Ethical Dilemmas in Sampling

Patrick Dattalo, MSW, PhD,
Virginia Commonwealth University

Journal of Social Work Values and Ethics, Volume 7, Number 1 (2010)

Abstract

This paper focuses on sampling as a nexus of ethical dilemmas experienced by social workers and other applied empirical researchers. It is argued here that social workers and other applied researchers have an ethical obligation to construct the smallest representative samples possible. Although random sampling is considered by many researchers as the gold standard methodological procedure for maximizing external validity and optimizing sample size, in practice, random sampling often is difficult to implement. Recommendations include using (1) deliberate sampling to balance a sample's composition in terms of typicalness and diversity; (2) randomization tests; (3) a broader perspective on external validity; (4) one-sided hypotheses; (5) sequential sampling; and (6) planned missing data designs.

Key Words: ethical dilemmas, external validity, unbiased sampling, optimal sample size

1.0 Introduction

The need to recognize and address ethical dilemmas at all stages of the research process is widely recognized and has been extensively discussed (cf. Engel & Schutt, 2009; English, 1997; Nelson, 1994; Rubin & Babbie, 2008). For example, ethical dilemmas may result from the researchable questions asked, the population and setting studied, the kind of information sought, and the methods used to collect data. Important ethical issues include voluntary participation and informed consent, anonymity and confidentiality, and accountability in terms of the accuracy of analysis and reporting. The need to identify and attend to the aforementioned ethical dilemmas has intensified with increasing emphasis on evidence-based practice, which is a process in which practitioners make decisions in light of the best empirical evidence available (Rubin & Babbie, 2008).

This paper focuses on sampling as a nexus of ethical dilemmas experienced by social workers and other applied empirical researchers. Emphasis will be on ethical dilemmas encountered during a quantitative research process in which the primary goal is to seek evidence about a characteristic or a relationship and to use statistical inference to generalize obtained results from a sample to a population. The National Association of Social Workers (NASW,) Code of Ethics is intended to serve as a guide to the everyday professional conduct of social workers, including researchers (NASW, 2008). At a minimum, sampling can present a conflict between social workers' ethical responsibilities for professional integrity (i.e., ensuring the rights of participants), and for evaluation and research (i.e., maximizing the generalizability of study results to the study population). The term "generalizability" is used here as a synonym for external validity, which is how well findings travel to other participants, times, and places (e.g., form a sample to a population) (Cook & Campbell, 1979).

Most social work research involves some risk or costs to participants. It is typically considered ethical if participants agree to take identified risks and to bear, or be reimbursed for, their costs. For example, risk to participants can occur if there is a new intervention, which does not address existing problems (i.e., no gain), or worse, exacerbates these problems. Costs to participants include, for example, travel and daycare expenses. Debate continues over an optimal balance between the adequate compensation of participants for their time and other expenses and their informed and voluntary decisions to be studied. Resolution of this debate seems especially difficult for participants who are vulnerable because they lack financial resources or the emotional or cognitive ability to make informed and voluntary decisions to participate in a study. Particularly vulnerable groups of study participants

include children, prisoners, persons who are terminally ill, persons who are economically disadvantaged, and persons with emotional or cognitive disabilities.

Accordingly, it is argued here that a fundamental ethical dilemma for social work researchers is to conduct studies in light of the need to maximize external validity while ensuring the rights of participants. Debate over the optimal balance between compensation and voluntary participation notwithstanding, at a minimum, social work researchers can help ensure participant rights by selecting the smallest representative sample necessary to obtain generalizable study results. For example, if a study seeks (1) sensitive information (e.g., from men in a study to compare the effectiveness of two residential substance abuse interventions); (2) information from a vulnerable population (e.g., children in a study to evaluate an intervention designed to reduce the psychosocial difficulties of children with diabetes), or (3) information during a crisis (e.g., from women seeking protective orders in cases of spousal abuse), it could be unethical to sample too many or too few people. If a sample is too small, a study could miss important effects, place unnecessary demands on participant privacy and time, or waste valuable resources. If a sample is too large, the study could make unnecessary demands on participants or misuse other resources.

Participant rights, sample size, and generalizability are interrelated issues. For clarity, the issues of ensuring participant rights and maximizing external validity and ensuring participant rights and minimizing sample size will be discussed in separate sections. Accordingly, the following discussion is focused on maximizing a sample's external validity through unbiased (i.e., random) sampling and optimal (i.e., the smallest necessary) sample size. First, the limitations of random sampling as a strategy to maximize external validity are presented. Next, three alternatives to random sampling to maximize external validity are proposed: (1) deliberate sampling for typicalness and diversity; (2) randomization tests; and (3) a broader perspective on external validity. Then, three alternatives strategies to obtain optimal sample size are presented: (1) using one-sided hypothesis; (2) sequential sampling; and (3) planned missing data designs. Please see Figure 1 for a summary of alternatives to random sampling and strategies for optimizing sample size.

Figure 1
Strategies to Maximize External Validity While Ensuring Participant Rights

Strategy	Purpose	Random Sampling Occurred?
Deliberate Sampling for Typicalness and Diversity	Reduce Sample Bias	No
Randomization Tests	Reduce Sample Bias	No
Broader Perspective on External Validity	Reduce Sample Bias	Yes and No
Use of One-Sided Hypotheses	Minimize Sample Size	Yes
Sequential Sampling	Minimize Sample Size	Yes
Planned Missing Data Designs	Minimize Time and Effort of Sample Participants	Yes

2.0 Maximizing External Validity through Unbiased Sampling

The ultimate goal of sample design is to select a set of elements from a population in such a way that descriptions of those elements accurately portray characteristics of the population (i.e., parameters) from which they were selected. Another important goal of sample design is to yield maximum precision (i.e., minimum variance) per unit cost. The sampling process begins with the identification of an appropriate population to study to answer researchable questions, and includes (1) the formulation of a sampling strategy, and (2) determination of sample size and composition to maximize the external validity.

2.1 What Random Sampling Does and Does Not Do

The term random sample, also called probability sample, is used to describe a sample that is free from systematic error. A sample is unbiased, then, if every element in a population has an equal chance of being selected. According to classical statistical sampling theory, if random selection from a known population is performed, characteristics of the sample can be inferred and tend to mirror corresponding characteristics of the population. If random sampling is not performed, there is no theoretical basis for statistical inference. Only information about a sample can be described. However, although random sampling for representativeness minimizes systematic error, sampling biases still can occur for the following reasons:

1. A complete randomization process is usually not implemented (Cook, 1993). Frequently, only units are randomized, which is only one of three different areas that define an event. The other two areas that define

an event are place or setting, and time (Cook & Campbell, 1979);

2. Random sampling does not minimize all error in a research design. There are other types of bias in the sample that may contribute to error, such as non-sampling bias (e.g., measurement error (Henry, 1990);
3. The sample may not be representative of the population because it is too small, and therefore, likely to be too homogeneous;
4. The representativeness of the sample may be impacted by attrition and refusal of the participants to take part in a study;
5. Random sampling permits causal generalization to a target population but not *across* multiple populations (Cook & Campbell, 1979). The latter is important for establishing an abstract principle of causality and is best done through multiple replications across units, setting, and time;
6. Conclusions drawn from random samples are based on inferential theory or the probability of the occurrence of an event. Random sampling, alone, does not guarantee accurate estimates of population parameters; and
7. Random sampling usually requires considerable resources compared with nonrandom sampling strategies. If resources are inadequate to enumerate a representative sampling frame and to draw sample size for adequate statistical power, the hoped for goal of random sampling (i.e., an unbiased sample), may not be achieved.

If random sampling is not possible social work researchers should consider a nonprobability alternative, such as deliberate sampling, to help to ensure generalizabiltiy.

2.2 Deliberate Sampling for Typicalness and Diversity

Cook and Campbell (1979) argue that deliberate sampling, also termed purposive sampling, may be useful if a sample is carefully constructed. Deliberate sampling is a type of nonprobability sampling in which elements are knowingly chosen based on a study's research questions (Cook & Campbell, 1979). Blankertz (1998) and Bull (2005) emphasize the following two variations on deliberate sampling: (1) deliberate sampling for diversity, which involves selecting a sample with a wide range of characteristics that are expected to influence results; and (2) deliberate sampling for typical instances, which involves selecting at least one instance of each class that is impressionistically similar to that class's mode.

Deliberate sampling can be used to achieve a variety of research goals. Potential applications of deliberate sampling are as follows:

1. Study a time-limited population (e.g., all clients being served by a department of social services);
2. Study a subset of a population (e.g., only clients being provided child protective services by a department of social services);
3. Primary data analysis; that is, deliberate sampling can be used to select clients for a pilot study that will be used to guide a larger scale study;
4. Secondary data analysis; that is, to select a sample from an existing data set for a secondary analysis; and
5. Descriptive analysis; that is, a researcher can select a small subsample and closely examine typical and unusual or extreme elements.

Deliberate sampling shares certain characteristics with stratified sampling. In a stratified sample, the sampling frame is divided into non-overlapping groups or strata (e.g., age groups, gender). Then, a random sample is taken from each stratum. Stratified sampling uses groups to achieve representativeness, or to ensure that a certain number of elements from each group are selected. Like stratified random sampling, deliberate sampling can be used to control the characteristics of cases being selected (cf. Armitage, 1947; Kott, 1986).

There is empirical evidence of the ability of stratified random sampling to increase precision when the strata have been chosen so that members of the same stratum are as similar as possible in respect of the characteristic of interest; the larger the differences between strata, the greater the gain in precision (cf. Armitage, 1947; Kott, 1986). Stratification (and deliberate sampling) can help to ensure that not only the overall population, but also that key subgroups of the population, are represented. For example, if the subgroup is small, and different sampling fractions are used to "over-sample the small

group" stratified random sampling will generally have more statistical precision than simple random sampling. The benefits of stratification are greatest when the strata or groups are homogeneous; that is when within-groups variability is lower than the variability for the population.

The following example seeks to achieve the first of the five aforementioned research goals (i.e., study a time-limited population) and follows a procedure described in Blankertz (1998). Note that this procedure is versatile and could be used to achieve any of these five research goals.

A researcher conducts a study to determine the effectiveness of a peer-led eating disorders prevention intervention in reducing eating-disorder risk factors in young women (18 to 21 years of age). The intervention is implemented, at the discretion of the school, in public four-year colleges and universities in a state as a part of new student orientation. First, a sampling frame is used to conceptualize a deliberate sample for diversity and typicalness of new students. Next, all new students in this deliberately constructed sampling frame are randomly assigned to either an intervention or a control group. The intervention consists of eight two-hour group sessions that were delivered by trained peer facilitators. Participants completed questionnaires that assessed eating-disorder risk factors pre and post treatment with higher scores indicating a greater risk of eating disorders. Demographic characteristics of participants, including age, gender, race/ethnicity, region (based on the home address of students), and BMI (Body Mass Index) score were also collected. Results consisted of a comparison of the intervention and control group means. Further analysis consisted of a comparison of the intervention and the control groups in three subsamples (i.e., two for diversity and one for typicalness).

Deliberate sampling for diversity involves selecting two subsamples, each chosen to "differ as widely as possible from each other" (Cook & Campbell, 1979, p. 78). In addition, these two subsamples samples should be selected to vary across several characteristics, including time and place. It is helpful to view each characteristic in each subsample as the endpoint on a continuum of a ratio, interval, or ordinal variable. For nominal variables, each characteristic represents a different category of that variable. Each subsample should contain clusters of elements that represent endpoints of ordinal, interval, or ratio variables, or the different categories of a nominal variable. That is, each subsample should contain values that represent opposing endpoints or categories.

Cook and Campbell (1979) explain that "given the negative relationship between 'inferential power' and feasibility, the model of heterogeneous instances (i.e., sampling for diversity)would seem most useful, particularly if great care is made to include impressionistically modal (i.e., typical) instances among the heterogeneous ones" (p. 78). Moreover, Cook and Campbell (1979) conclude that

> Practicing scientists routinely make causal generalizations in their research, and they almost never use formal probability sampling when they do. Scientists make causal generalizations in their work by using five closely related principles: (1) surface similarity, (2) ruling out irrelevancies, (3) making discriminations, (4) interpolation and extrapolation, and (5) causal explanation. Deliberate or purposive sampling for heterogeneous instances; and impressionistic or purposive sampling of typical instances are essential components of these principles (p. 24).

With the careful matching of sampling strategy to purpose, deliberate sampling can be a useful alternative to random sampling. If random sampling is not possible social work researchers should consider a nonprobability alternative, such as randomization tests, to help to ensure generalizabiltiy**.**

2.3 Randomization Tests

According to Howell (2007), randomization tests differ from parametric tests as follows:

1. There is no requirement that a sample is randomly drawn from a population;
2. There is no assumption about the population from which the sample is drawn (e.g., it is normally distributed), although as sample size increases, the distribution produced by permutations approaches the normal distribution;
3. Because there are no assumptions about a population, no sample statistics are used to estimate population parameters; and
4. Although test statistics are calculated, they are not utilized in the same way as they are in parametric hypothesis testing. Instead, the data are repeatedly

> randomized across groups, and test statistics are calculated for each randomization. Therefore, at least as much as parametric tests, randomization tests emphasize the importance of random assignment of participants to treatments.

A randomization test can be described as follows. A test statistic is computed for study data (e.g., a *t*-test), termed an obtained result. Then, these data are permuted repeatedly and the test statistic is computed for each of the resulting data permutations. When data are permuted, the sample is divided or rearranged by random assignment without replacement to fill the first group, and then to fill the second group until each group contains a new sample of the same size as the original group. These permutations, including the one representing the obtained result, constitute the reference set for determining significance. The proportion of data permutations in the reference set that have a test statistic values greater than, or for certain test statistics, less than or equal to the value for the obtained result, is the p-value.

For example (hypothetical), in a study of the effectiveness of a new treatment to increase empathy in a group of spouse abusers, participants are randomly assigned to either a treatment or a control group. One group is a control condition with scores of 25, 22,23,21,17 on an empathy scale, and the other group was the treatment condition with scores of 30, 27, 28, 29, 29. If the treatment had no effect on scores, the first number that was sampled (25) would be equally likely to occur in either group. With five observations in each group, and if the null hypothesis is true, any five of these 10 observations would be equally likely to occur in either group. These data are "exchangeable" between conditions.

After calculating all of the possible arrangements of the aforementioned 10 observations with five observations in each group (there are 252 possible arrangements), the relevant test statistic is calculated (independent groups *t*-test) for each possible arrangements, and compared to the obtained *t*-test value (4.9252) to test the null hypothesis of no difference in scores between the treatment and the control group. In this case, there are two arrangements of these data that would have a smaller mean for the control group and a larger mean for treatment group. For a one-tailed test, there are two data sets that are at least as extreme as these data. Consequently, a difference that is at least as large as the obtained *t*-value of 4.9252 would occur two times out of 252 for a probability of .006 under the null hypothesis. That is, this difference is statistically significant at $p < .01$. Edgington (2007), Erceg-Hurn and Mirosevich (2008), and Rodgers (1999) provide more detailed explanations and examples of the use of randomization tests.

Stata (http://www.stata.com/) is a commercial general purpose statistical software that can be use to perform a wide range of randomization tests. A free alternative is David Howell's program, Resampling.exe, which is available online from http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html. This software can be used to perform a limited range of randomization tests.

2.4 Toward a Broader Perspective on External Validity

Whether or not random sampling is possible social work researchers should consider a broad perspective on external validity. Reasoning from data points in a sample to an estimate of a population characteristic is an instance of induction. Hume, who was an 18th century Scottish philosopher, usually is credited with discovering "the problem of induction." As identified by Hume, the problem of induction is how to establish induction itself as a valid method for empirical inquiry. See, for example, Wood (2000) for a detailed explanation. According to Rosenberg (1993),

> Hume recognized that inductive conclusions could only be derived deductively from premises (such as the uniformity of nature) that themselves required inductive warrant, or from arguments that were inductive in the first place. The deductive are no more convincing than their most controversial premises and so generate a regress, while the inductive ones beg the question. Accordingly, claims that transcend available data, in particular predictions and general laws, remain unwarranted (p. 75).

To clarify the fundamental limitations of statistical, sampling-based generalization, consider the hypothesis, H_A that the average difference in the perceived effectiveness, which two groups of social workers (i.e., those working in a public social services agency and those working in a public mental health agency) associate with a particular intervention, is 3. In other words, a researcher does not know the numerical value of the average difference in perceived effectiveness between two groups, but hypothesizes it to be 3 (where effectiveness is measured, for example, on a scale from 1 to 5). The researcher then tests the H_0 hypothesis that there is no difference between the

average perceived effectiveness between the two groups by taking a random sample of social workers from each group, and uses the average of the sample from each group as an estimate of the average of the perceived effectiveness for that group.

If the legitimacy of inductive reasoning is unquestioned, then the researcher could reason that the sample average is generalizable to the population average. However, if the legitimacy of inductive reasoning is questioned and Hume's argument is applied, there would be no sound basis for making any statement about the value of the population average. This idea can be expressed as follows: Just because all differences between the two groups in past samples have an average of 3 does not mean that all or any differences between the two groups in future samples will have an average of 3.

Significance tests based on probability sampling, at best, provide very specific information about a population based on a sample's characteristics. In statistical significance testing, the p-value is the long-run probability of obtaining a result (e.g., differences in perceived effectiveness between two groups) at least as extreme as the given result, assuming the null hypothesis. As Cohen (1994) pointed out, what researchers and consumers of research want to know is the population parameter, given the statistic in the sample and the sample size. Unfortunately, the direction of the inference is from the population to the sample, and not from the sample to the population (Thompson, 1997). That is, the logic of hypothesis testing assumes the null is true in a population, and asks: given this assumption about the parameters of a population, what is the probability of the sample statistic?

Campbell and Stanley (1966) eloquently call attention to the "painful" limitations of inductive reasoning when they state:

> Whereas the problems of internal validity are solvable within the limits of the logic of probability statistics, the problems of external validity are not logically solvable in any neat, conclusive way. Generalization always turns out to involve extrapolation into a realm not represented in one's sample. Such extrapolation is made by assuming one knows the relevant laws. Thus, if one has an internally valid [design], one has demonstrated the effect only for those specific conditions which the experimental and control group have in common, i.e., only for pretested groups of a specific age, intelligence, socioeconomic status, geographical region. . . Logically, we cannot generalize beyond these limits; i.e., we cannot generalize at all. But we do attempt generalization by guessing at laws and checking out some of these generalizations in other equally specific but different conditions. In the course of the history of a science we learn about the "justification" of generalizing by the cumulation of our experience in generalizing, but this is not a logical generalization deducible from the details of the original experiment. Faced by this, we do, in generalizing, make guesses as to yet unproven laws, including some not even explored (p. 17).

Campbell and Stanley (1966) conclude that "induction or generalization is never fully justified logically" (p. 17), and they argue that a sample can, at best, offer only limited support for generalization.

Evidence of result generalizability is critical to the accumulation of knowledge, and should be provided by authors. Accordingly, social work researchers should always provide a detailed description of a study's sample. A detailed description is necessary to understand the population being studied and to judge whether the extent of generalizing results seems appropriate. Also, when possible, a comparison of study participants and information about the population should be provided to enable readers to evaluate a sample's representativeness in terms of the larger population from which it was drawn.

The ability to generalize from one situation to another depends on the ability to understand underlying principles and to recognize which underlying principles apply in a given situation. According to Mook (1983), there is no alternative to thinking through, case by case (1) what conclusions are desired; and (2) whether the specifics of a sample or setting prevent these conclusions (p. 386). Mook argues that any generalization to a population of interest must be made on other than statistical grounds.

A broader perspective on generalization recognizes that it requires a series of inferences and judgments regarding the appropriateness of applying findings, concepts, or theories to new or different settings or phenomena. Generalization, therefore, involves identifying similarities and differences between research participants and between research contexts to assess whether a finding or theory is relevant to a new domain (Audi, 2003). Lee and Baskerville (2003) propose a framework of four different types of generalizability built upon the distinction between empirical and theoretical statements as the inferential content. Empirical statements refer to data from and descriptions of

empirical phenomena; theoretical statements refer to phenomena that cannot be directly observed and therefore can only be theorized from empirical data or other theories (p. 232). A second distinction that forms this typology is the distinction between "generalizing *from"* and "generalizing *to"* (p. 232).

Generalization is usually considered to be the ultimate goal of quantitative research. However, an expanding acceptance of the complementary and supplementary roles of qualitative and quantitative approaches to social work should serve as a reminder of the need to recognize the tension between the particular and the general throughout the research process, and of the potential contributions of both random and nonrandom sampling strategies. This tension suggests the importance of thinking more deeply about the content, function, and ethical implications of result generalizations.

3.0 Maximizing External Validity through Optimal Sample Size

Sample size influences the quality and accuracy of empirical research. In general, increased sample size is associated with decreased sampling error. The larger the sample, the more likely the results are to represent the population. However, the relationship between sampling error and sample size is not simple or proportional. There are diminishing returns associated with adding elements to a sample. The relationship between sample and accuracy may be clarified by the use of the concept of statistical power. The notion of statistical power is attributed to Neyman and Pearson (1928), although Fisher (1925) addressed similar issues in his discussions of test and design sensitivity and popularized in the behavioral sciences by Jacob Cohen (c.f., Huberty, 2002, for a detailed discussion).

Power is the probability of rejecting the null when a particular alternative hypothesis is true. More simply, statistical power is the probability of detecting a pre-specified effect size (e.g., a minimally important one). An underpowered study is one for which the projected scientific or clinical value is unacceptably low because it has less than 80% chance of resulting in statistical significance at an a priori set a level (usual $p < .05$). Researchers should avoid conducting studies that are ''underpowered.'' Conversely, researchers should avoid conducting studies with too large a sample size. Studies with samples that are too large may needlessly place respondents at risk, waste their time, and misuse other resources, such as professional time and scarce research dollars. Accordingly, researchers should focus on determining the smallest necessary sample size.

Bacchetti, Wolf, Segal, and McCulloch (2005a; 2005b) discuss how sample size influences the balance that determines the ethical acceptability of a study. That is, the balance between the burdens that participants accept and the clinical or scientific value that a study can be expected to produce. The average projected burden per participant remains constant as the sample size increases, but the projected study value does not increase as rapidly as the sample size if it is assumed to be proportional to power or inversely proportional to confidence interval width. This implies that the value per participant declines as the sample size increases and that smaller studies therefore have more favorable ratios of projected value to participant burden. Bacchetti et al. (2005a; 2005b) provocatively conclude that their argument ''does not imply that large studies are never ethical or that small studies are better, only that a small study is ethically acceptable whenever a larger one is'' (p. 113).

Analysis by Bacchetti et al. (2005a; 2005b) addresses only ethical acceptability, not optimality; large studies may be desirable for other than ethical reasons. The balance point between burden and value cannot be precisely calculated in most situations because both the projected participant burden and the study's projected value are difficult to quantify, particularly on comparable scales. Bacchetti et al. (2005a; 2005b) provided a service by encouraging researchers to think of value and burden on a per-participant basis and by arguing that the expected net burden per participant may often be independent of sample size (Prentice, 2005). Institutional review boards are becoming more sophisticated regarding power and sample size issues, and, consequently, there could be fewer studies with inappropriate (too large or too small) sample sizes in the future. If random sampling is possible, social work researchers should consider testing a one-sided hypothesis to minimize sample size.

3.1 One- Versus Two-Sided Hypotheses as Determinants of Sample Size

The estimation of the minimum sample size requires the specification of the minimal difference in outcome (effect size or δ) that would be practically important to be detected. In addition, researchers must specify (1) an acceptable α-level ($1 - \alpha$ is the probability of detecting a significant difference when the treatments are really equally effective), (2) an acceptable β-level ($1 - \beta$ is the probability of not detecting a significant difference when there really is a difference of magnitude δ or larger), and (3) the

standard deviation of the hypothesized effect size in the population of interest. Finally, a researcher should explicitly choose between two-sided or one-sided statistical testing. As Knottnerus and Bouter (2001) suggest, the importance of this last criteria decision often is neglected. A one-tailed hypothesis specifies a directional relationship between groups. That, is the researcher not only states that there will be differences between the groups but specifies in which direction the differences will exist. Anytime a relationship is expected to be directional (i.e., to go one specific way) a one-tailed hypothesis is being used. This is the opposite of a two-tailed hypothesis. With a two tailed hypothesis the researcher would predict that there was a difference between groups, but would make no reference to the direction of the effect (Bland & Altman, 1994).

Knottnerus and Bouter (2001) argue that a research hypothesis expresses scientific uncertainty regarding a plausible, potentially practically important effect. Consequently, a research question is often hypothesis-driven and typically "one-sided." Accordingly, if a new intervention is compared with no treatment, the one-sided approach would be adequate. Moreover, for example, assuming $\alpha = 0.05$, $\beta = 0.80$, a moderate effect size of Cohen's *d*, and equal numbers in the intervention and treatment groups of a study, each group needs a minimum sample size of 88 in case of one-sided testing; and 105 per group in case of two-sided testing. This means that the two-sided approach requires an additional 34 or 19% more participants than the one-sided approach.

As Moye and Tita (2002) explain, however, there are important limitations to the one-tailed test in a clinical research effort. A major difficulty is that the one-sided testing philosophy reveals a potentially dangerous level of investigator consensus that there is no possibility of participant harm produced by the intervention being tested. Although the two-sided hypothesis test can complicate experimental design, increasing sample size requirements, this approach is ultimately more informative and potentially prevents subsequent exposure of research participants and the general population to harmful interventions. Although two-sided tests are only capable of establishing a difference (rather than a difference and direction), researchers may explore their data and determine in which direction any significant difference lies. In fact, researchers should routinely report a confidence interval around an outcome measure, such as a mean.

When deciding whether a one- or two-sided hypothesis approach is most appropriate for a study that they are planning, researchers may consider prior evidence and the practice implications of the intervention being studied (Enkin, 1994). This body of existing studies, including meta-analyses may provide evidence in support of these methodological choices. For interventions not previously studied or about which few studies have been conducted, a one-sided view seems reasonable if the comparison is between that intervention and no intervention. Regardless of which hypothesis testing approach is selected, the researcher should formulate the research hypothesis *a priori*. If random sampling is possible, another strategy to minimize sample size is sequential sampling (Dunnett & Gent, 1996; Posch & Bauer, 2000; Whitehead, 1997).

3.2 Sequential Sampling

Sampling strategies, whether probability or non-probability, can be categorized as either single, (also termed fixed), or multiple, (also termed sequential) (Stephens, 2001). With a sequential sampling strategy, after a first sample is tested, there are three possibilities: accept, reject, or make no decision about a hypothesis. If no decision is made, additional samples are collected and each sample is analyzed to determine whether to accept or reject a hypothesis, or to proceed and collect another sample. More specifically, in a sequential sampling design, data are analyzed periodically, and sample size is not a single fixed number. An appropriate schedule for interim analyses is established together with a stopping rule, which defines the outcomes that lead to early termination of a study.

The classical theory of hypothesis testing is based on a sample of fixed size (Neyman & Pearson, 1928). In this sample, the null hypothesis H_0 is tested against an alternative hypothesis H_1. A significance level α is defined a priori (i.e., in advance of data collection), which is the probability of the null hypothesis being falsely rejected. Consequently, in a classical fixed sample design, the sample size is set in advance of data collection, and hypothesis testing occurs after all observations have been made. The main design focus is on choosing a sample size that allows a study to discriminate between H_0 and $H_1$and answer the research questions of interest.

In fixed sample design, then, together with practical considerations, a study's sample size is determined a priori by setting up null and alternate hypotheses concerning a primary parameter of interest (θ), and then specifying a Type I error rate (α) and power (1-B) to be controlled at a given treatment effect size ($\theta = \Delta$). Usually, traditional values of α and B are used (i.e.., $\alpha = .05$, $B = .20$); however, there can be considerable debate over the choice of the effect size (Δ). In general, the smaller the effect size, the larger the sample size needed to

detect it. The choice of Δ is crucial because, for example, reducing a selected effect size by 50% leads to a quadrupling in the sample size for a fixed sample. Using a sample size that is small relative to a selected effect size can result in a study that is underpowered (i.e., unlikely to detect a smaller, but possibly still important, effect). Consequently, Cohen (1988) and others (cf. Adcock, 1997; Orme & Hudson, 1995; Stolzenberg & Relles, 1997) have proposed the use of a sample big enough to detect the smallest worthwhile effect. A disadvantage of all fixed sample designs is that estimated sample size is the same regardless of the magnitude of the true intervention effect. Accordingly, one approach to increasing the congruence between estimated and true effect sizes is to perform interim analyses with sequential sampling.

With a sequential sampling strategy, after a first sample is tested, there are three possibilities: accept, reject, or make no decision about a hypothesis. If no decision is made, additional samples are collected and each sample is analyzed to determine whether to accept or reject a hypothesis or to proceed and collect another sample (Jennison & Turnbull, 2000). More specifically, in a sequential sampling design, data are analyzed periodically, and sample size is not a single fixed number. An appropriate schedule for interim analyses is defined together with a stopping rule, which defines the outcomes that lead to early termination of the study. For example, sequential sampling allows consecutive testing, with possible rejection of the null hypothesis, after each set of observations in a pair of groups (e.g., intervention and control).

With sequential sampling, for ethical and practical reasons, results can be monitored periodically and, if sufficiently large or small effects are observed, data collect may be stopped early. Evidence suggests that sequential designs require fewer participants than fixed sampling designs (Jennison & Turnbull, 2000; Whitehead, 1997). Tests of sequential samples have been developed that allow for early stopping to reject or accept the null hypothesis while preserving the integrity of the test; that is, maintain desired Type I error and power.

Sequential sampling design parameters include (1) power; (2) sample size; (3) number and timing of analyses; (4) criteria for early stopping (i.e., evidence against the null hypothesis, the alternative hypothesis, or both); and (5) stopping rules (i.e., the relative ease or conservatism with which a study will be terminated at the earliest analysis versus later analyses). A sequential sampling plan consists of two or more stopping rules. Data are monitored at interim time-points and the process is terminated early if, for example, a difference between two interventions in terms of an outcome can be established statistically at any one of the interim looks. Since the data will be tested repeatedly in a group-sequential study, the burden of proof must be more stringent at each of the interim looks than without interim monitoring. Otherwise, there is an increased that chance fluctuations in the data will be misinterpreted as demonstrating a real underlying effect. This increasing stringency is accomplished by establishing a stopping boundary at each interim look (Pampallona & Tsiatis, 1994; Proshan & Hunsberger, 1995).

In summary, sample size estimation is a key component of empirical research. A sequential sampling strategy may be most useful when appropriate effect sizes and estimates of variability necessary for sample size calculations are not known. In addition to saving time and resources, sequential sampling can reduce study participants' exposure to an inferior intervention. Sequential sampling also may be useful when conducting a pilot study. Sequential sampling helps determine whether the researcher has taken a large enough pilot sample to properly evaluate different sampling designs, and to use the standard deviation from the pilot sample to calculate sample size for a larger scale study.

A limitation of group-sequential sampling is an increased probability of Type I error because of repeated significance testing. Unadjusted, repeated significance testing of the accumulating data increases the overall significance level beyond the pre-specified nominal significance level. Consequently, interim analyses and their interpretations need to be done judiciously. To reduce the probability of Type I error, a study's protocol should contain a formal rule for stopping the study early. The decision to conduct an interim analysis should be based on sound scientific reasoning. Researchers should avoid the use of vaguely defined and misunderstood terms and phrases such as "administrative looks," "administrative interim analyses," "interim analysis for safety, and "interim analysis for sample size adjustment" (Sankoh, 1999). If random sampling is possible, a third strategy that social work researchers should consider using is planned missing data designs to minimize measurement instrument length, and consequently, costs to both participants and investigators (Graham et al., 2006).

3.3 Planned Missing Data Designs

When researchers design measurement instruments for a study, they universally must balance a desire to seek information from participants against participants' costs of providing this

information. Graham et al. (2006) described a planned missingness design called two-method measurement (also see Allison & Hauser, 1991, who describe a related design). For example, social work researchers typically (1) collect demographic and other background information data; and (2) and administer at least one standardized scale of moderate length. The two-method measurement design may allow the researcher to collect complete demographic data, and partial data (on a random sample of participants) for the standardized scale(s). A possible limitation of this type of design is that it requires the use of structural equation modeling (Muthe´n, Kaplan, & Hollis, 1987).

Another, design described by Graham et al. (2006) for maximizing information while minimizing participant costs is the three-form design. In its generic form, the three-form design allows researchers to increase by 33% the number of questions for which data are collected without changing the number of questions asked of each participant by dividing all questions asked into four items sets. One set (X) contains questions most central to the study outcomes, and is asked of all participants. Three additional sets of questions (A, B, C) are constructed, with each set containing one-third of the remaining question. Sets A, B, and C are rotated, such that one set is omitted from each of the three forms (i.e., X and two of the A, B, C sets).

An advantage of planned missing data designs is that less data are required, and therefore, less data needs to be collected. However, it is not clear how patterns of incomplete data should be structured and incorporated into research designs, particularly in longitudinal designs. In cross-sectional research designs, the use of a reference variable has been shown to be effective in terms of obtaining the correct estimates in the context of planned incomplete data structures (McArdle, 1994; Graham, Hofer, & MacKinnon, 1996). In such designs, a reference variable refers to obtaining complete data for a given variable, or for one variable within each factor of the research design. A reference variable is used as a baseline measure, which should aid the imputation process since full information is provided for all participants in relation to other variables being studied. The use of a reference variable also would seem to be an attractive option in some longitudinal research designs, since it could be incorporated across administrations. The efficacy of this design was tested by Bunting and Adamson (2000) through a series of simulations, and the result suggested that parameter estimates are both precise and efficient.

4.0 Conclusions

The need to recognize and address ethical dilemmas at all stages of the social work research process is widely recognized. This paper has focused on sampling as a nexus of ethical dilemmas. It has been argued that social workers and other applied researchers have an ethical obligation to construct the smallest representative samples possible. Random sampling is considered by many researchers as the gold standard methodological procedure for maximizing external validity and optimizing sample size. However, in practice, random sampling often is difficult to implement. Although assembling the smallest representative sample possible may seem daunting at times, recommendations include (1) deliberate sampling to balance a sample's composition in terms of typicalness and diversity; (2) randomization tests; (3) a broader perspective on external validity; (4) use of one-sided hypotheses; (5) sequential sampling; and (6) planned missing data designs. Moreover, fulfilling the aforementioned ethical obligation to construct the smallest representative samples possible usually will benefit from a mix of strategies to maximize external validity and minimize sample size.

References

Adcock, C. J. (1997). Sample size determination: A review. *Statistician, 46* (2), 261–283.

Allison P. D., & Hauser, R. M. (1991). Reducing bias in estimates of linear models by remeasurement of a random subsample. *Sociological Methods & Research. 19*, 466–492.

Armitage, P. (1947). A comparison of stratified with unrestricted random sampling from a finite population. *Biometrika*, *34*, 273-280.

Audi, R. (2003). *Epistemology: A contemporary introduction to the theory of knowledge.* New York: Rutledge.

Bacchetti, P., Wolf, L. E., Segal, M. R., & McCulloch, C. E. (2005a). Ethics and sample size. *American Journal of Epidemiology*, *161*, 105–110.

Bacchetti, P., Wolf, L. E., Segal, M. R., & McCulloch, C. E. (2005b). Bacchetti, et al. respond to ‘‘Ethics and sample size—Another view.’’ *American Journal of Epidemiology*, *16*, 113.

Bland, J. M., & Altman, D. G. (1994). One and two sided tests of significance. *British Medical Journal*, *309*, 248.

Blankertz, L. (1998). The value and practicality of deliberate sampling for heterogeneity. *American Journal of Evaluation*, *19*, 307-324.

Bull, B. (2005). Exemplar sampling: Nonrandom methods of selecting a sample which characterizes a finite multivariate population. *The American Statistician, 59*, 166-172.

Bunting, B. P., & Adamson, G. (2000). Assessing reliability and validity in the context of planned incomplete data structures for multitrait-multimethod models. *Developments in Survey Methodolgy*. Available online http://mrvar.fdv.unilj.si/pub/mz/mz15/bunting.pdf

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them* (pp. 39-83). San Francisco: Jossey-Bass.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation – Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Dunnett C., & Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine, 15*, 1729–1738.

Edgington, E. S. (2007). *Randomization tests*. New York: Marcel Dekker.

Engel, R., & Schutt, R. K. (2009). *The Practice of Research in Social Work.* Thousand Oaks, CA: Sage.

English, B. (1997). Conducting ethical evaluations with disadvantaged and minority target groups. *Evaluation Practice, 18*, 49-54.

Enkin, M. W. (1994). One and two-sided tests of significance: One-sided tests should be used more often. *British Medical Journal, 309*, 873-874.

Erceg-Hurn, D.M., & Mirosevich, V.M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*, 591-601.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323–343.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*, 197-218.

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Howell, D. (2007). *Randomization tests on correlation coefficients*. Available online at http://www.uvm.edu/~dhowell/StatPages/Resampling/RandomCorr/randomization_Correlation.html.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 6*, 227-240.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall.

Knottnerus, J.A., & Bouter L.M. (2001). The ethics of sample size: Two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology, 54*, 109-110.

Kott, P. S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika, 73*, 485-491.

Lee, A. S., & Baskerville, R. L. (2003). Generalizing generalizability in information systems research. *Information Systems Research, 14*, 221–243.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Mutivariate Behavioral Research, 29*, 409–54.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379–387.

Moye, L. A., & Tita, A. T. N. (2002). Defending the rationale for the two-tailed test in clinical research. *Circulation, 105*, 3062-3065.

Muth´en, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431–462.

NASW. (2008). The NASW code of ethics. Available online at http://www.socialworkers.org/pubs/code/code.asp

Nelson, J. C. (1994). Ethics, gender, and ethnicity in single-case research and evaluation. *Journal of Social Service Research, 18*, 139-151.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for

purposes of statistical inference. Part I and II. Biometrika, *20*, 174-240, 263-294.
Orme, J. G., & Hudson, W. W. (1995). The problem of sample size estimation: Confidence intervals. *Social Work Research*, *19*, 121–127.
Pampallona, S., & Tsiatis A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, *42*, 19-35.
Posch, M., & Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics*, *56*, 1170–1176.
Prentice, R. (2005). Invited commentary: Ethics and sample size—Another view. American *Journal of Epidemiology*, *161*, 111–112.
Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, *51*, 1315–1324.
Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, *34*, 441-456.
Rosenberg, A. (1993). Hume and the philosophy of science. In D. Norton (Ed.), *The Cambridge companion to Hume* (pp. 64-89). New York: Cambridge University Press.
Rubin, A., & Babbie, E. R. (2008). *Research methods for social work*. Belmont, CA: Brooks/Cole.
Sankoh, A. J. (1999). Interim analyses: An update of an FDA reviewer's experience and perspective. *Drug Information Journal*, *33*, 165-176.
Stephens, K. M. (2001). *The handbook of applied acceptance sampling*. Milwaukee, Wisconsin: ASQ Quality Press.
Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, *62*, 494–507.
Thompson, B. (1997). Statistical significance testing practices in *Journal of Experimental Education*, *66*, 75-83.
Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. New York: John Wiley.
Wood, A. (2000). Hume: The problem of induction. Available online at http://www.stanford.edu/~allenw/Phil102/Hume%20-%20Induction.doc