4-22-2019

Methodological Issues With Coding Participants in Anonymous Psychological Longitudinal Studies

Lillian M. Audette
Tennessee State University

Marie S. Hammond
Tennessee State University

Natalie K. Rochester
Tennessee State University

Recommended Citation

L.M. Audette, M.S. Hammond, N.K. Rochester "Methodological Issues With Coding Participants in Anonymous Psychological Longitudinal Studies" Educ Psychol Meas. 2020 Feb; 80(1): 163–185. https://dx.doi.org/10.1177%2F0013164419843576

Article

Educational and Psychological
Measurement
2020, Vol. 80(1) 163–185
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164419843576
journals.sagepub.com/home/epm

Methodological Issues With Coding Participants in Anonymous Psychological Longitudinal Studies

Lillian M. Audette[1], Marie S. Hammond[1] and Natalie K. Rochester[1]*

Abstract

Longitudinal studies are commonly used in the social and behavioral sciences to answer a wide variety of research questions. Longitudinal researchers often collect data anonymously from participants when studying sensitive topics to ensure that accurate information is provided. One difficulty gathering longitudinal anonymous data is that of correctly matching participants across waves of data collection. A number of methods have been proposed for using nonidentifying codes to match anonymous participants; however, currently there is no consensus on the most effective method. This article reviews and analyzes the literature on nonidentifying codes and provides recommendations for researchers interested in using these types of codes in conducting anonymous longitudinal studies.

Keywords

research methodology, experimental design, anonymous, longitudinal, best practices

Researchers make numerous decisions when conducting a research study. They must choose their hypotheses, sample, type of data, and data collection procedures. In addition, they must determine effective data labels/codes to ensure differentiation and confidentiality among sample participants. This labelling becomes more

[1] Tennessee State University, Nashville, TN, USA
*Natalie K. Rochester is now at South Texas Veterans Health Care System

Corresponding Author:
Lillian M. Audette, Department of Psychology, Tennessee State University, 3500 John A. Merritt Boulevard, Nashville, TN 37209, USA.
Email: laudette@my.tnstate.edu

important the greater the number of instances of data gathered from the same individual, generally called ‘‘waves.’’ Collecting data longitudinally is defined as instances in which data are collected from the same participants at multiple timepoints (Heiman, 2000). In addition to gathering data longitudinally, one can gather data either anonymously or nonanonymously. Collecting data anonymously occurs when researchers do not ask for any ‘‘personally identifying information’’—such as participants’ names, full birth dates, email addresses—anything unique to participants across waves, and/or which could also be used to discover their personal identity (Fisher, 2013). Anonymous data collection increases participant trust that their answers cannot be connected with a specific person. When researchers choose to conduct a longitudinal study that entails the collection of data anonymously, there is a unique difficulty: how can researchers accurately connect participants to their data across the longitudinal waves, while maintaining anonymity? This article reviews the different strategies that have been proposed to address this problem and provides recommendations concerning their use in a research study.

Longitudinal Research

In longitudinal studies, multiple waves of data gathering are used to understand how variables might change as a function of time or as a function of an intervention. Longitudinal data provide a fuller and more complex picture of the topic of interest. This article specifically focuses on methods for accurately connecting participant data between two waves of a study, although the presented approaches apply equally to longitudinal studies with more than two data collection waves.

Although some studies start with a large pool of participants from which they draw for subsequent waves of data gathering, many longitudinal studies add participants at successive waves. Two common reasons are to obtain information about cohort differences or to increase the number of participants and power of a study. The reverse may also true—many longitudinal studies lose participants across successive waves. Participants may move, die, stop, or be unable to participate for a myriad of reasons. As a result of these factors, researchers using longitudinal studies must keep thorough records of a changing participant pool, across multiple waves, to ensure reliable and valid data. Maintaining accurate records is made more difficult when researchers are collecting their data anonymously.

Thus, the correct matching of participants at each wave in the data collection process allows the researcher to reduce the likelihood of low statistical power and high Type II error rates (Bedeian & Feild, 2002). Furthermore, incorrect participant matches wastes researcher’s time and resources—time spent creating incorrect matches, and time spent analyzing and interpreting incorrect data. Accurately matching participants across data collection time points is, therefore, a critical problem for longitudinal researchers, which, similar to maintaining good records, becomes more difficult when researchers collect their data anonymously.

Anonymity in Longitudinal Research

There are three reasons researchers might want to collect anonymous data when conducting a longitudinal study: (a) participant bias, (b) researcher bias, and (c) legal/ ethical compliance issues. These reasons are defined and discussed below.

First, data collected anonymously will differ from data that is collected nonanonymously. Research has shown that participants often provide different answers to the same self-report measures, depending on whether the participants know that the survey is anonymous versus nonanonymous (Grube, Morgan, & Kearney, 1989; Kearney, Hopkins, Mauss, & Weisheit, 1984; Olson, Stander, & Merrill, 2004; Stander, Olson, & Merrill, 2002; Thomas, Wright, Adler, & Bliese, 2004). Research indicates that when participants trust that they cannot be identified in any manner, they respond more honestly. This can be particularly important for research into potentially sensitive topics. Anonymous participants, compared with nonanonymous participants, report higher rates of childhood sexual abuse (Stander et al., 2002), higher rates of mental health distress symptoms (Thomas et al., 2004), and are more likely to self-report a sexual minority orientation (Baldwin et al., 2017).

Second, anonymous data reduces the likelihood of confirmatory bias, which strengthens a study's validity (Heppner, Wampold, Owen, Thompson, & Wang, 2016; Nickerson, 1998). When participants provide personally identifying information, there is always the possibility that those analyzing the data might be able to identify participants. For instance, the data analyst might know or recognize that a child participant is a family member—by their unique combination of birthday, gender, and school—and the data analyst may then be unconsciously influenced when analyzing the data. If no personally identifying information is collected and the data are always anonymous, this potential threat to validity is better controlled.

Third, anonymous data collection may be used to help researchers comply with various local, state, or national laws. For example, the Health Insurance Portability and Accountability Act (HIPAA) and the Family Educational Rights and Privacy Act (FERPA) are two federal laws in the United States which specify stringent requirements for protecting personal health information and student education records, respectively (Fisher, 2013). These types of laws and regulations often focus on ''personally identifying information,'' as previously defined. Anonymous data collection therefore allows researchers to best comply with HIPAA, FERPA, or other local or national legal requirements. Thus, for any or all the above reasons, researchers may wish to conduct their longitudinal research anonymously.

Current Options for Coding Participants in Longitudinal Studies Anonymously

Four different methods were observed in reviewing the research literature, including (a) Collecting nonanonymous data that is later de-identified, (b) Using preexisting unique identification codes, (c) Using an electronic anonymizing system, and (d) Using self-generated identification codes (SGICs). Each of these methods is presented in order of increasing level of benefits. These methods are described below.

Nonanonymous Data Collection Later De-Identified. The first method is to collect nonanonymous data, and then both de-identify and anonymize the data subsequent to data collection, and prior to the data being shared with or used by researchers/data analysts. In this method, a team conducting a longitudinal study designate an external researcher who reviews and de-identifies the nonanonymous data. Ideally, the external reviewer would have no other role in the study and no contact with the study's participants.

During data collection, all participants would have provided personally identifying information. The designated external researcher views the collected (nonanonymous) data, de-identifies and anonymizes the data, then provides the de-identified dataset to the research team for their analyses (Kadison, Pelletier, Mounib, Oppedisano, & Poteat, 1998; Murray, 1992; Tenhiälä & Lount, 2013; Udry & Bearman, 1998). Furthermore, the same external researcher would match participants across the different waves based on the identifiable information provided by the participants.

Preexisting Unique Identifiers. The second method of collecting anonymous longitudinal data consists of using preexisting unique identifiers. Preexisting unique identifiers are extant combinations of numbers, letters, or both, that uniquely identify an individual participant. Examples of preexisting unique identifiers include student identification, Social Security, or driver's license numbers.

In choosing this option, the researcher(s) would select a stable, relevant identifier(s) based on the appropriateness for the proposed survey. They would not ask participants for any other identifying data except that of the chosen preexisting unique identifier(s). At each wave of data collection, participants would be asked again to provide their preexisting unique identifier(s), which would be used to match the participant's data across waves.

Electronic Anonymizing System. A third method of collecting anonymous longitudinal data is through the use of an electronic anonymizing system. These systems are typically tied to online data-gathering systems which provide anonymous identification codes primarily through two methods. First, through the use of an online data gathering system that assigns a random identification code (Kiesner, Mendle, Eisenlohr-Moul, & Pastore, 2016; Williams & Guerra, 2007), or second, through a mobile device application which provides researchers with anonymized datasets (Tregarthen, Lock, & Darcy, 2015). As a full review of the many currently available electronic anonymizing systems is beyond the scope of the present article, we will just note that there are many potential software programs that may facilitate the collection of anonymous longitudinal data.

Self-Generated Identification Codes. A fourth method of collecting anonymous longitudinal data is through the use of SGICs. Usually, the SGIC is created from the answers to a number of personally salient questions. The answers are combined in a predetermined order to create the SGIC.

To provide an example, Yurek, Vasey, and Havens (2008) used a four-question SGIC. They asked participants to report their mother's first initial of their first name, their number of older brothers, the month in which they were born, and the first letter of their own middle name. Participants were asked to give these details at Wave 1 (0 months), Wave 2 (6 months), and Wave 3 (12 months). Thus, a participant who indicated that their mother's name was Anne (A), that they have one older brother (01), were born in July (07), and whose middle name is Drew (D) would generate the identification code A0107D. This self-generated identification code could then be used to link the participants' data at Wave 1 to their data at Wave 2 and Wave 3.

Researchers who use this method generally allow the codes to be fault–tolerant, such that exact agreement of participants' SGICs across waves is not expected. Generally, participants are allowed to have at least one question not match between waves (one-off) or two questions not match between waves (two-off). With fault–tolerant methods, data loss of 20% to 30% is not unusual, although less than 10% data loss is possible if the choice of questions is greatly simplified (DiIorio, Soet, Van Marter, Woodring, & Dudley, 2000; Kearney, 1982; Schnell, Bachteler, & Reiher, 2010). It should be noted that, when exact agreement between codes is required, losses of up to 50% of matched pairs are common (Schnell et al., 2010).

Research Questions

Over the years, researchers have created a number of different solutions to address the problem of effectively and accurately collecting anonymous, longitudinal data. To strengthen the quality of future anonymous, longitudinal research, the following research questions were addressed:

Research Question 1: What are the advantages and disadvantages of the four methods of ensuring anonymity of data in longitudinal research?

Research Question 2: How do methodological decisions, such as time between waves and self-generated identification code (SGIC) length, impact researchers' ability to link data across waves when using a particular SGIC method?

Research Question 3: Which combination of elements is most effective in creating SGICs to link data across waves?

Method

Below, we report our process for obtaining and analyzing information on anonymizing participant data. In addition, we review how we evaluated the four methods found in the literature for coding participants in anonymous longitudinal research.

Sample Description

To understand best practices in data anonymization, a literature review was conducted using EBSCOhost, ScienceDirect, and Sage Psychology & Counseling Collection databases. Where possible, to increase the likelihood that all relevant articles were identified, searches were conducted using the earliest publication date available in each database. That date was specified for EBSCOhost (1979) while ScienceDirect and the Sage database did not allow for the selection of dates for their searches. Terms used in the searches included ''panel study,'' ''longitudinal study,'' and ''anonymous.'' Articles were identified that were (a) psychological in nature, (b) self-identified as presenting longitudinal research, and (c) contained details on the anonymizing methodology used. Exclusion criteria at this level focused on ensuring that the sample included anonymous longitudinal research studies: ''Alcoholics Anonymous,'' ''Delphi'' (for the ''Delphi Method'' or ''Delphi Expert Panel Method''), ''Narcotics Anonymous,'' and ''anonymous reviewers'' (for ''thanks to anonymous reviewers''). In addition, articles not published in English were also excluded.

The creation of the sample used for analysis was conducted in six stages that are described here. Stage 1 is the initial search. The initial literature search produced 514 citations. Stage 2 eliminated citations that did not use an anonymous, longitudinal methodology, based on a reading of the citations' abstracts. Of the 514 studies, 488 studies were removed because they did not use an anonymous, longitudinal methodology, and 6 studies were removed because they were duplicative, resulting in a sample of 20 studies. Stage 3 eliminated studies that did not describe their anonymizing method sufficiently for categorization and/or analysis. Of the 20 studies, one study was removed due to insufficient information about their methodology, resulting in a sample of 19 studies. Stage 4 removed articles (n = 9) that did not use the self-generated identification code methodology, resulting in a sample of 10 studies. Stage 5 removed studies that did not provide specific enough information for statistical analyses. In Stage 5, four criteria were used: (a) providing information on all elements used in the code, (b) providing perfect and/or one-off match rate, (c) providing number of participants, and (d) providing information on the time frame between waves. Of the 10 studies, 1 study was removed, resulting in a sample of 9 studies. Stage 6 reviewed the references of all 20 articles produced by Stage 2 to ensure that all relevant studies were identified for inclusion in the sample. An additional 11 studies were identified that met the criteria from Stages 2 through 5 above. These 11 studies were combined with the 9 studies produced by Stage 5, resulting in a final sample of 20 studies (see Table 1 for literature search results).

Data Coding

Data were reviewed and coded prior to analysis. For Research Question 1, the four anonymous, longitudinal methods were coded in response to three criteria (Rochester, 2015):

Table 1. Summary of Literature Search Results.[a]

Database	Search terms	Total studies found	Studies meeting inclusion criteria
EBSCOhost	Anonymous, panel study	4	0
EBSCOhost	Anonymous, longitudinal study	147	14
ScienceDirect	Anonymous, panel study	2	1
ScienceDirect	Anonymous, longitudinal study	30	0
Sage	Anonymous, panel study	22	0
Sage	Anonymous, longitudinal study	309	5

[a]While a total of 514 citations were found, eliminating studies that did not utilize an anonymous, longitudinal methodology and eliminating duplicate citations resulted in a total number of unique studies identified of 20.

- Was the method truly anonymous?
- What was the method's match rate between waves?
- What is the method's utility?
    - Participants' trust in the anonymity of their data?
    - Risk of confirmatory bias?
    - Whether the method meets legal definitions of anonymity?
    - Amount of required investment of time or funds?

Results and Discussion

> **Research Question 1:** What are the advantages and disadvantages of the four methods of ensuring anonymity of data in longitudinal research?

The four methods of conducting anonymous, longitudinal research were analyzed to identify the benefits and disadvantages of each (see Table 2 for overview of benefits vs disadvantages).

Nonanonymous Data Collection Later De-Identified

Two primary benefits of collecting nonanonymous data, followed by de-identification, were identified. First, this method increases accurate identification across data waves over anonymous data collection. Second, it can be used regardless of whether the data are collected via paper or electronic format, requiring low investment.

Four primary disadvantages of collecting nonanonymous data, followed by de-identification were observed. First, the contradiction between the claim that the data would be anonymous while collecting personally identifying information is likely to have introduced the biases inherent in nonanonymous responses (Catania, Gibson, Chitwood, & Coates, 1990; Grube et al., 1989; Kearney et al., 1984; Olson et al., 2004; Stander et al., 2002). Second, collecting identifying data that is later

Table 2. Comparing Anonymizing Systems' Benefits and Disadvantages.

		Issue with . . .						
Anonymizing system	True anonymity	Participant trust	Confirmatory bias	Anonymity requirements	Participants' accurate recall	Limited to paper or electronic format	Poor matching accuracy	Count of benefits/ disadvantages
Nonanonymous data collection later de-identified	yes	yes	yes	yes	—	no	no	2/4
Preexisting unique identifiers	no	yes	no	yes	yes	no	no	4/3
Electronic anonymizing systems	no	no	no	no	—	yes	no	5/1
Self-generated identification codes	no	no	no	no	yes	no	no	6/1

anonymized increases the likelihood of confirmatory bias. To minimize impact, we recommend that those involved in such a study ensure that the designated external researcher does not interact with participants, and that the designated external researcher de-identifies the data in such a way that subsequent data analysts cannot reverse-identify the participants. When the data collectors and the data analyzers are the same people, the risk is heightened as well (MacCoun & Perlmutter, 2017). Third, collecting nonanonymous data may make the study be subject to different laws and regulations, as mentioned before with HIPAA and FERPA, even if the data are later de-identified (Fisher, 2013). Collecting nonanonymous data could put a greater burden on researchers to ensure legal compliance. Fourth and finally, collecting nonanonymous data is not truly ‘‘anonymous’’ and thus does not meet the present study’s second criteria. Thus, although collecting nonanonymous data and anonymizing the data afterward has the highest rates of correctly matching participants across data collection waves (Davis-Kean, Jager, & Maslowsky, 2015; Schnell et al., 2010), this method still has a number of important limitations.

Preexisting Unique Identifiers

Four primary benefits to the use of preexisting, unique identifiers were observed. First, preexisting unique identifiers can be considered a truly anonymous method of data collection. Second, from the researcher’s perspective, preexisting unique identifiers facilitated matching across waves. Use of a self-relevant identification code, theoretically, facilitates participant recollection across time due to the use of information the participant is less likely to forget. Third, without identifying information, this method decreases the likelihood of researcher identification of individual participants, and thus minimizes the chance of confirmatory bias. Fourth, preexisting, unique identifiers can be used regardless of data collected method (paper or electronic format).

Three primary disadvantages in using preexisting unique identifiers were observed. First, pre-existing unique identifiers may not uniquely identify all participants across all waves, compromising match rates. This can occur in three ways (Schnell et al., 2010): (a) incorrect recollection of the preexisting unique identifier on the part of the participant; (b) changes in the preexisting, unique identifier during the course of the longitudinal study, causing a mismatch (e.g., changes in drivers’ license number due to a move); or (c) the preexisting unique identifier(s) does not uniquely or accurately separate one participant from another (e.g., the recycling of student identification numbers). Second, while the use of preexisting unique identifiers is more private than the use of non-anonymous data collection methods, it is dependent on the relevant legal requirements. This disadvantage is more likely to occur as researchers comply with external demands for data sharing (National Science Foundation, 2017). Instances of supposedly anonymous data being re-identified as coming from a particular university have occurred through a combination of such information as the total number of students in a class year, the nationalities represented, and the college majors offered (Zimmer, 2010). Third, as discussed in the previous method, the use of any personally identifying information is likely to affect the participant's confidence in the anonymity of the data they provide (DiIorio et al., 2000). For instance, Widrich and Ortlepp (1994) used full birthdate as a preexisting unique identifier. One would expect individuals to perfectly remember their birthdates, however, 3.2% of their participants provided different birthdates in Wave 2 then they provided in Wave 1.

Electronic Anonymizing System

There are benefits to using electronic anonymizing systems, which vary depending on the specific electronic anonymizing system chosen. Five primary benefits were identified in the research literature. First, if set up and used optimally, these systems will prevent identification of participants, which leads to the second benefit—that of reducing or preventing confirmatory bias. Third, participants are likely to view it as truly anonymous and respond in accord with that belief (Catania et al., 1990; DiIorio et al., 2000; Kearney, 1982; Schnell et al., 2010; Tregarthen et al., 2015). Fourth, if set up and used in consultation with technology and legal experts, electronic anonymizing systems can also facilitate researchers' ability to comply with various legal requirements to appropriately protect patients' privacy and confidentiality (Fisher, 2013). Fifth and finally, depending on the system chosen and the population of interest, some electronic anonymizing systems can match participants across data waves with near-perfect accuracy (Kiesner et al., 2016).

Disadvantages in the use of electronic anonymizing systems include those that are attributable to this particular method and those that are unique to a specific system. Potential disadvantages depending on the specific system may include participants' distrust of the level of anonymity, correct matching issues, confirmatory bias issues, and legal issues (Fisher, 2013; Kraut et al., 2004). A consistent disadvantage across electronic anonymizing systems are the challenges in implementation. Regardless of the system and its specifics, using electronic anonymizing systems is likely to incur substantial investments of both time and money in contrast to those for nonelectronic methods for anonymizing longitudinal data (Kraut et al., 2004). Depending on the system and researcher needs, access to computers, and the internet for both researchers and, more importantly, participants may become a barrier. Furthermore, an electronic anonymizing system can be more complex to implement than other electronic data collection methods, such as Survey Monkey or Qualtrics. For example, in a study by Tregarthen et al. (2015) it took more than 18 months to create and refine a mobile device application for anonymous longitudinal data collection about eating disorder self-monitoring. Electronic anonymizing systems are potentially costly in terms of time, or money, or both. If researchers are considering the use of an electronic anonymizing system, they will have to decide if it is first, possible, and second, worthwhile.

Self-Generated Identification Code Benefits

There are six main benefits to using SGICs to match participants when collecting anonymous longitudinal data. First, SGICs are truly anonymous, as they are not preexisting identifiers and are created using data with sufficient variability and impersonality to prevent individual identification. Second, this increase in degree of as well as perception of, anonymity has been shown to increase the quality and quantity of responses from participants (DiIorio et al., 2000; Kearney, 1982). Third, the use of salient personal information to create the SGIC increases the potential for improved recall of the SGIC to facilitate matching across waves (DiIorio et al., 2000). Three benefits of the use of SGICs relate to the quality of the research: the reduction in the potential for confirmatory bias as insufficient data to produce a chance re-identification will be gathered; increased compliance with laws and regulations as a result of not collecting personally identifying information, and utility across data formats (paper and electronic formats). SGICs, therefore, have a number of distinct advantages over the three previous methods of collecting anonymous data in longitudinal studies.

There is one disadvantage in using SGICs. Participants must remember and accurately self-report the information used to create the SGICs at each wave. Thus, some data loss is expected when using SGICs. The literature supports the necessity of carefully choosing the personal information used to create the code, which requires more time and effort of researchers than picking a pre-existing unique identification code (DiIorio et al., 2000; Schnell et al., 2010).

Summary

To summarize, four methods of anonymous longitudinal data collection were examined resulting in varying patterns of benefits and disadvantages. First discussed was the nonanonymous data collection (later de-identified) with two benefits (higher and more accurate matching and useful regardless of format) and four disadvantages (low participant trust in confidentiality, re-identification/confirmatory bias issues, impact of legal requirements, and not truly anonymous data collection). Second, the use of preexisting, unique identification codes with four benefits (truly anonymous data collection, increased matching across waves due to use of easily remembered self-relevant code, reduction in likelihood of confirmatory bias or re-identification, and useful across formats) and three disadvantages (failure to uniquely identify all participants across all waves, the impact of re-identification of participants through data sharing and legal requirements, and compromise of participant confidence in confidentiality). Third, the use of an electronic anonymizing system has five benefits (true anonymity encouraging accurate responding, prevention of re-identification of participants and confirmatory bias, increased compliance with legal requirements, higher match rates when set up correctly) and two types of disadvantages (general disadvantage in cost and implementation time as well as those unique to the particular system). Fourth, the use of SGICs has six benefits and one disadvantage. The advantages include truly anonymous data collection, increased appearance of confidentiality, facilitation of participant recall of code information, reduction in likelihood of confirmatory bias, increased legal/regulatory compliance, and useful across all data formats. Researchers can ameliorate the one disadvantage with the judicious choice of self-generated identification code questions.

> **Research Question 2:** How do methodological decisions, such as time between waves and self-generated identification code (SGIC) length, impact researchers' ability to link data across waves when using a particular SGIC method?

Multiple articles published data on multiple waves, so each Wave X to Wave Y is analyzed separately in the following analyses, and counted as one ''study'' for calculating sample size, for example, a study with data from Wave 1, Wave 2, and Wave 3 is *n* = 3, counting separately Wave 1 to Wave 2, Wave 1 to Wave 3, and Wave 2 to Wave 3. ''Perfect match rate'' is when no elements were allowed to vary between waves in an SGIC match. ''One-off match rate'' is when one element in the code was allowed to vary between waves in an SGIC match. All averages and correlations are unweighted. Due to unique element choices, driven by a unique population of study, Wilson et al. (2010) is not included in the remaining analyses.

Thirty-nine code elements were used, 11 of which were used by just one study (see Table 3 for specifics). The average number of elements in an SGIC was 5.85 (*SD* = 1.66, *n* = 33) ranging from 3 to 9. The most common SGIC length was seven elements (*n* = 11) followed by five (*n* = 5), see Table 4 for more details.

For all studies, the perfect match rate was *M* = 65.3% (*SD* = 15.8%, *n* = 25) ranging from 42.0% to 94.3%. The one-off match rate was *M* = 80.87% (*SD* = 11.76, *n* = 23) ranging 51.30% to 98.71%. Sample sizes were heterogeneous, ranging from 78 to 8,136 (*M* = 1,506.8, *SD* = 1623.9, *n* = 33), as were time between waves, from 0.25 to 24 months (*M* = 8.6, *SD* = 6.4, *n* = 33). See Table 5 for details.

Overall, there were no clear relation between number of elements and match rate (perfect match rate *r* = 0.02, *n* = 25; one-off match rate *r* = 0.46, *n* = 23). Average perfect match rate (see Table 6) was highest for six elements (*M* = 92.0%, *SD* = 3.2%, *n* = 2), both study lengths of 0.25 months long, followed by five elements (*M* = 81.0%, SD = 8.4%, *n* = 4), studying lengths ranging from 0.5 months to 4 months. Note the element lengths with the highest average perfect match rate also had the shortest time frames. Average perfect match rate was lowest for three elements (*M* = 55.4%, *SD* = 7.4%, *n* = 4), with study lengths ranging from 6 to 18 months. There were similar findings for one-off match rates (see Table 6), with five to six elements having the highest match rate but again the shortest time frames (*M* = 92.7%, *SD* = 10.2%, *n* = 8), study lengths ranging from 0.25 to 0.5 months, followed by seven elements (*M* = 84.2%, *SD* = 6.3%, *n* = 14), study lengths ranging from 1 to 20 months. Average one-off match rate was lowest for four elements (*M* = 58.9%, *SD* = 7.7%, *n* = 4), study lengths ranging from 6 to 18 months, though no one-off study used three elements. The contribution of the shorter timeframe to the variation in match rates between SGIC lengths was unclear, due to incomplete information.

Table 3. Names and Frequencies of Self-Generated Identification Code (SGIC) Elements' Use.

General element	Frequency of general element	Number of Studies using subelement (exact subelement)					Total frequency
Birthday (including day, month, and year)	4	8 (Birth day, 1-31)	7 (2nd digit, Birth Day)	16 (Birth Month)	1 (Birth Year)	2 (Odd/Even Birth Year)	38
Own first name (full)	—	12 (1st letter)	2 (2nd letter)	9 (3rd letter)	2 (If 1st initial A-M or N-Z)		25
Own last name (full)	—	2 (2nd letter)	9 (3rd letter)	4 (Last letter)			15
Own middle name (full)	—	9 (1st letter)					9
Mother's first name (full)	—	18 (1st letter)	9 (2nd letter)	7 (4th letter)			34
Father's first name (full)	—	10 (1st letter)	2 (2nd letter)	7 (3rd letter)			19
No. of siblings	1	7 (no. of older siblings)	6 (no. of older brothers)	3 (no. of older sisters)	1 (no. of sisters)	1 (no. of brothers)	19
Sex	10						10
Race/ethnic category	6						6
First letter, street residing	3						3
School name	2	1 (last primary school 1st initial)					3
Birth order within family	—	1 (only, oldest, youngest, middle)	1 (If twin and twin order)				2
Class within school	2						2
Age	1						1
Father's profession	1						1
Elements not included in analyses							
Favorite pet's name	3						3
City of all-time favorite sports team	3						3
First best friend's name from high school	3						3
Mascot of last-attended high school	3						3

Note. All elements not included in analyses are from Wilson et al. (2010), as noted in the article.

Table 4. Frequencies of Self-Generated Identification Code (SGIC) Lengths.

Number of elements in SGIC	*N* (studies)
9	1
8	2
7	11
6	2
5	5
4	7
3	4
2	0
1	0

Note. *N* (studies) counts number of studies; one article can have multiple studies as each Wave X to Wave Y is counted separately. Excludes Wilson et al. (2010) data.

Table 5. Descriptives From Self-Generated Identification Code (SGIC) Studies.

	Minimum	Maximum	Average	*SD*	*N* (studies)
Count of elements	3	9	5.85	1.66	33
Match rate (perfect)	42.00%	94.31%	65.27%	15.78%	25
Match rate (one-off)	51.30%	98.71%	80.87%	11.76%	23
N (sample size)	78.00	8136.00	1506.79	1623.86	33
Month count	0.25	24	8.61	6.43	33

Note. For *N* (studies) one article can have multiple studies as each Wave X to Wave Y is counted separately. Excludes Wilson et al. (2010) data.

Counterintuitively, there was also no clear relation between time between waves and match rate. Overall, the relation was negative for perfect match rate and time between waves ($r = -0.40$, $n = 25$) and one-off match rate and time between waves ($r =-0.43$, $n = 23$), as one would expect. However, looking at the relation between match rate and time between waves more closely, perfect match rates did not present a clear pattern across lengths: 18 to 24 months ($r = 0.93$, $n = 3$), 12 to 13 months ($r = -0.08$, $n = 7$), 6 to 9 months ($r = 0.9$, $n = 5$), <6 months ($r = -0.56$, $n = 10$). Nor did one-off match rates produce a clear pattern: 14+ months ($r = 0.89$, $n = 4$), 12 months ($r = 0.62$, $n = 8$), 6 to 9 months ($r = 0.96$, $n = 4$), <6 months ($r = -0.23$, $n = 7$). In addition, there was not a clear relation between number of participants and match rate (perfect match rate $r = 0.11$, $n = 25$; one-off match rate $r = 0.003$, $n = 23$).

Research Question 3: Which combination of elements is most effective in creating SGICs to link data across waves?

Table 6. Statistics on Self-Generated Identification Code (SGIC) Length for Perfect Matches and One-Off Matches.

Count of elements	Average match (perfect)	*SD* match (perfect)	*N* (studies)	Months range
3	0.554	0.074	4	6-18
4	0.564	0.085	4	6-18
5	0.810	0.084	6	0.5-4
6	0.920	0.032	2	0.25
7	0.568	0.136	6	1-12
8+	0.580	0.069	3	12
Count of elements	Average match (one-off)	*SD* match (one-off)	*N* (studies)	Months range
4	0.589	0.077	3	6-18
5-6	0.927	0.102	3	0.25-0.5
7	0.842	0.063	14	1-20
8+	0.757	0.072	3	12

Note. For *N* (studies); one article can have multiple studies as each Wave X to Wave Y is counted separately. Excludes Wilson et al. (2010) data.

An examination of individual elements' contribution to match rates was possible for a subset of data, in which the element-by-element match rates were reported (n = 6). Of the 25 elements with such data, 19 elements were used only once or twice, and, therefore, are not included on this report; complete data can be obtained from the first author. The six remaining elements, presented in Table 7, produced variable error rates. Errors were recorded if a participants' data could still be matched between waves, but the particular element of interest was either (a) left blank or (b) mismatched between waves. Personally salient elements obtained the lowest average error rates: birth month (M = 1.5%, n = 4) obtained the lowest percentage of errors by a wide margin, with error rates ranging from 0.0% to 3.1%. Middle initial obtained the next lowest (M = 8.9%, n = 4), with error rates ranging from 2.8% to 15.9%, followed by race (M = 13.5%, n = 3), with error rates ranging from 4.3% to 20.5%. The elements obtaining greater percentages of errors were family-type elements: number of older siblings (M = 16.4%, n = 3; 8.0% to 22.7%); first initial of father's first name (M = 16.7%, n = 3; 13.2% to 20.5%); and lastly, first initial of mother's first name (M = 18.8%, n = 4; 6.6% to 48.5%). The large ranges and low number of studies is a limitation in the generalization of our element-specific findings. Due to the variety of elements and low number of studies using each element, further statistical analyses on these elements' contributions to match rate were not conducted.

Of note: out of 514 results found in our literature review just 1.8%, or 9 studies, used a self-generated identification code and provided sufficient information for their data to be included in our analyses. Even when we consider only the 81 studies that remained after the initial abstract review; those included in analyses are still low, at just 9.9%. Our findings are thus limited by the studies that included a thorough discussion of the methodological and statistical information.

Considerations When Choosing Questions for Self-Generated Identification Code Elements

Before applying our literature review and analyses to formulate final best-practices recommendations, we review specific criteria needed to evaluate the inclusion or exclusion of code elements. In order that the code is accurate, unique, and consistent across data collection waves, researchers must carefully choose which questions are used to produce the self-generated identification code. There are five factors to be considered. The questions should ask for elements that will be (a) salient, (b) constant, (c) nonsensitive, (d) easy to consistently format the same, and (e) difficult to decode. We will review each of these requirements in turn.

Regarding salience, the questions should ask for personally relevant information (salience) that the participant is likely to know. This general recommendation was confirmed by our literature review and the results of the analyses suggesting that personal information was more accurate than familial information (Table 7). For instance, the street on which a participants' elementary school was located is unlikely to be salient enough to be consistently remembered; whereas, asking for the name of their first pet is likely to be more salient and, thus, remembered. Regarding constancy, the information should remain constant over time. For instance, a participant's age changes every year, as might the reporting of the number of siblings; however, a participant's birth month is constant. Regarding nonsensitivity, the information should not be so sensitive that participants will avoid responding to the question. For instance, a participants' sexual orientation, while nonidentifying, might be sensitive enough that a participant would avoid responding to the question; however, a participants' sex is also nonidentifying and less likely to be as sensitive.

Regarding formatting, the information should be easy to answer consistently in the same format. High schools, colleges, and even cities often have abbreviations (Louisiana State University vs. LSU or New York City vs. NYC) that are duplicative (UM: University of Missouri, University of Mississippi, University of Minnesota). Choose information that is easy to answer accurately and in the same format across waves. For example, the formatting of a participants' high school may vary significantly (Woodrow Wilson High School vs. Wilson High vs. WWHS), whereas the formatting of a participants' birth state will stay the same (Pennsylvania is always PA). Thus, the clarity of instructions for responding—''full state name'' versus ''state abbreviation''—further facilitates accuracy of consideration.

Finally, regarding non-identifiability, the information used to create the self-generated identification code should be sufficiently nonidentifiable that participants are confident in the anonymity of their data. Confidence that the code minimizes accidental identification by a researcher is paramount to increasing the likelihood

Table 7. Percentage of Errors per Element.

Citation	*N* (participants)	Middle initial (%)	Birth month (%)	Mother initial (%)	Father initial (%)	No. of older siblings (%)	Race (%)
Kearney (1982)	410	10.70	2.40	8.00	16.30	22.70	20.50
Yurek et al. (2008)	163	15.90	3.10	48.50	—	—	—
Kearney et al. (1984)	383	6.30	0.00	6.60	13.20	18.50	15.60
McGloin, Holcomb, and Main (1996)	601	—	0.40	—	—	—	4.30
Dilorio et al. (2000)	2538	2.80	0.00	12.20	20.50	—	—
Schnell et al. (2010)	293	—	—	—	—	8.00	—
Average error percentage	731.33	8.90	1.50	18.80	16.70	16.40	13.50
N (Studies)	6	4	4	4	3	3	3

Note. "Errors" counts (a) when elements were left blank or (b) elements did not perfectly match between time points but the participant could still be matched. Dilorio et al. (2000) required Birth Month to perfectly match between waves, for a match to be recorded, hence the 0.0% match rate. It is not included in the average or *N* (studies) for birth month.

that participants will answer all the questions honestly and accurately. For instance, participants may not feel comfortable giving their full birthdate because it may lead to the individual's identification; however, they may feel more comfortable providing the month in which they were born, which is less identifiable.

As can be seen, the questions chosen for the self-generated identification code are very important, and each of the aforementioned five factors must be carefully considered for each selected question. Wilson et al. (2010) provide an example of issues that might arise if the chosen questions are not carefully selected. This study was a survey of U.S. military personnel; thus, a primary consideration was that the answers to their self-generated identification code questions not appear in any of the participants' official military records. The researchers then elected to use four unusual pieces of personal information: (a) favorite pet's name, (b) city of all-time favorite sports team, (c) first name of best friend from high school, and (d) mascot of last attended high school. The self-generated identification code thus created is relatively difficult to decode—addressing the critique above. However, these questions are nonsalient (school mascot), nonconstant due to changes in perceptions that occur over time (favorite sports team, favorite pet's name, best friend from high school), and difficult to consistently format (city of favorite sports team). As a result, Wilson et al. (2010) obtained relatively low match rates, with a maximum match rate of 24.69% (occurring between the 3- and 6-month waves), using the fault-tolerant method. When the self-generated identification code was required to match perfectly, match rates of 1% or less between waves were obtained. In contrast, other studies with SGICs with more carefully selected questions reported match rates ranging from 61.2% (DiIorio et al., 2000) to 94.7% (Kearney, 1984). Thus, the Wilson et al. (2010) study highlights the importance of careful selection of the questions used to create SGICs.

Recommended Best Practices

After examining the different methods for coding participants in longitudinal studies anonymously, the use of self-generated identification codes appears to most effectively minimize the disadvantages and maximizes the benefits across a broad spectrum of research with humans. To ensure the high-quality use of this method, considerations and recommendations are provided below.

Overall, previous research and the present analyses demonstrates that personal information is more reliable over time than family information for self-generated identification code elements. The most frequently reported personal elements included birth month, first initial of their first name, sex, and own middle initial. Utilization of the first initial of the participant's first name is not recommended as it fails to meet Criteria 2, constancy over time. Birth month, sex, and own middle initial meet all five of the code element criteria and are recommended for use based on our literature review and analyses.

Table 8. Example of Best Practices Recommendation for Self-Generated Identification Codes.

	Question stem: What is the . . .					
	Month you were born?	Sex you were assigned at birth?	First initial of your first middle name?	First initial of your mother's first name?	Number of older siblings	Self-generated identification code
Example response	January	Female	Katherine	Mary	2	
Code created	01	F	K	M	02	01FKM02

Our analyses found that five (5) or more code elements in an SGIC was related to higher match rates. Accordingly, we recommend the use of the following two nonpersonal questions, to increase the number of elements used and correspondingly increase match rate. Specifically, we recommend using first initial of mother's first name, the most commonly used element in our literature review (see Table 3), and number of older siblings, the most accurate nonpersonal element in our analyses (see Table 7). Both these elements meet all five of the code element criteria.

Based on the research reported herein, we recommend asking participants for the month in which they were born, their assigned sex at birth, the first initial of their first middle name (i.e. Lillian Katie Carin Hammond would enter K), their mother's first initial of their first name, and their number of older siblings. If participants do not have a middle name, researchers could, as an alternate, ask for either a participants' first initial of either participant's first or last name, depending on the characteristics of their population of interest (i.e. if studying children, last name would be preferable given many children have nicknames which are unstable). Similarly, for participants who do not have any older siblings, researchers could ask participants to use ''0'' or a non-number symbol like ''X.'' Each of these pieces of information—birth month, assigned sex at birth, first initial of first middle name, first initial of mother's first name, and number of older siblings—could be formatted in a manner of the researcher choice depending on the needs of their study (see Table 8 for an example).

Conclusion

When data are collected—either by paper or electronically—there are difficulties in connecting participants across longitudinal waves while maintaining anonymity. This article reviewed four methods to address this difficulty. Overall, the optimal method for matching participants across data waves is through the use of self-generated identification codes. In descending order, the remaining reviewed options include the electronic anonymizing system, preexisting unique identification code, and collecting nonanonymous data—the last was evaluated as the least optimal. If using self-generated identification codes, based on this analysis, the present authors recommend using birth month, assigned sex at birth, first initial of first middle name, first initial of mother's first name, and number of older siblings as the questions to create the self-generated identification code.

Authors' Note

An earlier version of this article was presented on July 25, 2018, at the 29th International Congress of Applied Psychology, in Montreal, Quebec, Canada.

Acknowledgments

The authors would like to thank Carin K. Smith for their contribution to this project.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by grants from the National Science Foundation (NSF #1238778 & #1623145), awarded to Marie S. Hammond, PhD, Professor, Counseling Psychology, Tennessee State University, Nashville, TN 37209.

ORCID iD

Lillian M. Audette https://orcid.org/0000-0002-9928-8846

References

*References marked with an asterisk were included in the analysis.

Baldwin, A., Schick, V. R., Dodge, B., van Der Pol, B., Herbenick, D., Sanders, S. A., & Fortenberry, J. D. (2017). Variation in sexual identification among behaviorally bisexual women in the Midwestern United States: Challenging the established methods for collecting data on sexual identity and orientation. *Archives of Sexual Behavior*, *46*, 1337-1348. https://doi.org/10.1007/s10508-016-0817-0

Bedeian, A. G., & Feild, H. S. (2002). Assessing group change under conditions of anonymity and overlapping samples. *Nursing Research*, *51*, 63-65.

*Carifio, J., & Biron, R. (1978). Collecting sensitive data anonymously: The CDRGP technique. *Journal of Alcohol and Drug Education*, *23*, 47-66.

*Carifio, J., & Biron, R. (1982). Collecting sensitive data anonymously: Further findings on the CDRGP technique. *Journal of Alcohol and Drug Education*, *27*, 38-70.

Catania, J. A., Gibson, D. R., Chitwood, D. D., & Coates, T. J. (1990). Methodological problems in AIDS behavioral research: influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin*, *108*, 339-362.

Davis-Kean, P. E., Jager, J., & Maslowsky, J. (2015). Answering developmental questions using secondary data. *Child Development Perspectives*, *9*, 256-261. doi: 10.1111/cdep.12151

*DiIorio, C., Soet, J. E., Van Marter, D., Woodring, T. M., & Dudley, W. N. (2000). An evaluation of a self-generated identification code. *Research in Nursing & Health*, *23*, 167-174. Retrieved from https://www.jstor.org/stable/2748630

*Faggiano, F., Richardson, C., Bohrn, K., Galanti, M. R., & EU-Dap Study Group. (2007). A cluster randomized controlled trial of school-based prevention of tobacco, alcohol and drug use: The EU-Dap design and study population. *Preventive Medicine*, *44*, 170-173.

Fisher, C. B. (2013). *Decoding the ethics code* (3rd ed.). Thousand Oaks, CA: Sage.

*Grube, J. W., Morgan, M., & Kearney, K. A. (1989). Using self-generated identification codes to match questionnaires in panel studies of adolescent substance use. *Addictive Behaviors*, *14*, 159-171. https://doi.org/10.1016/0306-4603(89)90044-0

*Haberman, P. W., Josephson, E., Zanes, A., & Elinson, J. (1972, March). High school drug behavior: A methodological report on pilot studies. In *Proceedings of the First International Conference on Student Drug Surveys*. New York, NY: Baywood.

Heiman, G. W. (2000). *Understanding research methods and statistics: An integrated introduction for psychology* (2nd ed.). Houghton Mifflin.

Heppner, P. P., Wampold, B. E., Owen, J., Thompson, M. N., & Wang, K. T. (2016). *Research design in counseling* (4th ed.). Boston, MA: Cengage Learning.

*Honig, F. (1995). When you can't ask their names: Linking anonymous respondents with the Hogben number. *Australian Journal of Public Health*, *19*, 94-96.

*Josephson, E., & Rosen, M. A. (1978). Panel loss in a high school drug study. In D. B. Kandel (Ed.) *Longitudinal research on drug use: empirical findings and methodological issues* (pp. 115-l33). New York, NY: Wiley.

Kadison, P., Pelletier, E. M., Mounib, E. L., Oppedisano, P., & Poteat, H. T. (1998). Improved screening for breast cancer associated with a telephone-based risk assessment. *Preventive Medicine*, *27*(3), 493-501. https://doi.org/10.1006/pmed.1998.0313

*Kearney, K. A. (1982). *Collecting longitudinal data anonymously: Compensating for respondent errors in self-generated identification codes*. Pullman: Washington State University.

*Kearney, K. A., Hopkins, R. H., Mauss, A. L., & Weisheit, R. A. (1984). Self-generated identification codes for anonymous collection of longitudinal questionnaire data. *Public Opinion Quarterly, 48*(1B), 370-378.

Kiesner, J., Mendle, J., Eisenlohr-Moul, T. A., & Pastore, M. (2016). Cyclical symptom change across the menstrual cycle: Attributional, affective, and physical symptoms. *Clinical Psychological Science*, *4*(5), 882-894. https://doi.org/10.1177/2167702616635031

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the internet. *American Psychologist*, *59*, 105-117. doi:10.1037/0003-066X .59.2.105

MacCoun, R. J., & Perlmutter, S. (2017). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological* *science under scrutiny: Recent challenges and proposed solutions* (pp. 589-612). New York, NY: Wiley.

*McAlister, A., & Gordon, N. P. (1986). Attrition bias in a cohort study of substance abuse onset and prevention. *Evaluation Review*, *10*, 853-859.

*McGloin, J., Holcomb, S., & Main, D. S. (1996). Matching anonymous pre-posttests using subject-generated information. *Evaluation Review*, *20*, 724-736.

*Moberg, D. P., & Piper, D. L. (1990). An outcome evaluation of Project Model Health: A middle school health promotion program. *Health Education Quarterly*, *17*, 37-51.

*Morgenstern, M., Sargent, J. D., Engels, R. C. M. E., Scholte, R. H. J., Florek, E., Hunt, K., . . . Hanewinkel, R. (2013). Smoking in movies and adolescent smoking initiation: Longitudinal study in six European countries. *American Journal of Preventive Medicine*, *44*, 339-344. https://doi.org/10.1016/j.amepre.2012.11.037

Murray, S. (1992). Turning an elite cross-sectional survey into a panel study while protecting anonymity. *Journal of Conflict Resolution*, *36*, 586-595.

National Science Foundation. (2017). *Proposal & Award Policies & Procedures Guide* (XI.D.4). Arlington, VA: Author. Retrieved from https://www.nsf.gov/pubs/policydocs/pappg17_1/pappg_11.jsp#XID4

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175-220.

Olson, C. B., Stander, V. A., & Merrill, L. L. (2004). The influence of survey confidentiality and construct measurement in estimating rates of childhood victimization among Navy recruits. *Military Psychology*, *16*, 53-69. doi:10.1207/s15327876mp1601_4

*Ripper, L., Ciaravino, S., Jones, K., Jaime, M. C. D., & Miller, E. (2017). Use of a respondent-generated personal code for matching anonymous adolescent surveys in longitudinal studies. *Journal of Adolescent Health*, *60*, 751-753. doi:10.1016/j.jado health .2017.01.003

Rochester, N. K. (2015). *Methodological issues with coding participants in psychological longitudinal studies* [White Paper].

*Schaalma, H. P., Kok, G., Bosker, R. J., Parcel, G. S., Peters, L., Poelman, J., & Reinders, J. (1996). Planned development and evaluation of AIDS/STD education for secondary school students in the Netherlands: Short-term effects. *Health Education Quarterly*, *23*, 469-487.

*Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the use of self-generated identification codes. *Evaluation Review*, *34*, 391-418. doi:10.1177/0193841X10387576

Stander, V. A., Olson, C. B., & Merrill, L. L. (2002). Self-definition as a survivor of childhood sexual abuse among Navy recruits. *Journal of Consulting and Clinical Psychology*, *70*, 369-377. doi:10.1037//0022-006X.70.2.369

Tenhiälä, A., & Lount, R. B. (2013). Affective reactions to a pay system reform and their impact on employee behaviour. *Journal of Occupational and Organizational Psychology*, *86*(1), 100-118. https://doi.org/10.1111/joop.12002

Thomas, J. L., Wright, K. M., Adler, A. B., & Bliese, P. D. (2004). *Reporting psychological distress: Anonymity versus non-anonymity*. Heidelberg, Germany.

Tregarthen, J. P., Lock, J., & Darcy, A. M. (2015). Development of a smartphone application for eating disorder self-monitoring. *International Journal of Eating Disorders*, *48*, 972-982. doi:10.1002/eat.22386

Udry, J. R., & Bearman, P. S. (1998). New methods for new research on adolescent sexual behavior. In R. Jessor (Ed). *New perspectives on adolescent risk behavior* (pp. 241-269). Cambridge, England: Cambridge University Press.

Widrich, L., & Ortlepp, K. (1994). The mediating role of job satisfaction in the work stress-marital interaction relationship. *South African Journal of Psychology*, *24*, 122-130. doi: 10.1177/008124639402400303

Williams, K. R., & Guerra, N. G. (2007). Prevalence and predictors of internet bullying. *Journal of Adolescent Health, 41*(6 Suppl.), S14-S21. doi:10.1016/j.jadohealth.2007.08.018

*Wilson, A. L. G., Hoge, C. W., McGurk, D., Thomas, J. L., Clark, J. C., & Castro, C. A. (2010). Application of a new method for linking anonymous survey data in a population of soldiers returning from Iraq. *Annals of Epidemiology*, *20*, 931-938. doi:10.1016/j .annepidem.2010.08.008

*Winchester, L., Dobbinson, S., Rissel, C., & Bauman, A. (1996). Anonymous record linkage using respondent-generated identification codes—A tool for health promotion research. *Health Promotion Journal of Australia*, *6*, 52-54.

*Yurek, L. A., Vasey, J., & Havens, D. S. (2008). The use of self-generated identification codes in longitudinal research. *Evaluation Review*, *32*, 435-452. doi:10.1177/019 3841X08316676.

Zimmer, M. (2010). Subject privacy and the release of the tastes, ties, and time dataset. In: *Workshop on Revisiting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research, Computer Supported Cooperative Work Conference*. Savannah, GA.