SURVEYS · ETHICS AND LAW

Privacy, Anonymity and Confidentiality in Survey Data

Survey respondents share personal information on the understanding that it will be protected. This article explains the differences between privacy, confidentiality and anonymity, why removing names is not enough to prevent identification, the main methods of statistical disclosure control, and how privacy concerns affect survey participation.

ModQu research article · 2,738 words · 27 references (APA 7th edition)

Every survey asks people to share information about themselves, and many ask about sensitive matters such as income, health, sexual behaviour or political views. Respondents agree on the understanding that their answers will be protected. Keeping that promise is an ethical duty, often a legal one, and a practical necessity: if people believe their answers might be exposed, they refuse to take part or give less honest answers. Yet protecting survey data has become harder as more data about people are available elsewhere and as computing power makes it easier to match records. This article explains the main concepts and the methods used to protect respondents.

Three related ideas

The terms privacy, confidentiality and anonymity are often used interchangeably, but they mean different things. Figure 1 summarises the distinctions.

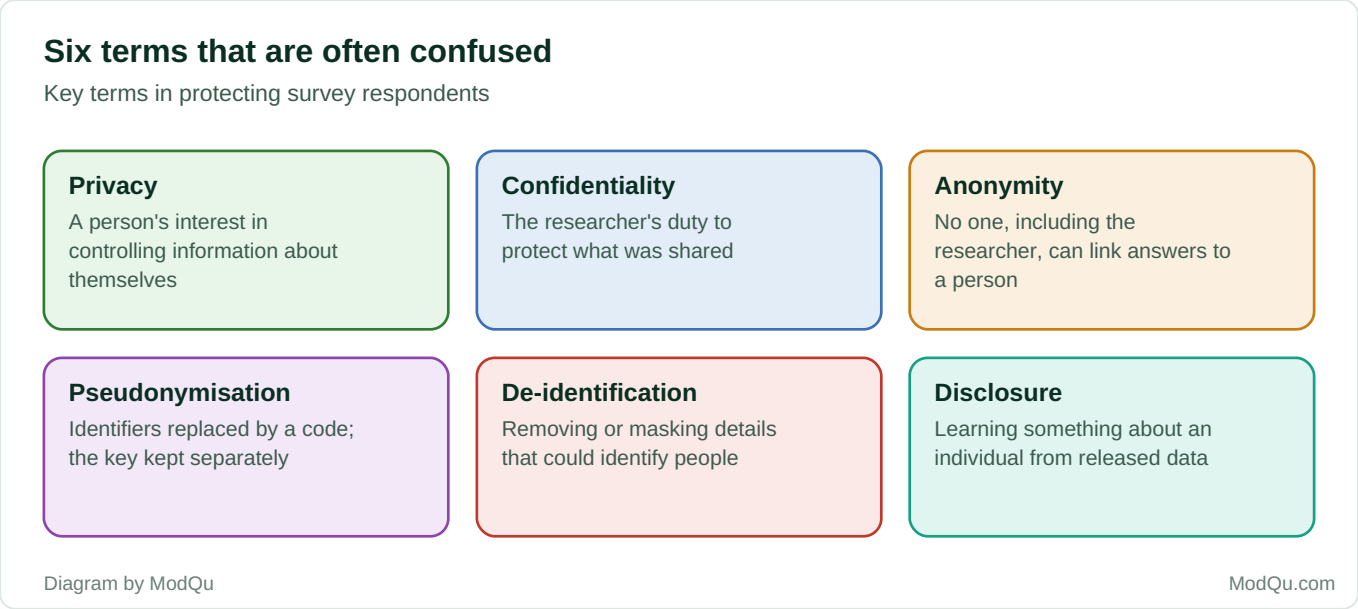


Figure 1. Key terms in protecting survey respondents. Diagram by ModQu.

Privacy is a person's interest in controlling information about themselves and in being left alone. Westin (1967) defined it as the claim of individuals to decide when, how and to what extent information about them is communicated to others. Solove (2006) set out a taxonomy of privacy problems, from collection and processing to dissemination and intrusion. Nissenbaum (2004) argued that privacy is best understood as **contextual integrity**: information flows are appropriate when they match the norms of the context in which the information was shared. A respondent who tells a researcher about their health expects the information to be used for research, not passed to an employer.

Confidentiality is the obligation of the researcher to protect information that has been shared. Confidential data may identify the respondent, but the researcher promises not to reveal identities or to use the data in ways that could harm the respondent.

Anonymity means that the researcher cannot link the data to the individual who provided them. Truly anonymous surveys collect no names, addresses, email addresses or other identifiers, and no information that could be combined to identify someone. Many surveys described as anonymous are in fact confidential, because the researcher holds identifying information, such as an email address used to send reminders, even if it is stored separately from the answers (Wiles et al., 2008).

Pseudonymisation replaces direct identifiers with a code, while the key linking codes to identities is kept separately. Pseudonymised data are still personal data under many data protection laws, because re-identification is possible with the key.

Why privacy matters to survey research

Effects on participation

Concerns about privacy and confidentiality affect whether people take part. In studies of the 1990 United States census, people who were more concerned about privacy were less likely to return their census forms (Singer et al., 1993; Couper et al., 1998). Experiments have shown that people's willingness to take part depends on how likely they think disclosure is and how much harm it would cause, and on the sensitivity of the topic (Couper et al., 2008, 2010).

Assurances of confidentiality help, but only in some circumstances. A meta-analysis found that stronger confidentiality assurances increased response for sensitive topics, but had little effect or even a slightly negative effect for non-sensitive topics, where elaborate assurances may raise suspicion (Singer et al., 1995).

Effects on honesty

Respondents are more likely to report sensitive behaviour when they feel their answers are private. Self-administered questionnaires, which do not require respondents to tell an interviewer, produce higher reports of drug use, sexual behaviour and other sensitive matters than interviews (Tourangeau & Yan, 2007). Ong and Weiss (2000) found that students were more likely to admit to cheating when answers were anonymous than when they were confidential. Special techniques such as **randomised response**, in which a random device determines whether a respondent answers the sensitive question or an innocuous one, protect respondents further, because no individual answer reveals anything certain about the respondent (Warner, 1965).

Privacy during data collection

Privacy matters at the moment of answering as well as afterwards. In face-to-face interviews, other household members may be present and may hear the answers. Studies have found that the presence of a spouse or parent can change answers to questions about relationships, alcohol use or political views. Interviewers are therefore trained to seek a private setting where possible, and sensitive sections are often handed to the respondent to complete on a tablet or laptop, with answers hidden from the interviewer. In telephone interviews, respondents may be overheard, and yes-or-no formats can help by allowing answers that reveal little to listeners. In online surveys, respondents may answer on shared devices or at work, where others may see the screen, and the survey should not leave sensitive answers visible on a device after completion.

Why removing names is not enough

A common belief is that data are safe once names and addresses are removed. This is often false. Many people can be identified from a few ordinary characteristics. Sweeney (2000) estimated that 87 per cent of the population of the United States could be uniquely identified by the combination of five-digit postal code, sex and full date of birth. Rocher et al. (2019) showed that 99.98 per cent of Americans could be correctly re-identified in any data set using 15 demographic attributes, and that even heavily sampled data sets offered little protection.

Figure 2 illustrates how quickly respondents become unique as characteristics are combined.

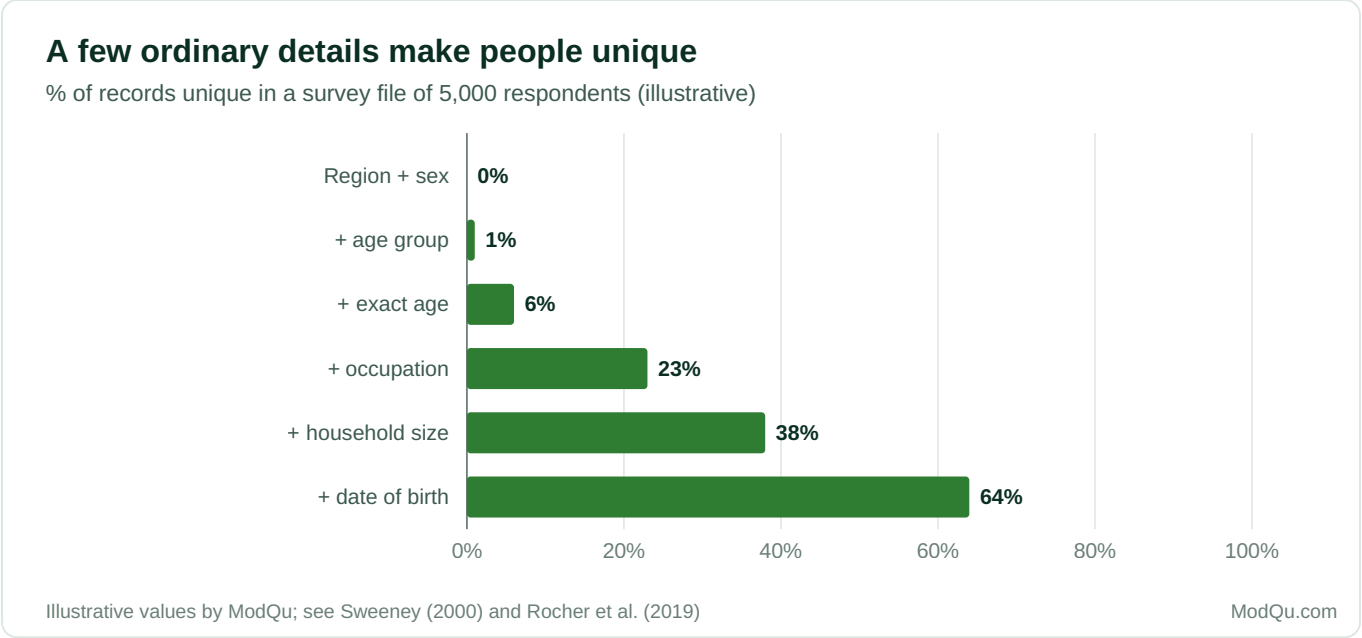


Figure 2. Share of respondents unique in a survey file as more characteristics are combined (illustrative values). Chart by ModQu.

Re-identification can also use data from outside the survey. Narayanan and Shmatikov (2008) showed that users in a large data set of anonymous film ratings could be identified by matching a few of their ratings with public reviews on another website. Ohm (2010) argued that such results undermine the legal assumption that anonymised data are harmless.

For survey researchers, the lesson is that a data set without names can still identify people, especially when it includes detailed geography, exact dates, rare characteristics or free-text answers.

Balancing protection and usefulness

Every method of protection reduces the detail or accuracy of the data to some degree. Grouping ages into bands prevents analysis of single years of age; suppressing small areas prevents local analysis; adding noise reduces precision. The aim is not zero risk, which would require releasing nothing, but an acceptable balance between the risk of disclosure and the usefulness of the data for legitimate research. That balance depends on the sensitivity of the data, the likely motives and resources of anyone trying to identify respondents, the consequences of identification and the safeguards around access.

Statistical disclosure control

Statistical disclosure control is the set of methods used to reduce the risk that individuals can be identified in published tables or released microdata, while keeping the data useful (Willenborg & de Waal, 2001; Hundepool et al., 2012; Duncan et al., 2011).

Assessing risk

The first step is to assess risk. Analysts identify **key variables**, those that an intruder might know from other sources, such as age, sex, region, occupation and ethnicity, and count how many records are unique or rare on combinations of them. Records that are unique in the sample and likely to be unique in the population pose the greatest risk.

Methods for microdata

Methods that reduce the detail in data include:

- **Removing direct identifiers**, such as names, addresses and contact details.
- **Recoding**, such as replacing exact age with five-year bands or detailed geography with larger regions.
- **Top-coding**, replacing very high values of income or wealth with a ceiling.
- **Suppression**, removing particular values for records that would otherwise be unique.

Methods that alter data include:

- **Adding noise**, small random changes to numerical values.
- **Swapping**, exchanging values of some variables between similar records.
- **Microaggregation**, replacing values with averages of small groups of similar records.

A widely used criterion is **k-anonymity**: each record should be indistinguishable from at least k−1 others on the key variables (Sweeney, 2002). If k is 5, no combination of key variables should be shared by fewer than five records.

Methods for tables

Published tables can also disclose information. A table of income by occupation and small area may contain cells with one or two respondents, whose values are thereby revealed. Rules suppress cells below a minimum count, and in business statistics, cells dominated by one or two large firms. Additional cells are then suppressed so that hidden values cannot be calculated by subtraction (Hundepool et al., 2012).

Synthetic data

Synthetic data replace real records with artificial ones generated from a statistical model of the real data. Rubin (1993) proposed releasing fully synthetic data, generated by multiple imputation, so that no real respondent appears in the released file. Partially synthetic data replace only the most sensitive values. Synthetic data can preserve many statistical relationships, but analyses that the model did not anticipate may give misleading results, so they are often used to let researchers develop their analyses before running them on the real data in a secure setting (Drechsler, 2011).

Differential privacy

Differential privacy is a mathematical definition of privacy protection. A method is differentially private if its results are almost the same whether or not any single individual's data are included, which limits what anyone can learn about an individual from the output. This is usually achieved by adding carefully calibrated random noise to statistics (Dwork et al., 2006). The United States Census Bureau adopted differential privacy for the 2020 census, a decision that improved formal protection but raised concerns among some users about the accuracy of figures for small areas (Abowd, 2018).

Special risks

Some kinds of survey data carry particular risks.

Open-ended answers. Free-text answers often contain names of people, employers, places and events that identify the respondent or others. A respondent describing their job may name their employer and position, making them easy to identify. Text answers must be reviewed and edited before release, and automated tools can help flag names and places but rarely catch everything.

Recordings. Audio and video recordings of interviews or of respondents' own answers contain voices and faces, which are directly identifying. They require explicit consent, secure storage and usually transcription with identifying details removed before any sharing.

Paradata and device data. Web surveys automatically capture information such as IP addresses, device types, browser details and the time spent on each question. IP addresses can reveal approximate location and, in some cases, identity. Researchers should decide in advance which paradata are needed and delete or truncate the rest.

Location data. Precise geographic coordinates, whether recorded by an interviewer's device or by a smartphone app, can identify where a respondent lives or works. Location data are usually released only at the level of larger areas, or replaced by characteristics of the area, such as its level of deprivation.

Linked data. Linking survey answers to administrative records or other sources increases the amount of information about each person and therefore the risk of identification. Linked files are usually available only through secure access arrangements.

Panels and longitudinal data

Panel surveys pose particular challenges, because information about the same people accumulates over many years. A characteristic that is harmless in a single wave, such as a change of job, can become identifying when combined with dates of moves, births and other events across waves. Researchers must assess disclosure risk across the whole history of each record, not wave by wave. Panels also hold contact details for many years in order to trace respondents, and these must be protected throughout. When panel members die or withdraw, decisions must be made about whether and how their past data continue to be used, in line with what they were told when they joined.

Sharing and archiving

Data archives play a major part in protecting survey data while making them available for further research. Before depositing data, researchers prepare documentation, assess disclosure risk and apply suitable treatment. Archives typically offer several levels of access: open files that anyone can download, files available to registered users who agree to conditions of use, and controlled files available only to approved researchers in secure settings. The conditions of use usually forbid attempts to identify respondents and require users to report any accidental identification. Sharing in this way greatly increases the value of survey data, but it depends on the consent that respondents gave and on the promises made to them at the time.

Controlled access

Not all protection comes from changing the data. Many organisations provide detailed data only in controlled settings. The **Five Safes** framework, widely used by statistical agencies and data archives, considers five dimensions of access (Desai et al., 2016; Ritchie, 2017):

- **Safe projects**: is the use appropriate and in the public interest?
- **Safe people**: can the researchers be trusted to use the data properly?
- **Safe settings**: does the access environment limit unauthorised use?
- **Safe data**: has the data been treated to reduce disclosure risk?
- **Safe outputs**: are results checked before release to ensure they do not reveal individuals?

Under this approach, less altered data can be made available when the other safeguards are strong, such as in a secure data laboratory where researchers' outputs are checked before they leave.

Practical protections in survey projects

Beyond formal disclosure control, good practice in everyday survey work includes:

- **Collecting only what is needed.** Every identifying or sensitive item collected adds risk.
- **Separating identifiers from answers.** Contact details used for fieldwork should be stored separately from survey responses and deleted when no longer needed.
- **Securing data.** Encryption, access controls and secure transfer protect data during collection, processing and storage.
- **Training staff.** Interviewers, coders and analysts should understand confidentiality obligations and sign agreements.
- **Checking free text.** Open-ended answers often contain names, places and details that identify respondents or others, and should be reviewed before release.
- **Caution with small groups.** Results for small groups, such as small teams in employee surveys, can identify individuals, so minimum reporting sizes should be set.
- **Accurate promises.** Respondents should be told exactly what protection they will receive. Describing a confidential survey as anonymous is misleading.

When things go wrong

Despite precautions, breaches can occur: a laptop with survey data may be lost, a file may be sent to the wrong person or a system may be hacked. Organisations should have a plan for such incidents, including how to contain the breach, assess its seriousness, inform those affected and, where the law requires, notify the data protection authority. Reviewing incidents and near misses helps to strengthen procedures. Respondents' trust, once lost, is hard to regain, and a single well-publicised breach can reduce participation in surveys well beyond the organisation responsible.

Legal frameworks

Many countries regulate the processing of personal data. The European Union's General Data Protection Regulation applies to any information relating to an identified or identifiable person, which includes pseudonymised survey data. It requires a lawful basis for processing, limits data to what is necessary, requires security and gives individuals rights of access and erasure, with some exceptions for research (European Parliament & Council of the European Union, 2016). Data that are truly anonymous, so that individuals can no longer be identified by any means reasonably likely to be used, fall outside its scope. Statistical agencies are also bound by statistics laws that forbid the release of identifiable information and make breaches a criminal offence.

An example

A research team plans to release the microdata from a national survey of 12,000 adults about health and employment. A risk assessment finds that, using region, age, sex, occupation and ethnicity, 9 per cent of records are unique in the sample, most of them older people in rare occupations in small regions. The team recodes age into five-year bands, groups small regions together, merges rare occupations into broader categories and top-codes income. After these changes, fewer than 0.5 per cent of records are unique, and those few are treated by local suppression. The public file is released through a data archive, while the detailed file, with full geography and exact ages, is made available only in a secure setting for approved projects, with outputs checked before release (illustrative example).

Conclusion

Privacy, confidentiality and anonymity are distinct ideas, and confusing them can lead to promises that cannot be kept. Removing names does not make data anonymous, because ordinary characteristics can identify people. Statistical disclosure control, synthetic data, differential privacy and controlled access each reduce the risk of identification, at some cost to the usefulness of the data. Choosing the right combination, and being honest with respondents about the protection they will receive, keeps the trust on which survey research depends.

References

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (p. 2867). Association for Computing Machinery. https://doi.org/10.1145/3219819.3226070

Couper, M. P., Singer, E., & Kulka, R. A. (1998). Participation in the 1990 decennial census: Politics, privacy, pressures. *American Politics Quarterly, 26*(1), 59–80. https://doi.org/10.1177/1532673X9802600104

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics, 24*(2), 255–275.

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2010). Experimental studies of disclosure risk, disclosure harm, topic sensitivity, and survey participation. *Journal of Official Statistics, 26*(2), 287–300.

Desai, T., Ritchie, F., & Welpton, R. (2016). *Five Safes: Designing data access for research* (Economics Working Paper Series No. 1601). University of the West of England.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: Theory and implementation*. Springer. https://doi.org/10.1007/978-1-4614-0326-5

Duncan, G. T., Elliot, M., & Salazar-González, J.-J. (2011). *Statistical confidentiality: Principles and practice*. Springer. https://doi.org/10.1007/978-1-4419-7802-8

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Eds.), *Theory of cryptography* (pp. 265–284). Springer. https://doi.org/10.1007/11681878_14

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union, L 119*, 1–88.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., & de Wolf, P.-P. (2012). *Statistical disclosure control*. Wiley. https://doi.org/10.1002/9781118348239

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy* (pp. 111–125). IEEE. https://doi.org/10.1109/SP.2008.33

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*(1), 119–157.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review, 57*, 1701–1777.

Ong, A. D., & Weiss, D. J. (2000). The impact of anonymity on responses to sensitive questions. *Journal of Applied Social Psychology, 30*(8), 1691–1708. https://doi.org/10.1111/j.1559-1816.2000.tb02462.x

Ritchie, F. (2017). *The "Five Safes": A framework for planning, designing and evaluating data access solutions* [Conference paper]. Data for Policy 2017, London. https://doi.org/10.5281/zenodo.897821

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications, 10*, Article 3069. https://doi.org/10.1038/s41467-019-10933-3

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 461–468.

Singer, E., Mathiowetz, N. A., & Couper, M. P. (1993). The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. Census. *Public Opinion Quarterly, 57*(4), 465–482. https://doi.org/10.1086/269391

Singer, E., Von Thurn, D. R., & Miller, E. R. (1995). Confidentiality assurances and response: A quantitative review of the experimental literature. *Public Opinion Quarterly, 59*(1), 66–77. https://doi.org/10.1086/269458

Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review, 154*(3), 477–564. https://doi.org/10.2307/40041279

Sweeney, L. (2000). *Simple demographics often identify people uniquely* (Data Privacy Working Paper No. 3). Carnegie Mellon University.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*(309), 63–69. https://doi.org/10.1080/01621459.1965.10480775

Westin, A. F. (1967). *Privacy and freedom*. Atheneum.

Wiles, R., Crow, G., Heath, S., & Charles, V. (2008). The management of confidentiality and anonymity in social research. *International Journal of Social Research Methodology, 11*(5), 417–428. https://doi.org/10.1080/13645570701622231

Willenborg, L., & de Waal, T. (2001). *Elements of statistical disclosure control*. Springer. https://doi.org/10.1007/978-1-4613-0121-9