Midwestern University IRB

Creating Non-Identifiable Codes for Research Participants

In certain situations, investigators need to be able to link research participants' data (across time, between settings, etc.). While this may be easily accomplished in studies approved under a full or expedited review process, being able to link data in studies qualifying for an exempt status can be difficult since collecting the necessary information to link data can actually prevent a study from being exempt. There are ways to create codes that allow investigators to link participants' data without identifying the participant. Below are a few examples of procedures to create these non-identifiable codes. These are by no means exhaustive. If you have ideas for other methods, please notify the IRB chair for potential inclusion in this document.

Example 1: Create a code made from several pieces of information that are readily known to the research participant but are not known to the investigator

There are various pieces of information (e.g., elements of dates of birth, phone numbers, make and model of one's first car) that a research participant can readily provide and will not change over time. Box 1 provides examples of these elements. The listed elements listed are not meant to be exhaustive, as there are certainly many other examples that could have been included. Selecting several elements from the table below allows participants to create their own code that allows investigators to link their data without having their identity revealed. Investigators can select several of the elements below to allow participants to create their own unique codes. The number of elements that are selected depends on the size of the sample involved. For example, with a sample of 100 patients, it may be sufficient to select only two of the following elements. When working with a sample of 1000 students, however, 4-5 elements may be needed. Selecting several elements also increases the potential that the investigator will be able to link data even when the participants provide incorrect information for one element or if there are several participants with identical information for a particular element.

In addition to the concatenation of elements, it is possible to have the participant perform mathematical operations on various numerical elements. For example, multiply their two-digit month of birth with their two-digit date of birth. Multiplication or division will produce fewer duplicate results than addition or subtraction. This can avoid the potential for the participant's code containing any actual raw information (e.g., month of birth). The possibility of mathematical errors, however, cannot be ignored. If this approach is used, ensuring that participants have access to calculators is beneficial. It is also recommended that at least two letter-based elements (e.g., mother's and father's first name initials) be used to enable linking data when mathematical operations on different numbers produce the same result.

A few caveats must be mentioned when using this method. Investigators must not use the participant's month *and* date of birth together in the raw format (i.e., not the product or sum of the two) as this is too readily identifiable. Collecting the month and year can avoid this, as can using another individual's date of birth (e.g., mother's or father's date of birth) or performing a mathematical operation on the month and date of birth (e.g., product of the month and date). Collecting the full five-digit ZIP code is not advisable since it may identify a participant if the sample is small or if only one or two participants come from a given ZIP code. This can be avoided by using the first or last three (3) digits.

Once the elements are selected, it is important to provide explicit instructions so it is clear to participants what they are providing and in what format (see Box 2). These instructions should be provided to participants at each data collection session to ensure they remember the construction of their unique code. If data are being collected electronically (e.g., in an online survey), it may be beneficial to create a separate data entry field for each character of the code (or at least sections of the code, such as the two characters for initials) and to use data validation rules to ensure that characters vs. numbers are used when appropriate and that the entry is of the appropriate length (e.g., two-digit date).

Box 1 – Sample elements for code generation

Initials	Other letters (Using the full word could be acceptable)	Dates (or pieces of dates)	Other numbers
• Mother's first and/or middle initials • Mother's maiden name initial • Father's first and/or middle initials • Spouse/partner/significant other's first and/or middle initials • Sibling's first and/or middle initials	• First letter of the make of your first car (e.g., H for Honda) • First letter of the model of your first car (e.g., A for Accord) • First letter of the street on which you grew up (e.g., S for 123 Sesame Street) • First letter of your first pet's name	• Mother's month of birth • Mother's date of birth • Father's month of birth • Father's date of birth • Your month of birth[a] • Your date of birth[a] • Your year of birth • Spouse/partner/significant other's month of birth • Spouse/partner/significant other's date of birth	• Your area code • First 3 digits of your phone number (i.e., your 3 digit exchange; 555 in 630-555-1212) • Last 4 digits of our phone number (e.g., 1212 in 630-555-1212) • Model year of your first car • ZIP code (first or last 3 digits)[b]

[a] Using both month and date of birth together is not acceptable; also any raw elements of dates directly related to an individual other than year (e.g., birthdate) are protected elements under HIPAA

[b] Using the full five-digit ZIP code is not advisable as it may identify an individual depending on the size of the research sample; also this is a protected element under HIPAA

Box 2 – Sample instructions

As part of this study, the investigators need to link your responses across time. To do this, you will create your own unique code. This will not allow the investigators to determine your identity. It is only to facilitate linking your responses. Your code will be 7 characters long. The first two characters are your father's first and middle initial. The third character is your mother's maiden name initial. The last four are your 2-digit date of birth and your 2-digit year of birth. As an example, Sally Thompson is participating in this project. She was born on February 20, 1984. Her father's name is James Edward Thompson. Her mother's name is Nancy Jane Thompson (maiden name is Smith). Sally's identification code would be JES2084.

Example 2: Create a unique code for each participant that is managed/maintained separately from the data collection process

Creating a unique code in this manner may reduce the burden on the participant and reduce the potential of not being able to link participants because of incorrect codes. In this method, a unique code is created for each participant. The codes are linked to the participant's identity (i.e., name) before the first data collection period. At the first collection period, a participant receives his or her code at random. The codes are then used at subsequent data collection sessions without the investigator knowing which code is associated with any given participant.

An example of how this can be implemented may be helpful. An investigator is conducting a project where students complete a questionnaire at the beginning of their academic program and then again at the end of each of their four years in the program, so five (5) data collection points total. The investigator creates a set of self-adhesive labels for the maximum number of students who will be participating in the project. There are no student names attached to any codes at this point. The number of labels for each student is the same as the number of data collection points. At the first data collection session, the sets of labels are given to each student along with a blank envelope. Students are instructed to affix one label to the questionnaire. The remaining labels are placed in the envelope where the student writes their name on the front, seals the envelope, and then returns the envelope to the investigator (possibly in a separate box in order to keep the completed questionnaire separate from the envelopes). At subsequent data collection sessions, the sealed envelopes are returned to the respective students along with another blank envelope. As before, they affix one label to the new questionnaire and place the remaining labels in the envelope on which they write their names, seal, and then return. This is repeated until the last data collection session where the final label is used. The envelopes remain sealed until the next data collection session, so the investigator does not know which code goes with which student. It is also possible for an individual other than the investigator to keep the sealed envelopes so that they are physically separated and not in the possession of the investigator and anonymity is further ensured.