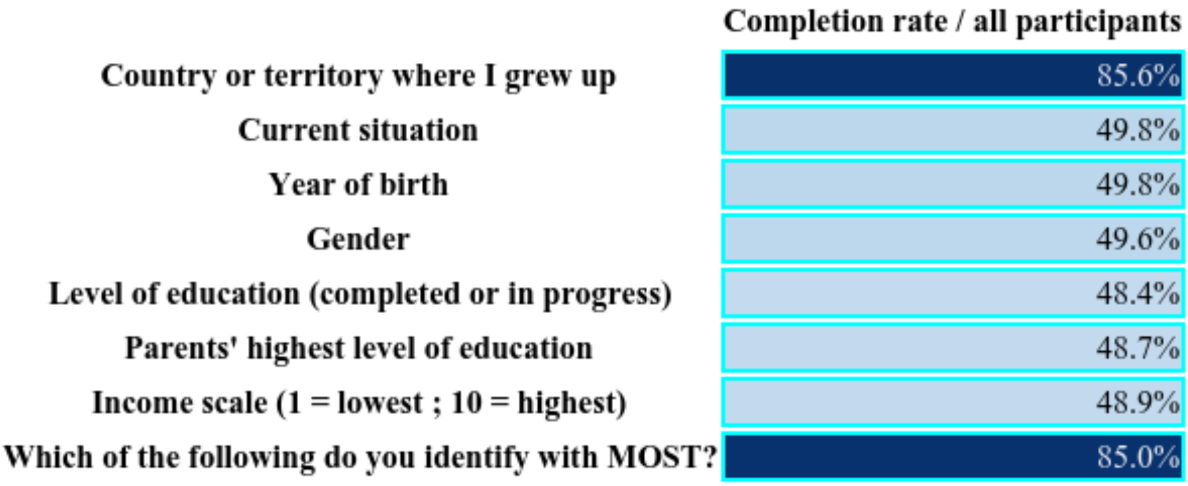
The origin of each question is denoted by a colored frame: cyan for the 8 segmentation questions, yellow for the 11 open questions, and magenta for the 16 empathy questions.



85.6%	53.4%	48.2%	49.6%	37.2%	36.5%	36.0%
85.0%	55.0%	47.3%	48.4%	37.0%	36.4%	35.9%
64.2%	50.7%	48.0%	48.7%	36.9%	36.1%	35.8%
61.3%	47.1%	49.8%	48.9%	36.8%	36.2%	35.7%
55.5%	42.8%	49.8%	37.4%	36.6%	36.1%	35.7%

5.2 Segmentation questions

The completion rates have been computed upon the 41,753 participants. We observe that the segmentation questions have a completion rate around 50%, except questions "Country or territory where I grew up" and "Which of the following do you identify with MOST?" which are around 85%.

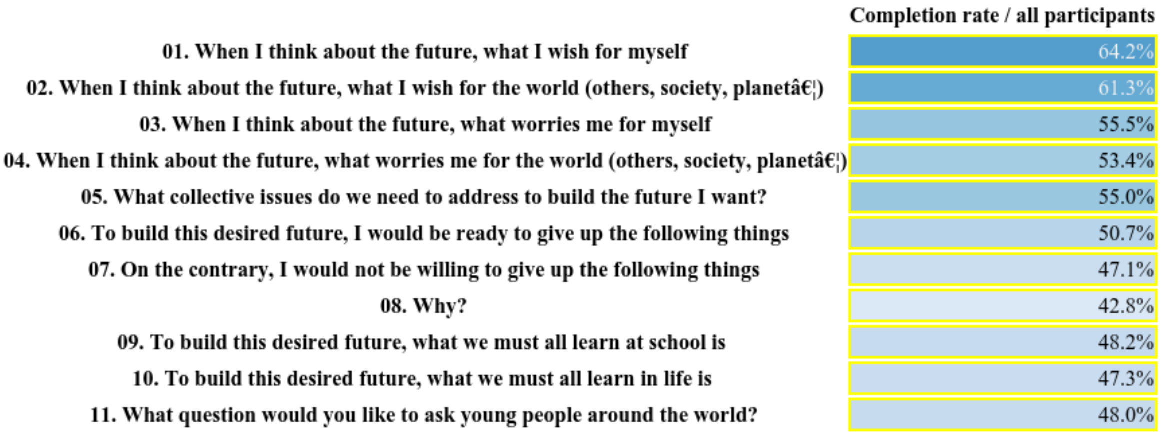


	Completion rate / all participants
Country or territory where I grew up	85.6%
Current situation	49.8%
Year of birth	49.8%
Gender	49.6%
Level of education (completed or in progress)	48.4%
Parents' highest level of education	48.7%
Income scale (1 = lowest ; 10 = highest)	48.9%
Which of the following do you identify with MOST?	85.0%

5.3 Open questions

5.3.1 Completion rates

The completion rates have been computed upon the 41,753 participants. We can observe that the open questions have a completion rate in the range 47% to 64%, except question 8 ("Why?").

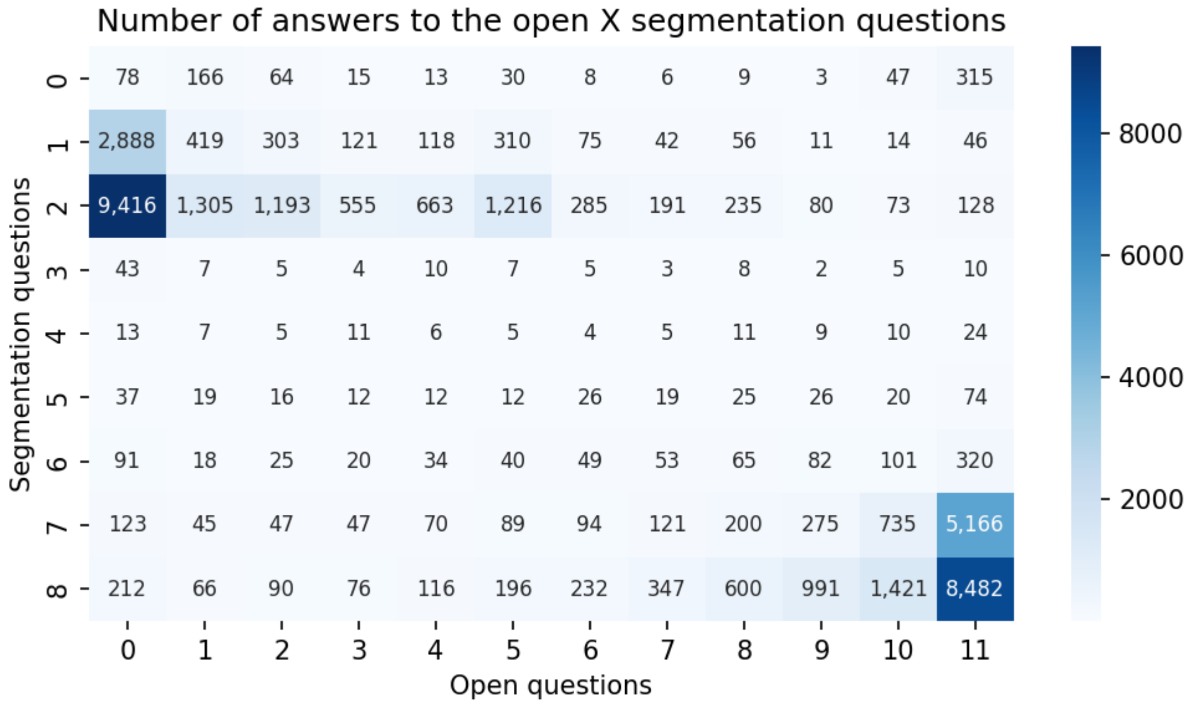


	Completion rate / all participants
01. When I think about the future, what I wish for myself	64.2%
02. When I think about the future, what I wish for the world (others, society, planetâ€¦)	61.3%
03. When I think about the future, what worries me for myself	55.5%
04. When I think about the future, what worries me for the world (others, society, planetâ€¦)	53.4%
05. What collective issues do we need to address to build the future I want?	55.0%
06. To build this desired future, I would be ready to give up the following things	50.7%
07. On the contrary, I would not be willing to give up the following things	47.1%
08. Why?	42.8%
09. To build this desired future, what we must all learn at school is	48.2%
10. To build this desired future, what we must all learn in life is	47.3%
11. What question would you like to ask young people around the world?	48.0%

5.3.2 Cross tabulation with segmentation¶

The table below displays the number of participants according to the number of answers to segmentation x open questions (counting only 1 answer per open question).

For instance, bottom right, 8,462 participants did answer to all open questions (11) and to all segmentation questions (8).

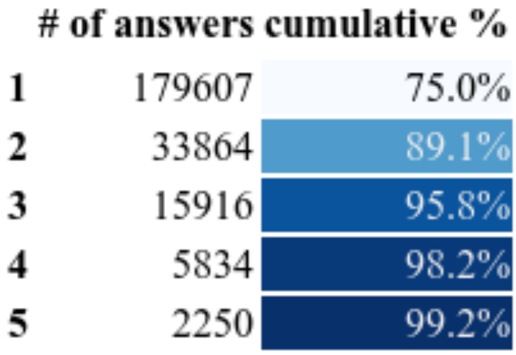


	0	1	2	3	4	5	6	7	8	9	10	11
0	78	166	64	15	13	30	8	6	9	3	47	315
1	2,888	419	303	121	118	310	75	42	56	11	14	46
2	9,416	1,305	1,193	555	663	1,216	285	191	235	80	73	128
3	43	7	5	4	10	7	5	3	8	2	5	10
4	13	7	5	11	6	5	4	5	11	9	10	24
5	37	19	16	12	12	12	26	19	25	26	20	74
6	91	18	25	20	34	40	49	53	65	82	101	320
7	123	45	47	47	70	89	94	121	200	275	735	5,166
8	212	66	90	76	116	196	232	347	600	991	1,421	8,482

5.3.3 Other insights

The average number of answers per question is 1.44

By keeping a maximum of 3 answers per question in the tabular datasets, 95.8% of answers are returned.

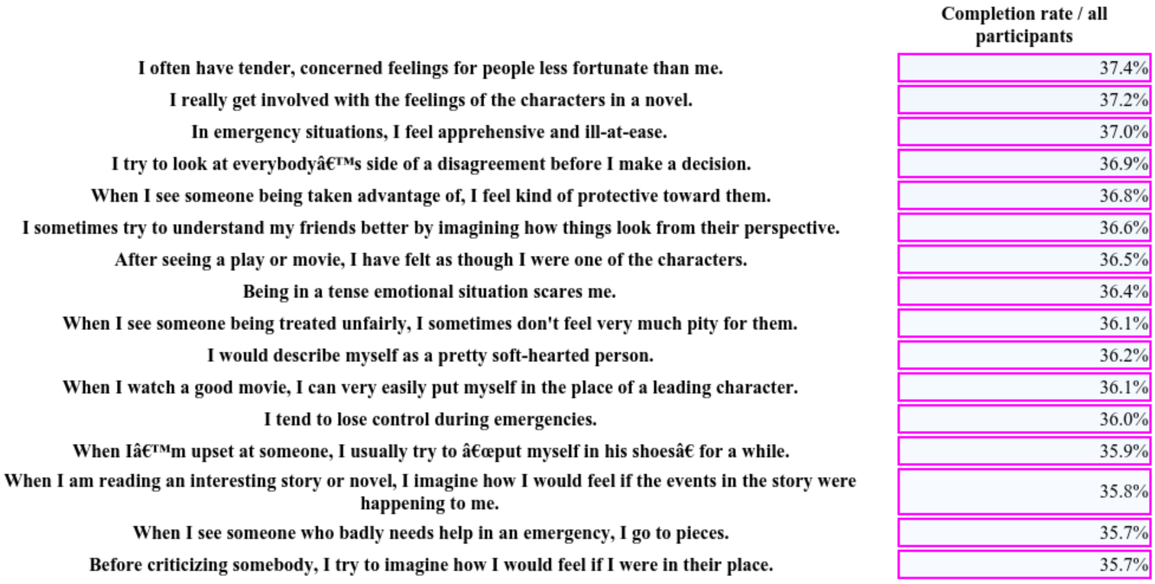


	# of answers	cumulative %
1	179607	75.0%
2	33864	89.1%
3	15916	95.8%
4	5834	98.2%
5	2250	99.2%

5.4 Empathy questionnaire

5.4.1 Completion rates

The completion rates have been computed upon the 41,753 participants. The completion rate of Empathy questions is in the range 35%-37%, except for the question "When I see someone being treated unfairly, I sometimes don't feel very much pity for them" see the **Warning** in part 2.3).

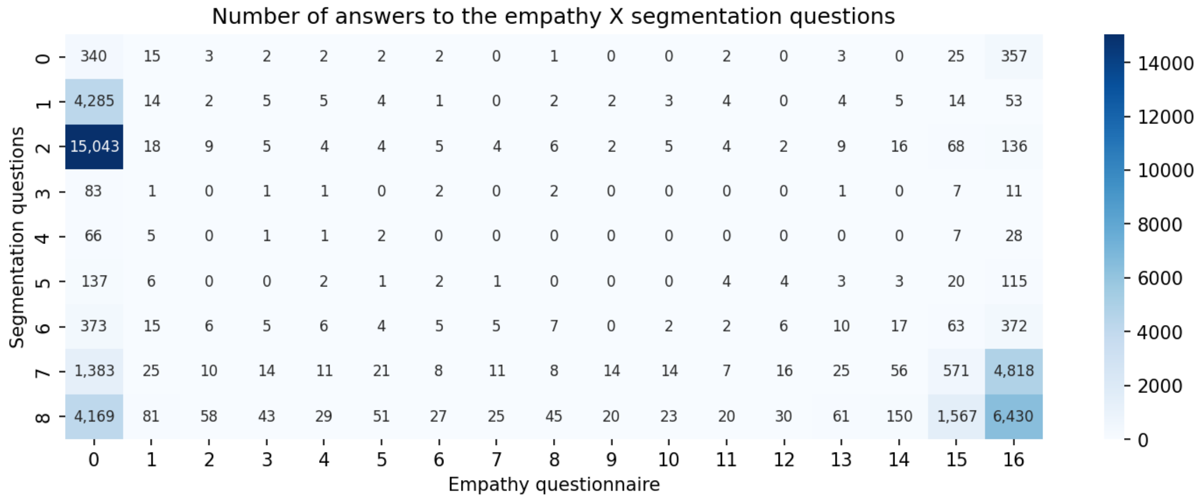


	Completion rate / all participants
I often have tender, concerned feelings for people less fortunate than me.	37.4%
I really get involved with the feelings of the characters in a novel.	37.2%
In emergency situations, I feel apprehensive and ill-at-ease.	37.0%
I try to look at everybodyâ€™s side of a disagreement before I make a decision.	36.9%
When I see someone being taken advantage of, I feel kind of protective toward them.	36.8%
I sometimes try to understand my friends better by imagining how things look from their perspective.	36.6%
After seeing a play or movie, I have felt as though I were one of the characters.	36.5%
Being in a tense emotional situation scares me.	36.4%
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	36.1%
I would describe myself as a pretty soft-hearted person.	36.2%
When I watch a good movie, I can very easily put myself in the place of a leading character.	36.1%
I tend to lose control during emergencies.	36.0%
When Iâ€™m upset at someone, I usually try to â€œput myself in his shoesâ€ for a while.	35.9%
When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.	35.8%
When I see someone who badly needs help in an emergency, I go to pieces.	35.7%
Before criticizing somebody, I try to imagine how I would feel if I were in their place.	35.7%

5.4.2 Cross tabulation with segmentation

The table below displays the number of participants according to the number of answered empathy $\times$ segmentation questions.

For instance, bottom right, 6,430 participants did answer to the whole empathy questionnaire (16) and to all segmentation questions (8).



Number of answers to the empathy X segmentation questions

Segmentation questions \ Empathy questionnaire	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	340	15	3	2	2	2	2	0	1	0	0	2	0	3	0	25	357
1	4,285	14	2	5	5	4	1	0	2	2	3	4	0	4	5	14	53
2	15,043	18	9	5	4	4	5	4	6	2	5	4	2	9	16	68	136
3	83	1	0	1	1	0	2	0	2	0	0	0	0	1	0	7	11
4	66	5	0	1	1	2	0	0	0	0	0	0	0	0	0	7	28
5	137	6	0	0	2	1	2	1	0	0	0	4	4	3	3	20	115
6	373	15	6	5	6	4	5	5	7	0	2	2	6	10	17	63	372
7	1,383	25	10	14	11	21	8	11	8	14	14	7	16	25	56	571	4,818
8	4,169	81	58	43	29	51	27	25	45	20	23	20	30	61	150	1,567	6,430

6. Acknowledgement

Author: Francis Wolinski (Yotta Conseil)

Date: 9 November 2023

Version: 1.1