Guidance for Ethical Use of Data in Research

There are many regulatory and ethical issues to consider when conducting research in the primary and secondary education context. This guidance addresses research involving students, educational records, and educational settings.

1. Why is it important to think about research data as an ethical issue?

One of the most complicated aspects of protecting research participants from harm entails understanding risks present in research data collection. We often may consider a particular type of data safe or harmless without being aware of additional risks created by research design. We may also consider the protections we create around research data sufficient for a specific research study without being aware of additional ethical issues, regulatory requirements, or best practices.

In order to craft an effective and ethical research protocol, it is necessary to first understand what types of data may be collected and how these data should be handled and stored. The criteria for approval (45 CFR 46.111) used by IRBs specifically mandate that "*there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data*" any time research is conducted with human subjects. Before completing your IRB application, it is important to have a specific plan in mind for how you will be protecting participants by properly handling their data.

2. What are different types of research data?

There are standard ways to describe types of data and methods of research data collection. Understanding this terminology is the first step in crafting an ethical research design. These concepts are organized below by how identifiable data might be, what type of data are collected, and when data are collected.

- By Identifiability:
    - *Individually Identifiable*: According to Lindenwood University policy, identifiable means that the identity of a participant may be readily ascertained by the investigator or others. In this policy, a number of conditions and data points are listed as elements of identifiable data. This definition is consistent with OHRP, which generally considers private information or specimens to be individually identifiable as defined at 45 CFR 46.102(f) when they can be linked to specific individuals by the investigator(s) either directly or indirectly through coding systems.
    - *Directly Identifiable*: A code or element of a data set which singly or in combination can be connected directly to a participant. Examples of direct identifiers are name, address, or birthdate.
    - *Indirectly Identifiable*: Any information or an element of research design which may permit accidental identification of a research participant. Examples include combinations of geographical and other personal information or unique information

present in a small data set. Data can also become indirectly identifiable due to the location or timing of a research activity.
  - *Deidentified*: Data which was once identifiable, but any directly or indirectly identifiable information has been removed.
  - *Coded*: Data which have been deidentified, and a code containing numbers, letters, or other elements has replaced the identity of each participant. An individual called an honest broker may retain the link between this code and each participant to permit the re-identification of a specific participant.
  - *Anonymized*: Data from which all directly or indirectly identifiable information has been removed, and no code is retained that might connect a participant with their data. The researcher or others are not able to identify any participants in the data set.

- By Type or Context:

  - *Private Information*: Information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information that has been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public (for example, a medical record). Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects (45 CFR 46.102(f)).
  - *Sensitive*: Information that can cause harm to a participant if disclosed. Data can be considered sensitive in one context and not in another, given differences in personal, ethical, and cultural perspective. Sensitivity can also be a function of research design and sample size. Lindenwood University policy describes specific types of data as sensitive.
  - *Confidential:* Data which are identifiable, but collected and stored in such a way that only the researcher has access to any connection between the data and the participant who provided it.
  - *Publicly Available*: Information obtained through public sources of data. Examples include census data or data sets made available by government agencies for research purposes. If a data set is not available without additional approval or authorization, it cannot be considered public.
  - *Protected Health Information (PHI):* Individually identifiable health information recorded in any form or medium that is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse and relates to the past, present, or future physical or mental health or condition of an individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual. PHI is subject to regulation by HIPAA and typically may not be released without a patient's authorization.
  - *Limited Data Set*: HIPAA regulation describes a type of PHI which may be released without a patient's authorization. These data are limited to sets which only include

dates, location information, and age. A limited data set may only be released if a Data Use Agreement has been signed.

- *Educational Record*: Data from any record related to a student which is maintained by an educational agency or institution. Education records are subject to regulation by FERPA and PPRA and typically may not be release without individual authorization.
- *Directory Information*: Data contained in the education records of a student which would not generally be considered harmful if disclosed for the purposes of research. Lindenwood University has a specified set of data elements considered directory information. Directory information is subject to regulation by FERPA.

- By Source or Timing of Collection:
  - *Primary Data*: Data collected directly from participants, sources, or observations. These data are typically collected by the researcher and the research team during the conduct of researcher through surveys, interviews, and structured observation or interaction.
  - *Secondary Data*: Data that have already been collected for another purpose, prior to the research activity. Common examples of secondary data are census data, state or district level educational data, or data collected during the routine operations of an organization. Secondary data are often obtained upon request from an organization or honest broker for the purposes of research.

3. What is the difference between Deidentified and Anonymized data?

The terms “deidentified” and “anonymous” are often confused. Knowing the difference between the two is essential in developing practices for ensuring the privacy of research participants. Deidentified data are data from which any identifiable information has been removed, and a code linking participants with their data remains. Deidentification permits the possibility that a research may re-identify a participant to collect further data or confirm the accuracy of data collected. In contrast, anonymized data no longer retains any connection between the data and participants. The researcher or others would not be able to reconstruct the identity of a participant from the data set.

4. What is the difference between Confidential and Anonymous?

Keeping data confidential and anonymizing data are distinct practices, even though these terms are commonly used in a similar way. Anonymous simply means that the research team is not able to link any data with the identity of a participant. Anonymous data can be created in several ways. Identifiers are not collected during primary data collection. Identifiers are removed from a data set and no code is retained to connect participants and their data. In addition, many survey platforms can automate the anonymizing of survey results.

Confidential means that a link between a participant and their data exists, but only the research team has access to this link. Maintaining confidentiality is an effective way to protect participants when they are providing sensitive information. Common ways to maintain confidentiality include conducting recruitment and consent processes in a private setting, storing research data in encrypted formats, and creating a code to replace identifiers in data sets.

5. What is the difference between a Direct and Indirect Identifier?

A direct identifier is a specific element of data, such as a name, social security number, or address, which directly connects a participant and their data. Lindenwood University policy lists several conditions and elements which render data identifiable. HIPAA policy lists 18 unique identifiers which are important to consider if conducting research involving Protected Health Information (PHI).

An indirect identifier is any element or condition of data which potentially allows someone to connect a research participant and their data. There are cases in which a data set may not contain any specific direct identifier, but a combination of different data elements may lead to the identification of a participant. Given the increasing power of computation, there are many cases in which seemingly anonymized data sets permit the identification of participants. There are also elements of research design which may create indirect identifiers. For example, if a research activity is conducted in a specific location, people might identify people as research participants when they enter that location.

6. What is an Honest Broker?

An Honest Broker is an individual responsible for maintaining the link which makes coded data possible. Many institutions will use an individual to manage data sets and ensure that researchers only have access to deidentified data. This Honest Broker often has specific training to ensure transmission of data meets different regulatory requirements, such as those defined in FERPA or HIPAA. Under HIPAA regulations, an honest broker may create and release a deidentified or limited data set under certain conditions. FERPA regulations also describe conditions under which a data set could be considered deidentified.

7. What are common practices used to create Coded Data?

Data are considered Coded when all identifiers in a data set have been replaced with unique study IDs, code numbers, or pseudonyms. In some research designs, an Honest Broker, the investigator, or designated members of the research team retain a key which maintains a link between a participant and their coded data. This key is then kept secure to ensure the privacy of participants. Coded data can be considered anonymous if no key is retained, or deidentified if a key is retained. Common coding methods include:

- Assigning participants a unique study ID and maintaining a separate key (e.g. in an Excel spreadsheet) linking a participant and their code.
- Coding data to only retain the minimum necessary identifiable data, such as creating age ranges to replace birth dates and geographical labels for participant addresses. The key to these codes are retained by the researcher.
- Programming software to randomly shift elements of data (such as all participants' date of birth) that can later be reversed to regenerate the original identifiable set.

- Providing participants a unique study ID or name that they will use on follow-up surveys. This code would be provided at an initial visit or with an initial survey and then used to match survey responses when multiple surveys or interviews are planned.

8. What are some common methods used to Deidentify data?

Data are considered deidentified when a participant's data and their identity cannot be readily ascertained by the researcher or others. There are many strategies commonly used to deidentify data, such as the following:

- *Masking* is a general way to describe any technique used to replace identifiable elements in a data set with different elements. For example, a birth date may be altered based on a predetermined algorithm, which allows the process to be reversed at a later time. As this algorithm would be applied in the same way to all instances of the data point, the data can still be analyzed in an effective way. In other cases, identifiable data may be replaced by inauthentic equivalents, such as alternate geographical locations or age ranges.
- *Perturbation* is a type of masking involving making minor changes to a data set to ensure participants cannot be identified. This is a common procedure in research involving rare conditions or participants with unique characteristics. In order to ensure statistical validity, this technique uses probability distribution or value distortion to create "noise" in a data set while retaining its analytic value.
- *Top-coding* is a simple method used when data sets include test scores. As test scores at the upper and lower limits can often become incidental identifiers, top-coding creates a defined maximum or minimum for test scores outside of a predetermined boundary.
- *Redaction* involves strategically removing identifiable data from a data set prior to public use or sharing.
- *Generalization* is useful when identifiable data points can be consistently transformed into abstract representations or ranges. For example, specific address or location information can be turned into ZIP codes or geographical categories. A five-digit ZIP code could be generalized to a three-digit ZIP code. Specific ages can be transferred into age range categories.

9. What are common research data protections to consider?

Before a researcher begins to collect research data, it is essential that they have a detailed plan for research data protections. It is helpful to think of each study as a narrative, a story of how a certain set of data are collected, handled, and stored. Please refer to the "Research Data and Materials Security" guidance for detailed description of best practices for research data protections.