The Role of Data Science in Pursuit of Long-Term Relationship Between Respondents and Social Researchers

Kamil Wais

Data Science For Social Good Conference, Chicago, USA
September 29, 2017

The urgent need for declarative data for solving social problems

In order to solve many social problems there is often a need to conduct an ad-hoc social study and gather new quantitative data from population of interest. Gathering additional information is often desirable even if secondary data already exist. Existing secondary data can be not comprehensive enough or detailed enough or not up-to-date. The solution is to ask people relevant questions directly and gather enough declarative data of acceptable quality that can be helpful in answering our research questions and solving a particular social problem.

Traditionally, to gather such declarative data we could conduct a survey using one of off-line or on-line research techniques. But inevitably, sooner or later, we will face common barriers that can prevent even most important social research projects from being conducted: surveying people is expensive, respondents are hard to recruit, response rates are declining, and methodological concerns are rising. The problems are growing even more, if we want to conduct repeated measures studies within longitudinal panels, when we want to return with our questions to the same respondents at least once after some time.

Traditional model of respondent-researcher relationship

If our research projects aim at solving important social problems we often ask ourselves why it is so difficult to conduct the research cost-effectively and with full cooperation between researchers and respondents. One of the reasons is that the process of gathering declarative data is designed from the perspective of a researcher, not from the perspective of a respondent. In a traditional approach respondents' perspective and needs are completely neglected.

The traditional model of Respondent-Researcher Relationship is based on the following heuristic:

1. Find potential respondents within time window that is convenient primarily for a researcher and not necessarily convenient for respondents.
2. Convince the potential respondents to devote their time to work through a questionnaire, which is often boring and of low quality.
3. Convince the respondents to share some valuable, also often quite private and sensitive information with someone who will benefit in some way from their answers.
4. Do not share any valuable information with the respondents.
5. Give the respondents nothing or as little as possible in return in order to be as much

cost-effective as possible.

6. Repeat the whole process described above as many times for as many respondents as necessary.

As a consequence, participation in subsequent surveys is a burden to respondents and is often badly compensated. From this perspective, it is not surprising why respondents do not want to be respondents any more. In the long run this model is clearly not sustainable but social researchers act like they made an assumption that the general population of respondents is infinite or it is an easily renewable resource. These assumptions are not only wrong but also harmful for future studies based on any interactions with respondents.

Such traditional approach was efficient at the onset of survey research and is still preferred when only short-term efficiency is considered. It is safe to assume, that most researchers and research agencies are more interested in instant outcomes than in long-term responsiveness of general population. This situation can be seen as the classic "tragedy of the commons" describing a situation where a shared resource is spoiled and depleted by collective actions of all the actors driven by their individual self-interest which is at odds with the long-term interests of the common good.

Old paradigms in on-line research techniques

Additionally to the respondent-researcher relationship problem, old research paradigms originated from off-line research is still dominant even in on-line survey research techniques. Thus, contemporary on-line research techniques are not natively on-line—they simply mimic off-line techniques. They are mainly off-line questionnaires converted into more or less advanced HTML forms with some additional functionalities (like randomization, skip logic, new question types) but they still inherent off-line characteristics. For example, they are:

1) **repetitive**—many questions from one survey is repeated in other questionnaires; respondents are forced to answer the same questions again, and again (for example social-demographic questions);
2) **time-consuming**—on-line survey need to be completed in one, relatively long, block of time; in paid surveys even if the payment increases non-linearly with the duration of the survey, the marginal utility decreases quickly and the burden of a respondent in longer on-line surveys is often too high, which leads to drop-outs that cannot be prevented with the payments;
3) **linear**—the survey questions need to be answered one by one in linear fashion; there should not be any interruption to this process; non-linear behavior, which is natural for Internet users, is often forbidden in on-line surveys;
4) **based on non-equivalent information exchange**—respondents, after sharing valuable information with researcher, do not get any valuable information in return; in rare cases respondents can access the results of the survey only after a long period or in a form of a non-interactive and non-customized report.

Towards new paradigms in Social Science research techniques

With the use of the full potential of Internet technologies, we should be able to implement a new

approach to on-line declarative data collection that also support long-term respondent-researcher relationship. We need to develop a new model for important social science research, so that:
1) People are eager to become new panelists and respondents.
2) Panelists are willing to stay in research panels and answer new questions.
3) Panelists are willing to help recruiting new panelists among their social networks.

In this framework we need to prioritize the long-term partnership among respondents and social scientists and make answering questions much easier and much more valuable for a respondent. We can do that by introduction a fundamental principle—in the process of information exchange both sides should receive something valuable for them. The process should be more similar to a equivalent conversation between partners about important topic than an interrogation of one side by another.

To streamline the declarative data collection process, we should develop an appropriate on-line research tools that let go of old off-line paradigms in social research. We should be able to ask people the same question once without repeating it. Answers to selected questions should be updated after established expiration period or at respondent's will at any time if needed. Respondents should be able to answer a single question at a time or small set of questions. Respondents should be able to choose their own device (mobile or desktop) and best time to answering our questions without worrying about consequences of interruptions to the survey.

What is perhaps the most important principle for establishing long-term relations with respondents is that they should be able to access feedback relevant to their answers immediately. The feedback should be provided at the end of a question set if not right after each answered question. The value of the feedback will be much higher for a respondent, if the feedback will be customized to this particular respondent. This can be achieved by pre-programming feedback templates which are customized to a particular respondent, using his or her previous answers to this or other questions. The simplistic but real-world example of this approach is a salary survey, which asks you about your current salary and compares it to salaries of people similar to you.
The feedback for a given respondent can be based on different data sources:

- the answer of the respondent to the given question,
- previous answers of the respondent to the given question,
- answers of the respondent to other questions,
- answers of other respondents to given or other questions,
- external (research or administrative) open data,
- outcomes from reference research (aggregated or summarized), and
- other sources.

Necessary components of the new model and the role of Data Science

The transformation of traditional surveys into modern on-line applications for collecting declarative data will not be successful unless three necessary components are provided:

1) **reference data** – as the empirical base for question feedback obtained from various sources (research, administrative, non-governmental);
2) **analytical scripts** – preparing valuable customized feedback based on respondents' answers and the reference data;

3) **Internet technologies** – to fully utilize native Internet environment (mobile, reactive, responsive, nonlinear).

It is obvious that building such on-line data products requires programming skills and deep knowledge of Internet technologies, and behavior patterns of Internet users. However, the most important skills are Data Science skills, which are crucial for data manipulation, preliminary analysis, preparing syntax for question feedback, preparing analysis and data visualizations for question feedback.

The role of Data Science here is to use statistical inference, machine learning algorithms, and interactive data visualization, among others, to provide highly customized, comprehensive and truly valuable feedback to the respondents. At the same time, the feedback must be visually attractive and easily understandable for wider, non-expert audience.

This approach calls for close collaboration among researchers, data scientist, and programmers. Choosing appropriate reference data, preparing them for feedback purposes and providing appropriate interpretation will need collaboration with social and data scientists. Data scientist will need to closely collaborate with programmers to prepare data visualization that can be easily accessible without distortion on mobile devices. One of the main concerns for researchers, data scientists and programmers will be security and confidentiality of sensitive data shared by the respondents and saved within the research tool.

Possible applications and future developments

There are many possible applications for on-line data products with instant feedback. Some of them can be used to provide new declarative data that are necessary for solving chosen social problems. We can use such data products for continuous collections of opinions from local communities, conducting satisfaction surveys of health services or other services provided by private institutions or governmental agencies, assessing performance of local authorities, rating educational and cultural institutions, making forecasts of future social and political events based on aggregated respondents' predictions, and more.

By developing data products that are natively dedicated to on-line declarative data collection and providing instant question feedback to the users, we can collect unique paradata for further examination. With the same methods, we can implement continuous, detailed evaluation of questions and feedbacks provided by the respondents. Additionally, knowing historic data of respondents' answers and their interactions with the research app, we can perform detail survival analysis of panelists in order to improve user experience and the quality of our question-feedback content. User experience can be additionally improved by applying gamification techniques, which still are not popular in traditional on-line research methods.

Data provided by respondents can be used not only for a particular question feedback but also for more comprehensive, detailed and highly customized individual reports, which can be offered to the respondents after answering some larger sets of questions.

Summary

There is an urgent need for new declarative data that can help solve important social problems. However, such data are more and more often difficult to obtain even if the research project is non-for-profit and aims at solving some social problem of great importance.

The main reason for this situation is the perseverance of traditional model of respondent-researcher relationship. This model is harmful to social science research in general, and often under-founded socially important research projects in particular. Additionally, traditional on-line research techniques which collect declarative data are obsolete. They do not fully take advantage of Internet technologies and specificity of the needs of Internet users.

In order to advance declarative data collection for social good, we need to implement new model of long-term respondent-researcher relationship. In this model there is a need for close collaboration between social scientist, programmers and data scientists. This collaboration is necessary for the transformation of old social science research techniques into modern on-line data products for collecting declarative data and providing instant customized feedback for the respondents.

The main goal of these new tools is to support stable on-line panels of respondents willing to participate in important social research projects in exchange for valuable content provided instantly by data scientists via the same research tool.