

DeZIM Afghans in Turkey Survey

Method & Data Report

Version 1.0.0 | June 2025



Editor

Deutsches Zentrum für Integrations- und Migrationsforschung (DeZIM) e.V.

Forschungsdatenzentrum DeZIM.fdz | Mauerstraße 76 | 10117 Berlin

+49 (0)30 804 928 93 | fdz@dezim-institut.de

Responsible

Ramona Rischke

Zeynep Yanaşmayan

Authors

Ramona Rischke (DeZIM-Institut)

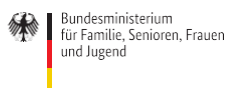
Sefa Sahin

Roya Jahanbakhsh

Citation

Rischke, Ramona; Sahin, Sefa & Roya Jahanbakhsh (2025): Afghans in Turkey Survey. Data and Method Report (Version 1.0.0). Deutsches Zentrum für Integrations- und Migrationsforschung (DeZIM). Berlin.

Funding through BMFSFJ



Content

1	Introduction	1
2	Overview	2
3	Research tools	3
3.1	Questionnaire content.....	3
3.2	Pretests	3
4	Sample design	5
4.1	Sample design.....	5
4.2	Recruitment Process.....	5
4.3	Field phase.....	6
5	Data preparation.....	7
5.1	Respondent roster	7
5.1.1	Processing of raw data	8
5.2	Household and family roster	9
5.2.1	Combining data stored in different raw datasets.....	10
5.2.2	Constructing variables of interest	10
5.3	Coding of missing values.....	14
5.4	Coding of open-ended data	15
6	Anonymization	19

List of Tables

Table 1: Project overview	2
Table 2: Questionnaire contents	3
Table 3: Variables and value labels in the household and family roster.....	12
Table 4: Overview of observations, responses and filterings.....	14
Table 5: Coding of missing values.....	15
Table 6: Overviews codes for quali_1_categ.....	17
Table 7: Overview codes for quali_2_categ	18
Table 9: Anonymization.....	19

1 Introduction

Within the project “Beyond onomastics – New methods for reaching small, hard-to reach and/or hidden groups” of the German Centre for Integration and Migration Research (DeZIM) the “DeZIM Afghans in Turkey Survey” collected data that allows investigating migration and integration experiences of Afghan refugees in Turkey.

The survey instrument of the study was developed by the German Centre for Integration and Migration Research (DeZIM) in collaboration with researchers from the TRANSMIT project (Transnational Perspectives on Migration and Integration) of the DeZIM Research Community and in consultation with the practical implementation partners. The practical implementation of the study was carried out by YÖNTEM Research Consultancy Ltd. Co.

Within the framework of the project funded by the German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth, the data has been processed, anonymised, and documented in collaboration with the project staff for the purpose of data re-use and made available as Scientific Use Files (SUF). Access to DeZIM data is regulated by law. Prerequisites for the use of a SUF are a scientific research purpose, employment at a scientific institution and the conclusion of a data use contract. In addition to the data sets, documentation materials on the data sets are provided. This method and data report is part of the documentation for the present study, along with the survey instrument that is provided in a separate document.

In the second chapter of this method and data report, the central data and working modalities of the study are summarised in tabular form. This is followed by a discussion of the questionnaire content in chapter 3, the sampling design and field phase in chapter 4, and the data processing in chapter 5. A description of the individual steps of data anonymisation can be found in chapter 6.

2 Overview

For quicker orientation, the report is preceded by a brief overview of the most important data and working modalities of the study.

Table 1: Project overview

Title	DeZIM AFGHANS IN TURKEY SURVEY
Responsible institution	DeZIM-Institute
Surveying institutes	YÖNTEM Research Consultancy Ltd. Co.
Funding	German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth
Project team	Ramona Rischke, Zeynep Yanaşmayan, Roya Jahanbakhsh, Sefa Sahin, Mariel McKone Leonard
Total population	Afghans in the pre-selected cities of Turkey
Survey method	CAPI interview
Survey period	15 March 2022 – 7 April 2022
Sample	774 individuals
Selection procedure	Purposive high variance sampling in selected public locations known to be frequented by Afghans and regulated network sample
Survey documents	Standardised questionnaire
Interviewers	19 interviewers
Citation of the data	Rischke, Ramona; Yanaşmayan, Zeynep; Jahanbakhsh, Roya; Sahin, Sefa; Leonard, Mariel M. (2024): DeZIM Afghans in Turkey Survey. Data set, Version: 1.0.0. SUF download Berlin: German Centre for Integration and Migration Research (DeZIM).
DOI – Data access route	https://doi.org/10.34882/dezim.transmit2.download.1.0.0 https://doi.org/10.34882/dezim.transmit2.remote.1.0.0 https://doi.org/10.34882/dezim.transmit2.onsite.1.0.0

3 Research tools

3.1 Questionnaire content

The questionnaire was developed by the project team at the DeZIM-Institute (see Table 1) in collaboration with researchers from the TRANSMIT research team at the Humboldt-University of Berlin (Dr. Simon Ruhnke & Dr. Nader Talebi) and finalised in consultation with the survey company. In addition to socio-demographic information, the questionnaire includes the sections that are listed in Table 2.

Table 2: Questionnaire contents

Sections	Content
Section 1: Introduction & consent	Date and logistics
Section 2: Household and family roster	Demographics of respondent, household members and selected family members outside the household
Section 3: Respondent well-being and employment situation	Household income, relative wellbeing and employment situation of respondent
Section 4: Respondent's education	Respondent educational attainment and aspirations
Section 5: Ethnicity, identity & life satisfaction	Ethnic background(s), identities and life satisfaction
Section 6: Family displacement and respondent migration history	Migration trajectories, reasonings, detentions
Section 8: Discrimination experiences	Discrimination & solidarity
Section 9: Migration aspirations	Migration aspirations
Section 10: Qualitative open-ended questions	Discrimination and hope
Section 11: Social integration	Belonging and social interactions
Section 12: Psychological items	Risk, uncertainty, locus of control
Section 13: Respondent's health	Health & vaccination access
Section 14: Neighbourhood characteristics & internal migration	Neighborhood characteristics and internal migration
Section 15: Telephone use	Telephone & internet usage
Section 16: Post-interview questions for enumerators	Post-interview questions for enumerators

3.2 Pretests

A pretest took place during the interview training under field conditions. The main purpose of the pretest was to evaluate the fluidity of the interview process and the understanding of the questions. Yöntem reported¹ that questions related to the migration history have been answered with some hesitations:

¹ Technical Report provided by Yöntem and summarized here.

“For instance, during fieldwork visits in Izmir by the supervisor team, it was observed that one of the respondents did not even mention their stay in Germany while answering the questionnaire. However, while chatting with them after the questionnaire, they mentioned their stay in Germany only in a friendly chat context. When it was asked why they did not tell anything about their stay in Germany during the questionnaire, it is understood that the family stayed illegally in Germany for a while and they would like to go back to Germany from Turkey again within a few years, they are in Turkey for transition, therefore, he would not like to mention about their stay. It was observed that the questions which are related to the Turkish community lead to hesitations in answering them as well. The respondent would not like to give answers for them to ensure their security status in Turkey. During the field team members chats with the participants after interviewing, the participants always asked about the reasons of asking that much different aspects and detailed questions.”

As a result of the pretest, several questions were revised. All questions related to family members as well as the Turkish host society were critically reviewed to keep them to an absolute minimum. In addition, more extensive introductory statements were provided to increase the transparency towards respondents in terms of explaining *why* we are collecting detailed information.

4 Sample design

The aim of the survey is to shed light on Afghans in Turkey, with a particular focus on Afghan refugees. Due to the oftentimes unsecure legal status that many Afghan refugees are exposed to in Turkey, they need to be considered a hard-to-reach population. To the best of our knowledge, there are no population statistics available for the target group that could have been used as a sampling frame.

4.1 Sample design

In close collaboration with the survey company, the sample of the present study used a purposive sampling to reach Afghan refugees. To increase the chances of reaching the targeted population, cities were pre-selected such that they feature a relatively large known presence of Afghan refugees. These cities included: Istanbul, Izmir, Hatay, Gaziantep, Şanlıurfa, Bursa, Van, Ankara. The population information has been collected by the survey company based on qualitative information provided by knowledgeable local informants. The targeted districts have been determined similarly, again to increase the potential for reaching Afghan refugees. Yöntem summarized the process as follows²:

“Before the study, 8 cities have been determined while Şanlıurfa has been eliminated after one and a half-day screening of Afghan refugees. In Şanlıurfa; firstly, some streets with a high population of refugees have been screened, and secondly, after meetings with local authorities like mukhtars, the migration office, it is understood that the only region that the field team could recruit refugees is Ceylanpınar which is a border place and which is risky in terms of security for the field team. Hence, it is decided to move Şanlıurfa’s N=90 target to the other 4 cities already in the sample. Those cities were chosen based on their potential of having Afghan groups.”

4.2 Recruitment Process

For recruitment within the selected cities and districts, snowball sampling techniques have been employed and the initial recruitment was either done in public spaces as well as through networks given that the team of enumerators comprised Afghans. To minimise potential network biases in the sample, some restrictions have been put in place before the study. In particular, each respondent and enumerator could refer to a maximum of 3 other Afghan refugees who were neither from their household, school, nor workplace.

Yöntem summarized the recruitment process as follows³:

“Recruitments have been done mainly in the streets as it was thought that Afghan refugee households could not be reached through the screener [comment: e.g. through random walks]. As expected, it was observed that conducting surveys on street has some special challenges like the current weather

² Technical Report provided by Yöntem and summarized here.

³ Technical Report provided by Yöntem and summarized here.

conditions, the presence of strangers around, and the length of the questionnaire compared to household surveys which could be more comfortable in terms of taking the participants time and privacy needed. The very initial plan of the sampling regions has been discussed in summer times while the fieldwork has been done in winter conditions, it is observed that recruitment processes have been also affected by this in terms of reaching target people in the streets.

On the other hand, it was also seen that using the snowball technique allowed the team to make some household visits for study as well. Going to the refugee's house with a reference – through other participants who they know – has solved the trust issue by the participant. It is observed that the Gaziantep is the most welcoming province in the sample in terms of allowing household visits.

Furthermore, especially in Van, another recruitment process has been successful in reaching the target population; pre-recruitment and inviting them to a central place on a given day and time to survey them consecutively. They were called not by our enumerators but through one of their acquaintances, and no hesitations for participating have been experienced by participants. The person who called them to the central location was also an Afghan refugee who lived in Van for many years and always support the participants in their daily lives through his social networks.

Yöntem team has conducted some extra callbacks, especially to those participants to assure the quality of the data collected via this methodology and the accuracy of the responses has been confirmed. The variety in demographic of those groups is also confirmed; there were young female respondents as well as older male respondents etc..."

4.3 Field phase

The field phase started on 15 March 2022 and ended on 7 April 2022. A total of 774 interviews were conducted for the survey that went through an initial quality check by the survey company Yöntem. The interviewer training as well as the supervision and monitoring during the field phase was organised and provided by Yöntem.

5 Data preparation

Survey answers are gathered into two different datasets which are 1) **respondent roster**, and 2) **household & family roster**. Unique observations in the **respondent roster** are individual respondents (unique observation identifier: “hhid”) while the **household and family roster** includes information provided by the respondent on the age, and gender composition of their household and family, as well as the whereabouts of selected family members. The unique observation for the **household and family roster** are individual households and family members (unique observation identifier: “hhid” + “indid”) that can be linked to the **respondent roster** through the unique respondent identifier (“hhid”).

Special codes and filters

In the data, filtering is used for two reasons. First, some questions are asked only for respondents who meet specific criteria that can be viewed in the questionnaire. For example, the question how many years respondents have attended school is asked only among those who previously reported that they have attended school at some point in their lives. As it can be seen from the questionnaire, such criteria are always based on former questions asked to respondents in the survey. Secondly, filtering is used in questions that can have more than one answer. The number of answers to these questions vary among respondents. Since each answer provided is stored in a separate variable, this implies that for some respondents, there are more variables than answers provided, and these are filtered, which can be thought of as a “non applicable” response. For example, the number of answers provided to the question of which countries respondent have lived more than 3 months vary among respondents. There are five variables to save answers separately since respondents gave a maximum of five country names. However, there are less than five answers given by most of the respondents such that a number of variables are “not applicable” to them and coded as filtered.

5.1 Respondent roster

The **respondent roster** includes all questions in the survey except specific questions used to construct the **household and family roster**. Specifically, information on the age, gender and residence of child(ren) and spouses of respondents are provided in the **household and family roster**. The **respondent roster** follows the same structure as the questionnaire. The order of the variables follows the questionnaire, and the headings are present as variables to indicate sections of the questionnaire that can be seen in Table 2 above.

The questionnaire (column “dataset”) provides information on where each variable can be found (“resp” for **respondent roster**, and “hhold/fam” for the **household and family roster**).

5.1.1 Processing of raw data

Handling of multiple response answers

In the first step, variables that allowed multiple answers were transformed. The main motivation for this is to provide categorical answers similar to the way in which the questionnaire shows response codes. In the raw data, however, multiple answers are coded in de facto dummy variables for each possible response. In the processed data, the fact that respondents do have the option to choose more than one answer is accounted for by defining multiple categorical variables (with suffix *_1* to *_max number of responses provided*). If only one answer to - e.g. languages spoken - was given by the respondents, then the response to the remaining second and third language spoken is coded as -3 (filtered). To clarify with another example, respondents received the question what their most important reasons were when they moved to Turkey in particular (variable *"migleb_why"*). The interviewers were instructed to ask this question openly and asked to code the response(s) given to them in 27 possible response codes including other (specify). In this specific example, answers given by respondents range from one to fourteen responses. The data transformation process entails creating 14 categorical variables (*"migleb_why_1"* to *"migleb_why_14"*) that each include the same 27 valid response codes. For each respondent, the number of responses is translated into an equal number of non-missing *"migleb_why_**" variables: If a respondent reported three reasons to have moved to Turkey, *"migleb_why_1"* to *"migleb_why_3"* will range from codes 1 to 27, whereas *"migleb_why_4"* to *"migleb_why_14"* are coded -3, i.e. as filtered.

Recoding of questions asking for countries (e.g. country of birth)

Some of the questions are related to countries. Respondents received questions such as which citizenship(s) they have and where they would like to stay permanently. All variables that include country-related answers are coded with 3 digits ISO code numbers of countries and country names are provided in the labels. For the questions concerning the country of last residence (*"migleb_start"*) and the country(s) where people have previously lived for more than 3 months aside from their country of birth (*"mig_dest"*), in addition to ISO codes, another 4 digit code was defined for the option of "never lived outside of Turkey or country of birth".

Generating additional variables of interest

In the raw data, information on spouses and children were provided in separate datasets and have been integrated into the **respondent roster** and **household and family roster**, respectively. The variables included in the **respondent roster** refer to how many child(ren) of respondent live together with them and how many of them live outside of their household. These are named as *"children_inside"* and *"children_out"*.

An additional variable has been generated to indicate the time spend in Turkey (*"duration_g"*) until the reference year 2022. The suffix of *"_g"* is used to indicate that this variable is generated rather than asked to

the respondents directly. The information used for creating this variable is the time since last arriving in Turkey (*"mingleb_lastentry_y"*) as well as the age of respondents (*"age_resp"*) if the respondents indicated that they have not lived elsewhere. Due to a substantial number of missing information regarding the month of last entry, *"duration_g"* is only constructed using the information on the year of last entry. Finally, *"longitude_2"* and *"latitude_2"* variables are generated that contains imputation for missing variables in the original *"longitude"* and *"latitude"* variables. Unfortunately, it is not possible to reconstruct exact locations for missing observations. Therefore, points within the neighbourhood where the survey is taken with the individuals were chosen as the location for them and indicated in the generated *"longitude_2"* and *"latitude_2"* variables. The interviews were predominantly taking place in public spaces and the geo-coordinates refer to the place of the interview.

5.2 Household and family roster

In addition to the **respondent roster**, a **household and family roster** is prepared based on the information provided by the respondent. The main feature of this data is that every individual in the household and the family of the respondent is a unique observation. While we acknowledge that there is no uniform definition of the term *"family"*, it was within the scope of our data collection to enquire about a small number of selected (nuclear) family members, i.e. spouse(s), child(ren) and parents in case that the respondent is below age 26. For brevity, we will refer to them as *"family"* in what follows.

The **household and family roster** consists of 4.351 observations in total and 3.819 observations for household members including respondents themselves. Respondents and the people who share the same house with the respondents are considered as household members. Every household is coded with a unique household id *"hhid"* that is also used as the unique respondent identifier. Household members can either include nuclear family members of respondents, other relatives or people who are not related to the respondent. As can be seen from the questionnaire, respondents were asked separately about the size and general composition of their households (in terms of their relationship and gender composition) as well as the age/gender/whereabouts of selected family members. In the **household and family roster**, information from different variables and raw datasets are combined.

The way in which the questionnaire was collecting information on the age, gender and whereabouts of different households and family members was informed by the knowledge interest to learn more about the age and gender distribution of Afghan households in Turkey, as well as the prevalence of separated (nuclear) families. At the same time, we were keeping to a minimum the information we enquired about respondents regarding other individuals. This is why we are only asking about the age categories of household members, for instance.

5.2.1 Combining data stored in different raw datasets

Answers for male respondents who have been asked about their spouses have originally been provided in a distinct dataset that was separate from the remaining respondent roster. This dataset contained responses by male respondents regarding if they have a spouse, the residence of their spouse(s) as well as their age should they live within the same household. These answers are transferred to the **household and family roster**. In total, 292 spouses of male respondents are present in the dataset. The age for 19 female spouses (living outside of respondent's households) is coded as -3(filtered) since their age was not enquired. In total, 73 spouses of female respondents are placed in the **household and family roster** as unique observations with the knowledge of their gender, age, and residence. Since respondents received question about the age of their spouses only if their spouse shares the same household, the age of 5 male spouses to female respondents was coded as filtered (-3). Finally, "*cowives_f*" and "*cowives_fnb*" variables showed whether female respondents' husbands have other spouse/s and how many they are. There is not any female respondent whose husband has another wife or wives. We subsequently dropped the variable "*cowives_fnb*" from the dataset.

All respondents have received questions about how many children they have, the age of their child/ren and where their child/ren live if they have any. Answers were provided in a distinct dataset separate from respondent roster. These answers are transferred to the **household and family roster**. In total, 1,071 children of respondents are recorded in the **household and family roster**.

Respondents received questions about where their father and mother live **only** if the respondent is younger than 26 years while not having a spouse or partner. The background of this design decision was to allow us constructing measures of family separation from spouses and children, while acknowledging that young respondents should also be considered separate from their "nuclear" family if they do not yet have formed a family (i.e. have a spouse/partner) on their own. Irrespective of these questions related to the whereabouts of their parents, all respondents answer whether their father and mother are part of their current household (variables "*hhcomp_3*", "*hhcomp_4*"). From the knowledge above, a total number of 286 mothers and 272 fathers are coded as unique observations in the **household and family roster**. Since there is no question about the age of mothers and fathers asked to respondents, the age variable is coded -3 (filtered) for all mothers and fathers of respondents.

5.2.2 Constructing variables of interest

Age group and gender of respondents

Except for 9 individuals from one household (with "*hhid*" variable "*TU1_4_78*"), the relation of all individuals with the respondent were provided. For those 9 members, -1 (refuse to answer) is used to indicate knowledge is not provided by respondents when it was asked. Those 9 members are either relative or nonrelative household

members. However, from the information provided by the respondent we cannot reconstruct which individual corresponds to which age-group and gender. Further details about the composition of the family and household for this specific case is given in the Table 4.

While the question about the household size and the questions about the family structure are consistent with one another for almost all respondents, there are two exceptions for the respondents with household id “TU1_4_105” and “TU1_4_8 “. To start with the first respondent, while the reports household size was 14, there are only 4 members that can be inferred from questions about structure of the household, namely a spouse, two children of the respondent and respondent himself. Consequently, there are either 10 individuals missing whose information cannot be inferred, or they are reported in excess. We decided to recode the household size as 4 considering this to be the most realistic scenario. In the second case (household id “TU1_4_8”), the respondent stated her household size to be 4, only 2 members can be inferred from the questions about the structure of household, which are one relative and the respondent herself. The respondent stated that she has 4 kids but refused to answer (-1) any questions about their age and specific whereabouts. However, she also said that she has children inside household. Therefore, we assume that 2 out of 4 kids are residing inside the household (based on the information of her household size) and record the whereabouts of the third and fourth child as well as details related to their age and gender as refused (-1).

Overall, information on the individual gender (variable “*gender*”) of household and family members cannot be reconstructed for 326 individuals. As it can be seen from the gender column of Table 4, in 188 out of 300 observations, the code is -3 (filtered) (instead of the code for refusal is used) since some information was indeed provided by the respondents, yet it is not possible to simultaneously infer the gender and age of non-nuclear family relatives as well as non-relatives in the household: Respondents received questions on how many relatives and non-relatives form one household with respondent, how many of them have which gender identity and how many of their relatives and non-relatives respectively, fall within which age intervals. Again, the purpose of this was to keep to an absolute minimum the information that was asked from respondents that related to other persons. We explained that we wanted to learn about the demographic composition of their households, yet it was not essential for us to collect information about every individual in the household separately. Consequently, it is only possible to infer both gender and age group of all household members at the same time if all relatives or non-relatives happen to share the same gender (such that the gender distribution for each age group is clear) or the same age group (such that the age groups for each gender category is clear). For those respondents who share a household with relatives or non-relatives from different genders, trade-off is made in favour of reporting age groups. Therefore, the gender of 188 relatives and non-relatives are coded as -3 (filtered).

Except for 749 observations (see Table 4), the age group of all other observations is included in the age variable. 585 out of 749 are coded -3(filtered). 585 of them are the mothers, fathers and spouses of respondents

who are not the member of household such that their age group was not enquired. The remaining 164 out of 745 observations are coded –1, since respondents chose to not provide responses to questions related to the age group of the respective household or family members.

Whereabouts of family members (variable “*residence*”)

A variable was constructed to indicate where members of the family live (variable “*residence*”), and in which country they live in if they reside outside of Turkey (variable “*residence_d*”). Except for 41 observations, knowledge of residence is known for all observations (see Table 4). If the individual household or family member resides in Turkey, the “*residence_d*” variable is coded –3(filtered). For mothers and fathers who live in a different country than the respondent, “*residence_d*” variable also gets –3(filtered), since the country was not asked in these cases.

Relationship to the respondent

In the **household and family roster**, the variable “*relation_g*” is constructed to indicate how a particular family or household member is related to the respondent. For the respondent him/herself, the relation variable takes the value “respondent”.

The variables and the associated labels used in the **household and family roster** are provided in the Table 3 below.

Table 3: Variables and value labels in the household and family roster

Variable	Explanation	Values (Labels)
hhid	unique respondent id	TU1_#_+ (such as TU1_2_23)
indid	unique individual id for every family or household member within each respondent's household	hhid_# (# is unique number to indicate observation, ranging from 1 to number of household members)
hhmem	dummy variable which shows whether an individual observation refer to a person residing inside or outside of the respondent's household	1 (if observation is household member) 0 (otherwise, i.e. family member outside the household)
relation_g	shows the relationship of individual household and family members to the main respondent	1 (respondent) 2 (spouse) 3 (child) 4 (mother)

		5 (father) 6 (other relative) 7 (non-relative)
gender	gender of individual household or family member	1 (Male) 2 (Female) 3 (Other)
age	age group of individual household or family members	1 (5 or under 5 years old), 2 (more than 5 and less or equal to 10 years old), 3 (more than 10 and less or equal to 17 years old), 4 (more than 17 and less or equal to 35 years old), 5 (more than 36 and less or equal to 50 years old), 6 (more than 50 and less or equal to 65 years old), 7 (more than 65 years old)
residence	shows where the individual household or family member currently lives	1 (same house) 2 (same neighborhood) 3 (same city) 5 (another region in Turkey) 6 (abroad) 7 (deceased) 8 (other)
residence_d	shows the county where the individual family member resides if they reside outside of Turkey	3 digits ISO code of individual countries
Note: Bold type indicates value labels in Stata		

Table 4: Overview of observations, responses and filterings

hhid	Household members	relation	gender	age	residence	residence_d
12 (TU1_4_78)	10 (hh members) 2 (Outside hh)	1 (respondent)	1 (male)	1 (18-35)	1 (same house)	1 (filtered)
		1 (mother)	1 (female)	2 (filtered)	2 (same neighborhood)	2 (filtered)
		1 (father)	1 (male)	9 (refuse to answer)	9 (same house)	9 (filtered)
4.339 (Others except household above)	3.809 (hh members) 530 (Outside hh)	773 (respondent)	1.181 (female)	193 (less than 5)	3.312 (same house)	42 (Afghanistan)
		365 (spouse)	2.838 (male)	455 (6-10)	4 (same city)	4 (Australia)
		1.071 (children)	3 (other)	503 (11-17)	19 (same city)	3 (France)
		271 (father)		1.945 (18-35)	28 (another region in Turkey)	6 (Germany)
		285 (mother)		265 (36-50)	61 (abroad)	5 (Iran)
		375 (relative)		50 (51-65)		1 (Iraq)
		1.199 (non-relative)		10 (older than 65)		3.359 (filtered)
					1 (refuse to answer)	1 (refuse to answer)
				583 (filtered)	167 (same house)	18 (Afghanistan)
					4 (same neighborhood)	1 (Canada)
					4 (same city)	3 (Germany)
					6 (another region in Turkey)	2 (Iran)
					335 (abroad)	523 (filtered)
					31 (deceased)	
					36 (refuse to answer)	36 (refuse to answer)
				18 (refuse to answer)	17 (same house)	1 (Iran)
					1 (abroad)	17 (filtered)
			188 (filtered)	9 (less than 5)	179 (same house)	179 (filtered)
				20 (6-10)		
				33 (11-17)		
				82 (18-35)		
				26 (36-50)		
				8 (51-65)		
				1 (older than 65)		
				9 (refuse to answer)	9 (same house)	9 (filtered)
			138 (refuse to answer)	1 (18-35)	1 (same house)	1 (filtered)
				128 (refuse to answer)	124 (filtered)	124 (filtered)
					4 (refuse to answer)	4 (filtered)

5.3 Coding of missing values

In general, the following codes were used for differentiating different missing values: “-1” indicated that a respondent refused to provide an answer, “-2” was used if the respondent indicated that they do not know the response to a given question (which was an option that was not available throughout the questionnaire, but reserved for cases where this response seemed potentially relevant from a content perspective, e.g. the highest

level of parental education). “-3” was used to indicate that there is missing information due to filtering, i.e. the question was either not asked or did not apply.

Table 5: Coding of missing values

Code	Wertelabel
-1	Not replied/ refused to answer
-2	Don't know
-3	Filtered

5.4 Coding of open-ended data

The questionnaire contains a number of open-ended questions, from which the following categorical variables were manually constructed:

Current and past occupation (variables “occ_catag”)

The following open-ended question has been asked: “What is your current occupation? Please give the exact name of the job or work you do, e.g. 'Construction worker' or 'harvest worker' rather than 'worker', or 'mechanical engineer' rather than just 'engineer', or 'manufacturing clothes' rather than 'textile'.”

Although it is explicitly stated in the question that the aim is to know the exact job name, there are some incomplete answers (e.g. “worker”). In addition, several different names have been used for the same jobs by different respondents. We recode all responses to the best of our knowledge into the following pre-defined categories, meaning that some of them might be empty⁴. Examples for responses provided by respondents are provided in parentheses.

-1= refused to answer

-3= filtered

0= no job / never worked in Turkey

1= Architecture and Engineering Occupations (e.g. repairman, master in automotive, automotive worker, master in automotive, exhaust master)

2= Arts, Design, Entertainment, Sports, and Media Occupations (e.g. translator, photographer and computer works, home decore)

3= Building and Grounds Cleaning and Maintenance Occupations (e.g. cleaning, car wash, garbageman, junkman, collecting paper, sorter in recycling workshop, I collect cardboard, garbageman, paper collector, scraper sorter)

4= Business and Financial Operations Occupations

⁴ The categories were informed by example categories provided <https://www.mymajors.com/career-list/> and <https://www.recruiter.com/careers/> (last access 31 July, 2023)

- 5= Community and Social Services Occupations
- 6= Computer and Mathematical Occupations
- 7= Construction and Extraction Occupations (e.g. insulation, welding, carpenters, dyer, gypsum master, coal company worker or gardening, I work with the elevator master, help with plumber, infrastructure piping, tile worker, plumber, metal job worker)
- 8= Education, Training, and Library Occupations
- 9= Farming, Fishing, and Forestry Occupations (e.g. agricultural worker, fish cleaning, sheep herder, grass, horticultural worker, orchard man, I am dealing with irrigation works, rose house,)
- 10= Food Preparation and Serving Related Occupations (e.g. cooks, coster, waiters, dishwasher, butchers, nutrition, bakery, döner, headwaiter, dishwasher, pide worker, confectionary worker, confection)
- 11= Healthcare Practitioners and Technical Occupations (e.g. doctor)
- 12= Healthcare Support Occupations
- 13= Installation, Maintenance, and Repair Occupations
- 14= Legal Occupations
- 15= Life, Physical, and Social Science Occupations (e.g. non-governmental organization)
- 16= Management Occupations (e.g. managing)
- 17= Military Specific Occupations
- 18= Office and Administrative Support Occupations (e.g. lawyer clerk, coster, public employee)
- 19= Personal Care and Service Occupations (operator porter, barbers, porters, hairdresser, travel agency employee)
- 20= Production Occupations (e.g. textile, leather industry and tailor, garment worker, spinning company worker, in the paper factory, shoe worker, worker at a garment manufacturing facility, worker in the factory, filature worker, scaffolder in construction, worker at vending machine factory, sewing master, clothing staff)
- 21= Protective Service Occupations
- 22= Sales and Related Occupations (e.g. traders, ironmongery, greengrocer and chicken shop, grocery store worker, phone store, sheep trading, paysage, rice trade, hawker, wholesale market, craftsman shop owner, working at a flower shop, I have a glassware shop, pastry shop, packer at the factory, pitchman)
- 23= Transportation, Distribution and Logistics (e.g. drivers, at the gas station, I'm porter, auto package, carrier, I carry goods)
- 24= worker (e.g. basic worker, normal worker, I am a worker in a private company, drying worker)
- 25= Daily worker (e.g. self-employed in daily work)
- 99= missing information (e.g. private sector, freelancer)

Qualitative Questions (variables “*quali_1*”, “*quali_2*”, “*quali_3*”, “*quali_4*”)

Each respondent received one randomly administered open ended question. If the question was not asked, the responses are coded -3.

Responses to the qualitative questions are provided in their original version, i.e. open-ended text responses, and additionally recoded as dummy variables and/or categorical variables, respectively.

Discrimination of different migrant groups in Turkey (variable “quali_1”)

The question reads: “We spoke about discrimination earlier. In your opinion, are there any groups of migrants in Turkey that are in a better position compared to others? If so, which are they?”

Based on the responses provided, we create two additional variables: a dummy variable (variable “quali_1_dummy”) and a categorical variable (“quali_1_categ”). The dummy variable is coded as follows:

- 0 = No, there are not.
- 1 = Yes, the are.
- 2 = Don’t know
- 3 = Filtered

The codes for the categorical variable is provided in Table 6 below. Some responses that would potentially fit more than one category are included in “other”. This concerns, for instance, the following responses: “The ones who are not illegal and the rich” or “Immigrants from Europe and Syrian refugees receiving aid”.

Table 6: Overviews codes for quali_1_categ

Code	Category	Example responses
-1	Refused to answer	“Don’t say anything”, “Don’t want to talk about it”
-2	Don’t know	“I do not know, I can’t think of anything”, “I dont have an idea about this topic”, “I dont know, what shall I say?”, “No answer for this topic”
-3	Filtered	Question was not asked
0	No/No one	“Every in one situation, I think none”, “No”, “No afghan group looks better”, “No group”
1	Syrian	“Syrians. They used to be able to come and open shop and get help, Syrians are in good condition”
2	Uzbek	“Uzbek”, “Uzbeks”
3	Iranian	“Iranians are better”, “Iranians are more comfortable”
4	Tajik	“Tajiks”
5	Africans	“Africans are more comfortable”
6	Afghans	“All Afghans”
7	Turkmen	“Turkmens are better than the others”
8	Rich people	“Rich ones”, “Yes, for richer people”
9	Legal ones	“those who are not illegal”
11	Kurd	“Kurds”

10	Other	"Afghans who are wealthy member of parliments", "Immigrants from Europe and Syrian refugees receiving aid", "Iranians and Syrians", "There are differences", "Uzbek Turkmen", "İstanbul ankara"
----	-------	---

Hope in the future (variable "*quali_2*")

The question reads "What are the things around you that give you hope for a good or better future?"

Based on the responses provided, we create a categorical variable ("*quali_2_categ*") using the codes provided in Table 7 below.

Table 7: Overview codes for *quali_2_categ*

Code	Categorise	Examples
-1	Refused to answer	"Did not response"
-2	Don't know	"I dont know"
-3	Filtered	Question was not asked
1	A better life, peaceful life	"A quiet life away from war and turmoil", "A calm and beautiful life gave me hope"
2	Obtaining asylum	"acceptance of asylum"
3	Safety and Security	"It gives me hope to know that the country will not fight", "peace and security"
4	Health	"End of infectious disease"
5	Freedom	"Freedom"
6	Family, Friends and relatives	"Our children's living in safe conditions made us happy and gave us hope", "My life, my relatives", "My grandchildren"
7	Immigration to another country	"Leaving from this country", "Living in a country in Europe"
8	Improvement in education, work and financial situation	"My university", "My work and life", "I'm saving money and I'm going to start my own business. it gives me hope"
9	The government of Turkey	"The leader of the country", "Ak parti", "State"
10	No hope	"There is nothing", "Unfortunately I have no hope for the future", "Nothing"
11	Other	"The future", "The God"

6 Anonymization

Anonymization DeZIM.fdz offers the collected data as anonymised Scientific Use Files (SUF) in the framework of the study DeZIM Afghans in Turkey Survey. The degree of anonymisation depends on the respective access methods. Overall, the SUFs are offered via three different access paths: via download, via remote access and on-site at the DeZIM Institute. The more access to the data is technically controlled by the access channel, the less the variance of the data has to be reduced by aggregation and the greater its analysis potential remains. Direct identifiers are not included in any data set, as the data was pseudonymized in advance. The following variables were anonymized in order to ensure anonymization:

Table 8: Anonymization

	variables	Download	Remote	Onsite
Locational information	province, district, neighborhood, latitude, longitude, location_oth	no release	no release	release
Open response categories	hhinc_src_oth, label_self_oth, migleb_why_oth, emo_leaveself_oth, emo_stayself_oth, discrleb_hier_oth, discrleb_hier_oth_eth_scl, stay_reas_oth, mig_desire_oth, quali_1, quali_2, quali_3, quali_4, support_oth, dwelling_oth, ethnicity_oth, lang_other_oth, plans_reas_oth, occ, occ_tur_last	no release	no release	release
Variables that contain country information	origin_c, citizenship_1, citizenship_2, citizenship_3, mig_dest_1, mig_dest_2, mig_dest_3, mig_dest_4, mig_dest_5, residence_partner_female_abroad, edu_etry, national_ident_1, national_ident_2, migleb_start_1, migleb_start_2, migleb_start_3, nb_afgh, plans_etry_1, plans_etry_2, mig_desire	no release	release	release

Ethnicity	ethnicity_1, ethnicity_2, ethnicity_3, ethnicity_fellow, ethnicity_enumerator	no release	release	release
Language	Lang_other_1, lang_other_2, lang_other_3, lang_other_4, lang_other_5, lang_other_6, lang_other_oth	no release	release	release
Year information	migleb_lastentry_y, migleb_lastentry_m, duration_g, edu_year, workstart_age	no release	release	release
Additional survey information	datestart, timestart, timeend	no release	release	release
Small categories	hysize_camp, othrel_total, othrel_male, othrel_fem, othrel_5, othrel_10, othrel_17, othrel_35, othrel_50, othrel_65, othrel_99, nonrel_fem, nonrel_5, nonrel_10, nonrel_17, nonrel_35, nonrel_50, nonrel_65, nonrel_99, quali_1_categ, quali_2_categ	no release	release	release
Aggregations	Age_resp, hysize_dwelling, children, children_inside, children_out, nonrel_total, nonrel_male, years_school, years_univ, resid_time_Year	Release with aggregations	Release without aggregations	Release without aggregations

Note. In addition to the respondent roster, the onsite access also includes the family roster and a longitude latitude supplementary dataset. The remote access contains the respondent and family rosters, while the download package includes only the respondent roster.