Chapter 3 - Designing research

David Shipworth
d.shipworth@ucl.ac.uk
University College London
Energy Institute
14 Upper Woburn Place
London
WC1H 0NN, UK

Gesche M. Huebner
g.huebner@ucl.ac.uk
University College London
Energy Institute
14 Upper Woburn Place
London
WC1H 0NN, UK

Abstract. The aim of this chapter is to set out a process that researchers can follow to design a robust quantitative research study of occupant behavior in buildings. Central to this approach is an emphasis on intellectual clarity around what is being measured and why. To help achieve this clarity, researchers are encouraged to literally draw these relationships out in the form of a concept map capturing the theoretical model of the cause and effect between occupant motivations and energy use. Having captured diagrammatically how the system is thought to work, the next step is to formulate research questions or hypotheses capturing the relationship between variables in the theoretical model, and to start to augment the diagram with the measurands (things that can actually be measured) that are good proxies for each *concept*. Once these are identified, the diagram can be further augmented with one or more methods of measuring each measurand. The chapter argues that it is necessary to carefully define concepts and their presumed relationships, and to clearly state research questions and identify what the researcher intends to measure before starting data collection. The chapter also explains the ideas of *reliability*, *validity*, and uncertainty, and why knowledge about them is essential for any researcher.

1. Introduction

The aim of this chapter is to set out a process that researchers can follow to design a robust quantitative research study on occupant behavior in buildings. The material is introductory, and is intended to provide an overarching framework for thinking about the research design process. It is not sufficient in itself, but refers to other chapters in this book and to other more detailed sources of information on specific elements. Whilst this chapter is necessarily highly abridged and incomplete in many areas, it should steer the reader away from some of the main errors and misunderstandings to which the field as a whole is prone.

It is important to note that this chapter takes a broadly quantitative social and physical realist approach to researching occupant influences on energy demand in buildings. This arises from this book's origin—to improve the representation of building occupants within building energy simulation models, which are themselves quantitative and realist in their representation of the world. The aim is therefore to establish relationships (ideally causative ones) between the external environment, the building and its internal environment, occupant behavior, and building energy consumption.

Taking a realist approach means that there are occupants' actions that directly affect energy demand in buildings, and that these actions can be explained in part through the use of concepts that are independent of the researcher and the individual occupants themselves. Concepts are central to research design and will be discussed throughout this chapter. A concept can be thought of as an abstract idea that captures the central elements of what it refers to. Examples of concepts include temperature, comfort, glare, environmental attitudes, financial costs, etc. A realist approach takes the view that while these concepts can never be measured perfectly (i.e., without any error) they can be measured and used to—imperfectly and incompletely—predict occupant behavior. The chapter therefore does not take a solely social constructivist approach of saying that occupant influences on energy use in buildings are purely a *construct* of human social processes with no meaning or existence outside of the individual occupants engaged in them. It thus places this work more in the context of such academic disciplines as physics, psychology, quantitative social science, and behavioral economics, and less in such academic disciplines as qualitative sociology, anthropology, and ethnography.

Central to the approach taken here is an emphasis on intellectual clarity around what is being measured and why, and literally drawing this out in the form of a theoretical model of the cause and effect relationships between occupant motivations and energy use in the form of a concept map. Having captured diagrammatically how the system is thought to work, the next step is to formulate research questions or hypotheses capturing the relationship between variables in the theoretical model, and to start to augment the diagram with the *measurands* (things that can actually be measured) that are good proxies for each concept. Once these are identified, the diagram can be further augmented with one or more methods of measuring each measurand. In research adopting a realist approach, be it qualitative or quantitative, it is necessary to carefully define concepts and their presumed relationships, and to clearly state research questions and what the researcher intends to measure before starting data collection. Some higher forms of analysis are only applicable to quantitative research approaches, such as quantifying reliability;

however, an awareness of the ideas of reliability, validity, and uncertainty is essential for any researcher.

This chapter is a synthesis of descriptions of the research and model building approaches used in both the physical and social sciences. Casti (1992) defined a mathematical model of a system as "...the specification of observables describing such a system and a characterization of the manner in which these observables are linked" (p. 2). This is a useful starting point for a discussion on the process of designing a program of research to understand how occupant behavior influences energy demand in buildings. This definition has two key elements that relate to designing research: firstly, specifying what the measurement will be, the observables (how they are measured is the realm of research methods); and secondly, determining if the variables are causally related (this is one of the functions of research design). While Casti's definition is a useful starting point, it is insufficient. Another essential element is theory. As noted by Ruttkamp (2002), "The only way in which we can have scientific contact with the world…is through actions involving selection, abstraction, and generalization, which are always executed within some theoretical framework or disciplinary matrix…." (p. 17). This third element, theory, allows making sense of the observables, and the relationships (links) between them, within an explanatory (i.e., theoretical) framework that permits transferring these insights between instances. These three elements, methods, research design, and theory, need to be brought together in order to design any program of research.

2. Why do the research (research aims and questions)

Determining a good research aim or research question is an essential and often neglected first step in the research process. Bouma (2000) distinguishes between when research *aims* are appropriate, and when research *questions* and/or hypotheses are. Where research is exploratory or descriptive, then a research aim is appropriate. When research is more explanatory or seeking to establish causation, then research questions and hypotheses are appropriate. In the context of occupant behavior in buildings, both forms of research are common, although descriptive research predominates. Unlike a research question, which specifies relationships between two or more concepts, a research aim describes a more general area of enquiry, and leaves open greater scope for exploratory data analysis through looking for patterns between different elements of the data collected. It needs to be remembered, however, that such descriptive or exploratory work can only be used to describe correlations between the concepts measured. If establishment of causation is desired, subsequent, more experimentally based work needs to be conducted.

Examples of research questions in the context of occupant behavior:

"Do occupants open windows more frequently as CO_2 levels rise?" or "Do occupants tilt blinds when sun shines directly on their computer monitor?"

As Bouma (2000) notes, a good research question postulates relationships between two concepts and facilitates the process of designing a research study to answer that question. Two separate aspects need to be addressed. The first aspect relates to the things that are measured (for example, CO_2 levels and window opening). These concepts need to be operationalized in ways that allow measuring them—for example, in the case of CO_2 and window opening, using appropriate sensors or through observations. The second aspect relates to the nature of the relationship between the concepts. The capacity to determine whether the relationship is correlational or causal is determined by the choice of research design. This is discussed in detail below.

A well-framed research question makes the construction of hypotheses far easier. Hypotheses are appropriate in cases where a quantified measure of confidence in the answer is desired, and take the form of a statement (a declarative sentence) of what the researcher expects to happen.

Possible research questions and resulting hypotheses: A research question such as, "Do occupants open windows more frequently as CO_2 levels rise?" may give rise to a range of hypotheses such as, "As CO_2 levels rise (hypothesized cause) occupants will open windows more frequently (hypothesized effect)"; "As CO_2 levels rise occupants will open windows for longer periods of time"; and/or "As CO_2 levels rise occupants will open windows wider".

It is typical for one research question to give rise to many hypotheses, as hypotheses need to be sufficiently specific to be measurable without ambiguity. This usually takes the form of a measure of *statistical confidence* between the data gathered and the theoretical model of occupant behavior and energy outcomes being explored. In such hypotheses testing, it is usually a measure of the lack of fit between the measured data and the inverse of the hypotheses—the null hypothesis—that is used. The null hypothesis is the embodiment of scientific skepticism; it assumes that there is no relationship between the things being measured, here CO_2 levels and window opening, and only accepts that there is one if there is enough evidence to reject this conservative assumption. Many introductory textbooks have been written about statistical hypotheses testing and with this has come a degree of standardization of key parameters, such as the levels of confidence needed (frequently a *p-value* of 0.05, i.e., 95% confidence, is cited). Statistical confidence is a measure of how confident one is that the study can correctly determine if an intervention failed to work. To be 95% confident means that there is only a one in twenty chance of a false positive finding, i.e., saying the intervention worked when in fact it did not. This is also called a *type I* error. Statistical *power* is the converse of this. It is a measure of how confident one is that the study can correctly determine if an intervention worked. To have 80% statistical power means that there is only a one in five chance of a false negative finding, i.e., saying the intervention did not work when in fact it did. This is also called a '*type II*' error. In the energy in buildings area both of these are important. Neither incurring the costs of energy savings measures that are ineffective (type I error), nor discarding interventions that work (type II error) would be a good outcome.

For research on the influences of occupant behavior on energy demand in buildings it is important to realize that such high levels of statistical confidence may or may not be appropriate, and the reader is referred to the recent pronouncements by the American Statistical Association (Wasserstein and Lazar 2016) for a more rounded discussion of this topic.

Example of the differing effect of statistical confidence: A 1:20 (5%) chance of having incorrectly counted occupants leaving a building in the event of a fire is inappropriately low (in a building of 100 occupants this could leave 5 trapped inside)—whereas requiring only a 1:20 (5%) chance of incorrectly identifying the number of people in a room for the purposes of estimating fresh air volumes is inappropriately high (only an approximate estimate of occupancy is needed to adjust fresh air volumes appropriately). In each case, the appropriate levels of statistical confidence and statistical power need to be assessed against the risks of making each type of error and the costs involved in reducing them.

3. Identifying the concepts to measure and how they link together (theory)

One of the most important elements in the process of identifying concepts to measure and how they link together is the drawing out of a theoretical model. This step should be undertaken both when doing more descriptive work based on exploratory data analysis and when seeking to understand cause and effect using research questions and hypotheses. It should take the form of a diagram of concepts and links showing how they are related. There are many software packages in which such a theoretical model can be drawn, where one of the most useful is Cmap, a free, dedicated concept mapping software package (Novak and Cañas

2006). The advantage of using concept mapping software is that it allows the labeling of concepts, as well as links between concepts, thus creating a map of how different factors interact. This theoretical model can either be one that reflects a *Theory* (i.e., an established theory that is to be tested in a specific context like the Theory of Planned Behavior (Ajzen 1991)) or a *theory* (i.e., the researcher's own mental model of how occupant behavior influences energy use in buildings).

In constructing this map, it is important that as many causal steps as possible be included. For example, the model might link occupant thermal comfort to home energy demand. This could be as simple as: occupant (cold) thermal discomfort ——(leads to)——> occupants turning up the thermostat ——(leads to)——> greater energy use. This is intuitively reasonable, but it makes a lot of assumptions: how occupants will respond to cold thermal discomfort (through adjusting the thermostat); the home (that changing temperatures at the thermostat changes the temperature where the occupant is); the thermostat (that it is connected and working); and the boiler and heating system (that it can deliver the heat output necessary to raise the temperature where the occupant is). It is important that as many of these assumptions and causal links as possible be expanded in the theoretical model to allow the researchers to decide what to measure along the causal chain and to understand if they do not find a relationship between their primary variables of interest (say, thermal comfort and home energy use) that they are aware that the breakdown in the causal chain can be anywhere along it, and it is not just that occupants do not act as expected.

Ideally, the model would go from occupant motivations through to energy use. This roots the model in psychological, social, or physiological drivers, and explains how these are translated through occupant behavior and interaction with (or in reaction to) elements of the building to changes in energy and power use. Such rooting of the causal model in occupant motivations helps in identifying potential points of intervention with occupants to change how they respond to (or interact with) the building, while the modelling of the building's response to this interaction allows testing of whether the assumptions about the building controls and physics are as imagined.

An example of such a theoretical model represented in the Cmap software is provided in Figure 3.1. This is based on the Theory of Planned Behaviorwhich is one of the most well-known and tested theories in psychology to understand the antecedents of behavior. It postulates that the attitude toward a behavior, subjective norms, and perceived behavioral control shape an individual's behavioral intentions and, ultimately, their behavior.

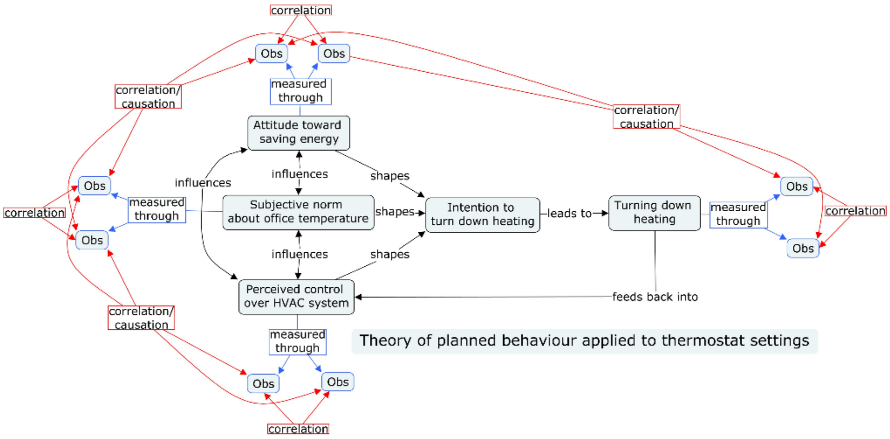


Figure 3.1 Graphical representation of the Theory of Planned Behavior. Black boxes and links represent the established theory. Blue boxes and links represent measurable properties. Red boxes and links represent analytical relationships for testing validity.

3.1. Concepts

In both the physical and social sciences, concepts play the important role of being the thing that researchers are frequently trying to measure. In the area of occupant influences on energy use in buildings, concepts range from social norms of behavior, through thermal/aural/visual comfort, to temperature/sound pressure levels/illuminance levels, to building management systems and heating, ventilation, and air conditioning (HVAC) systems, and to energy, power and carbon emissions. It may seem alien to link together things as seemingly disparate and diffuse as social norms with things as apparently concrete as temperature and energy—but this is only because the latter have been reified (i.e., made concrete through an agreed process of measurement) and their methods of definition and measurement so widely accepted that it has been forgotten that they were once as ill-defined and vague.

Put simply, concepts are those things researchers are usually *trying* to measure. They are not, however, usually the things that are actually measured because it is usually only possible to measure proxies to concepts. This is why concepts and variables are not the same thing. This will be discussed further in Section 6. The theoretical model should have concepts as its nodes, with such concepts connected by a series of links indicating the relationships between the concepts.

3.2. One to one relationships (links)

In the theoretical model, links describe how it is thought that concepts influence each other. While not necessary, it is often useful to attach signs to these links indicating whether the relationship is thought to be positive (+), negative (-), or unknown/variable (?). Doing this makes constructing research questions and hypotheses simple, as they are then just verbal descriptions of the relationships between concepts in the model. In the example discussed above regarding window opening behavior, the theoretical model may have a link between CO_2 levels and window opening. This could be translated into the research question, “Do occupants open windows more frequently as CO_2 levels rise?” If the theoretical model labeled the relationship with a +, this would give rise to the hypotheses: “As CO_2 levels rise (hypothesized cause) occupants will open windows more frequently (hypothesized effect)” and “As CO_2 levels rise occupants will open windows wider”. Thus, the sign on the link in the theoretical model indicates the expected relationship between the concepts.

It is important to note that there can be many more links in a model than nodes (concepts), as each node can have links to many others. That said, it is important not to end up linking each node to every other node, as that conveys little information—it merely says, “everything is connected to everything else”. Believing that each node should be connected to every other node can arise from two issues. Firstly, this can arise through including temporal relations (i.e., feedback loops) in the model. One variable can influence another in the short term, but the second variable can then influence the first, either directly, or indirectly, at a different timescale. It is often important to define the time scale of interest and exclude feedback processes that occur over longer or shorter periods. The second reason “everything is connected to everything else” models arise is because the concepts used are not defined precisely enough--this is a question of scope, and it may be that for the purposes of a study intermediary steps are out of scope.

3.3. One to many relationships (hierarchies)

Another major form of relationship to be aware of when constructing the theoretical model is hierarchies. This is where concepts have a natural nested structure. When studying occupants in buildings such hierarchies are rife—people within offices, offices within premises, premises within buildings, buildings within companies, etc. Identification of these relationships is important, as it will influence the unit of analysis (the entity on which data are collected) and the definition of the target population (the group from which the sample is drawn), and will inform the sort of analysis run on the data. If, for instance, the assumption is that the actions of occupants are strongly shaped by the building they are in, then it would not make sense to draw a sample of 1,000 people from only four buildings and think they constitute a representative sample of the population. One thousand people is usually sufficient to be a statistically representative sample of a large population, but only where those people are independent and sampled at random from the whole population. In this case the sample consists just of four occupant-building combinations. If assuming that buildings do not influence occupant actions then such a sample of 1,000 is fine, but if assuming that buildings strongly shape occupant actions then the population should be of buildings, and a sample size of four buildings is very inadequate. This illustrates why having a representation of hierarchy in the theoretical model matters for the research design.

4. Units of analysis, populations, and scope

Having established the theoretical model, the next step is to delineate the scope of its applicability. This will require a clear statement of population of interest, i.e., the population of units of analysis the theoretical model is supposed to represent. This is likely to be as constrained by the resources available for the study as by what the researcher would theoretically like to represent. The geographical scope of applicability, along with the temporal scope need to be defined, as well—both will tell those using the study where and when the results are no longer applicable. Finally, the required degree of *precision* needs to be decided, which will determine the sampling strategy and the sample size required. Each of these concepts is discussed in turn in the sections that follow.

4.1. Units of analysis

The unit of analysis is the thing that data are collected about. In the context of occupant behavioral impacts on energy demand in buildings this can be quite a range of units: from companies, through campuses, to buildings, premises, floors, individual offices, down to individual occupants. For domestic buildings, the unit of analysis may be homes, rooms, or individual occupants. The challenge in this area is that there are strong hierarchical relationships between these levels, and so the behavior of the same individuals in different buildings may vary more than the behavior of different individuals in the same building. Where this is the case, then it is probably more appropriate to think of the building as being the unit of analysis.

4.2. Population of interest and scope

Where the building is the unit of analysis, then those characteristics of the building that shape occupants' influence on energy demand help define the population of buildings that the findings of the study apply to. For example, if occupant behavior in naturally ventilated buildings with high thermal mass is studied, the population may be these buildings, and thus the sampling strategy needs to sample from a population of such buildings to generate generalizable results. The limits of *where* (geographical scope) and *when* (temporal scope) the findings would apply need to be defined. The geographical scope could be determined by external conditions ranging from climate regions to the extent of external pollution (a factor influencing window opening behaviors). Hence, the findings may be restricted just to naturally ventilated buildings in temperate climates with low levels of external pollution. There may also be temporal limitations, either seasonal (results only applying in spring, summer and autumn) or in terms of a specific longevity (results only applying for the next decade due to expected changes in technology or society).

4.3. Descriptive or inferential statistics

The overwhelming majority of statistical work in the buildings field is classed as *descriptive statistics*. Descriptive statistics report on the statistical characteristics of the data gathered. If the study is of 100 buildings, then descriptive statistics describe those 100 buildings. Examples are reports on frequencies (e.g., counting how many double-glazed windows are present) or correlations between variables (e.g., windows opened for longer when ambient temperatures increase). The common element being that the findings only relate to the units of analysis studied, and nothing can be said about whether the findings apply more generally.

Inferential statistics, on the other hand, seek to make statements about things that have not been studied directly. Inferential statistics, also called inductive statistics, describe the statistical characteristics of similar unobserved buildings, such as the population from which the units of analysis (e.g., occupants/buildings) were drawn. To be able to say something about a population by studying a sample it is necessary to know how well the sample represents the population. This is the field of sampling and sample size calculations. Whilst not specific to experimental research, sampling and sample size calculations are crucial in any experimental research design.

To recap, inferential statistics is the method to make inferences from the collected data to more general conditions. It is what is commonly described when using statistical measures such as *confidence intervals* and p-values for research findings.

Confidence intervals are a function of the sampling error (also known as "*standard error*") and depend on the size of the sample—the bigger the sample the smaller the sampling error. Confidence intervals express the range of values within which the parameter of interest (e.g., the mean) of the population from which the sample is drawn can be said to fall, based on that same parameter in the sample (e.g., the sample mean).

Modern statistical software has many advantages, but one of its disadvantages is that it will provide answers to questions without first testing whether the assumptions on which those answers are based have been met.

For example, before calculating and reporting confidence intervals for findings can be meaningful, specific assumptions must hold. These include that the standard deviation of the population is known (not just that of the sample); that each member of sample was randomly and independently selected from the population; and that the sample is (or can be transformed to be) approximately normal. Where these assumptions do not hold, calculation of confidence intervals is still possible, but requires changing the default settings in most statistical software. For example, if the standard deviation of the population cannot be determined from other published statistics and if the sample size is small, then the Student's t- distribution can be used instead of the z-distribution. This widens the confidence intervals and helps account for the uncertainty arising from using the sample standard deviation rather than that of the population. In the context of built environment studies these assumptions are frequently violated and non-standard approaches are needed.

It is important to remember that confidence intervals only represent one particular aspect—and frequently a fairly minor aspect—of the uncertainty that is inherent in the research process. Issues such as instrument *accuracy* and precision (discussed above) are not captured in the calculation of confidence intervals. It is an unfortunate reflection on contemporary academia that to quantify is to reify; the capacity to quantify one element of uncertainty (sampling error) is somehow thought to make it more real than other forms of uncertainty which, while less easy to quantify, are no less real and frequently far more important.

Determination of sample sizes for inferential statistics in a building occupancy study is challenging because of the hierarchical structure of the problem as discussed above. In order to understand this, a brief recap on some of the fundamental concepts of statistics is required. Inferential statistics, of which sample size calculations are a part, is about making the statements about a population based on a measured subsample of that population. All calculations of sample sizes are predicated on the assumption that there is a well-defined population, and that an unbiased sample from that population can be selected through a random selection process in which each member of the population has an equal probability of being selected into the sample. In practice, this is virtually impossible to do, and so judgment is called for in assessing the extent to which the way with which units of analysis were selected into the sample may bias the outcome.

The aim of any research should be to match the underlying assumptions of the statistical methods used; thus, the researcher should seek to clearly define their population of interest, and, wherever possible, to draw members from that population with equal probability. It is common to see comparisons of the descriptive statistics of a sample (i.e., reporting on house type, household size, income, other demographics) compared to those of a nationally representative survey, like a census, with authors reporting that because the sample looks like the census (usually through visual comparison of histograms) that the sample is representative. While this provides some reassurance, it is not strictly speaking correct—particularly in energy in buildings work. Usually such demographic factors explain only a limited share of the observed *variance* between households' energy consumption, and so some measure of demographic similarity does not necessarily translate into similar patterns of energy consumption. It is also worth noting that reporting values such as confidence intervals is also not meaningful or necessary when all members of the population are surveyed (i.e., in a census).

The choice between descriptive and inferential statistics is an important one that will fundamentally shape the research and the conclusions that can be drawn. While most researchers would like their findings to apply more generally, the work involved in doing so is considerable and so the decision to do so should not be taken lightly.

4.4. Required precision

Of particular importance in this context is defining the precision with which the outcome variables need to be known for the findings to be relevant to the substantive problem being addressed. Precision, often called “reliability” in the social sciences, is a measure of how much spread there would be in the data if exactly the same thing were to be measured with the same instrument many times. It is different from accuracy, which is a measure of how well these measurements correspond with the true value. Most instruments have some level of imprecision (say, ± 1˚C on a thermistor), which puts a fundamental limit on how precisely a measurement can be specified.

Precision is important because most interventions in buildings will be subject to some form of cost benefit analysis, with the intervention implemented if it can be shown that the benefits outweigh the costs. In this context, it is important to know in advance the likely costs, thus providing a prior estimate of the size of the benefits (energy savings, indoor air quality improvements, etc.) required for the intervention to be deemed worthwhile.

Implications of precision: If the intervention is only expected to change, say, internal temperature by 1˚C, and temperature can only be measured to ± 1˚C, then it is unlikely to detect an effect of the intervention with that level of imprecision—a different instrument would need to be used (e.g., one that measures temperature to ± 0.1˚C).

Similarly, it is important to determine the statistical confidence required of the findings. This will vary with context. If the objective is to publish in refereed journals, then 95% statistical confidence is frequently expected. If the objective is to decide between two alternate courses of action incurring similar costs, then statistical confidence greater than 50% (i.e., on the balance of probabilities) may be all that is required, depending on the balance of risks associated with false positive (type I) and false negative (type II) errors for each option.

To recap, false positive (type I) errors occur where the intervention being trialed did not actually work, but the study concluded that it did. The risk here is of implementing an intervention that does not work, thus wasting time and money. The more worried one is about this, the higher the level of statistical confidence needed. False negative (type II) errors occur where the intervention being trialed actually did work, but the study concluded that it did not. The risk here is of throwing out a good idea and missing out on potential improvements to the building. The more worried one is about this, the higher the level of statistical power needed.

In general, the higher the precision with which the results need to be known, the more expensive the trial will be. High costs could arise from the need to measure things more precisely or the need to reduce the uncertainty in generalizing to the population of interest, which will require a larger sample of the chosen units of analysis.

5. Sampling and sample size

5.1. Sample frames

As discussed above, each study should specify the population to which the findings are thought to apply. Once this is specified, then if generalization from a sample to a population (inferential statistics) is to be used, a *sample frame* is needed from which to draw a sample. Factors identified in the theoretical model as influencing the outcome variable(s) of interest will need to be addressed (exemplified or nullified) in the construction of a sample frame. The sample frame is (ideally) a list of all units of analysis in the population. In some cases, depending on the unit of analysis, this may be difficult to obtain. Where such a list (sample frame) is available, then the sample is drawn from this list using the sampling strategy. Where such a list is not available, less statistically correct methods will need to be used such as quota sampling—for example, choosing a certain number of buildings in each of a range of categories that the theoretical model says will be important.

5.2. Sampling strategies

There is a wide range of sampling strategies. These broadly divide into probability-based methods, which are needed for generalizing from the sample to the population, and non-probability sampling methods, which are often used for pragmatic and costs reasons.

Of the probability-based methods, the "gold standard" is pure *random sampling*. This is the ideal case, as every member of the population (as represented in the sample frame) has an equal probability of being included in the sample. It would amount to drawing the sample purely randomly from the sample frame, all chosen units consenting to being monitored and then monitoring them all with no missing data. It needs to be stressed that all inferential statistics are based on the assumption that the sample is drawn at random from the population and any deviation from this is a compromise of this most basic assumption on which inferential statistics is based.

Because pure random sampling is often both very difficult and very expensive, a range of alternative methods have been developed that are still statistically generalizable. A full description of such methods is beyond the scope of this chapter; examples include systematic sampling (sampling every *n*th member of the sampling frame, but starting at a random point between 1 and *n*, so each member has an equal probability of being sampled); stratified random (where a random sample is drawn from different strata of interest, e.g., low-, medium-, and high-rise buildings, or urban, sub-urban, and rural buildings, but with the proportions of the population reflected in the strata of the sample); and cluster sampling (where groups of co-located members of the population are selected, e.g., ten buildings in each of five cities).

The best of non-probabilistic sampling methods is quota sampling, where a set of important criteria drawn from the theoretical model are identified and units of analysis selected on a first come first served basis until a quota is reached in each cell of the sample frame. For example, in a study of occupants and their adaptive responses to thermal comfort, Gauthier and Shipworth (2015) used a sample frame of age, weight and gender, and recruited people (her unit of analysis) to populate that frame.

Second most robust is purposive sampling, in which population members are recruited based on certain characteristics considered useful to the study. This may vary from deliberate selection of extreme cases to get a sense of the breadth of possible responses; to heterogeneous sampling, i.e., taking a spread of participants to cover the whole range of possible responses; to homogenous sampling in which some forms of variance are deliberately excluded through selection of a sample; to critical case, or typical case sampling. Other, less robust forms of sampling include snowball (where participants recommend others they know to participate); self-selection (the widely used practice of allowing people to volunteer, or opt-in to a trial); and convenience (where trial participants are based on whoever is to hand—hence the proliferation of studies of people and buildings on university campuses!). Each of these methods carries significant "health warnings" to the robustness of the trial, with all three methods having the potential to introduce significant biases into the results.

5.3. Spatial sampling

Spatial sampling varies from the geographic dispersal of research subjects with the population ranging from local to global, through to the spatial density of deployment of sensors collecting environmental variables in an occupied space. In both cases, the required density of sampling depends on the rate of change of the variable of interest in space and on the sensitivity of the other variables in the theoretical model being used for the research design to changes in those variables. In many instances, existing standards or established models will provide guidance on such sensitivities. For example, thermal comfort, as represented in the *predicted mean vote* (PMV) model, is far more sensitive to changes in ambient temperature than it is to changes in relative humidity. Thus, even if both ambient temperature and relative humidity were to change at equal rates in the space, it would not be necessary to sample relative humidity as frequently. Spatial sampling is conceptually similar to any other form of sampling (population or temporal), where the factors driving the size of the sample are the *effect size* the researcher is trying to measure (what magnitude of change is considered worthwhile detecting) and the variance in the space (how much different locations vary from each other). If measuring a variable that varies a great deal, or if the theoretical model is thought to be very sensitive to that variable, or if trying to detect a small change in the outcome variable of interest of the model, then a larger sample is needed.

Sample size calculators can be used to determine spatial sample sizes, but this is seldom done for a range of reasons. Firstly, spatial data are usually highly spatially auto-correlated, i.e., the value of a variable in two adjacent points in space is likely to be pretty similar. Secondly, usually there is good prior knowledge of how a variable is likely to change in space both inside and outside buildings—particularly environmental variables such as temperature and light levels. Thirdly, the units of analysis (people, buildings, etc.) are seldom randomly distributed within the geographic scope of the study. All of this, coupled with the expense and impracticality of monitoring a large number of physical locations, makes purposive sampling both more acceptable and more pragmatic. For most studies, the aim is to measure variables experienced by the unit of analysis; hence, environmental parameters are best measured where the units of analysis (e.g., people or buildings) are located. Doing this reduces the uncertainty that arises from having to estimate these values from data collected in another time and place. This is the basis of the so-called “right here right now” approach to gathering thermal comfort data. While sampling at the unit of analysis is ideal, there are often times when it is not practical, and instead sampling is done at fixed points in the environment. This could be by using secondary weather station data, or by monitoring values inside buildings at fixed heights and locations away from people.

Qualitative rules in determining a sensor strategy: firstly, it is important to estimate the accuracy and precision with which each variable needs to be known (see “required precision” above). Accuracy differs from precision in that it refers to any systemic bias in the readings. In physical monitoring an example would be a poorly calibrated sensor which is always reading above or below the “true” value. In psychology, it may arise from a psychological trait such as centrality bias (where people tend to avoid picking the end values of scales). A sensor located away from the unit of analysis may well record values that are consistently different from those at the unit of analysis. It is important to think through how large a difference is tolerable before the findings are no longer fit for purpose. Secondly, as discussed above, it is important to consider how much imprecision is acceptable. The greater the imprecision in the measurements, the less likely it is to find statistically significant results. Imprecision clouds data with noise, making the signal harder to detect. If trying to find a small signal (for instance, a weak influence of occupant behavior on energy use in buildings), then as much precision as possible is needed in the measurements. The final issue to consider is under-specification of the measurand. This is addressed in Section 6.2.

5.4. Temporal sampling

The principles of temporal sampling are similar to those of spatial sampling, except applied in the time dimension. Again, rate of change is the key determinant, along with the sensitivity of the variables of interest to that change, and the response-time of the system. It may not be necessary to frequently sample a variable that changes rapidly, where it is acting on a system that changes slowly or where the outcome variables of interest in the theoretical model are comparatively insensitive to that variable. Conversely, if the variable changes rapidly, and the system and its outcome variable of interest is responsive to that change, then sampling at high frequency may be required.

As with spatial variability, temporal variability can be highly auto-correlated, i.e., values of a variable sampled closely in time can be very similar. For this reason, temporal sampling rates will primarily be driven by the rate at which the variable is thought to change. An additional element to add to the concept map of the theoretical model is an a priori estimate of the rate of change of each variable to be measured. This can be based on previous studies or preliminary fieldwork/pilot studies. The second factor that determines the sampling rate is the characteristic timescale of change of the system. Nicol (2012) argues that there is no point in taking comfort votes from people at intervals of less than half an hour because for practical purposes their comfort state does not change sufficiently between such intervals to warrant it. While this may or may not be true, if the objective of the study and the theoretical model are consistent with this, then there would be little point measuring data at higher temporal frequencies unless assessing this claim was part of the objectives of the study.

While such rules of thumb can be used to determine regular temporal sampling rates for most studies, there are instances where different temporal sampling strategies are appropriate. This becomes particularly apparent in wireless sensor networks where minimizing energy use by sensors can be critical. Here, more sophisticated sampling rates can be used such as variance-based sampling. Such approaches vary the rate of sampling in proportion to the rate of change of the variable of interest. When the variable is static or changing slowly, then sampling can be quite infrequent. When the variable is changing rapidly, then the sensor can increase the rate of collecting and transmitting data to capture the additional information when it is useful. These sampling strategies are currently under development in computer science—those considering using them would need to liaise with their sensor developers to implement such strategies (see also Chapter 4). It is also necessary to determine the thresholds at which the sampling rates should change; this is frequently expressed as a change in the variable relative to the historical observed range of variance for each variable.

Other bases for determining sampling rates include matching or replicating other studies in the field to ensure comparability, sampling as frequently as battery/memory/financial constraints will allow (a conservative strategy given it is always possible to down-sample to lower frequencies, if desired), and adaptive designs in which sampling is initially high, but is reduced after preliminary data analysis if the rate is in excess of requirements.

5.5. Sample size calculations

One of the most frequently asked—and unfortunately most difficult to answer—questions in research design is, “How large should my sample be?”. This is important, because if no relationship (descriptive design) or causation (experimental design) is found, it could be for a range of reasons. Firstly, there could simply be no effect; secondly, it could be because variables were not measured precisely enough; and thirdly, it could be because the study was underpowered. An underpowered study is one in which too few participants have been tested to detect an effect with the desired level of statistical confidence and power.

Hence, in order to determine an adequate sample size, sample size estimations are essential. In the following sections, methods for calculating sample sizes will be discussed, as will concepts such as confidence intervals, p-values, types of statistical errors, and statistical power. Calculation of sample sizes is a significant research area in its own right, and one addressed extensively in the quantitative social sciences and psychology fields. For the purposes of this chapter, the focus is on two of the main areas for which sample sizes are calculated: *internal validity* and inferential statistics.

Internal validity refers to the extent to which the findings from the study can be correctly attributed to the interventions being experimentally tested. While it is not exclusively the case, people gathering data through surveys are frequently more concerned about questions of inferential statistics, while people conducting experiments are frequently more concerned about questions of internal validity. As these are all complicated topics in their own right, the reader is referred to standard texts in the field such as (Groves et al. 2004).

5.6. External validity

External validity is the assessment of the extent to which the findings from the sample can be considered to apply to similar, but not identical, units of analysis. These units of analysis can be considered as forming a range of similar but distinct populations that the findings can be said to hold for. These should not be confused with the general population (say all people or buildings in the country). These populations of similar units of analysis are defined by how closely the units of analysis are to those in the study. For example, a study of the thermal comfort of sixth grade school children may involve a sample of 200 students in ten schools. External validity arguments could be made that the same findings would hold for other students (e.g., the grades above or below) in those schools, or that they may even hold for students in other (similar) schools. Such arguments ultimately rest on qualitative arguments and citations of other studies' findings to support such claims. Citing confidence intervals and p-values for other (related) populations is inappropriate, as the argument for the external validity of these findings to these other groups is ultimately not a statistical one.

In this context, the above discussion about hierarchies, inferential statistics, external validity, units of analysis, and sample frames needs to be borne in mind. The sample frame needs to represent the population of the units of analysis, whether occupants in a building or buildings in sector of the building stock. Once having found or developed such a sample frame, a random subsample can be drawn of the size needed to achieve a certain level of statistical confidence (see note below on calculation of sample sizes). It is important to remember here that "random" is a well-defined term, and a suitable random number generator should be used to draw the sample. Then, the members of the chosen sample should be approached and recruited into the study. If not enough units of analysis (e.g., people or buildings) are willing to participate, it is not acceptable to simply draw more potential participants from the sample frame, as this simply serves to drive up the nonresponse rate (as discussed below). Whilst it is tempting to conduct "opt-in" trials, where volunteers are sought to participate in the project, this immediately violates the underlying assumption that each member of the population has an equal probability of being part of the trial, for by definition those who choose to participate are different from those who do not. The correct approach is to attempt to recruit all of those drawn at random from the sample frame, and then carefully note the percentage of those who accept to those who do not. This percentage, known as the response rate, needs to be as high as possible in order to minimize nonresponse bias. If only one in 10 people asked agrees to participate, then by definition 9/10 people have chosen not to—thus again violating the underlying assumption that the sample represents the population. There is a considerable literature in the quantitative social sciences about how to maximize participation rates in surveys and experiments. Amongst the best works in this area are those by Dillman, for example, "The Tailored Design Method" (2000). Whilst these methods are primarily designed for use in social surveys, they are also in many cases equally applicable to the recruitment of participants into field studies in buildings.

There are many situations where the above approach of using sample frames, random samples, and avoidance of self-selected samples is either unworkable or (arguably) unnecessary. Where the study is of something that self-selection is unlikely to influence, then it can be argued that any sample of sufficient size, random or non-random, can be generalized from. Where use of non-random samples is unavoidable, then the researcher is left balancing different forms of uncertainty. Using or increasing trial participant numbers through use of self-selection, snowballing, or other non-random methods increases precision by increasing sample sizes; however, it does not increase accuracy. Addressing this means either acknowledging that the findings only pertain to, say, building occupants who volunteered to participate, or arguing that the causal mechanisms at play are independent of the act of volunteering to participate. For example, where non-randomly sampled participants are then random allocated to experimental groups, conclusions can be robustly drawn about the outcome of the experiment on the participants—but these findings can only be inferred to apply to people likely to volunteer for such experiments.

Whilst there are instances in which the aim is to generalize from a sample to a population of people within an individual building, frequently the goal is also to try and generalize across a particular class of building within the building stock. As discussed above, this is enormously challenging, particularly in the non-domestic buildings area. The best global example of such a non-domestic building survey is the long-running Commercial Buildings Energy Consumption Survey (CBECS) in the USA. Although constructing such a survey may seem like an impossible task, it is important to note that if the theoretical model states that buildings shape users' responses to them, then it is very important to include a representative sample of such buildings in the study. Failure to do so means that no sensible statistical claims about the generalizability of the findings can be made. Effectively, the study is a conglomeration of case studies rather than a survey. It is for this reason that most of the reported confidence intervals from studies in this field do not make statistical sense, as they do not define the population to which they are claiming statistical generalizability—and if they do, they do not have a sufficiently large and representative sample drawn from that population to support such claims. It is important in this context not to disregard studies that fall short of the statistical requirements for generalizability. For logistical reasons, few studies in the buildings field achieve such requirements and, as mentioned above, sampling uncertainties are only a small proportion of the uncertainty in reported findings irrespective of sample sizes.

This covers some, but not all, of the range of issues identified in Figure 3.2 on threats to the validity of inferential statistical findings. A detailed description of the measures undertaken to address each of these threats is beyond the scope of these guidelines and is covered in standard undergraduate texts on social survey design, for example (Sarantakos 2012).

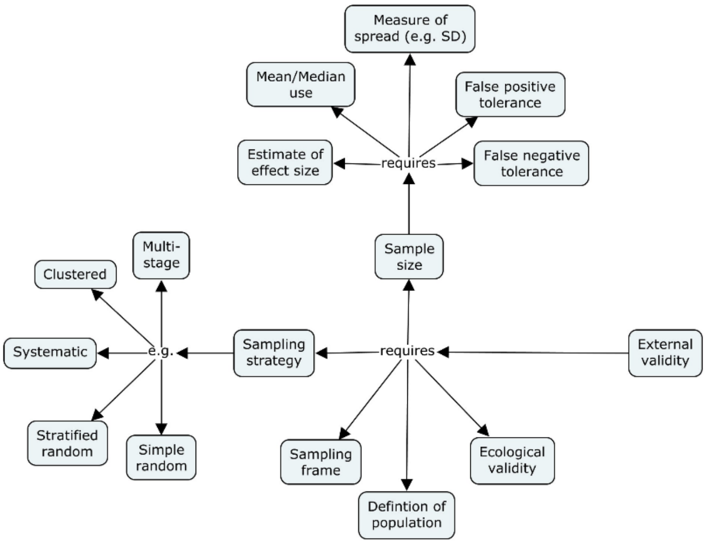


Figure 3.2 Threats to the validity of inferential statistical findings.

5.7. An illustrative example of sample size calculations

An example of how to calculate sample sizes for a trial is provided that is loosely based on the British **energy**wise project. Very simplified calculation methods are presented here for the purposes of exposition.

Project Summary: Energywise assesses how much electricity fuel poor customers in Great Britain will save if provided with a smart meter and some energy-saving appliances. The project uses a randomized control design. The unit of analysis is the home (house + household).

Sample size calculation for establishing inferential statistical validity.

Aim: To ensure that the findings observed in the sample will hold in the wider population with a given degree of statistical confidence.

Step 1: Determining the population size

This was set at 260,000 based on the estimate of the number of customers on the Priority Services Register (a proxy for fuel-poverty) in the UK Power Networks' distribution zones. While an underestimate of the number of fuel poor in Great Britain, for populations over 20,000 estimated sample sizes change little.

Step 2: Calculating the sample size

For sample size statistics for inferential statistical validity the following equation was used (PSU 2014).

$$n = \frac{\left[\frac{P[1-P]}{\frac{A^2}{Z^2}+\frac{P[1-P]}{N}}\right]}{R}$$

Where:	Inputs:
n = sample size required	
N = population size	N = 260,000 (see above)
P = variance in population	P = 0.5. Assuming 50% of participants save more than 6% and 50% less.
A = precision	A = 5%
Z = confidence level	Z = 1.6449 for 90%
R = Estimated Response rate	Adjusted after calculation

This produces a value of n of 271 survey participants required in the trial.

5.8. Internal validity

Internal validity is a key concern in experimental research designs, such as randomized control trials. A key mechanism for ensuring internal validity is the provision of intervention and control groups that are initially statistically identical, differing only in the application of the intervention to the intervention group. One key test of internal validity is the test of the likelihood that observed differences between the intervention and control groups are statistically significant, and at what level of statistical confidence.

The capacity to statistically distinguish between the intervention and control groups is only one of the issues to be considered with respect to internal validity. Sarantakos (2012) provides a good overview of the range of issues known as “threats to internal validity” that must also be considered when designing such trials, as well as descriptions of the measures that have to be undertaken to address such issues.

To illustrate the process of calculating sample sizes for internal validity an example is again provided based on the British **energy**wise project.

Step 1: Determining the level of statistical confidence and power needed for internal validity on the trial

The consortium members were asked:

“Are you more worried about:

a) mistakenly accepting an intervention that doesn’t work because the evidence wasn’t strong enough? or

b) mistakenly rejecting an intervention that does work because the evidence wasn’t strong enough?"

The first of these relates to false positive (type I) errors, and the second to false negative (type II) errors.

To properly assess these, a risks-based approach to costs and benefits is needed, i.e., the probability of the error needs to be multiplied by the magnitude of the consequences expressed in human or monetary terms. This is an area where the judgment of the researcher is called for.

In the **energy**wise project the following approach was adopted

“Tell me, in percentage terms, how sure you want to be that an intervention actually delivers the energy savings we measure?”

A) On the balance of probabilities (i.e., 50-65% confident)

B) Pretty confident (i.e., 65-80% confident)

C) Beyond reasonable doubt (80-95% confident)

D) Almost certain (>95% confident)

“Tell me, in percentage terms how sure you want to be that we don’t mistakenly reject an intervention that actually does work?”

A) On the balance of probabilities (i.e., 50-65% confident)

B) Pretty confident (i.e., 65-80% confident)

C) Beyond reasonable doubt (80-95% confident)

D) Almost certain (>95% confident)

The consensus amongst the project partners, on both the risk of false positives and false negatives, was that the group wanted to be “pretty confident” which was translated into a statistical confidence of 0.25 and a level of statistical power of 0.75.

It is difficult to overstate the importance of conducting this often overlooked step in sample size calculations. In many cases in energy use in buildings, occupant behavioral energy savings are only one element of the operational decision to install a given technology. They are frequently a “nice to have” benefit of, for instance, upgrading a building control system, or making a decision that incurs comparatively little additional cost. In this context, requiring 95% confidence of a trial is operationally inappropriate because the risks of failure are small (although, for academic publication purposes, it may be necessary).

Step 2: Determining the effect size

For the **energy**wise project, data on effect size was taken from the Energy Demand Research Project: Final Analysis report published by the UK energy regulator Ofgem (Raw and Ross 2011). This study, known as the EDRP, was the most up-to-date study on the effect size of smart meters available in Britain at the time. The following quote shows how uncertain the potential savings may be: “In the case of electricity consumption… a full range of 0-11% (energy savings) for some periods and customer groups" (p.4).

In light of this, and because of the nature and extent of the intervention in the **energy**wise trial, an energy saving of about 6% from the intervention group was used.

Step 3: Determining the mean and standard deviation of electricity consumption

The inputs of the mean and standard deviation were taken from government statistics, specifically the "Review of typical domestic consumption values" consultation document (Villalobos 2013).

The standard distribution of domestic electricity users in the UK was used (UK Profile class 1 electricity consumption). This provided the following values:

- Arithmetic mean: approximately 3,200 kWh/annum

No figure for standard deviation was provided, and so an estimate was made based on the inter-quartile range as follows:

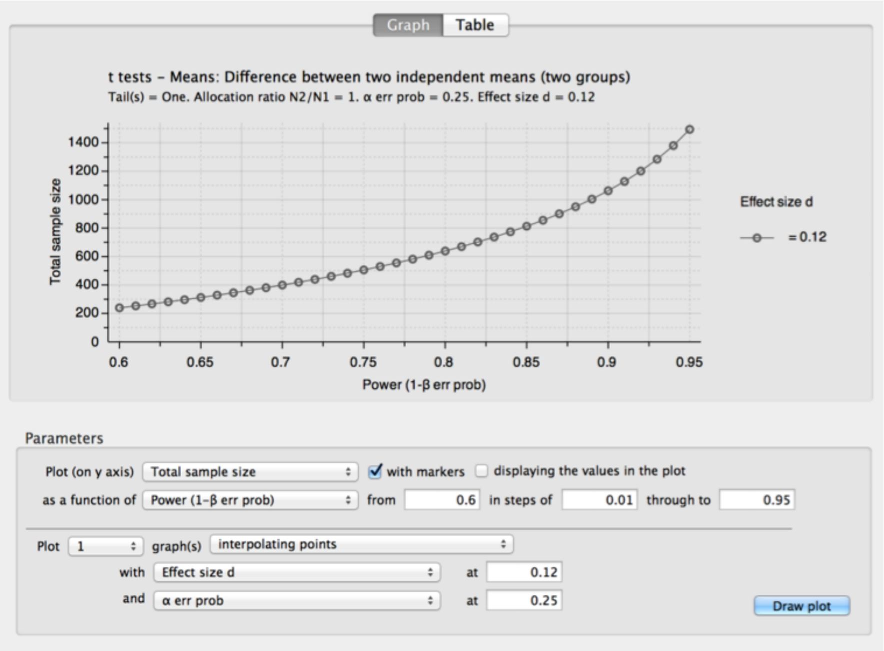
- Average inter-quartile: (1,200+1,600)/2=1,400 kWh/annum
- The ratio of interquartile range to standard deviation range is 34%/25%=1.36
- Estimate of standard deviation is therefore ~1.36*1,400=1,900 kWh

This, however, was for an average home, and needed to be adjusted for fuel-poor homes which were the subject of the study, as data on the mean and standard deviation is not available for this subpopulation. An adjustment was made based on the following logic: fuel poor customers are a subpopulation of all UK Electricity Profile Class 1 customers. They will, however, have a lower mean and a narrower standard deviation, as they are a more homogeneous group living in smaller homes. It was thus estimated that **energy**wise trial participants would have a mean electricity consumption of 3,000 kWh and a standard deviation of around 1,500 kWh. Note that these adjustments were merely educated guesses, as no further information was available on which to base these corrections.

Step 4: Sample size calculation for establishing internal validity

These sample size calculations were done using the G*Power 3.1.7 sample size calculation software as reported in Faul (2009; 2007).

The analysis presented here uses the simplest test possible: a one-tailed t-test comparison of the difference between two independent means (two groups) using the input parameters above. Figure 3.3 shows how sample size scales with the degree of statistical power desired. The value of 0.75 used in the calculation above corresponds to the estimated sample size of 506 on the graph.


Graph
Table
t tests – Means: Difference between two independent means (two groups)
Tail(s) = One. Allocation ratio N2/N1 = 1. α err prob = 0.25. Effect size d = 0.12
Total sample size
1400
1200
1000
800
600
400
200
0
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
Power (1-β err prob)
Effect size d
= 0.12
Draw plot

Parameters

Plot (on y axis) Total sample size ☑ with markers □ displaying the values in the plot

as a function of Power (1-β err prob) from 0.6 in steps of 0.01 through to 0.95

Plot 1 graph(s) interpolating points

with Effect size d at 0.12

and α err prob at 0.25

Figure 3.3 Effect of varying statistical power on the estimated sample size.

Note that this is a very rough initial estimation of the sample size needed to distinguish a 6% effect size between two equally sized groups with a statistical confidence of 0.25 and a statistical power of 0.75.

The sample size calculation for internal validity generates an estimated intervention group and control group size of 253 each, i.e., 506 in total. The sample size calculation for external validity generates an estimate of 271 in total. The test for internal validity is the larger of the two, and is therefore the factor determining sample size. In addition to this, an allowance for the estimated number of participants leaving the trial ("dropouts") needs to be made and added to the sample size.

5.9. Dropouts and response rates

There are two adjustments that need to be made to the sample size calculation in order to determine the number of participants that need to be recruited. These are the expected dropout rate and the expected response rate.

The sample size calculation is based on the number of participants needed to conduct the analysis of the data at the end of the study. However, dropouts are likely, and hence the initial sample needs to be increased by the expected number of dropouts. In shorter term experimental work, such as a week-long survey of occupant behavior in an office building, comparatively few people may drop out of the study. In contrast, however, if conducting a year or multi-year study of occupant behavior in homes, 30 to 50% of participants may either move house, or choose to leave the study. The number of dropouts will be a function of the respondent burden (i.e., the inconvenience that participants have to put up with) and the duration of the study. The higher the respondent burden and the longer the study, the greater the likelihood of dropouts. When estimating the number of dropouts, the most useful method is to look to similar studies and make adjustments based on what expectation of the respondent burden and duration from the dropout rates reported in those. The calculated sample size should be increased by the expected dropout rate. In the example used above, the sample size of 506 would be increased by 30% to reach the number of people to account for later dropouts (in this case to 723).

The estimate of the likely response rate, (i.e., the ratio of who was invited to participate to the number that accepted that invitation) will vary depending on the method used to recruit participants. It is worth noting that expectations around what is an acceptable response rate vary from field to field. In the quantitative social sciences, particularly at the level of national statistics, statisticians will frequently start to become concerned when response rates drop below around 70%. In contrast to this, it is not uncommon in building occupancy studies for response rates either to be unknown, or to be substantially below 10%. The critical issue here is that any reduction below 100% represents a certain degree of self-selection of the sample.

6. How to measure concepts (Methods)

Having looked at research questions, established the theoretical model, and determined the boundaries of the applicability of the study, the next step is to determine how to measure the concepts in the theoretical model. This is the realm of research methods. Other chapters in this book talk in detail about different specific research methods and these should be referred to as appropriate. This section is going to focus on issues of clearly defining what is being measured and ways of trying to quantify some of the uncertainty in the measurements.

6.1. Concepts and constructs

In research on occupants in buildings, relevant concepts include temperature, comfort, glare, productivity, and adaptive response. These are used to construct a theoretical model of how occupants respond to their physical environments.

It is useful to draw a distinction between concepts and constructs. Markus (2008) distinguishes between *concepts*, which he defines as the reification of all actual or potential instances of a set of experiences in the real world, and *constructs*, which are the instances of these in a specific population. Within a population, concepts and constructs are the same thing; however, the distinction becomes particularly important in international comparative work where concepts transfer between populations and constructs may not.

The benefit of such a distinction is that the area of occupant behavior in buildings is a highly international one in which researchers may frequently attempt to measure the same concepts, acknowledging that how those concepts are constructed and operationalized will necessarily need to take into account differences in climate and culture.

6.2. Operationalizing constructs into measurands

Operationalizing constructs is the process of determining how best to measure them. Sometimes they can be measured directly with a single instrument, for example, air temperature. Frequently, however, it is necessary to combine outputs from a range of instruments to measure the construct of interest. When multiple instruments are needed to measure a construct the term latent variables or hidden variables are frequently used to describe them.

Trochim (2006) captures this in one of his diagrams, reproduced in Figure 3.4.

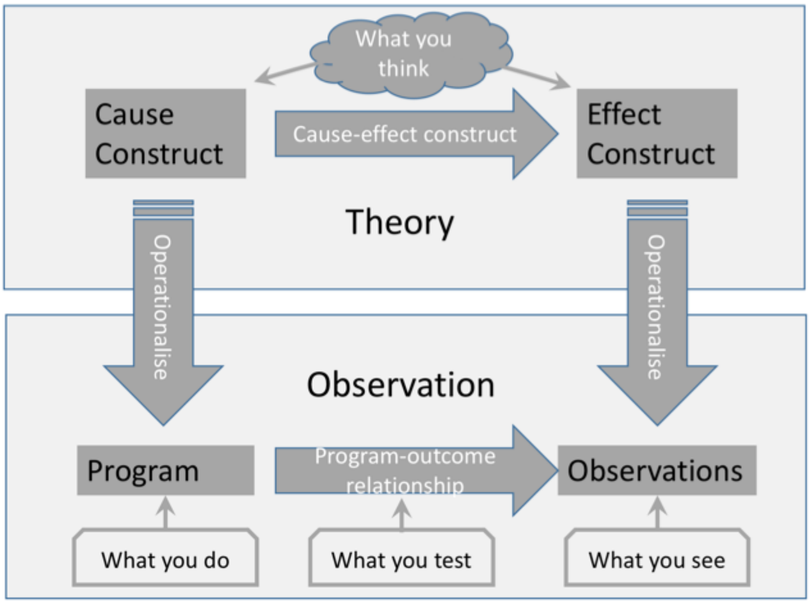


Figure 3.4 Theory -> observation relationship (after Trochim 2006).

In Figure 3.4, the theoretical model is represented in the top half of the diagram, while the translation of this into a concrete program of research is represented underneath. The aim of *operationalization* is to translate the theoretical model into measurable things as validly and as reliably as possible. Over time and multiple research programs, elements in the observation box will inform and change the theory box. In any individual research study, the observations of the research program reflect the theoretical model being evaluated.

Implicit in the construction of the theoretical model as advocated above is the need to clearly specify the study's outcome variable of interest. This must be done before the onset of the experiment to avoid "fishing" for significant effects after the experiments. How the outcome variable will be measured needs to be defined in detail. This will need to be done in the context of the research aim or question the study is designed to answer.

One of the key uncertainties that arises in the operationalization of constructs is what is called "underspecification of the measurand".

> The term "measurand" is used in the field of metrology (the science of measurement) and is defined by the International Bureau of Points and Measures (Joint Committee for Guides in Metrology (JCGM) 2008a) as "quantity intended to be measured"—in this case, the construct.

Underspecification of the measurand is the failure to specify *exactly* what it is that should be measured. For example, external temperature with respect to a building is often not specified exactly. External temperature can vary considerably around the envelope, and thus any measurement of the concept of external temperature is subject to considerable error as each researcher will operationalize the concept differently. If specified more precisely—say, external ambient temperature measured within a Stevenson screen at 1.5 meters above ground level one meter away from the building envelope at each compass point with the arithmetic mean value taken—there would be far less (but still some) leeway to measure differently. Underspecification of the measurand is not a problem of measuring; it is a problem of operationalizing concepts—and leads to uncertainty in comparing the results of different studies and in replicating studies

6.3. Latent variables

Latent variables—also known as "hidden variables" or "hypothetical constructs"—are variables that cannot be measured directly. Some authors distinguish between the terms, using the term "hidden variable" as something that physically exists and could therefore in principle be measured directly, but for cost or other reasons may not be, and "hypothetical variables" as those that do not physically exist, but are useful explanatory tools, for example, attitudes or inflation.

> Latent variables are common in all fields of research including in building occupancy studies. They vary from things like the volume of a room (which is constructed from a series of individual linear measurements and knowledge of geometry), to operative temperature (which requires measuring both air and radiant temperature), to psychological variables such as environmental attitudes or perceived control, which are usually measured through a set of questions.

The construct of interest is, for practical or other reasons, not directly observable and must be measured by combining the outputs from multiple individual instruments. In psychology there is a considerable methodological literature about how scales (i.e., sets of questions) should be developed and a considerable body of statistical science behind their evaluation.

6.4. Instruments

Various aspects of instrument selection, development, and placement will be discussed in detail in other chapters, particularly Chapter 4 through Chapter 8. One general point that should be made is that, conceptually, social research methods (participant observation, social surveys, interviews, focus groups) are also instruments in that they are designed to measure specific things that are subject to the same forms of uncertainties (imprecision, inaccuracy, etc.) as their physical counterparts. Thinking of physical, physiological, psychological, and social instruments in the same way is useful in supporting cross-disciplinary collaboration and establishment of a common vocabulary of measurement in this highly interdisciplinary and socio-technical area of study.

6.5. Quantifying uncertainty

The International Bureau of Points and Measures emphasizes the fact that any quantitative measure consists of three components. The first part consists of some multiplication of the number of base units (for example, a home might use 2,000 kWh of electricity per annum). The second part stipulates an error margin around that value (e.g., ± 100 kWh per annum). The third part stipulates the probability that the “true” value lies within that error margin (e.g., 0.9). Any quantitative assessment that fails to clearly identify each of these three elements for each measurement is incomplete and makes the result difficult to interpret. This ideal is one that is frequently hard to achieve in practice, but the ideas that it entails are important for researchers to understand. In particular, the third component is a reminder that instruments never perfectly capture the true value that is intended to be measured (i.e., the measurand). Accepting that all measurements are approximate and never perfect has two consequences. Firstly, that it is necessary to estimate the degree of precision required in order for findings to be useful. This is a function of the purpose of the study and can be established before any considerations of methods is undertaken. Secondly, that it is necessary to decide whether the measurements taken and models used allow making statements that fall within this required degree of precision. Without the quantification of the uncertainties surrounding the study answers it cannot be judged whether the measurements and models are suitable for any given purpose.

Uncertainty quantification is a complicated and specialist field that is beyond the scope of this book. An excellent introductory reference on instrument error and error propagation in the physical sciences is Taylor (1997) and an authoritative guideline on error propagation using Monte Carlo analysis is provided by the Joint Committee for Guides in Metrology (JCGM) (2008b). Interestingly, in the area of instrument validity and reliability, the social sciences have developed better frameworks for assessment, e.g., the Multi-Trait Multi-Method (MTMM) approach (Campbell and Fiske 1959).

7. How to measure relationships (Research design)

Once having identified concepts, turned them into constructs, and operationalized them into things that can be measured, the challenge remains of determining the nature of the relationship between the concepts in the research question. There are essentially three types of relationships that could exist between the concepts measured: there could be a *causal* relationship, the concepts could be *correlated*, or they could be entirely *independent*. It is the role of research design to determine the nature of the relationship between the concepts. Research design is the process of devising a process that directly satisfies a brief, in this case, the research question or research aim.

Broadly speaking there are two forms of research design: descriptive (or correlational) research designs, and experimental (or causative) research designs.

It is important to note that both descriptive and experimental research designs use the same research methods. For example, a sensing campaign supported by occupant surveys can support analysis that is either descriptive of the relationships between the variables or shows causal relationships between variables. In order to establish causation, all other possible explanatory factors (all confounding variables) need to be eliminated implying that nothing else could have caused this observed relationship. This is conventionally and best done using experimental designs[1]. Such designs look to isolate the effect of one variable on another by holding all others constant in a controlled environment. This is a powerful and

[1] There are some methods of analysis that some analysts argue can establish causation outside of an experimental context. Lead amongst these is Judea Pearl and his application of statistical graphical modelling methods such as Bayesian networks (Pearl 2000). This is both a highly advanced field of statistical analysis, and a hotly contested topic that is beyond the scope of this book.

valuable approach, but not without limitations. The primary critique of such methods is their potential lack of ecological validity, i.e., that the findings from such studies do not reflect "real world" conditions and so what is observed in the lab or experimental field trial may not be observed in uncontrolled conditions. The more naturalistic the environment is in which the occupant experiences the experiment, the greater the ecological validity (see also Chapter 7). However, a more naturalistic setting makes control of confounding variables more difficult.

While it may seem intuitive that two variables are causally related, it is all too often the case that they are linked through a third variable which causes them to vary simultaneously.

A good example of confounding variables is the relationship between CO_2 levels and thermal comfort in a room. As occupants come into a room CO_2 levels will rise alongside temperature. If the relationship between CO_2 and thermal comfort is measured, it would show that they are highly correlated. There are also valid metabolic arguments as to why CO_2 may change metabolic rate and cognitive function and consequently impact on thermal sensation. In standard field monitoring conditions, it is very difficult to disentangle the rise in CO_2 with the associated rise in temperature, and thus to determine whether it is the CO_2, the temperature rising, or both that is impacting on people's thermal sensation. Therefore, standard monitoring field studies are not a good research design to try and answer this particular research question. Here the experimental precision of laboratory conditions is preferable, allowing independent variation of CO_2 levels from temperature levels in order to isolate the effect of one variable from the other.

A concept map illustrating some of the key concepts in both descriptive (correlational) and experimental (causative) research designs is provided in Figure 3.5.

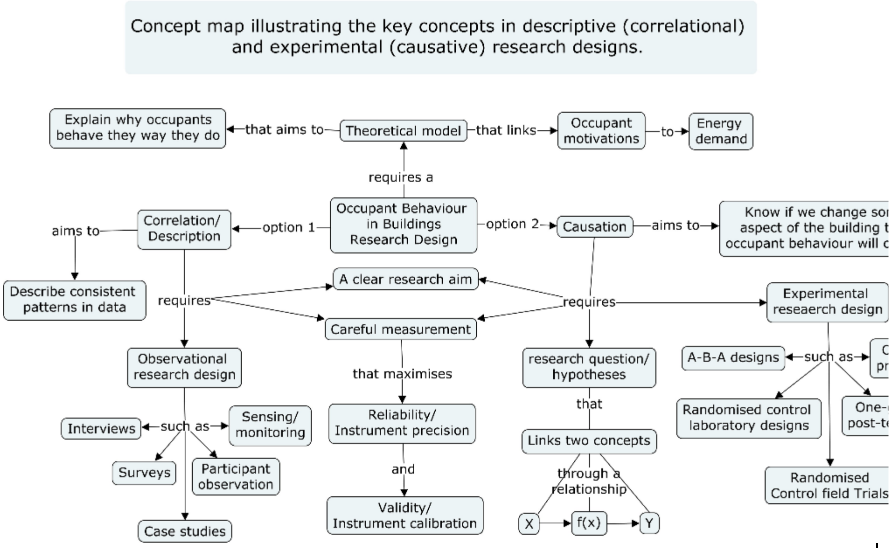

Concept map illustrating the key concepts in descriptive (correlational) and experimental (causative) research designs.
Explain why occupants behave they way they do
that aims to
Theoretical model
that links
Occupant motivations
to
Energy demand
requires a
Occupant Behaviour in Buildings Research Design
option 1
option 2
Correlation/ Description
aims to
Causation
aims to
Know if we change sor aspect of the building t occupant behaviour will c
Describe consistent patterns in data
requires
A clear research aim
Careful measurement
requires
Experimental reseaerch design
Observational research design
that maximises
research question/ hypotheses
A-B-A designs
such as
C pr
Interviews
such as
Sensing/ monitoring
Reliability/ Instrument precision
that
Randomised control laboratory designs
One- post-t
Surveys
Participant observation
and
Links two concepts
through a relationship
Randomised Control field Trials
Case studies
Validity/ Instrument calibration
X
f(x)
Y

Figure 2.5 Concept map illustrating some of the key concepts in both descriptive (correlational) and experimental (causative) research designs.

Deleted: *3*

7.1. Descriptive (correlational) designs

Descriptive (correlational) research designs are the mainstay of studies into the impact of the occupant behavior on building energy demand. This would classically take the form of gathering data through installed sensors, virtual sensors, or data gathered for other purposes (frequently termed ‘administrative’ data), potentially augmented with some occupant surveys delivered either on paper or electronically through smart-phones or computers. The data would then be analyzed for correlations between the variables. Such a study design allows to understand relationships in the data, but not to say that a change in one observable causes a change in the other. There are times when this seems counterintuitive. This is usually where the theoretical model or mental model feels like the only possible explanation for an observation. For example, it is tempting to interpret window opening behavior as always being related to regulating the thermal or indoor air quality environment within the building, particularly where this is the purpose of the study. However, alternative reasons can also explain why occupants may be opening windows—for example, out of habit, or in a residential setting to talk with people outside, or to listen to the birds in the garden. The sun coming out would correlate both to a rise in internal temperature and to increased bird activity in the garden. It is very easy when interpreting data from an energy perspective to mislabel a correlation (here between internal temperature and window opening behavior) as causative. Drawing conclusions of causation in instances where other potential mechanisms have not been controlled for can easily lead to wrong conclusions. There are a wide range of descriptive research designs which are covered in detail in many textbooks (e.g., Bryman (2015); Saunders (2015)). Three of the main types of design are covered briefly here.

7.2. Case studies

One of the most widely used designs in the research on occupants in buildings is the case study. As the name suggests, a case study focuses on one individual instance (say, an individual building, campus, or community) and applies multiple methods to understand the workings of that particular case. Case studies offer no capacity for generalization because they are a sample of one. An excellent reference on the use of case study research is Yin (2013).

Some argue that if the case is in some senses archetypal, then lessons learned can be translated to similar cases. This is intuitively reasonable, but scientifically indefensible, as studying one case can say nothing about whether other similar cases work in the same way. It is tempting to assume that if other cases share similar characteristics and those characteristics are found to be explanatory of the behavior of the individual case, then the results must surely apply more broadly. This assumption only holds, however, under a certain theoretical or mental model of those factors which are important across the set of similar cases—an assumption that seldom holds true in practice.

Case studies are enormously powerful for identifying factors the commonality of which can then be explored using more sample-based research designs. One approach which seeks to span the gap between individual case studies, and a population-based sample, is the Qualitative Comparative Analysis method developed by Ragin (1987). This approach has now developed into a suite of methods which seek to systematically draw out commonalities between a small set of case studies. The approach is widely used in international comparative analysis and frequently is based on numbers of case studies ranging from 10 to 50.

7.3. Cross-sectional design

A cross-sectional design is one that gathers data at a particular point in time from a range of units of analysis (occupants, buildings, etc.). A one-off social survey is a classic example of the approach. When correctly designed, such approaches can support generalizations from the sample to the population. The design and construction of social surveys is covered in Chapter 8.

Cross-sectional designs can either be conducted once or at multiple points in time, thus creating a repeat cross-sectional research design. Repeat cross-sectional design does not measure the same people at each point in time, but rather generates a new representative sample from the population each time the survey is conducted. This distinguishes them from longitudinal surveys in which the same people are measured repeatedly through time. Repeat cross-sectional designs have the advantage that it is easier to draw a cleaner representative sample at each time point, than it is to try and maintain a panel of the same participants through time. If the aim of the research is to understand changes at the population level, then repeat cross-sectional designs are most appropriate.

Classic introductory texts on social survey designs include: Sarantakos (2012), which covers nearly all aspects of social research to a good undergraduate level of understanding, and Foddy (1993), which is excellent for details on survey and interview questions.

7.4. Longitudinal surveys

Longitudinal surveys are ones that measure the same units of analysis through multiple points in time. There are many different kinds of longitudinal surveys, including panel surveys, where "panel" is the name given to the sample of units of analysis (people, buildings, etc.) being drawn to represent the population and then followed through time with repeated surveys; and cohort studies, where a group of units of analysis sharing a common characteristic (say, a sample of buildings of a certain type built in the same year) are followed through their lifetime. Again, further details on such designs are provided in Chapter 8.

7.5. Causative (experimental) designs

An experiment is a procedure to test a hypothesis. The main difference between experimental research and other types of research is the aim of establishing causality, i.e., insight into cause-and-effect relationships, by testing what happens to an outcome variable if a specific factor is manipulated. The outcome variable is usually called the "*dependent variable*". The *independent variable*, also called the "treatment or the intervention", is under direct control of the researcher and is used for creating experimental conditions.

An example to illustrate the experimental approach and its variables: It might be interesting to know whether a pop-up window displayed on the screen at the end of the working day with a prompt to turn off the computer before leaving leads to a higher number of turned off computers (intervention group) than when providing no such prompt (control group).

The dependent variable would be the number of computers turned off at the end of the day, monitored over specific time period (e.g., two weeks) and averaged over that period. The pop-up window constitutes the independent variable.

Extraneous variables are factors not of interest to the researcher, but that need to be controlled for as they can also impact on the dependent variable and their effect can be confounded with the effect of the independent variable (hence they are also called "confounding factors"). The age and type of computer might be confounding variables in that people with old computers that take a long time to boot-up, or that do not permit being shut down with programs still open, make it less likely that someone will shut a computer down. Random assignment of participants to groups is one method of eliminating the effect of extraneous variables. If the sample is large enough, one would expect the same distribution of the extraneous variable in the group receiving the intervention and the one not, e.g., the same number of older computers, in both groups. However, in relatively small groups, randomization might not work. Where this is the case, another method is to control for the effect of those variables in the analysis of the data. By including them as variables in the statistical analysis the effect of the intervention can be tested while holding the extraneous variable constant. This allows the effect of the extraneous variable to be analyzed and accounted for. Extraneous variables are more of a problem when they are not obvious and when randomization cannot be relied on to ensure an equal distribution across all groups, e.g., because the sample is too small.

Random assignment is a critical feature of experimental work; it ensures that the groups are the same in important characteristics and that differences in the outcome measures are attributable to the intervention and not differences between the groups per se. The **energy**wise trial example given above is an example of one of the most common (and best forms of) randomized experimental design: a randomized design comparing control and intervention group in a post-test.

Pre-tests can be used in experimental studies. In the example given, the number of computers turned off before the intervention might be counted to establish a baseline. Since this could easily be done after all employees have left in the evening there would not be any concern that, in doing so, employees' attention would be drawn to the need to switch off computers and hence influence the trial's outcome. This is, however, a concern in other settings where, by including a pre-test, a topic is made salient to trial participants (e.g., making them more aware of energy use), and thus the pre-test could impact on the post-test. Pre-tests are also associated with higher costs, time, and effort, and hence are not necessarily advisable. However, they can be useful in other respects: in the example, a pre-test might reveal that all computers are switched off anyway, and hence, that there is no point in running the study!

Two other forms of experiment exist. The first is the quasi-experiment. It has the same elements as a true experiment, but lacks the crucial aspect of randomization, i.e., participants are not randomly assigned to conditions. Instead, assignment to conditions is via self-selection. This poses a serious problem because the assumption that groups are equal no longer holds, and hence there might be confounding variables. While the extent to which groups differ on certain easily measured variables (age, gender, income, etc.) can be assessed, it is quite plausible that there remain confounding variables which are hard to assess because they are difficult to measure. Ultimately, it is not known what made participants decide to choose one intervention over another, or to be in the control group. Despite this significant disadvantage, quasi-experiments are common in applied settings because they avoid a lot of the logistics of establishing a true experiment, or allow analysis of things that would be impossible or unethical to conduct experiments on. For an excellent example of a quasi-experimental design see the recent thermal comfort study by Luo (2016).

The third main type of experiment is the natural experiment, where a naturally occurring condition is contrasted with a comparison condition. Here the cause cannot be manipulated, i.e., the independent variable is not set by the researcher. For example, an earthquake might destroy several high-rise buildings in one city, and so a study might test if inhabitants of that city are less likely to buy flats in high-rise buildings over the next two years than inhabitants of a city of a similar size (and ideally, similar in other characteristics such as wealth, presence of industry, etc.) that was not affected by an earthquake. The big advantage of natural experiments is that they allow the study of the effect of phenomena that otherwise could not be studied; however, groups are not necessarily equal (or even similar), were not randomly assigned, and there might be a wide range of confounding variables.

8. Pre-analysis plans

One of the key points that Wasserstein (2016) notes on behalf of the American Statistical Association in their article on good practice in the use of tests of statistical significance is that, "Proper inference requires full reporting and transparency." They emphasize that

> *Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. ...Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis.* (p.10)

This is mirrored in an article by Simmons et al. (2011) article in which they argue that researchers have a lot of degrees of freedom to make decisions during the data collection and analysis that distort the research process and artificially inflate the probability that they will find positive results. To combat this, Taubman (2010) and many others have argued for development and publication of a data analysis plan prior to conducting the research, also called a "*pre-analysis plan" (PAB*). They note, "by planning and disclosing the hypotheses to be tested and specifications to be used in advance of seeing the data, the plan should avoid (or at least minimize) issues of data mining and specification searching." (p.3).

An analysis plan will usually include the following sections:

- Overview of the study (including: aim; research/experimental design; outcome measure; sample)
- Ethical considerations (including in experimental research ethical aspects arising from things like withholding intervention from one group, negative effects of an intervention, and privacy aspects).
- Statement of hypotheses to be tested (including: expected average effects; causal chain of process and mechanisms; heterogeneous effects on sub-groups)
- Estimating equations to be used (including: stating the spatial and temporal sampling frequency to be used; estimating average treatment effects; estimating treatment effects using interaction terms; what predicts the outcome variable of interest)
- Testing for balance if experimental design is used (including: randomization/balance checks)
- Procedures for addressing missing or low quality data, covariate imbalance and questions with Limited Variation (including: item non-response; covariate imbalance; questions with limited variation)
- Variable construction (including how each variable is to be constructed from the raw data)

Such analysis plans should be prepared in advance of the study and, in the ideal case (and as required by some journals), published online to ensure full accountability of analysis and so editors can check that no additional analysis has been conducted to “massage” the data to achieve desired outcomes.

The other issue which analysis plans serve to improve is statistical conclusion validity. Statistical conclusion validity refers to the extent to which statistics are used properly and appropriate conclusions drawn from analysis. It relies on other forms of validity that extend to the choice of analysis methods, with a particular emphasis on whether the underlying assumptions of these analysis methods (frequently normality of distributions) hold in the case of the analysis conducted (see (Sackett et al. 2007)).

As with many aspects of best practice in research design, production and publication of such analysis plans is often not done in building occupancy research. This risks leading to high levels of cherry picking of favorable findings by running multiple analyses and publishing only those “of interest” (i.e., frequently those with positive relationships between variables). Such skewing of the research process makes both interpretation and replication of findings difficult or impossible and undermines the quality of work in the field. It should be stressed that performing exploratory data analysis by conducting tests not outlined in the original analysis plan is an entirely acceptable form of scientific practice—however, it is one that should only be used to generate hypotheses for testing in future well-designed studies in which such forms of analysis are written into original analysis plan.

9. Conclusions

Research design is an essential, but often misunderstood and overlooked component of the research process. This chapter lays out a systematic approach to research design centered on the construction of a concept map diagrammatically representing the theoretical cause-effect model that the research is seeking to test. Making this explicit through concept mapping requires representing the concepts being explored as nodes, and the relationships between those concepts as links. Once the theoretical model is mapped out, then research questions are easily articulated as the relationships between the concepts in the model. Hypotheses can be drawn from the research questions that the research can be designed to test. This approach also provides a framework for the writing of pre-analysis plans, which help researchers clearly articulate their proposed methods of analysis prior to collecting their data, thus helping to guard against malpractice, such as searching for statistically significant relationships between variables that were not the original intent of the study.

Occupant behavior in buildings research must be fit for purpose. To be fit for purpose, the purpose must be known and the findings of the research must fall within acceptable margins of error for that purpose. Therefore, to be useful, research must not only produce findings, but also quantify the uncertainty in those findings to show they lie within the acceptable margins of error for that purpose. To achieve this requires both quantifying uncertainty, but more importantly designing-out enough uncertainty to fall within required error margins. The procedures outlined in this chapter address both these elements. Accepting that things cannot be measured perfectly, mapping the theoretical model, choosing an appropriate research design, and selecting and applying appropriate methods all help in reducing uncertainty.

The procedures outlined in this chapter constitute best practice in research design and may seem intimidating to many new and established researchers in this field. Indeed, many of these methods represent the cutting edge of best practice in research in the more pure-science fields such as the social sciences, psychology, physics, and metrology. Studying the actions and influences of occupants on energy use in buildings is a theoretically and scientifically challenging task as scientifically demanding as any in the pure sciences. It is all too easy for the influences of occupants to become lost in a sea of confounding influences on energy demand, ranging from the impact of the weather, through the performance of the building fabric, to the behavior energy producing and consuming technologies and their control systems and the complex temporal interdependencies of all of these. To disentangle these influences and isolate the influence of occupants requires theoretical clarity and rigorously designed and conducted research in order to establish the foundations and significant findings of the field.

References:

Ajzen I (1991) The theory of planned behavior. Organizational Behavior and Human Decision Processes 50:179-211

Bouma G (2000) The Research Process. 4th edn. Oxford University Press, Melbourne, Australia

Bryman A (2008) Social Research Methods. 3rd edn. Oxford University Press, New York

Campbell D, Fiske D (1959) Convergent and discriminant validity by the Multi-Trait Multi-Method matrix. Psychological Bulletin, 6:81-105.

Casti J (1992) Reality Rules: Picturing the world in mathematics: The fundamentals, the Frontier. John Wiley & Sons, New York

Dillman D (2000) Mail and internet surveys: the tailored design method. Wiley, New York

Faul F, Erdfelder E, Buchner A, Lang A-G (2009) Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41:1149-1160.

Faul F, Erdfelder E, Lang A-G, Buchner, A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39:175-191.

Foddy W (1993) Constructing questions for interviews and questionnaires: theory and practice in social research. Cambridge University Press, Cambridge

Gauthier S, Shipworth D (2015) Behavioural responses to cold thermal discomfort, Building Research and Information 43(3):355-370.

Groves R, Fowler F, Couper M, Lepkowski J, Singer E, Tourangeau R (2004) Survey Methodology. John Wiley and Sons, Hoboken, New Jersey.

Joint Committee for Guides in Metrology (JCGM) (2008a) 200:2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM)

Joint Committee for Guides in Metrology (JCGM) (2008b) Evaluation of measurement data — Supplement 1 to the Guide to the expression of uncertainty in measurement — Propagation of distributions using a Monte Carlo method. BIPM: 90.

Luo M, de Dear R, Ji W, Lin B, Ouyang Q, Zhu Y (2016) The Dynamics of Thermal Comfort Expectations. Building and Environment 95:322–329.

Markus K (2008) Constructs, Concepts and the Worlds of Possibility: Connecting the Measurement, Manipulation, and Meaning of Variables. Measurement: Interdisciplinary Research and Perspectives 6(1-2):54-77

Nicol F, Humphreys M, Roaf S (2012) Adaptive Thermal Comfort: Principles and Practice. Routledge, London.

Novak J, Cañas A (2006) The Theory Underlying Concept Maps and How to Construct and Use Them, Institute for Human and Machine Cognition (IHMC).

Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, England

PSU (2014) An Equation for Determining Final Sample Size from Program Evaluation Tipsheet #60 - How to Determine a Sample Size. URL:http://extension.psu.edu/evaluation/pdf/TS60.pdf. Accessed 2014-10.

Ragin C (1987) The Comparative Method: Moving beyond qualitative and quantitative strategies. University of California Press, Berkley

Raw G, Ross D (2011) Energy Demand Research Project: Final Analysis. Ofgem, London. http://www.ofgem.gov.uk/Pages/MoreInformation.aspx?docid=21&refer=Sustainability/EDRP.

Ruttkamp E (2002) A Model-Theoretic Realist Interpretation of Science. Kluwer, Dordrecht

Sackett P, Lievens F, Berry C, Landers R (2007) A Cautionary Note on the Effects of Range Restriction on Predictor Intercorrelations. Journal of Applied Psychology 92(2):538–544

Sarantakos S (2012) Social research. 4th edn. Palgrave, Basingstoke

Saunders M, Lewis P, Thornhill A (2015) Research methods for business students. 7th Edn. Trans-Atlantic Publications

Simmons J, Nelson L, Simonsohn U (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22

Taubman S, Allen H, Wright B (2010) The short-run impact of extending public health insurance to low income adults: evidence from the first year of The Oregon Medicaid Experiment - Analysis Plan.

Taylor J (1997) An Introduction to Error Analysis: The study of uncertainties in physical measurements. University Science Books, Sausalito

Trochim W (2006) The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006)

Villalobos (2013) Review of typical domestic consumption values. Consultation document. Ofgem Ref: 113/13. 3 July 2013.

Wasserstein R, Lazar N (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician: 00-00.

Yin R (2013) Case Study Research: Design and Methods. 5th edn. SAGE Publications, Thousand Oaks