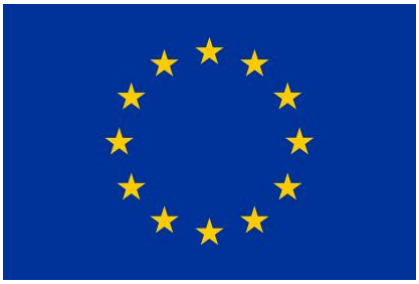
**Development, piloting and dissemination of new methods in data collection on discrimination against Roma in public services (ADinPS)**

project funded by the European Union
(REC-AG-2019/REC-RDIS-DISC-AG-2019)

# How to use "mystery shopping" to measure discrimination in a public service

*A guideline for NGOs*

by Nikolett Papdi and Agota Scharle



This publication was funded by the European Union's Rights, Equality and Citizenship Programme 2014-2020.

Project **"New methods in data collection on discrimination against Roma in public services (ADinPS)"** was implemented between May 2020 and February 2022 by a consortium led by the Centre for Policy Studies of the Central European University (CEU, Hungary) and including Budapest Institute (BI, Hungary), Center for Interethnic Dialogue and Tolerance – Amalipe (Bulgaria), ROMEA (Czechia) and Autonomia Foundation (Hungary).

The project aimed at contributing to fighting discrimination that Roma face in accessing public services through developing, piloting and disseminating methods of data collection that can be easily implemented by NGOs and used to monitor discrimination in a regular and systematic way.

During the project three grassroots NGOs, Amalipe (Bulgaria), ROMEA (Czechia) and Autonomia Foundation (Hungary) piloted the mystery shopping methodology in six experiments (two in each country):

- Bulgaria: enrolment in school, access to municipal social housing,
- Czechia: enrolment in school, renting municipal premises,
- Hungary: enrolment in school, registration of car,

and obtained data on discrimination that Roma face in accessing public services both accessed via online gateways and personal contact with providers.

Based on the piloting the methodology by the three NGOs, the present guideline was developed.



The ADinPS project was funded by the by the European Union's Rights, Equality and Citizenship Programme (REC) 2014-2020.

The content of this publication represents only the views of the authors and is their sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

# Contents

# How to use “mystery shopping” to measure discrimination in a public service

A guideline for NGOs

## 1. Purpose of the guide

This guide aims to support NGOs engaged in advocacy for the rights of Roma, or other discriminated groups. The guide explains why it may be useful to collect convincing evidence on discrimination and provides practical advice on how to plan and implement experiments that may generate such evidence. The focus of the guide is on mystery shopping, a method that is relatively cheap and easy to use.

The mystery shopping method is based on using avatars who act as users of a service, and we collect information on their experience. Discrimination is detected by comparing the experience of avatars belonging to the majority group with the experience of avatars belonging to the discriminated group.

The method includes three main stages: setting up the experiment, implementation, and data analysis, and each stage requires some expertise. The design stage requires some knowledge of statistics, the basics of scientific observation, and a good understanding of the service to be tested. The implementation stage requires some knowledge of data collection techniques and management skills. The analytical stage requires experience in cleaning data and a knowledge of counterfactual impact evaluation tools.

Some NGOs may have these skills in-house, others may be able to access them by teaming up with partner organisations. Some of these skills may best be obtained through contracting an external expert. However, if you think you might need to engage an external expert for the analytical phase, it is important to engage them already at the planning stage. This is because with this method, careful planning can make the analysis much easier and the evidence much more reliable, while mistakes made at the planning stage are often impossible to correct later.

## 2. Why and how to measure discrimination

### 2.1. Why is it useful to measure discrimination?

NGOs working with discriminated groups usually experience discrimination in their daily work: they know it happens and also know what forms it may take. For them, it may not seem obvious why it may be useful to measure it.

Reliable evidence can be very useful when NGOs want to convince government officials, politicians or the general public about the need for policy action. It may also be useful in convincing donors about the need for NGO activities. Many people have little personal experience with the Roma, and are not aware how often and in what ways they may face discrimination, so they tend to underestimate the problem.

Showcasing examples of discrimination in the media or court cases may be very effective for raising awareness. However, to convince government officials to spend money and reform existing policies, it is better to have reliable numbers: statistics and impact analyses that prove discrimination. If NGOs can only present individual cases, policymakers can always dismiss them, saying that the problem is not systemic[1] and that such cases are only individual failures of a few people.

Collecting information on discrimination may also be useful in improving the design of NGOs' activities. They can check if their own perceptions of where and how discrimination happens are confirmed by the scientific analysis and adapt their activities if necessary.

Another argument for NGOs to conduct such experiments is that, in many countries, nobody else will do this for them. Government agencies responsible for ensuring equal opportunities often lack the motivation and/or capacity to collect evidence or only focus on detecting discrimination in the private sector. Research may invest in a few in-depth studies but have no incentive (nor capacity) to regularly monitor discrimination across government services.

In this guide we will show examples of measuring discrimination in the public sector; however, the method can be adapted to the private sector.

## 2.2. Alternative methods of measurement

Mystery shopping is not the only method for measuring discrimination. Alternative methods include:

- surveys that directly ask people about their perceptions of discrimination
- statistical analysis of actual access to schools, healthcare, jobs, or other things that are important for equal opportunities and welfare.

These tools are very useful and can provide reliable evidence, but they have some drawbacks. Surveys are costly and may not capture genuine attitudes among people who know that discrimination is not socially acceptable, or may underestimate discrimination where people are less aware or less sensitive to notice it. The analysis of actual access to posiitons or services requires high levels of expertise and depends on the availability of good quality information on access, group identity and other qualities that may affect outcomes, but are not directly linked to discrimination (such as place of residence).

Situation testing is a similar tool, but is used in a small scale, with the purpose of proving discrimination in legal cases (Rorive 2008).

Compared to these other methods, an important limitation of mystery shopping is that it is best suited for testing discrimination in the first phase of accessing a service, as the risks involved tend to be higher in the later phases of the client journey.

[1] As we have limited space, let us mention two possible root causes of discrimination. First, prejudices may be just as common among public officers as in the private sector. Second, discrimination may take an idirect, often unintended form in practices that are equally applied but affect Roma (e.g. offering some public services conditional on having a registered address in a country where many Roma lack such registration) in a differential way, or the differential enforcement of policies that are seemingly neutral (e.g. police stopping Roma for ID checks more often than they check non-Roma passers-by).

# 3. Mystery shopping and ethical concerns

## 3.1. What is “mystery shopping”

Mystery shopping (or audit study / discrimination testing) is a special form of participant observation. The participant (or an avatar) acts as a service user and secretly observes the service provider; in other words, the method involves deception. In our case, the focus will be on how accessible the services are, and we are mainly interested in whether Roma experience any discrimination in this regard. Participants follow a pre-defined scenario that specifies their actions and also responses to possible behaviours and questions of the service provider.

- Regarding ethical concerns, the public officials and the service users are both considered participants in the research projects, but you only have control over the service user's actions and reactions.
- Your organization, as the facilitator of the exchanges between the two parties, is responsible for enabling the service user to execute the pre-agreed scenario.
- It is extremely important that the service user follow the scenario as closely as possible to minimize ethical risks and to be able to observe potential discriminatory behaviour.

## 3.2. Ethical risks and mitigation strategies

This section outlines the ethical and data protection issues that may come up while conducting mystery shopping experiments (MSEs) and provides recommendations as to how to mitigate the associated risks step by step in order to establish best practice for NGOs.

The table below shows the ethical risks associated with each communication channel of MSEs. Once you have decided on your channel(s), make sure that you read carefully and understand the risks involved and the ways in which you can minimize them.

**Table 1. Risks involved in using various communication channels**

| **Communication channels** | **Potential ethical risks** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Wasting service time** | **Privacy & confidentiality** | | **Consent issues** | | **Distrust** | **Vulnerable group** |
| | | **SU** | **PSE** | **SU** | **PSE** | | |
| **In person** | Y | Y | Y | Y | Y | Y | Y |
| **Phone** | Y | N | Y | Y | Y | Y | Y |
| **Phone by avatar** | Y | N | Y* | N | Y | Y | N |
| **Facebook/ messenger** | N | N | N | N | N | Y | N |
| **Email/letter** | N | N | Y* | N | Y | Y | N |

*Abbreviations: SU = Service user; PSE = Public service employee; Y = yes, there is a risk; N = no or small risk; * Only in case of poor data protection measures*

## 3.3. Wasting service time

To observe service providers' regular and usual attitudes and behaviours, you need to design scenarios that simulate real life situations. This requires the use of some **service time**, which can be costly. The ethical concern is higher if the service is in short supply or costly to provide.

You can reduce this risk if you:

- *Exclude* services that are especially costly and exclude all emergency services.
- Plan scenarios to be as *brief* as possible.
- Focus on the *exchange of information* rather than the actual usage of a service.
- Do not expect service providers to answer questions that are not closely related to their roles and responsibilities. Restrict the questions to their competencies.
- Reduce the complexity of cases by ensuring that the interaction is *between two people only.*

## 3.4. Risking privacy and confidentiality of participants

Even if you use fake identities for the clients, in face-to-face interactions the service provider may remember the client at a later date, which could negatively affect their usage of the service in future. If the interaction has negative consequences for the official (for example, if it is reported to their superior), and they get punished for their conduct, then it can generate negative feelings in them towards the group that has been discriminated against in general. There is also a small risk that service users and officials might already know each other.

The fake clients can remember the names of officials and recognize them in the future, as the officials have no choice but to take part in the experiment with their real identities. Lastly, public service employees can face negative consequences if the information about their behaviour is disclosed to their employer (or to others), such as not being given a salary increase or being examplarily punished.

You may reduce these risks if you:

- avoid face-to-face interactions, if possible. Use alternative channels, such as telephone, email or regular mail.
- do not place research subjects in locations where they can be identified by officials or vice versa (e.g. avoid their birthplace and their place of residence).
- anonymize the data submitted by the research subjects before analyzing it.
- assign pseudonyms to service users to ensure anonymity.
- limit the number of people involved in handling any sensitive information to the necessary minimum.
- do not share any sensitive information via untrusted platforms.
- keep track of who accessed the data and when.
- do not disclose any individual-level personal information to third parties.

- do not further investigate the individual performances of public employees and do not publish or share with the service provider any information about individual performance. If and when you inform the service provider of the research project, make it clear to them in advance that you will not be identifying to them who has committed what kinds of behavior.

## 3.5. Obtaining informed consent and using deception

In experiments that involve human subjects, it is a general requirement to fully inform participants about the research, but the MSE is an exception to the general rule, because it is based on deception.

Using deception in the mystery shopping method is absolutely necessary for ensuring the validity of the results. Firstly, public service officials cannot know about the experiment, because if they knew about it, they would probably change their behaviour, which would defeat the purpose of the research. Secondly, service users cannot know that the experiment focuses on ethnic discrimination. This is because the experiment relies on their observations, and if they know what they are supposed to observe, it is more likely that they will find it. This is called confirmation bias. Because of this, the real objective of the study, which is measuring discrimination, must be concealed from service users.

The potential risks are that (1) service users might feel distressed when they find out about the real objective of the study and (2) PSEs might feel deceived in case they find out about their unwitting involvement in the research.

You may reduce these risks if you

- where possible, avoid the involvement of natural persons as clients by opting for online channels where avatars can be used.

- inform service users about the main aspects of the project, such as its nature (covert observation by the method of mystery shopping), their role as observers, the objectives of the research which is to measure public sector service performance. But, do not tell them that they are also observing potential racist behaviours (to avoid confirmation bias).

- once the experiment is over, debrief service users, in other words, make them aware of the real objective of the study, which is measuring discrimination . Because you have previously informed them about the main conditions, it is very unlikely that they will react with anger or feel distressed. Informing them only *after the experiment* also ensures that the data you collected will not be biased (i.e. this helps avoid confirmation bias).

- offer multiple opportunities to withdraw from the experiment at different stages.

- design the experiment in such a way that it mirrors everyday encounters between service users and providers. You want to observe 'normal' & 'usual' behaviour, not to alter it.
- do not expose public employees to any unusual or inappropriate experiences which could not be considered as an inherent part of their role and daily routine.

## 3.6. Creating distrust at the societal level

Though it may be necessary and justified to use deception in a research experiment, we should be aware of and mitigate the negative side effects. Deception might damage the public's general trust in research, which has various negative consequences, e.g., it can jeopardize future research programmes. It might generate distrust of public service employees if/when their discriminative behaviours are publicized. Lastly, it can weaken trust between public bodies, the research community,and NGOs, especially if the level of trust is already relatively low.

You may reduce these risks by using a dissemination strategy that

- avoids confrontative language and ensures that public bodies do not feel threatened.
- highlights potential positive reactions (e.g. possibilities for learning and cooperation in reducing discrimination).
- acknowledges any positive findings and initiatives on the part of public bodies.
- portrays the problem as a common challenge (rather than framing it as an "us" and "them" issue).
- focuses on the affected public services, rather than making general claims about discrimination in public services.

It may also help if you inform the agencies targeted by your experiment (or inform the central level, if administration is very decentralised) before publishing any results and negotiate the timing of publication with them.

## 3.7. Involving a vulnerable group: minority ethnic background

If your experiment involves people with a minority ethnic background, the above risks are magnified, and additional precautions are necessary. If natural persons are involved in the experiment, they must be protected from any harm during the experiment and any later consequences of their participation. Also, the experiment should not strengthen or reproduce stereotypes. For example, if the experiment involves sending a request on email, and the authority responds positively asking for the sender to confirm their interest, there should be a follow up from the avatar explaining that they no longer need the service (and provide a legitimate reason), otherwise this may reproduce the stereotype of Roma being not responsible.

You may need to consider

- avoiding the use of face to face contact,
- using pseudonyms and avoiding their birthplaces and places of residence, to ensure the participant's anonymity & confidentiality,
- preparing the service users for the mystery shopping encounters very thoroughly and including exit-scenarios,
- involving professional actors rather than lay volunteers as service users (if possible),
- offering debriefing to service users to discuss any bad experiences.

# 4. How to set up a test

This chapter guides you through the most important decisions in designing an experiment. Once you have made your decisions, it is useful to double check if you can improve your design in order to minimise the risks listed in Chapter 3. The Appendix includes a concrete example of designing an experiment.

## 4.1. Choosing a service to test

Choosing the service depends partly on your advocacy goals, and partly on practical and ethical considerations.

First, consider your field experience and advocacy agenda:

- What are the services where you have come across discrimination?
- What are those where limited access has far-reaching negative consequences?
- Is the government aware of / do they acknowledge discrimination in these services?
- If you find evidence of discrimination in this service, how will you use that information?

Second, consider the feasibility of testing in the selected services. Exclude services that are:

- very costly or life-saving (e.g. the ambulance)
- compulsory or typically initiated by a public body (e.g. vaccination or regular health checks of children, or lung cancer screening for adults)
- available to most people who need it (unlike, e.g. social housing)
- provided automatically, where public officers have very little or no room to decide how they respond to a client and thus cannot discriminate.

Third, consider some practical issues that will make it easier to implement the testing:

- Is the service frequently used (will a sudden increase in the number of inquiries seem suspicious)?
- Can it be used by people not living in the locality? Is it relevant for newcomers?
- Can clients ask for information about the service without personal contact (e.g. by email, Facebook or other web interfaces), so that you can choose which channel works well for testing?
- If the service is for children, can you implement the test without involving children in the experiment?
- Does discrimination usually happen when people first approach the service provider, asking about the service? If discrimination mainly happens at a later stage (e.g. in social housing or other services where the evaluation process is long and there is a waiting list), this will be much more difficult to test using mystery shopping.
- Will you need in-depth knowledge of the locality to make a credible inquiry about the service? If so, this may increase the time and money spent on the testing.
- Does the service provider have many separate offices? If they have at least 60 local offices, that helps us in generating enough observations without becoming suspicious.

## 4.2. Choosing a communication channel

Once you have an initial selection for each possible service, you need to consider how these are usually accessed by the general public, as well as by the discriminated minority. Do people usually make a phone call, or write an email or via FB Messenger? Or do they write a comment to the provider's Facebook page? Once you know which channels are commonly used by the provider and by client groups, you need to weigh two main factors: the cost of using the channel, and the risk of the contact looking suspicious/unrealistic.

If the channel is unrealistic (i.e., if members of communities that usually have little access to such a channel begin using it), the service provider may either get suspicious, or will not recognise the request as coming from a disadvantaged person and may not respond in their usual way. This risk can ruin your experiment. At the same time, some of the usual channels can be very costly and also carry more ethical risks, as shown in Table 1.

When weighing these factors, it is important to consider that discrimination is likely to affect a broader group than just the community you are working with. So, for example, it may be that Roma people coming from a segregated community never use email, but there may be other Roma living in more integrated settings who do often use this channel when contacting service providers and who experience discrimination through it. Also, the use of online platforms is rapidly expanding, so providers may be increasingly open to inquiries coming from these channels.

**Table 2. The practical advantages and disadvantages of the main channels**

| | in person (walk-in) | phone | Facebook, messenger | email | letter |
|---|---|---|---|---|---|
| Real persons must be involved (costly) | yes | no | depends | no | no |
| Easy to create avatar | no | yes | no | yes | yes |
| Easy to include strong markers* | yes | no | yes | yes | yes |
| Real time interaction with client/official: costly | yes | yes | depends | no | no |
| Discrimination is very likely to happen | yes | yes | no | no | no |

**markers are used to make sure the service provider recognises the avatar as belonging to the discriminated group.*

As the table shows, there is no one channel that works best in all cases. In-person or telephone inquiries may be the best to capture discrimination, as they are more spontaneous and officers are less aware that responses may be recorded or traced afterwards. The downside is that these channels need real-person avatars, who need to be paid (if they are professional actors) or thoroughly trained (if they are volunteers), which is costly. Also, you need much more time for collecting the same number of observations,[2] compared to email messages, which can be sent to many recipients all at once. Emails have further advantages: written responses are easier to analyse in automated methods (which reduces the costs) and produce a clear-cut outcome (did they respond, or not?).

## 4.3. Developing the scenario

Once you decide on the service and the right channel, you need to define **a situation** where you can observe the interaction between the client and the service provider. Ideally, the situation should be:

- relatively simple, so that we can still compare the situations initiated by different clients, and measure any differences in how the provider handles the situation, but:
- there should still be room for discretion (discrimination)
- responding should not require much effort on the part of the service provider

This last consideration is important for avoiding ethical risks and also makes it easier to analyse the results. It is best if the situation relates to a single service, focuses on getting information about the

[2] You could reduce the time needed by involving many avatars, but this will make it more difficult to interpret your results as your avatars will have different personality traits. Ideally, your avatars should differ only in one aspect: belonging to the discriminated group or not.

service (rather than actually attempting to use a service), and it focuses on issues that commonly arise.

In most cases it is useful to choose 3 to 4 questions to ask, and to use questions where the official has some choice in how they respond. For this, it is important to understand if there are any laws or rules that prescribe how they must respond.

When you start working on the scenario, it can help a lot if you can find a trustworthy insider (someone working for the service provider) who can tell you about the typical interactions and questions they encounter on a daily basis. As there is a risk that insiders may discover your experiment while it is underway, it is best to approach them via informal networks for such information and, if possible, do not tell them about the experiment, or at least avoid mentioning the testing method.

Once you have a draft scenario, you need to write up the actual words to be used by the avatars. It is important to get the language and tone right so that your avatars seem as real and as ordinary as possible, otherwise they will be discovered as impostors or will not trigger the usual forms of discrimination that you want to measure.

For personal or telephone conversations you also need to plan how your avatar should respond, depending on the initial reactions of the service provider. In these scenarios, you need to plan an exit-option, i.e., describe how that avatar can get out of the situation if they feel bad about what happened, or if they feel they cannot continue playing the role of the avatar.

## 4.4. Recruiting and preparing or generating the avatars

The avatars should be realistic and represent a typical client. For each scenario, you need two types of avatars: one that represents the majority population and one that represents the discriminated group. It is important that the service provider should be able to see this difference very clearly. This can be ensured by using more than one "marker" or signs that "tell" the service provider which avatar belongs to the minority. A good marker can be:

- a name or a surname that is much more common among the minority than the majority population,
- a street address associated with an area well-known to be populated by the minority,
- a photo displaying a typical member of the minority (e.g. having visibly darker skin or wearing a typical cap or scarf),
- mentioning the name of an organisation that clearly works for/with the minority (e.g. a Romani dance group, or after-school tutors helping disadvantaged children),
- using words or an accent typical of the minority group and not common in the majority group.

Describe your avatar's main characteristics and make sure all the information to be shared with the service provider during the testing is consistent with these characteristics. So, if you describe your avatar as a young Roma woman with a primary school education, then all details about "her", the name, surname, home address, email account, photo, style of writing, grammatical mistakes, etc., should fit this character.

In all cases, it is very important that you choose or create avatars that are as similar to each other as possible, and the only important difference between them should be the marker of the feature you expect to trigger the discrimination. If avatars cannot be exactly the same (e.g. when you engage real persons), then it is important that they not differ in ways that may influence the service provider. For example, they must be the same gender and similar in age, looks, and clothing.

# 5. How to carry out a test

This section guides you through the practical aspects of implementing a test. In some of these aspects, you may need the support of external experts.

## 5.1. Choosing the sample size

Choosing the sample size and selecting the sample requires some statistical expertise, so it may be useful to ask for advice. When deciding about the sample size (how many service providers or their local units to include in the testing), you need to consider that a bigger sample improves the reliability of the outcomes but usually costs more, both in managing the testing process and also in the analysis of the data. In some cases the sample is limited by the organisational setup of the service you want to test: if, for example, the provider has 100 local units, you will not be able to have a sample of 500.

For a quantitative analysis, you will need a fairly large sample. Based on the existing empirical studies, a final sample of around 300 or more observations seems necessary for a reliable estimate of discrimination. The sample size can be slightly lower where discrimination is expected to be more widespread and easier to capture (e.g. in phone calls or in-person encounters), and needs to be higher if discrimination is expected to take very subtle forms (e.g. in written responses). Also, if you use emails or letters, you need to allow for non-responses, i.e., that the final sample of responses may be much smaller than the number of messages sent.

For a qualitative analysis, a smaller sample of at least 20 cases may already provide valuable insights.

## 5.2. Compiling the full list of service providers and getting contact details

To draw your sample, you need a full list of the local units of a service provider. If you are lucky, this list is available at an official website, such as the ministry supervising the given service. In some cases, the list is not published, but you can search a public database to find the service in particular locations, which you can use to compile the full list (if possible, using some automated solution, rather than manually copy-pasting each item). If the full list is not readily available, it may be worth contacting researchers, NGOs and/or private firms that are likely to have done this work for their own purposes and may be willing to share it or sell it at a reasonable price. Or, if you have good personal contacts in the government, you may also ask for the list from them.

You will need not only the name and location of each service unit, but also their contact information. Depending on which channel you chose to use, you need their street address, telephone number, email address, Facebook account, etc. When they have several public phone numbers or email addresses, you should use the one most likely to be used by real-life clients.

Depending on how you got the information, the contact details may not be up-to-date or may contain errors. It is important to check this, especially if you want to use emails, as a high share of bad addresses will reduce your final sample (and will also be troublesome to identify and omit from your data).

## 5.3. Why and how to randomise the control group

The mystery shopping method measures discrimination by comparing the experience of two groups of clients: the discriminated minority and the majority population (or subgroups within these). The method provides reliable evidence only if these two groups are matched with service providers randomly so that they meet similar providers. This ensures that any difference in their experience can be attributed

to ethnic background (or other protected characteristic) of the avatars, and not to the particular provider they met. For this randomisation, you first need a full list of service providers and some information about them, such as the location, its institutional form (if it is likely to impact the discrimination), and the percentage of the minority in the neighbourhood population or using the service.

To allocate avatars to services, you need to split the sample into two groups in a random way. Ideally, you may do this using statistical software, also ensuring that the subgroups are similar in their important characteristics: e.g. they contain a similar number of locations with a high share of the minority. If such a software is not available, you may use a simpler method, e.g. order the list of locations according to some characteristic that cannot have any link to discrimination, such as the first character in the street name of their address. The first half of the ordered list will be your subsample for the first avatar, and the second half will be the subsample for the second avatar. You can double check if the two subsamples are similar, e.g. by counting the number of large and mid-sized cities in each (see further explanation in the Appendix).

## 5.4. Timing

The timing of your experiment needs to be aligned to the service you wish to test. Some services can be accessed any time during the year or at any time of day, while others have a seasonal character. For example, parents usually want to enrol their children into kindergarten in September or January, while school enrolment in most countries starts in September, so inquiries about these services should be timed before these dates. Similarly, an amateur dance group might typically look for a municipal venue for their performance in the summer. By contrast, an inquiry about job search services may be initiated any time during the year.

Once you have worked out the details of the situation and matched the avatars with a list of service units, you also need to plan the timing of contacting the service providers. If you are using in-person avatars, it is important that both avatars time their visits or contacts roughly at the same time during the day. If you send emails, it is good to send your messages in bundles rather than in one go (to avoid being caught as spam), mixing the avatars within the bundles. That would mean sending the first 20% of the messages assigned to each avatar on a Monday and the second 20% on a Tuesday, rather than sending the first 40% of messages assigned to the first avatar on a Monday and then doing the same for the second avatar on a Tuesday.

## 5.5. Sending the avatars

When you engage real persons for the experiment, you need to train them: they need to understand the general purpose of the experiment, the scenario, and also what they need to observe and how to report it to you.

When you engage in-person avatars, you also need to prepare them for collecting information on their experience. This can be done by filling in a short observation sheet right after leaving the service provider. An example for an observation sheet is included in the Appendix.

When you use emails for the test, make sure to spend a few days using the newly created accounts before doing the testing: send and receive a few messages from them. This reduces the risk that your messages will end up in the spam folder.

## 5.6. Collecting the data

The methods of collecting information on the response of the service provider mainly depend on the communication channel that you use. For in-person or phone encounters, the information should be recorded by the avatar on an observation sheet. Technically, you could record the conversations using

a hidden microphone, but this would violate the privacy rights of the service provider's staff, so it is not recommended.

If you use written communication (social media, web-chat, emails, etc), the responses can be collected easily. However, the answers need to be coded, i.e. they need to be described in terms of measurable indicators, which can take a great deal of time if you have many responses. Designing the coding sheet also requires some analytical expertise.

Before you code the answers, it is important to first remove the ethnic markers (names, email addresses etc), so that the coders will not biased by that. This is especially important when coding the quality of the answer, which may involve some personal judgement. For similar reasons, it is useful to have two separate people code the subjective variables: this can be used to remove systematic bias coming from a coder's personality. If you involve an experienced expert in the analytical phase, it is best is to provide them with the codes of both (all) coders.

In the coding, it is advisable to use several indicators that describe the quality of responses: the number of responses (excluding error messages), the speed of receiving an answer (measured in work days), the number of questions answered, the helpfulness and kindness of the response, etc. Though some of these involve subjective judgement, in some cases it is only these indicators that will capture discrimination, and it can save time if you code several qualities in the first round, rather then return to the coding stage after finding no trace of discrimination in the "hard" indicators.

# 6. How to calculate and display the results

If your samples were properly defined and you matched avatars with the service units in a random way, discrimination can be easily calculated by comparing the averages of the outcome indicators between the two groups. For example, if you sent emails, and your Roma avatar received an answer from 76% of the providers, while your non-Roma avatar had a response rate of 83%, the 7% point difference (83-76) is due to discrimination. This section explains a few issues that may complicate this simple approach: you need to tackle these to make sure your results are not biased.

## 6.1. Factors that may bias the outcomes

The results will be biased or distorted if the sub-groups of providers responding to your avatars differ in ways that are related to discriminatory behaviour. This can happen if the matching of avatars to providers was not completely random, and also if there was a systematic difference between the avatars when implementing the tests. For example, if one of the avatars visited the provider in the morning while the other went there in the afternoon, this may introduce a bias. Or, if one avatar sent their emails on Thursday, while the other sent them on Friday, the latter may get a lower response rate.

## 6.2. Indicators of discrimination

Discrimination may take many forms. In practical terms, it is useful to focus on differences in the response that may hinder access to the service, such as waiting time, the amount of detail, the accuracy of the information provided, or the tone of voice that may discourage the client from claiming the service.

For example, when using email inquiries, you may use the following indicators:

- the number of responses (excluding error messages),
- the speed of receiving an answer (measured in work days),
- the number of questions answered,
- the helpfulness and kindness of the response

In in-person encounters, you may add further aspects of the officials' behaviour, such as:

- patience of the official
- going beyond the minimum response required
- being suspicious / assuming there is something fishy about the claim

## 6.3. A note on regression analysis

If, for some reason, the matching of avatars and service providers was not random, or the response rate was very low or uneven, regression analysis can be used to detect, and to some extent correct distortions in the results. If, for example larger local units were more likely to respond, and also less likely to discriminate, a simple comparison of outcomes between the avatars would paint an overly positive picture. This can and should be corrected by regression analysis. Such analysis requires statistical expertise.

## 6.4. Creating meaningful graphs

It is advisable to use graphs that show the difference between the outcomes for the avatars in an easily comparable way, such as a bar chart, where you can see the outcomes right next to each other. Using other visual formats such as infographs of videos can greatly increase the effectiveness of your dissemination efforts.[3] Creating clear and convincing graphs requires some expertise or experience, so it may be useful to engage external experts if you do not have such skills in-house.

[3] For example, this video presents an experiment on foreigners contacting football clubs (available at https://youtu.be/GGkQWiuTmXo ), which reached a much wider audience than the academic paper it is based on (Gomez-Gonzalez 2021).

# 7. Further reading


Adman, P. and Jansson, H. (2017) 'A field experiment on ethnic discrimination among local Swedish public officials.' *Local Government Studies*, 43:1, pp. 44-63, DOI: https://doi.org/10.1080/03003930.2016.1244052

Ahmed, A. and Hammarstedt, M. (2008) 'Discrimination in the Rental Housing Market: A Field Experiment on the Internet.' *Journal of Urban Economics* 64(2):362-372, September 2008, DOI: https://doi.org/10.1016/j.jue.2008.02.004

Bartoš, V., Bauer, M., Chyrilová, J. and Matějka, F. (2016) 'Attention discrimination: Theory and field experiments with monitoring information acquisition.' *American Economic Review*, 2016

Bertrand, M. and Duflo, E. (2017) 'Field experiments on discrimination.' In: *Handbook of economic field experiments*, 2017 – Elsevier, DOI: https://doi.org/10.1016/bs.hefe.2016.08.004

Bradbury, M. D. and Milford, R. L. (2016) 'Measuring Customer Service: Georgia's Local Government Mystery Shopper Program.' *State and Local Government Review*, 35, 3, pp. 206-213, DOI: https://doi.org/10.1177%2F0160323X0303500306

Butler M. D., and Crabtree C. (2017) 'Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions.' *Journal of Experimental Political Science* 4, pp. 57-67

Einstein, K. L. and Glick, D. M. (2017) 'Does race affect access to government services? An experiment exploring street-level bureaucrats and access to public housing.' *American Journal of Political Science* 61(1), pp. 100-116

Gaddis S. M. (2018) 'An Introduction to Audit Studies in the Social Sciences.' In: Gaddis S. (eds) *Audit Studies: Behind the Scenes with Theory, Method, and Nuance. Methodos Series (Methodological Prospects in the Social Sciences)*, vol 14. Springer, Cham, DOI: https://doi.org/10.1007/978-3-319-71153-9_1

Gomez-Gonzalez, C., Nesseler, C. and Dietl, H.M. (2021) 'Mapping discrimination in Europe through a field experiment in amateur sport.' *Humanities and Social Science Communications* 8, 95, DOI: https://doi.org/10.1057/s41599-021-00773-2

Hemker, J. and Rink, A. (2017) 'Multiple dimensions of bureaucratic discrimination: Evidence from German welfare offices.' *American Journal of Political Science* 61(4), pp. 786-803

Simonovits, G., Simonovits, B., Vig, Á., Hobot, P., Németh, R. and Csomor, G. (2021) 'Back to "normal": the short-lived impact of an online NGO campaign of government discrimination in Hungary.' *Political Science Research and Methods*, pp. 1-9

White, A. R., Noah N. L. and Faller, J. K. (2015) 'What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials.' *American Political Science Review* (February), pp. 1-14

# 8. Appendix

## 8.1. Example of an experimental scenario

This experiment tests discrimination in school enrolment, which is highly relevant for Roma communities. The service is provided by schools, and can be tested at the local level as school directors usually have some leeway in deciding who to admit to their school. In this experiment schools are contacted by email, which may not be the most typical channel that parents use, but it is not unusual (and probably became more widely used during the pandemic).

The sample (i.e. the schools that we choose to contact out of the total list of schools) includes schools where the proportion of disadvantaged students is below 40 % and excludes schools where there are very few disadvantaged children living in the catchment area of the school. The sample should include at least 500 schools (or more, if possible).

There are four avatars: primary educated Roma/non-Roma and secondary educated Roma/ non-Roma parents who contact the school for information. If low-educated Roma are very unlikely to use email, an alternative setup can be used: it is an informal mentor to the family who writes to the school (both for the Roma and non-Roma parent with primary education).

The roma markers are the surname of the mother, and the name of the child. The parent/mentor writes to the school asking six questions, some of which are needed to make the letter look credible, and some are needed to generate variation in the responses. In most of the questions asked, there is some government regulation but schools have their own rules of implementation. The last question concerns meal subsidies, which is probably provided by the municipality, but the school may have information about it. This is especially good for testing if the school makes an effort to be helpful.

The email message sent by middle-class avatars is this:

> *Dear Mr/Ms Principal [greeting adjusted to the gender of the principal]*
>
> *My name is [very common non-Roma / Roma middle class name]. I would like to move to your neighbourhood as my sister lives there, raising three children on her own. She clearly needs help and I thought we could manage more easily if I lived near her. Currently I work in an elderly care home (it's my fourth year here) and would like to find a similar job after the move. My older son will go to a boarding school in the autumn but my younger son will just turn 6 and I would like to find a school for him near our new home. This is why I am writing to you. Could you please help me by answering my questions. What is the application deadline? When and how can we apply to your school? What is the daily schedule of the school: in the afternoon, when can kids go home? Is it compulsory to stay until 4 PM? At which grade do you start teaching foreign languages? Lastly, can my kid get subsidised school meals (e.g. if mother is a lone parent)?*
>
> *Please help me find a good place for [common middle-class name for boys] and answer my questions.*
>
> *with best regards,*
>
> *[very common non-Roma / Roma middle class name]]*

The outcomes of the experiment can be measured in several indicators: the response rate, the number of questions answered, the tone of the response, or any barriers mentions, etc.

## 8.2. Example of an observation sheet

Observation sheet for visiting a government or municipal office in person.
*If possible, this sheet should be filled in by the avatar immediately after the visit*

OBSERVATION SHEET

avatar:
location:
date:
time of entry to government office:
time of exit from government office:
opening hours on the day of the visit:

**Aspects to be observed**

*Circumstances*
1. How many people are waiting in the office and how many seats are available for those waiting?
2. How warm is it?
3. How much time is left before closing time / lunch?
4. Any other circumstances that might affect the behavior of the administrators?
5. What is the usual procedure for the management of clients' queuing: clients draw a number as they arrive? there is person who allocates clients across officers? other?
6. Was there a security guard at the entrance? If so, where was he (watching the entrants at the gate, talking to cafeteria workers, etc.) exactly at the time of entry?
7. Were there any Roma employees (receptionist, security guard, administrators, cafeteria worker) in the office? If yes, who was it?

*Other clients*
8. At the time of the visit, are there any (other) Roma customers in the office, or outside the office? Is there any visible, noticeable characteristic of their presence (number, behavior, clothing, etc.) that may differ from non-Roma clients?
9. Was there any interaction between other Roma clients and any staff of the office? (This would require some time spent observing before the interaction started in the office.)

*Administrator /officer*
10. What was the administrator like? male or female, 20-30 years old, 40-50 years old, or over 50, did she/he appear to be a member of a discriminated minority (e.g. had a visible disability), or did she/he have a different hairstyle/dress from the usual official attire?

*Quality of your own interaction*
11. What or who helped you find your way around? Signs, security guard, other customer?
12. How would you describe the support provided by the office (Greeting, helpfulness, style of providing information, metacommunication, etc.)
13. How long did you have to wait?
14. Was what the administrator said understandable (i.e. comprehensible to a person not familiar with official procedures)? How did he or she react when you asked for further explanations?
15. Was there any reference to your Roma origin (if yes how) or was there a negative or positive difference in the way the administrator dealt with your case?

16. General impressions.

*Own interaction content*

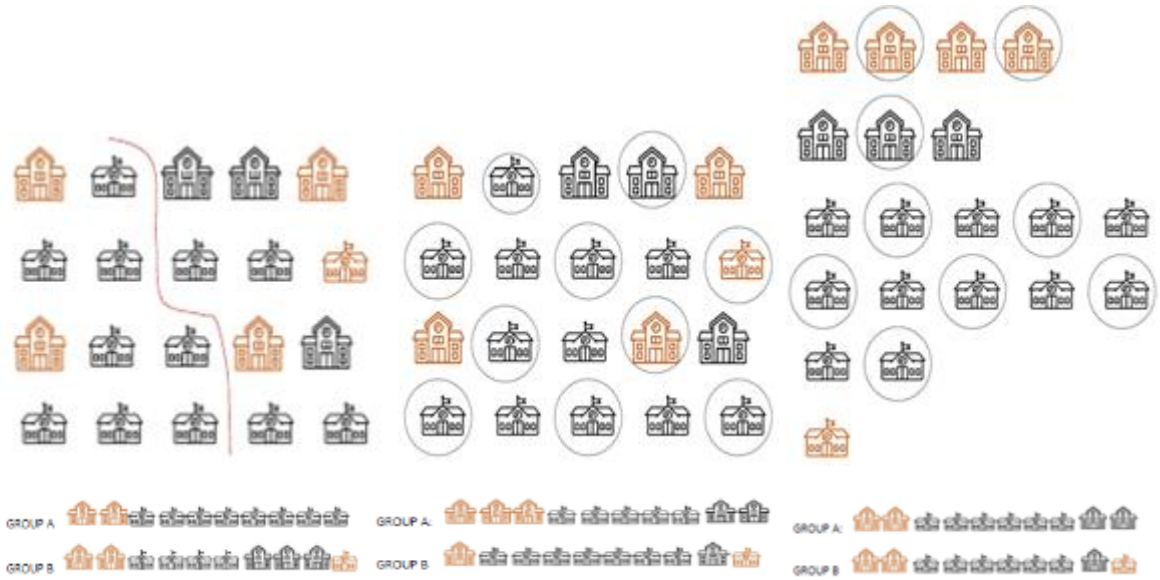
17. From whom and what information did you receive at the office (concerning the case):
(from whom:)..... (what:).....
(from whom:)..... (what:).....
(from whom:)..... (what:).....

18. Was there a question left unanswered by the officer?
19. Why did you not receive an answer?

## 8.3. Randomisation

To see how to split a sample in a random way, imagine an experiment about schools. You have small and big schools and some are public, some are private. How do you divide them into two groups (one to be contacted by a Roma avatar, the other by a non-Roma avatar)? Which of these methods give you a random sample?

**(a)** **(b)** **(c)**




The first method (a) looks "random", but in fact it is not: it does not ensure that schools located in any part of the list have an roughly equal chance to be selected in one or the other subgroup. The second method (b) ensures this, but does not work perfectly as the sample is very small. The last method (c), called stratified random sampling is the most reliable in this case.

Random sampling is important. If the two subgroups are different in a way that affects discrimination, the results of the experiment will not be reliable, they will be biased. E.g. if big schools discriminate less, the (a) type of grouping will show less discrimination if the Roma avatar is sent to the left subgroup and more discrimination if they are sent to the right subgroup (which includes more big schools). If you have a much bigger number of schools (or you know from existing research that school size has no impact on discrimination), (b) is also a reliable method for splitting schools.