AN ISTAT SURVEY ON CHILDREN AND YOUNG PEOPLE: ANALYZING FEEDBACK FROM RESPONDENTS[1]

Cinzia Conti, Gabriella Fazzi, Barbara D'Amen, Marco Rizzo

Abstract. Since 2014, Istat has been carrying out surveys in which young people are protagonists. The new edition of the Survey "Children and young people: behavior, attitudes, and future projects" (year 2023) was aimed at collecting information on the living conditions of Italian adolescents and adolescents from a migratory background aged between eleven and nineteen years. In an era of crisis and profound social change, such as the current post-pandemic period, there is a need to further enrich the statistical information collected and analyzed for the implementation of policies that enhance the human capital of the very young and improve their well-being. One initial goal of this contribution is to discuss the innovations introduced in the statistical process to reach young people. Specific strategies adopted by Istat include a web questionnaire optimized for any kind of device (including smartphones), the use of QR code to directly access the form and the availability of the questionnaire in Italian and nine other languages. The main objective of the contribution is to analyze, through textual analysis, the suggestions gathered in an open-ended question placed at the end of the survey. We will use statistical models like Analysis of Specificities, Sentiment Analysis and Cluster Analysis to examine term occurrences, identify the main concepts, and retrieve semantic relationships between them.

1. Introduction

Social science researchers are increasingly surveying young adolescents, but there is limited methodological knowledge on this age group. Much of the existing guidelines come from studies and theories focused on adults (Omrani et al., 2019; de Leeuw, 2011). In 2023 Istat carried out the survey on children and young people: forms of behavior, attitudes and future plans making them the primary respondents (without proxies). The literature, in fact, encourages the use of questionnaires for children starting from the age of eleven, with suitable precautions put in place (Borgers et al., 2000). Many innovations have been introduced in order to minimize the statistical burden on children and to maximize data protection and confidentiality.

[1] This article is the work of the authors. Paragraph 1 was written by Cinzia Conti, paragraph 2 was written by Cinzia Conti and Marco Rizzo, paragraph 3 was written by Marco Rizzo, paragraphs 4, 5, 6, were written by Barbara D'Amen and paragraph 7 was written by Gabriella Fazzi. Conclusions were written jointly by the authors.

The survey is based on a "light" questionnaire[2] that can be comfortably filled out even on a smartphone. The questionnaire includes sections on: demographic information; school life; citizenship and identity; social relationships; opinions about the future; and opinions about men and women. Given the innovative nature of the process, some questions about the respondents' opinions on the survey and the structure of the questionnaire were included at the end. In particular, there is a final open-ended question in which we asked, "*Do you have any suggestions for us for improving the questionnaire?*". Istat was indeed interested in gathering advice directly from their young respondents to enhance the data collection process. In this contribution, we wish to focus specifically on that open-ended question. We are interested in studying both the factors that can influence the propensity to answer (through regression models) and the content of the answers (textual analysis). The literature suggests that open-ended questions in surveys allow respondents to freely express their opinions, adding depth to the results and often providing more reliable measurements than closed questions. However, they require better cognitive and communication skills, which can affect response reliability. Open-ended questions also have a higher likelihood of "don't know" answers and non-responses, especially among young adolescents (Omrani et al., 2019; de Leeuw, 2011).

The first general hypothesis is that technological innovations may make surveys more appealing to young people. This paper aims to present the opinions of young people on the proposed questionnaire, the methods of data collection, and the topics introduced in the form.

The second – and more specific – hypothesis is that the socio-economic characteristics of the respondents and their families, as well as their relational networks, can influence both the propensity to respond to open-ended questions and the content of their answers.

2. Data and methods: a "smart" inclusive survey

Istat started conducting surveys on young people in 2015 with a survey on the integration of second generation migrants. In 2021, it conducted a survey focused on the issues faced by students aged eleven to nineteen during the COVID-19 pandemic. The new edition of the Survey on children and young people (year 2023) aims to collect information on behaviours, attitudes and living conditions of Italian children and children from a migratory background (both born in Italy and immigrants at a very young age) aged between eleven and nineteen (Istat, 2024). The aim of the survey is to highlight both weaknesses and the strengths of the new generations, to allow for a better valorisation of their energies and abilities in terms of policies and

[2] https://www.istat.it/fascicoloSidi/1542/Questionario%20italiano%20(Facsimile).pdf

actions. The research was set up, right from the early planning stages, with adolescents at the centre as active and participating subjects. Children represent a fundamental social subject for building the future and in an era of crisis and profound social change, there is a need to further enrich the statistical information collected and analysed for the implementation of policies that enhance the human capital of the young and to improve their well-being (Conti et al., 2024; UNECE, 2023).

The survey was conducted between 2 October and 20 December 2023 and involved a representative sample of almost 20,500 Italian young people and almost 18,000 young foreign residents in Italy[3].

The survey was carried out using the CAWI technique (Computer Assisted Web Interviewing)[4]. Among the innovations introduced by Istat in this data collection dedicated to young people it is essential to note that the questionnaire could be taken by scanning a QR code. It could also be submitted via PC, smartphone or tablet by accessing the link reported and typing the access code printed in the informative letter. Most of the sample preferred to fill out the questionnaire using a smartphone or a tablet rather than a PC. There is a difference between foreign nationals and Italians: 74.9% of Italians completed the questionnaire via smartphone or tablet compared to 83.1% among foreign nationals. The possibility of accessing via QR code favoured the use of these devices over PCs. Among those who answered through smartphone or tablet, 86.7% used the QR code. Instead, in the 2021 edition of the Survey on children and young people, where the questionnaire could only be accessed through a link, just 49 percent of users filled it out via smartphone or tablet. The introduction of the access via QR code could boost the number of respondents, especially for foreign nationals or for those who do not own a PC.

The survey is based on the consideration that the younger generations are "digital natives" (Sadiku et al., 2017). In the design of the survey internet, social media and smartphones play a central role. The questionnaire is moderately light: it takes about 23 minutes to fill out.

Another important aim of the survey was to be "inclusive" and also to collect information about the most vulnerable groups. To meet the language difficulties of young immigrants, for the first time in the history of Istat, the web questionnaire was available not only in Italian but in nine other languages: Albanian, Arabic, English, French, German, Mandarin, Romanian, Slovenian, Spanish and Ukrainian.

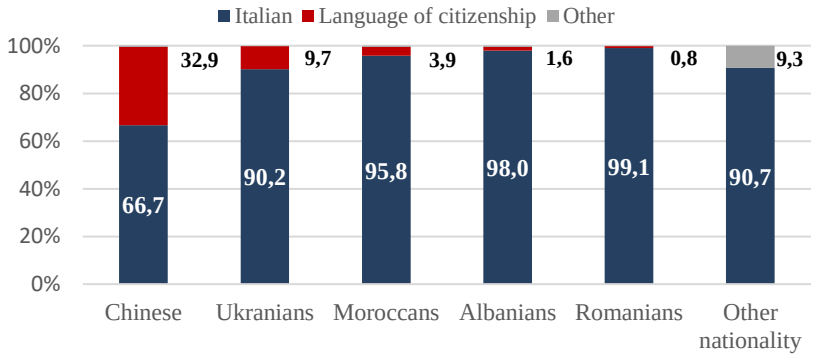
Apart from German that is spoken by national minorities in Italy, Mandarin was the most popular foreign language to be chosen, English was the second and the third

[3] The number of children selected based on optimal allocation was set, with some margin for over-sampling, at approximately 108,000. Stratification was defined by crossing the categories of region, age groups, and citizenship, resulting in a total of 588 levels (Istat, 2024).

[4] The decision to introduce the CAWI technique in population surveys makes it possible to contain the costs of public statistics and to exploit the potential offered by technology to capture segments of the population that are increasingly elusive compared to traditional techniques. (Istat, 2017)

was Ukrainian. About 1,700 foreign respondents filled out the questionnaire in a language other than Italian. Figure 1 shows the questionnaire language selected by respondents with respect to major citizenships. 32.9% of young people from a Chinese background decided to fill out the questionnaire in Mandarin Chinese (more than 800 questionnaires). This fact evidences low integration among young people from a Chinese background, even though more than 87% of the Chinese youth in the sample were born in Italy. Another interesting result is that just 9.7% of Ukrainian respondents selected Ukrainian. This fact is explained by the conflict in Ukraine. In our sample 301 Ukrainian children arrived in Italy for the first time in 2022, compared to an average of 73 new entrants per year for the previous nine years.

Figure 1 – *Distribution (%) of the respondents (11-19 year old) by selected language for the questionnaire and citizenship, year 2023.*



Istat, Survey on children and young people 2023

These results show us that giving respondents the opportunity to fill out a questionnaire in their preferred language facilitates the response of poorly-integrated communities. It also facilitates responses in emergency cases such as a forced migrant community in a country with a different language.

As previously mentioned, the propensity to answer was studied through logistic regression, where the dependent variable represents whether or not the respondents answered the open-ended question. The individual characteristics of the respondents were examined using structural variables: gender, territorial breakdown, citizenship, and age group. Socio-economic characteristics were assessed through the variables: shared bedroom; and mother's educational qualification[5]. Information regarding relational networks was measured using the variables: bullying indicator; school

[5] The mother's educational qualification is as an indicator of the family's socioeconomic status. We preferred this indicator over personal perceptions of their family's economic condition because the latter was not significant and to avoid issues with correlations between independent variables.

performance; frequency of seeing friends; and average hours a day on social media. All variables were derived from questions put to the respondents. The textual analysis, addressed in paragraph four, examines: the content of the answers; the techniques; lexical analysis; sentiment analysis; analysis of positive specificities; and cluster analysis.

3. Factors influencing the propensity to comment on the survey and the questionnaire

To support the textual analysis that will explore the suggestions shared through the above noted open-ended question, we investigated the factors that could most influence the respondents' willingness to answer the open-ended questions related to their opinions on the survey.

The analysis of the results, carried out via a binary regression model, focuses on the hypotheses mentioned in the first paragraph, considering the response variable of the logistic model as 'responds or does not respond' to the question "*Do you have any suggestions for us for improving the questionnaire?*"[6]. The odds ratios do not show values that are extremely distant from one. But all variables show a good level of significance, with p-values never above 0.01. The structural variables on individual characteristics included in the model show a higher propensity to answer among the younger age group: the older the respondents, the lower the propensity to answer the open-ended question. Interviewers with Italian citizenship are slightly more likely to share suggestions compared to foreign students. Gender and territorial breakdown do not have significant effects.

Among the socio-economic characteristics "sharing a bedroom" seems to positively affect the attitude to answer the question. Those who share a room with a relative are 17% more likely to answer than those who have a room for themselves. In the case of children with no brothers or sisters, it is more difficult for them to share a room with other people, but the variable is meant to measure the habit of relatedness and living conditions. As to the mother's educational qualification, we note that the higher the mother's level of education, the lower the propensity to answer the open-ended question.

To explore relational networks, we first examined the issue of bullying. The bullying indicator variable was constructed from seven questions in which the respondent was asked whether, in the past twelve months, he or she had been offended, threatened, ostracized, or defamed, either in person or online. If the respondent had experienced at least one of these aspects monthly, the "bullying

[6] Binary logistic model with logit distribution function. The independent variables are: citizenship, territorial breakdown, age group, gender, shared bedroom, mother's educational qualification, bullying indicator, school performance, frequency of seeing friends and average hours a day on social media.

indicator" is one. This aspect is the most important factor in the decision to answer the open question; youths who have experienced these forms of discrimination were 25% more likely to answer the open question than those who have not. The respondent's level of interaction on social networks and in person were also included in the study; these two aspects seem to go in opposite directions. The less often respondents meet their friends, the less likely they are to respond to the open-ended question. In contrast, those who spend only one or two hours on social media are more likely to respond to open-ended questions compared to those who spend more than three hours. Although the intensity of the estimates is very low, high use of social networks seems to be synonymous with slight isolation, rather than with strong friendships.

Table 1 – *Odds ratios, confidence intervals and p-values of logistic regression on the propensity to answer the open-ended question.*

Variable (ref.)	Odds ratio	Confidence interval	Pr(>|z|)
Mother university (middle school or lower)	0.91	0.86-0.96	***
Mother high school (middle school or lower)	0.95	0.91-0.99	*
Shared bedroom (not shared)	1.17	1.12-1.22	***
High bullying indicator (low)	1.25	1.19-1-32	***
Low school performance (high)	1.09	1.02-1.16	**
Sees friends less than once a week (everyday)	0.88	0.83-0.94	***
Sees friends several times a week (everyday)	0.89	0.84-0.94	***
4+ hours on social (< 1 hour)	0.85	0.78-0.92	***
3-4 hours on social (< 1 hour)	0.85	0.79-0.92	***
1-2 hours on social (< 1 hour)	0.91	0.84-0.98	*

*Note: levels of significance * p<0.05, ** p<0.01, *** p<0.001*

4. Textual analysis of open-ended question

The "Do *you have any suggestions for us for improving the questionnaire?*" open ended question included in the survey "Children and Young People" was analysed through a lexicometric approach via textual analysis.

The answers to the open-ended question were collected in a corpus that underwent a double stage analysis. In the first stage, the corpus was pre-processed and imported into the software TaLTaC2 to carry out a lexical analysis aimed at identifying positive specificities (Bolasco, De Mauro, 2013; Lafon, 1980; Lebart, Salem, Berry, 1997) and sentiment analysis. Following this first stage, the analysis of positive specificities allows for the detection of the main themes related to the different ages of the respondents, while the sentiment analysis revealed the presence of both positive and negative opinions expressed through the use of adjectives. In

the second stage, after lemmatization of the corpus, a cluster analysis was carried out.

According to this analytical strategy, in the first stage of the analysis, the corpus was prepared by removing special characters and replacing uppercase with lowercase letters. Hence, given that the software used for the analysis (TaLTaC2) allows the text to be processed in Italian, all the responses written in other languages were removed. Following this approach, the corpus analysed consists of 15,096 fragments (38.5% of respondents). This corpus was imported into TaLTaC2 and pre-processed by applying a procedure of parsing, normalization and lexicalization, in order to reduce the redundancy and to provide homogeneity among forms. After this preliminary procedure, the corpus consists of 86,980 word tokens and 6,568 word types (see table 2). In order to verify whether the textual data could be statistically processed, the type-token ratio lexical indicator was calculated (see table 2).

Table 2 – *Main characteristics of the textual corpus composed of the answers to the open-ended question.*

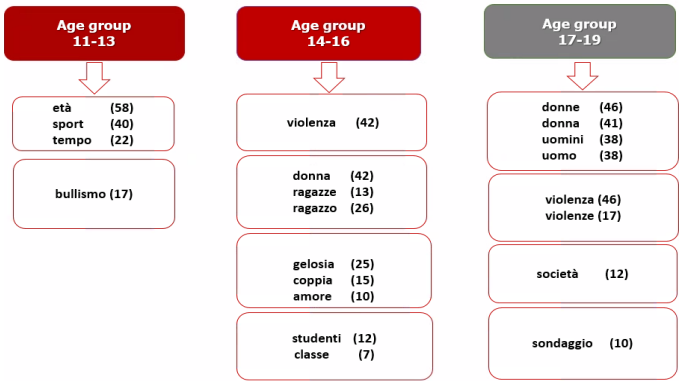
Lexicometric measurements	
Word types (V)	6,568
Word token (N)	86,980
Hapax (V1)	3,512
Type-token ratio index (V/N*100)	7.55%

According to the size of the corpus, the type-token ratio index, since it is 7.55% and < 20% (Bolasco, 1999, p. 203) shows that the corpus can be subjected to quantitative analysis.

5. The positive specificities

In order to investigate lexical differences related to the different ages of the respondents, we analysed the positive specificities for the age variable. In particular, specificity can be used to spot items that are both over- and under-represented in a corpus by applying a statistical test (Lafon, 1984, pp. 65-66). In our study, we are solely interested in positive specificities which highlight items that are over-represented in the corpus. Moreover, given that respondents were classified into three different age groups, we investigated positive specificities for the three modalities of the age variable. This analytical approach allows us to identify three different groups of words (positive specificities), as shown in figure 2.

Figure 2 – *Specificities for each age group from the textual analysis of the open-ended question. Year 2023.*



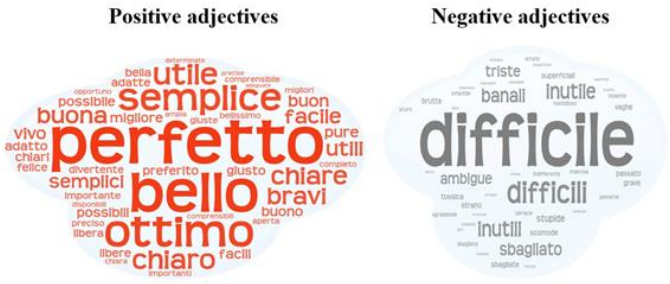
The first group characterizes the 11-13 age group of respondents that suggested improving the survey by introducing new themes such as bullying, cyberbullying, sports, free time, and road safety. Moreover, these respondents suggested using less direct questions applying age-appropriate language. The second group relates the 14-16 age group. The introduction of new themes to improve the survey is also suggested by these respondents including: video games, school life and geopolitical context, eating habits and relationship dynamics. They also expressed the need for a formal revision of the language in order to make it clearer. Finally, the third group characterizes respondents aged 17-19. In their opinions, the survey could have been improved by introducing new themes such as women's rights, men's mental health and themes related to current affairs. These respondents offered other suggestions for improving the survey, mainly related to the survey website and the survey technique. Regarding the website, they suggest modernizing it, in order to make it more suitable to the young age of the respondents. Moreover, they suggested completing the survey entirely online, without paper.

6. Sentiment analysis

Sentiment analysis was conducted using the software TaLTaC2 that automatically recognizes adjectives and classifies them as positive or negative (Bolasco, Della Ratta, 2004). Following this approach, the corpus underwent a grammatical tagging procedure in TaLTaC2 in order to identify the adjectives. This procedure allowed us to tag 429 adjectives, for a total of 2,479 word tokens. Regarding the adjectives, the sentiment analysis identified 61% positive adjectives and 39% negative. Thus, sentiment analysis highlighted a prevalence of positive

opinions of respondents. Figure 3 shows words clouds of both positive and negative adjectives.

Figure 3 – *Word cloud of positive and negative adjectives from the textual analysis of the open-ended question. Year 2023.*



Considering the positive adjectives, respondents defined the questionnaire as "perfetto" ("perfect", 437), "bello" ("beautiful", 109), "chiaro" ("clear", 106), "semplice" ("simple", 94) and "utile" ("useful", 51). On the other hand, for some other respondents the questionnaire was "difficile" ("difficult", 51), "inutile" ("useless", 22), "banale" ("banal", 10) and "sbagliato" ("wrong", 9). These adjectives provide a snapshot of the potential weaknesses of the survey, which will be explored with the following cluster analysis.

7. Cluster analysis

A cluster analysis was conducted (DHC- Descending Hierarchical Classification, Reinert method (1990), IRaMuTeQ software[7]) that allowed for the identification of eight groups of respondents, where the first three groups represent 98.4% of the sample. We will focus our analysis on three big groups, characterized by different socio-demographic aspects and the topics reported in their responses, identified through the lexical content of each cluster. Each cluster is characterized by the use of specific words. The first group (31.4%) has been named "the questionnaire methodologists" for their suggestions aimed at setting up specific actions on the questionnaire, such as making it shorter or improving the graphics. This group includes 33% of the respondents. They are mostly aged between eleven and thirteen, they tend to be male and foreign. Moreover, they use social media one or two hours a day. Economically, they are students from wealthier backgrounds. The second group (34%), named "the qualitative methodologists" asked for more opportunities

[7] We express our gratitude to Francesca Della Ratta for her invaluable assistance in conducting the IRaMuTeQ data analysis.

to express themselves freely: they would like open-ended questions, suggest replacing true/false responses with agree/disagree to better express their opinions, and they wanted more ways to respond to the closed-ended questions. In this analysis, we'll focus our attention on the third group (33.03%), named "the thematic" for their suggestions for introducing new topics. They are mostly female, aged between 14-19 years, and use social media less than an hour a day. They generally reported a low economic status. The words that characterize this group are primarily related to the everyday life of the young, such as family and school.

Table 3 – *Words that mostly characterize the "thematic" cluster in the textual analysis of the open-ended question. Absolute values, percentage values and Chi-square. Year 2023.*

Word	Eff. Tot.[8]	Perc.[9]	X^2[10]	Word	Eff. Tot.[3]	Perc.[4]	X^2[5]
Violenza	104	95.2	185.1	Sicuro	15	73.3	11.0
Bullismo	22	95.4	38.9	Lingua	31	93.5	51.6
Abuso	10	100.0	20.3	Straniero	33	90.9	50.2
Subire	7	100.0	14.2	Cittadinanza	28	92.9	5.5
Vittima	6	100.0	12.2	Razzismo	17	94.1	8.7
Discriminazione	6	100.0	12.2	Sessuale	28	92.9	5.5
Cyberbullismo	5	100.0	10.1	Lgbt	11	100.0	2.3
Mentale	37	100.0	75.5	Sessualità	8	100.0	6.2
Salute	30	96.7	55.2	Sport	52	71.1	4.4
Psicologico	20	100.0	40.7	Ambiente	27	96.3	9.0
Paura	13	92.3	20.7	Droga	14	100.0	8.4
Disagio	9	88.9	12.7	Guerra	12	100.0	4.3
Felice	13	76.9	11.3	Inquinamento	5	100.0	10.1

The analysis shows that the word "futuro" (future) stands out (Perc=95.6; X^2=244.4). Young people are asking more questions about their future projects, but they are also expressing their fears and insecurities.

"Ask questions that are also inherent to fear of the future, understood as climate fear, fear of an uncertain future because of the climate and the collapse of the planet. But also fear of the instability of Italian labor, fear of not being able to have a future."

They are worried about: violence, abuse, bullying, and cyberbullying. Additionally, they also express concerns about themes that usually are not asked to

[8] Number of text segments containing the cited word in the corpus at least once.
[9] Percentage of word occurrence on the text segments of this cluster for its occurrence in the corpus.
[10] Association between word and cluster.

this age group, such as mental health, fear, discomfort, racism, and sexuality. Taking a broader view, there is also a focus on collective issues, with words connected to the environment and war.

8. Conclusions

The study of the propensity to answer the open-ended question on suggestions for the survey shows that, among individual characteristics, age has the greatest effect: the younger the respondent, the greater the propensity to answer. Regarding socioeconomic characteristics, we note that a less comfortable situation, such as sharing a room, has a positive and significant effect on the propensity to respond. Finally, regarding relational networks, it is important to emphasize that adolescents experiencing bullying are more inclined to offer suggestions. The analysis indicated that the survey "Children and Young People 2023" conducted by Istat was largely appreciated by respondents, as highlighted by the sentiment analysis, which shows a higher frequency of positive adjectives in the responses of the youngest, expressing their positive attitude toward the survey. The use of adjectives, in the open-ended question, such as "perfect", "beautiful", "clear" and "useful" confirms respondents' positive opinions. This positive attitude was also confirmed by the low frequency of negative adjectives in the analyzed texts. These findings suggest that the survey should be repeated in the future, maintaining the innovations proposed in the 2023 edition, like the introduction of QR code. However, the analysis of the words included in the responses to the open-ended question identified some improvements, such as the inclusion of new topics, closer to the everyday life of a younger person and to geopolitical and social context. The introduction of these new topics could be added by taking into account the age of the respondents, in order to better represent their experiences and their sensibility.

References

BOLASCO S. 1999. *L'analisi multidimensionale dei dati*. Roma: Carocci Editore.

BOLASCO S., DELLA RATTA-RINALDI F. 2004. Experiments on semantic categorisation of texts: analysis of positive and negative dimension. In *Le poids des mots, Actes des 7es journées Internationales d'Analyse Statistique des Données Textuelles*. Louvain-la-Neuve: UCL, Presses Universitaires de Louvain, pp. 202-210.

BOLASCO S., DE MAURO T. 2013. *L'analisi automatica dei testi: fare ricerca con il text mining*. Roma: Carocci Editore.

BORGERS N., DE LEEUW E., HOX J. J. 2000. Children as Respondents in Survey Research: Cognitive Development and Response Quality. *Bulletin of Sociological Methodology/Bulletin de Mathodologie Sociologique*, Vol. 66, No. 1, pp. 60–75.

CONTI C., FANFONI L., FAZZI G. 2024. Giving children and young people a voice: the experience of the sample survey on children and young people in Italy. *Working Paper n.1*, Expert meeting on statistics on children Geneva, Switzerland, 4-6 March 2024.

DE LEEUW E. D. 2011. Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting. In Report prepared for the Annual Meeting of the Academy of Finland: Research Programs Public Health, Finland.

ISTAT 2024. *Nuove generazioni sempre più digitali e multiculturali*, Statistica Report, 20 maggio 2024.

ISTAT 2017. *L'utilizzo della tecnica cawi nelle indagini su individui e famiglie*, eBook letture statistiche.

LAFON P. 1980. Sur la variabilité de la fréquence des formes dans un corpus, *Mots*, Vol. 1, No 1, pp. 127-165.

LAFON P. 1984. *Dépouillement et statistiques en lexicométrie*. Paris: Slatkine Champion.

LEBART L., SALEM A., BERRY L. 1997. *Exploring textual data (Vol. 4)*. Dordrecht: Springer Science & Business Media.

LIU B. 2012. *Sentiment analysis and opinion mining (synthesis lectures on human language technologies)*. San Rafael: Morgan & Claypool Publishers.

OMRANI A., WAKEFIELD-SCURR J., SMITH J., BROWN N.. 2019, Survey Development for Adolescents Aged 11–16 Years: A Developmental Science Based Guide, *Adolescent Res Rev* 4, pp. 329–340

REINERT M. 1990. Alceste, une méthodologie d'analyse des données textuelles et une application: Aurélia de G. de Nerval, *Bulletin de méthodologie sociologique*, Vol. 28, pp. 24- 54.

SADIKU M., SHADARE A., SARHAN M., 2017. Digital Natives. International Journal of Advanced Research in Computer Science and Software Engineering.

UNECE. 2022. Guidance on statistics on children: spotlight on children exposed to violence, in alternative care, and with disabilities.

Cinzia CONTI, Istat, ciconti@istat.it
Gabriella FAZZI, Istat, fazzi@istat.it
Barbara D'AMEN, Istat, barbara.damen@istat.it
Marco RIZZO, Istat, marco.rizzo@istat.it