Matching survey responses with anonymity in environments with privacy concerns: A practical guide

Dominik Vogel

ORCID: [0000-0002-0145-7956](https://orcid.org/0000-0002-0145-7956)

This is the postprint version of the following article: Vogel, D. (2018). Matching survey responses with anonymity in environments with privacy concerns: A practical guide. *International Journal of Public Sector Management*, 31(7), 742-754. <https://doi.org/10.1108/IJPSM-12-2017-0330>.

Abstract

Purpose: In many cases, public management researchers' focus lies in phenomena, embedded in a hierarchical context. Conducting surveys and analyzing subsequent data requires a way to identify which responses belong to the same entity. This might be, for example, members of the same team or data from different organizational levels. It can be very difficult to collect such data in environments marked by high concerns for anonymity and data privacy. This article suggests a procedure for matching survey data without compromising respondents' anonymity.

Approach: The article explains the need for data collection procedures, which preserve anonymity and lays out a process for conducting survey research that allows for responses to be clustered, while preserving participants' anonymity.

Findings: Survey research, preserving participants' anonymity while allowing for responses to be clustered in teams, is possible if researchers cooperate with a custodian, trusted by the participants. The custodian assigns random identifiers to survey entities but does not get access to the data. This way neither the researchers nor custodian are able to identify respondents. This process is described in detail and illustrated with a fictitious research project.

Originality/value: Many public management research questions require responses to be clustered in dyads, teams, departments, or organizations. The described procedure makes such research possible in environments with privacy concerns; this is the case with many public administrations.

Keywords: matching; survey research; unique identifiers; data privacy; multilevel; custodian

Article Classification: Technical paper

Why is matching important?

The embeddedness of public administration in a multilevel structure comprising international, supranational, national, regional, and local organizations is a well-studied aspect of public administration (Benz, 2015; O'Toole, 1997). Hierarchical structures, however, are not only relevant at macro-level where organizations from different levels interact, but also at micro-level where actors within organizations interact. Aiming to provide clear levels of responsibility and hold members of public administration accountable for their actions, hierarchies are a central characteristic of public administration. Max Weber (1978) defined hierarchy as one of the building blocks of an ideal type bureaucracy. Hence, hierarchical structures are very relevant in public management research.

Such hierarchical structures result in a situation where research entities of interest for public management researchers are nested in a variety of contexts. Thus, the things we are interested in are not independent of each other. They share, instead, the same context factors (LaHuis et al., 2014; Snijders and Bosker, 1994, p. 6). Employees are nested in teams, departments, and agencies; government-owned enterprises in municipalities; pupils in classes and schools, and so forth. Figure 1 illustrates the clustered structure of many populations studied in public management research. In the displayed example, individual employees (level 1) are nested in teams with the same leader (level 2) and those teams are again nested in organizations (level 3).

Figure 1: Example of clustered structure of a population studied in public management research.¹

¹ Icons designed by Freepik from Flaticon

Such a nested or clustered structure becomes especially important when researchers choose to apply a quantitative approach to answer their research questions. In the last decade, this has become increasingly common in public management research. As Groeneveld et al. (2015) underline, this has become the dominant approach in recent years. They show that in 2010, a majority of empirical articles in the most important journals in the field used a quantitative research approach. Prominent research topics like public service motivation are even predominantly quantitative (Groeneveld et al., 2015; Ritz et al., 2016). Groeneveld et al. (2015) therefore agree with other scholars (e.g., Pitts and Fernandez, 2009) stating that public administration research is becoming increasingly quantitative. This trend can be seen as a broader development in the field, aiming to, on the one hand, emphasize the rigor of applied methods, and on the other make use of so far neglected methods like experiments, ethnography, or natural language processing (Grimmelikhuijsen et al., 2017).

With the increasing use of quantitative methods, currently predominantly based on survey data, the clustered nature of many research entities of interest gains importance for researchers in all parts of the public administration discipline. The clustering affects survey-based research efforts in different ways, depending on the study's purpose. In most cases, a study does not focus on the clustered structure itself. Nevertheless, it is important to consider the clustered nature of research entities, as ignoring it results in distorted statistical results. This is, for example, the case when employees from a limited amount of public organizations are surveyed (e.g., Kroll and Tantardini, 2017) or if scholars are interested in explaining students' test scores by using personal characteristics (e.g., Lubbers et al., 2010). Fitting a statistical model like an ordinary least squares regression without considering the clustering results in correlated standard errors and therefore violates one of the basic assumptions of such models (Bliese, 2002). Consequently, the probability of falsely concluding that there is an effect when actually there is no effect (Type I error) increases strongly (Steenbergen and Jones, 2002). This is even the case when standard errors are only slightly correlated (Barcikowski, 1981).

More and more researchers are interested in the hierarchical structure of public administration and nesting of actors within the public sector. Since employees are nested in teams and students in classes, they have something in common: the same context factors (LaHuis et al., 2014). Employees in the same team have the same supervisor and comparable working environments. Employees in the same organization are subject to the same top-level leadership, budgetary situation, human resources management efforts, and many more. Students in the same class share the same teacher, room, timetable, and so forth. These level-2 variables can be of great interest. Integrating context variables into theories can be one way of responding to the recent call for "[...] reflexivity about how the domains of theory and method interact [...]" (Pandey, 2017) and to reconsider the "big picture" (Pollitt, 2016; Roberts, 2017). Theorizing about public administration in its hierarchical context and testing

such theories by using hierarchical data can significantly contribute to the advancement of public management research. Data from different hierarchical levels also helps researchers to reduce common method bias (Podsakoff et al., 2003), thereby addressing a widely-discussed methodological issue in public administration research in recent years (George and Pandey, 2017; Jakobsen and Jensen, 2015).

Irrespective of whether researchers want to avoid biased estimations caused by clustered standard errors or address a multilevel research question, they are faced with the challenge to collect such data. This is especially difficult if individual-level, as well as level-2 data, are gathered from questionnaires addressing different hierarchical levels. Researchers need to reproduce the nested structure in the data to test a context variable's effect (level-2 variable) on an individual outcome (level-1 variable) or to cluster standard errors. They must be able to assign every level-1 observation to a level-2 observation. This is the matching procedure.

At first sight, this requirement hardly seems problematic. If it is important to know to which level-2 entity a respondent belongs, then corresponding information can be collected with pertinent data (e.g., Burtcher and Oostlander, forthcoming; Alimo-Metcalfe and Alban-Metcalfe, 2006), like asking participants explicitly which unit is theirs. However, such questions reduce or eliminate their anonymity, and they can refuse to divulge such private information (Singer et al., 1995).

In certain contexts, bypassing or reducing participants' anonymity can have even more serious consequences. For example, data privacy is a highly sensitive topic in many countries (Morey et al., 2015). Various organizations, especially in the public sector, are also more concerned with data privacy than others; this affects how their members respond to sensitive data collection efforts. In other cases, legal regulations or organization policies can prohibit the collection of identifying data; regulations also tend to impose restrictions on storing or processing such data. Even without formal prohibitions against collecting, storing, and processing this information, key stakeholders (e.g., organizational heads, data protection officers, work councils, single managers) could disagree and halt the study or prevent the organization's members from participating. In other cases, data collection raises ethical questions, which become the focus of attention of more and more public administration scholars (Van Thiel, 2014, p. 154ff.; Wessels, and Visagie, 2017). Several research questions require collecting data on sensitive issues like corruption, assessment of leaders, health issues, motivation, turnover intention, and many more. Researchers have an ethical obligation to protect answers to such questions. Losing such data or risking de-anonymization of respondents can have severe consequences for them. Collecting identifying information together with responses to sensitive issues therefore strongly increases ethical obligations for researchers. In many such cases, it is more ethical not to collect such data.

Consequently, many cases require a solution to identify clusters of participants, while preserving their anonymity and willingness to provide data. This article describes such a procedure and also includes practical advice for implementing a data collection strategy, which can pool participants without damaging their anonymity. It is based on the use of a custodian who assigns random identifiers to clusters of participants (e.g., teams or departments) and the separation of access to raw data from access to the assignment of identifiers. Although integrating a custodian into data collection is not a new idea and is frequently used in medicine (e.g., Mulligan, 2001) or public health research (e.g., Brown et al., 2017), it is rarely done to collect survey data in a multilevel context. Additionally, the described approach seems to be rather unknown in public administration research. I, therefore, outline the suggested procedure in this article, using a fictitious example from leadership research to illustrate the challenges, as well as necessary steps associated with such a process.

Matching

The discussed procedures aim at identifying which participants in a study belong to the same entity.² Specifically, public management research often seeks to learn which participants work in the same team, division, department, or organization (depending on the research question). Four different terms can describe this identification effort: matching, pooling, grouping, or linking.

For the study here, matching is the most appropriate term. The study seeks to match observations of the same unit and then use this information to analyze the data. Pooling and grouping are not appropriate terms, because they describe procedures for aggregating multiple observations into single data points, like in pooled regressions (Wooldridge, 2010, p. 191ff.) or when observations from the same unit are aggregated into single data points (Woehr et al., 2015). In both cases, information about data clustering is already available. Linking describes different procedures (e.g., record linkage), by which data about the same research object (person, organization, or country) are connected from different sources, and then analyzed (Herzog et al., 2007).

Matching involves identifying survey responses, belonging to the same (organizational) entity and using this information in subsequent analyses. These analyses could involve cluster information to correct standard errors (Binder and Patac,

² This research goal differs from aiming to identify the same participants in longitudinal research designs. In the latter case, self-generated identifiers are applicable (Kearney et al., 1984) cannot be applied to the problem posed in the present study without bypassing anonymity.

1994), hierarchical linear modeling (Bryk and Raudenbush, 1992), or building higher-level constructs (Woehr et al., 2015), like organizational climate (Pritchard and Karasick, 1973), leader–member exchange quality (Graen and Uhl-Bien, 1995), or organizational social capital (Tantardini and Kroll, 2015).

Different approaches for collecting matching-enabling data

Researchers can use different approaches, summarized in Table 1, to collect data, supporting matching respondents with entities. All have distinct advantages and disadvantages. In the first approach the questionnaire will *directly ask* the participants to provide the necessary information (e.g., team membership or department). Although it is the easiest way to ensure that data can be matched, it is problematic if participants do not want to give up their anonymity. In this case, they can resist from answering the necessary questions or refuse to participate at all. Since they could be confronted with their answers, it can also increase social desirability bias. That identifying information are missing or people refuse to participate can be avoided by using procedures, which *automatically collect* the necessary information. This can, for example, be done by marking printed questionnaires with an ID or using online survey tools, which generate and send individualized links to a survey and offer to connect the link to a participant's email address. This, however, does not solve the problem of refusal of participation and potential social desirability bias, because in line with various codes of conduct for research (e.g., American Psychological Association 2002; 2016) participants have to be informed about the collection of identifying information. It is tempting to *collect the identifying information secretly*, which solves the problems of potentially lower response rates and social desirability. However, it is not only unethical and a violation of the code of conduct for researchers, but also illegal in many countries. In the European Union, for example, the new General Data Protection Regulation (GDPR) forbids the collection of personal data without consent. Furthermore, participants can detect such hidden data collection quite easily. They can spot IDs on printed questionnaires, and know that unique hyperlinks contain unique elements. If they suspect that their anonymity is secretly annulled, they are much more likely to refuse participation and it could also severely damage the researcher's reputation. Researchers should, therefore, not follow this dubious path.

Table 1: Advantages and disadvantages of different approaches to collect matching-enabling data

Approach	Advantages	Disadvantages
Request identifying information	Easy and transparent	No/limited anonymity of participants, resistance to answer questions revoke anonymity, lower response rates, social-desirability bias
Automatically collect identifying information	Relatively easy, no missings for identifying information	No/limited anonymity of participants, lower response rates, social-desirability bias
Secretly and automatically collect identifying information	Relatively easy, no missings for identifying information, participants do not fear identification	unethical, illegal in many countries, no/limited anonymity of participants
Use custodian to anonymize data	Increases trust of participants, no missings for identifying information, no/limited anonymity of participants	High trust in custodian necessary, difficult, involves third party, custodian could act in collusion with researchers to disrupt anonymity
Use custodian to assign random identifiers	Increases trust of participants, no missings for identifying information	Difficult, involves third party, custodian could act in collusion with researchers to disrupt anonymity

The two remaining approaches involve a data custodian (Kelman et al., 2002; Schnell et al., 2013, p. 246f.), who functions as a link between the researchers and the organization to increase the participants' trust in the maintenance of their anonymity. The easiest approach would be the *combination of a custodian and the direct collection* of identifying information. In this approach, a custodian collects the data and replaces the identifying information with random IDs. The researchers are not able to bypass participants' anonymity. The participants need to trust the custodian implicitly because he or she receives all data and could check their answers or disclose the raw data to third parties. This problem can be avoided by separating raw data from the collection of identifying information. This can be achieved by asking a *custodian to assign random identifiers* to the participants. The approach is built on the principle that no one should have all the information, necessary to bypass

the respondents' anonymity. The researchers provide a list of random identifiers, and the custodian assigns a unique identifier to each organizational entity, included in the invitation to participate. The participants then use the identifier provided in the invitation to complete the survey. Alternatively, the identifier can already be included in a pre-printed questionnaire or used to generate unique links to an online survey. Although this procedure enables the custodian to obtain a list, revealing which random identifier has been assigned to which entity, the custodian cannot access the data. The researchers receive the data, including the random identifiers, but not the list, connecting the random identifiers with the entities. Each party keeps its information secret so that no single party has all the necessary information to disrupt anonymity. Figure 2 gives a detailed overview of all the steps for this approach.

Figure 2: Steps for a data collection process, which uses a custodian to assign random IDs.³

³ Icons designed by Freepik from Flaticon

Limitations and challenges of using a custodian to assign random IDs

A custodian who assigns random identifiers to the participants and who, therefore, separates the identifying information from the data, has many advantages. The participants' anonymity can be maintained, while researchers can match their responses. This, however, comes at a price (Kelman et al., 2002; El Emam et al., 2010). The procedure is time-consuming and requires much effort of the researchers. The surveyed organization(s) must be involved and must agree to participate in the process. Since the custodian needs a list of contact details to assign the random identifiers and participants need to trust the custodian, it is impossible to survey organizational members without previous consensus and agreement. Furthermore, due to the complexity of the process, it is difficult to research many organizations. The proposed procedure is, however, well suited for surveying most, if not all, members of a small number of organizations.

It is also not impossible to combine the raw data with the matching list. If the custodian hands the list with the assigned identifiers to the researchers or if the researchers provide the custodian with the raw data, anonymity could be impaired. Assuring the participants that such sharing will not happen, is therefore of utmost importance. The following two measures can help to establish the procedure's trustworthiness: (1) choose a custodian whom the participants trust, (2) sign a formal agreement with the organization. Trustworthy custodians are typically members of the works council, an organization's data protection officer, or someone in a comparable position and trusted by the employees (e.g., the monitoring officer). In certain cases, an external person could function as the custodian if all parties involved agree. A formal agreement concerning the data generation process could restrain the researchers from disclosing the raw data to third parties that should further increase the participants' trust in the research team. Asking the custodians to inform the participants before the procedure and give them an opportunity to ask questions can also be helpful. The researchers should also communicate that the organization's top management and work council support the study. Researchers should avoid the impression that participants are obligated to participate, as forced participation might result in lower data quality (Olson, 2013) and mischievous responders (Robinson-Cimpian, 2014).

The participants must ultimately trust the custodian because it remains possible to interfere with anonymity. Combining survey and archival data (e.g., turnover rates of teams) would also be more difficult because a custodian is involved. The data also have limited practical use for the surveyed organizations, because the entities' anonymity prevents the extraction of target-oriented measures from the results. For example, it is not possible to identify entities with low motivation, high dissatisfaction, or bad leadership.

The procedure may also introduce additional sources of error, like mistakes by the custodian when assigning the identifiers and communicating them to various entities. Other errors can occur when participants copy their identifiers from the invitation to the questionnaire; certain participants can also refuse to type in the identifier or intentionally type in a wrong one. There are two key procedures, which help to avoid some of the participants' errors. Use paper-based questionnaires and ask the custodian to write the identifier in beforehand or use questionnaires with identifiers already printed on them. Online survey tools can generate individual links for each entity, asking custodian to assign these links instead of the identifiers. However, the survey tool must provide a means to save the links participants have used, as well as their responses.

The following steps can be used as a guideline to minimize this procedure's limitation, ensuring successful data generation:

1. Reach consensus with surveyed organization(s) about data generation procedure and subsequent reporting.
2. Find a custodian whom participants trust.
3. Optional: Sign an agreement with the organization in terms of which data generation, custodian, and data protection are regulated.
4. Instruct the custodian thoroughly about the procedure.
5. Optional: Inform the participants about the survey and matching system. If possible, highlight the support of the top management, custodian, and data protection officer.
6. Generate a list of unique identifiers and provide it to the custodian.
7. The custodian assigns an identifier to each entity.
8. The custodian sends an identifier, as well as a survey link or paper-based questionnaire, to each member of the entities involved.
9. Optional: The custodian sends a reminder to the participants.
10. Analyze the data and write a report.

This procedure has major advantages, especially relevant in environments where participants are highly concerned about their anonymity or data privacy in general. The main advantage is that it offers a credible means of assuring them that their anonymity will be protected, while allowing the responses from each entity to be matched. In contrast to procedures whereby anonymity cannot be ensured, this procedure can potentially increase the response rate; it can even convince a previously resistant organization to participate in the survey. The following section offers a concrete example of a research project, using the proposed procedure.

Example: Employee survey on leadership and attitudes

To illustrate the proposed procedure one can imagine the following research project as a fictitious example. We imagine that we aim to investigate whether a good relationship between a manager and subordinate increases the follower's job satisfaction. The relationship is supported by the leader-member exchange theory (LMX) (Graen and Uhl-Bien, 1995).⁴ In a relationship, classified as "[...] high-quality LMX, leaders provide intangible and tangible resources to members [...]" (Erdogan and Enders, 2007, p. 321), resulting in higher job satisfaction.

To test the main hypothesis whether high-quality LMX positively affects the employees' job satisfaction, researchers rely on an employee survey at a large public agency. The questionnaire uses widely applied scales (e.g., the Job Descriptive Index by Smith et al. (1969) to measure job satisfaction and a scale to measure the quality of interactions between leaders and members by Bhal and Ansari (1996)). After initial contact, the organization's top management approves the research project (step 1). They signal that they are interested in the central question and believe that the study provides insights to the agency. To ensure top management's support, the researchers agree to write a report on the results. Although researchers can ask participants directly about which teams they belong to, they are concerned that the work council would not agree and would recommend that employees refuse to participate. The researchers also believe that employees would not participate if their answers were not completely anonymous. This is a critical phase in the development of such a research project because failing to address work council's, or employees' concerns can easily end a project. If a substantial part of the organization refuses to participate, the collected data is biased or very small and cannot be used.

To apply the proposed procedure, the researchers inform the work council about the study. After a presentation at a meeting, the council agrees to assign a member who would act as a custodian (step 2). It also agrees to inform the staff in the next staff newsletter about the study and the work council's support for it. To further allay the skeptics' fears, the researchers consent to sign a non-disclosure agreement with the agency, assuring that they will neither share the survey's raw data, nor participate in actions to de-anonymize the survey's participants (step 3).

Afterward, the custodian receives thorough instructions from the researchers, including detailed introduction of how to assign the identifiers to the organizational units, as well as the contact details of a person she could call any time to resolve possible problems or answer questions (step 4). This is very important as the custodian might not fully understand the procedure's purpose and might not be aware of the severe consequences small mistakes can have. It is therefore advisable to

⁴ Substantial literature confirms this relation; see Gerstner and Day (1997) for an overview.

thoroughly instruct the custodian and assure them that they can contact the researchers at any time. To assign the identifiers to the units, the custodian receives a file from the researchers, containing unique identifiers consisting of random combinations of four digits and letters (Step 6). The appendix provides an exemplary code for generating such a list in *R* (R Core Team, 2018). The custodian then uses the list to assign an identifier to each team of the organization (step 7).

In certain cases, it can help to have information about not only the organizational unit to which a response belongs, but also which part of the organization the unit belongs to. For example, the researchers could be interested in the effect organizational departments have. To collect this information, the custodian can assign an identifier with the same first character to all the units in the same department. However, this step partially challenges anonymity, because the number of units per department could provide information about the respective department.

Next, the custodian sends an e-mail invitation with the team's identifier and a link to the survey to each member of every team (step 8). The emails could be distributed team-by-team or by using a mail merge procedure. On the first page, the survey asks the participants to type in their assigned team identifier. This way, the identifier is stored with their responses. After a reasonable time, the custodian reminds the organization's staff to encourage more responses (step 9).

After the data collection period ends, the researchers analyze the data with hierarchical linear models or ordinary least square regressions with adjusted standard errors to answer the research question (step 10). Such data-analytical techniques are only possible if researchers are able to model the data's hierarchical structure. In this example, employees' job satisfaction is used as the dependent variable and LMX quality as the independent variable. Some socio-demographic variables like age, tenure, and gender are used as control variables. The team identifier is used to adjust standard errors in the regression. Without using the described procedure, the researcher would have to ask participants to indicate their team membership that most certainly would result in resistance by the work council and massive reduction of the response rate. Collecting data without any clustering information would leave researchers without any possibility to adjust the standard errors in a regression. This strongly increases the likelihood of a Type-I-error (Bliese, 2002; Steenbergen and Jones, 2002). Hence, the researchers might conclude that LMX quality increases employees' job satisfaction, although this is not the case. The conclusion would be based on a statistical model, which violates its basic assumption (i.e. independence of errors).

Conclusion

This article proposes a procedure for survey research that supports matching responses from the same organizational entity without revealing the respondents' identity. This option is especially helpful in contexts like the public sector, where respondents or organizations are highly concerned about data privacy or where surveys deal with sensitive issues, like well-being, leadership, health, coworker relations, or comparable topics. This is especially important in the case of public organizations as they are often monopolists and therefore easy to detect from data. Taking measures to sustain participants' anonymity is therefore essential for protecting them from, for example, media exposure.

The proposed procedure features a custodian, trusted by the potential participants (e.g., a member of the work council) and who assigns unique identifiers to organizational entities. Anonymity is ensured by separating the opportunity to match identifiers with entities (custodian) from the raw data (researchers). The procedure has already been used, with well-confirmed applicability, in an actual research project on antecedents of leadership behavior (Vogel, 2016).

The described approach contributes to an overall discussion about the rigor of public administration research (Del Rosso, 2015) and the call for increased integration of theoretical and methodological considerations (Pandey, 2017), as well as rising attention to research ethics (Van Thiel, 2014, p.154ff.; Wessels, and Visagie, 2017). A data collection approach similar to the one presented in this article enables scholars to develop and test theories which better fit the hierarchical structure of the things we study: employees, teams, managers, and organizations. The possibility of matching responses for the same entities should be considered as an opportunity to go beyond the simple correction of clustered standard errors. It offers opportunities to theorize about the effects of one hierarchical level on another and effects of contextual (level-2) variables. Recent research already seized the opportunity to develop and test multilevel theories in the field of public service motivation (Liu et al., 2017), organizational social capital (Kroll et al., 2017), and leadership (Jacobsen and Andersen, 2015). The described data collection approach, therefore, provides researchers with additional procedures to improve public administration theory.

References

- Alimo-Metcalfe, B. and Alban-Metcalfe, J. (2006), "More (good) leaders for the public sector", *International Journal of Public Sector Management*, Vol. 19 No. 4, pp. 293–315.
- American Psychological Association (2002), "Ethical principles of psychologists and code of conduct", *American Psychologist*, Vol. 57 No. 12, pp. 1060–1073.
- American Psychological Association (2016), "Revision of Ethical Standard 3.04 of the "Ethical Principles of Psychologists and Code of Conduct" (2002, as amended 2010)", *American Psychologist*, Vol. 71 No. 9, p. 900.
- Barcikowski, R.S. (1981), "Statistical Power with Group Mean as the Unit of Analysis", *Journal of Educational and Behavioral Statistics*, Vol. 6 No. 3, pp. 267–285.
- Benz, A. (2015), "European Public Administration as a Multilevel Administration: A Conceptual Framework", in Bauer, M.W. and Trondal, J. (Eds.), *The Palgrave handbook of the European administrative system*, Palgrave Macmillan, Houndmills; Basingstoke; Hampshire; New York, NY, pp. 31–47.
- Bhal, K.T. and Ansari, M.A. (1996), "Measuring quality of interaction between leaders and members", *Journal of Applied Social Psychology*, Vol. 26 No. 11, pp. 945–972.
- Binder, D.A. and Patak, Z. (1994), "Use of estimating functions for estimation from complex surveys", *Journal of the American Statistical Association*, Vol. 89 No. 427, pp. 1035–1043.
- Bliese, P.D. (2002), "Multilevel Random Coefficient Modeling in Organizational Research: Examples Using SAS and S-PLUS", in Drasgow, F. and Schmitt, N. (Eds.), *Measuring and analyzing behavior in organizations: Advances in Measurement and Data Analysis*, Jossey-Bass, San Francisco, CA, pp. 401–445.
- Brown, A.P., Ferrante, A.M., Randall, S.M., Boyd, J.H. and Semmens, J.B. (2017), "Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage", *Frontiers in public health*, Vol. 5, p. 34.
- Bryk, A.S. and Raudenbush, S.W. (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods*, SAGE Publications, Thousand Oaks, Calif.
- Burtscher, M.J. and Oostlander, J. (forthcoming), "Perceived mutual understanding (pmu)", *European Journal of Psychological Assessment*, available at: <https://doi.org/10.1027/1015-5759/a000360>.
- Del Rosso, S.J. (2015), "Our New Three Rs: Rigor, Relevance, and Readability", *Governance*, Vol. 28 No. 2, pp. 127–130.
- El Emam, K., Brown, A., AbdelMalik, P., Neisa, A., Walker, M., Bottomley, J. and Roffey, T. (2010), "A method for managing re-identification risk from small geographic areas in Canada", *BMC medical informatics and decision making*, Vol. 10, p. 18.
- Erdogan, B. and Enders, J. (2007), "Support from the top: Supervisors' perceived organizational support as a moderator of leader-member exchange to satisfaction and performance relationships", *Journal of Applied Psychology*, Vol. 92 No. 2, pp. 321–330.

- George, B. and Pandey, S.K. (2017), "We Know the Yin—But Where Is the Yang?: Toward a Balanced Approach on Common Source Bias in Public Administration Scholarship", *Review of Public Personnel Administration*, Vol. 37 No. 2, pp. 245–270.
- Gerstner, C.R. and Day, D.V. (1997), "Meta-analytic review of leader–member exchange theory: Correlates and construct issues", *Journal of Applied Psychology*, Vol. 82 No. 6, pp. 827–844.
- Graen, G.B. and Uhl-Bien, M. (1995), "Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective", *Leadership Quarterly*, Vol. 6 No. 2, pp. 219–247.
- Grimmelikhuijsen, S., Tummers, L.G., Pandey, S.K. and Tummers, L. (2017), "Promoting State-of-the-Art Methods in Public Management Research", *International Public Management Journal*, Vol. 20 No. 1, pp. 7–13.
- Groeneveld, S., Tummers, L.G., Bronkhorst, B., Ashikali, T. and van Thiel, S. (2015), "Quantitative Methods in Public Administration: Their Use and Development Through Time", *International Public Management Journal*, Vol. 18 No. 1, pp. 61–86.
- Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007), *Data Quality and Record Linkage Techniques*, Springer, New York.
- Jacobsen, C.B. and Andersen, L.B. (2015), "Is Leadership in the Eye of the Beholder? A Study of Intended and Perceived Leadership Practices and Organizational Performance", *Public Administration Review*, Vol. 75 No. 6, pp. 829–841.
- Jacobsen, M. and Jensen, R. (2015), "Common Method Bias in Public Management Studies", *International Public Management Journal*, Vol. 18 No. 1, pp. 3–30.
- Kearney, K.A., Hopkins, R.H., Mauss, A.L. and Weisheit, R.A. (1984), "Self-generated identification codes for anonymous collection of longitudinal questionnaire data", *Public Opinion Quarterly*, Vol. 48 No. 1B, pp. 370–378.
- Kelman, C.W., Bass, A.J. and Holman, C. (2002), "Research use of linked health data — a best practice protocol", *Australian and New Zealand Journal of Public Health*, Vol. 26 No. 3, pp. 251–255.
- Kroll, A. and Tantardini, M. (2017), "Motivating and Retaining Government Employees: The Role of Organizational Social Capital", *International Public Management Journal*.
- Kroll, A., DeHart-Davis, L. and Vogel, D. (2017), *Mechanisms of Social Capital in Organizations: How Individual and Team Perceptions Influence Commitment and Engagement. Paper prepared for the Public Management Research Conference (PMRC) 2017, June 8-10, Washington, D.C., USA.*
- LaHuis, D.M., Hartman, M.J., Hakoyama, S. and Clark, P.C. (2014), "Explained Variance Measures for Multilevel Models", *Organizational Research Methods*, Vol. 17 No. 4, pp. 433–451.
- Liu, B., Perry, J.L., Tan, X. and Zhou, X. (2017), "A Cross-Level Holistic Model Of Public Service Motivation", *International Public Management Journal*, Vol. 3 No. 1, pp. 1–26.

- Lubbers, M.J., van der Werf, M.P., Kuyper, H. and Hendriks, A.J. (2010), "Does homework behavior mediate the relation between personality and academic performance?", *Learning and Individual Differences*, Vol. 20 No. 3, pp. 203–208.
- Morey, T., Forbath, T.T. and Schoop, A. (2015), "Customer data: Designing for transparency and trust", *Harvard Business Review*, Vol. 93 No. 5, pp. 96–107.
- Mulligan, E.C. (2001), "Confidentiality in health records: evidence of current performance from a population survey in South Australia", *Medical Journal of Australia*, Vol. 174 No. 12, pp. 637–640.
- Olson, K. (2013), "Do non-response follow-ups improve or reduce data quality?: a review of the existing literature", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 176 No. 1, pp. 129–145.
- O'Toole, L.J. (1997), "Treating Networks Seriously: Practical and Research-Based Agendas in Public Administration", *Public Administration Review*, Vol. 57 No. 1, pp. 45–52.
- Pandey, S.K. (2017), "Theory and Method in Public Administration", *Review of Public Personnel Administration*, Vol. 37 No. 2, pp. 131–138.
- Pitts, D.W. and Fernandez, S. (2009), "The State of Public Management Research: An Analysis of Scope and Methodology", *International Public Management Journal*, Vol. 12 No. 4, pp. 399–420.
- Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y. and Podsakoff, N.P. (2003), "Common method biases in behavioral research: a critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol. 88 No. 5, pp. 879–903.
- Pollitt, C. (2016), *Advanced introduction to public management and administration*, Edward Elgar Publishing, Cheltenham, UK; Northampton, MA.
- Pritchard, R.D. and Karasick, B.W. (1973), "The effects of organizational climate on managerial job performance and job satisfaction", *Organizational Behavior and Human Performance*, Vol. 9 No. 1, pp. 126–146.
- R Core Team. (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Ritz, A., Brewer, G.A. and Neumann, O. (2016), "Public Service Motivation. A Systematic Literature Review and Outlook", *Public Administration Review*, Vol. 76 No. 3, pp. 414–426.
- Roberts, A. (2017), "What's Missing in Public Administration? The Big Picture", *PATimes*, Vol. 37 No. 2, p. 11.
- Robinson-Cimpian, J.P. (2014), "Inaccurate Estimation of Disparities Due to Mischievous Responders", *Educational Researcher*, Vol. 43 No. 4, pp. 171–185.
- Schnell, R., Hill, P.B. and Esser, E. (2013), *Methoden Der Empirischen Sozialforschung*, 10th ed., Oldenbourg, München et al.
- Singer, E., Thurn, D.R. von and Miller, E.R. (1995), "Confidentiality assurances and response: A quantitative review of the experimental literature", *Public Opinion Quarterly*, Vol. 59 No. 1, pp. 66–77.
- Smith, P.C., Kendall, L.M. and Hulin, C. (1969), *The Measurement of Satisfaction in Work and Behavior*, Rand McNally, Chicago.

- Snijders, T.A.B. and Bosker, R.J. (1994), "Modeled Variance in Two-Level Models", *Sociological Methods & Research*, Vol. 22 No. 3, pp. 342–363.
- Steenbergen, M.R. and Jones, B.S. (2002), "Modeling Multilevel Data Structures", *American Journal of Political Science*, Vol. 46 No. 1, pp. 218–237.
- Tantardini, M. and Kroll, A. (2015), "The role of organizational social capital in performance management", *Public Performance & Management Review*, Vol. 39 No. 1, pp. 83–99.
- van Thiel, S. (2014), *Research in public administration and public management: An introduction*, Routledge, New York.
- Vogel, D. (2016), *Führung im öffentlichen Sektor: Eine empirische Untersuchung von Einflussfaktoren auf das Führungsverhalten*, Universitätsverlag Potsdam, Potsdam.
- Weber, M. (1978), *Economy and society: An outline of interpretive sociology*, University of California Press, Berkeley, London.
- Wessels, J.S. and Visagie, R.G. (2017), "The eligibility of Public Administration research for ethics review: a case study of two international peer-reviewed journals", *International Review of Administrative Sciences*, Vol. 83 No. 1_suppl, pp. 156–176.
- Woehr, D.J., Loignon, A.C., Schmidt, P.B., Loughry, M.L. and Ohland, M.W. (2015), "Justifying aggregation with consensus-based constructs: A review and examination of cutoff values for common aggregation indices", *Organizational Research Methods*, Vol. 18 No. 4, pp. 704–737.
- Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., MIT Press, Cambridge; Mass.

Appendix: Generating unique identifiers using R

The following example creates an Excel file containing approximately seventy-five randomly generated unique identifiers. The identifiers have four characters (digits and capital letters), which can be modified by changing the number following `length <-`. The amount of identifiers can be adjusted by changing the number following `amount <-`. Please note that about 55 %–75 % of the initially generated identifiers are deleted in step 4, where identifiers with characters with a high likelihood of confusion are deleted. When executed, the script asks for a directory to save the file. The script's most recent version is also available at <https://github.com/DominikVogel/matching-ids>. A persistent version identical with the following code is available at <https://doi.org/10.5281/zenodo.1243175>.

```
# Title      : Generating random identifiers
# Date       : 2018-05-07
# Version    : R version 3.5.0 (2018-04-23)
# Release    : 1
# License    : MIT License

# 0. Requirements -----
-----
# The script requires the packages random and writexl
# uncomment the following lines to install the packages
# install.packages("random", dep=TRUE)
# install.packages("writexl", dep=TRUE)

# 1. Define amount of identifiers and their length -----
-----
amount <- 200 # amount of generated identifiers (max: 10,000)
length <- 4 # length of identifiers

# 2. Load necessary packages -----
-----
library(random)
library(writexl)

# 3. Generate identifiers -----
-----
ids <- randomStrings(n = amount, len = length, digits = TRUE,
                     uppervalpha = TRUE, loweralpha = FALSE, unique = TRUE, check = TRUE)

# 4. Exclude IDs with characters with high likelihood of confusion -----
-----
# Attention: Reduces amount of identifiers by 55%-75%
```

```

ids <- ids[!grepl("1", ids),]
ids <- ids[!grepl("0", ids)]
ids <- ids[!grepl("I", ids)]
ids <- ids[!grepl("J", ids)]
ids <- ids[!grepl("Q", ids)]
ids <- ids[!grepl("Z", ids)]
ids <- ids[!grepl("2", ids)]
ids <- ids[!grepl("7", ids)]

# 5. Export identifiers -----
-----
dir <- choose.dir(caption = "Select folder to save IDs")
write_xlsx(as.data.frame(ids),
  path=paste0(dir, "/ids.xlsx"),
  col_names=FALSE)

```