
Chapter Summary

The previous chapter discussed how a clearly defined goal helps the researcher or developer choose the type of study to perform. In this and the following chapter, it is assumed that an experiment, referred to as a user study, is to be executed. Different names are used to describe such studies depending on the discipline. For example, experiments, as they are called in psychology, are more often called user studies in informatics or randomized clinical trials in medicine. Regardless of the name used, the design of the study will influence whether any interesting results are found and the degree to which these results can be trusted and generalized beyond the study.

This chapter describes the different types of variables that one needs to understand and define when conducting a user study. The independent variable is the treatment or the intervention. In informatics, this is usually the new system or algorithm that needs to be evaluated. It is compared against one or more other conditions, systems or algorithms. The dependent variable is the outcome or the result that is important to the users, developers or researchers. In informatics, it is often an improvement in processes or decisions that can be attributed to the new system or algorithm. How these two types of variables are defined and measured will affect the trustworthiness of the study and also how well the results can be generalized to other situations. Confounded variables, nuisance variables and bias all affect the relationship between independent and dependent variables. By controlling these additional variables and choosing the best design, the researcher can ensure the best possible, honest results. A poor design can lead to spurious conclusions, but more often it will lead to missing existing effects and a waste of time, money and effort.

Independent Variables

The term *independent variable* means the same as *treatment* or *intervention* [1] and signifies a “causal event that is under investigation” [2]. The independent variable, manipulated by the researcher, describes what is expected to influence the outcomes. A treatment is a specific condition of this independent variable. The goal of a user study is to compare the outcomes for different treatments. The independent variable is connected to the dependent variable, which measures the outcome, by means of the hypotheses [3]. A simple hypothesis is a prediction of a causal effect of the independent variable on the dependent variable: depending on the condition of the independent variable a different outcome is predicted for the dependent variable.

A user study can have one or more than one independent variables and, in this case, each variable represents a treatment that can be controlled and systematically manipulated by the researcher. Studies with more than one independent variable are more complex to execute, analyze and interpret. The number of variables also affects the number of participants that need to be found for the study. Usually, more independent variables will mean that more subjects are needed. However, in some cases subjects can participate in multiple conditions.

In medical informatics, many studies will evaluate the impact of one independent variable only. This independent variable includes a new or improved information system that is to be compared with other, older approaches. For example, assume a researcher has designed a persuasive text messaging system that uses text messaging to encourage obese people to lose weight. The system sends messages a few times a day about possible activities that are suitable given the day of the week and the weather forecast. The goal is to help people lose weight by encouraging them to engage in physical activity. The study will test whether the persuasive messaging system is more effective than, for example, meetings with a dietician. However, it is possible to consider other independent variables. In this example, the researchers suspect that the system will be more suitable for younger users because most already love using their mobile phone. So the researchers also want to compare older and younger people, which can be defined as a second independent variable.

Types of Variables

The independent variables can be of different types, and these different types can be present in the same study. Understanding the types will help the researcher choose the levels to be used in the study. *Qualitative independent variables* describe different kinds of treatments. For example, a qualitative independent variable called “System Availability” could have two conditions: the presence (condition 1) or absence (condition 2) of the information system. Such a qualitative comparison also could be made between two types of systems, or between an information system and behavioral therapy, among others. In all these examples, there are two or more conditions for one independent variable. For the weight loss messenger system described above, it would be possible to compare weight loss of people who use

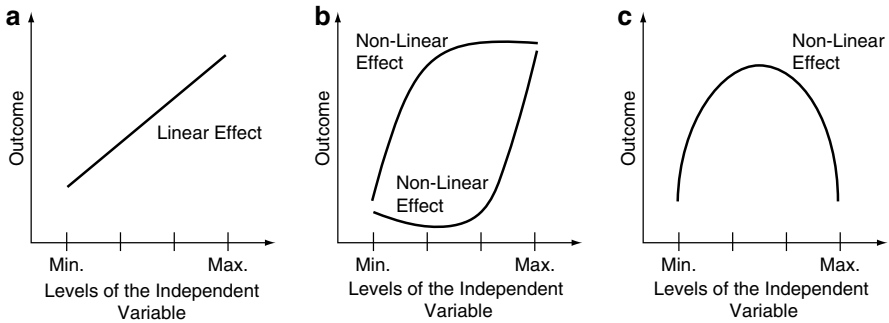


Fig. 2.1 Example effects

the system versus people who work with a dietician. In many cases, the current situation, and this can consist of an existing information system or a non-computer information system, serves as a baseline. For example, Gorini et al. [4] compare three conditions for treating generalized anxiety disorder: one condition consists of using virtual reality via a mobile phone with biofeedback, a second condition consists of the virtual reality via a mobile phone but without biofeedback and a third condition consists of no intervention.

Quantitative independent variables compare different amounts of a given treatment. For example, one could compare sending one message per day with sending one message during every daytime hour. When using a quantitative independent variable, it is important to carefully consider which levels to use. Especially when conducting an initial evaluation of a system, it is best to evaluate levels taken from a wide enough range so that the results will show as much of the impact as possible. If possible, include two extremes and at least one, but better two, levels in between. This will improve understanding the effect of the intervention.

For example, as shown in Fig. 2.1a, a linear effect may be present where the outcome is directly proportional to the input. Having only two levels will make it hard to show the type of relationship that exists. Figures 2.1b and 2.1c show other relationships where it is advantageous to have multiple levels of the independent variable. Figure 2.1b shows a relationship where the extremes are not the best values and where the effect of the independent variable levels off. Often this happens when the intermediate levels represent a more balanced treatment approach. Similarly, Fig. 2.1c shows how the extreme conditions do not present the best outcome. Worse, no effect at all would be noticeable if only the extremes were measured. With new systems, the extreme situation may still be unknown and a poorly designed study will lead to the conclusion that there is no effect, while the study only failed to measure it.

Use common sense when deciding on the levels of the independent variable. Consider theoretical reasons as well as ethical and practical limitations. Consider the following treatment levels of the text messaging system: the lower extreme value, the minimum, could be 1 message per day; the highest level, the maximum, could be 32 messages or 1 message every 30 min during the day. One message a day

may have no effect at all, while receiving a message every 30 min may be annoying and have adverse effects, such as people turning off their phones so they do not receive any more messages.

Note about Random Assignment

It is important to remember that random assignment of subjects to the experimental conditions is what makes a study a true experiment. By randomly assigning subjects to conditions, one can avoid systematic distortion of the results. A note of caution is needed however. Even though random assignment is critical to discover causal relations, it may introduce a new bias, especially in medicine. With many studies, especially clinical trials, patients or consumers will have a preference for a certain treatment; they usually prefer to receive the new treatment. For some patients, it is their last hope. Random selection does not take this preference into account, and this may influence enrolment, sample representativeness, attrition, adherence or compliance, and outcomes [5]. For example, only patients who are willing to be part of a placebo condition may participate, or patients may drop out when they suspect they are not receiving the new treatment.

Dependent Variables

A *dependent variable* is also called an *outcome* or *response variable* [1, 6] and represents the outcome of a treatment. The dependent variable should be chosen so that one can make a conclusion about the treatment in relation to the professed goal. It is expected that this dependent variable will show different outcomes for the different experimental conditions. If the goal is to develop an information system that helps people with weight loss, the dependent variable should reflect this goal and allow the researchers to draw conclusions about losing weight with help from the information system that is being evaluated. For example, the weight lost after 1 month could be the dependent variable. For the persuasive text messaging system described above, it is expected and hypothesized that participants will lose more weight with the text messaging system than without.

A good evaluation will have complementary measures to assess the impact of a treatment. When the outcomes of complementary measures point in the same direction, for example, that a system is user friendly, this provides a much stronger evaluation and the researcher can be much more confident about the conclusion. Moreover, such additional measures often are useful to help explain unexpected results of using the system. Keep in mind that each analysis will evaluate the impact of the conditions, the independent variable, on one outcome measure, the dependent variable, at a time.

When choosing a set of evaluation metrics, it is important to include existing metrics decision makers are already familiar with when possible. Regardless whether the decision makers are the future users, the buyers of the software or

fellow researchers, metrics used for many years or in many evaluations are more likely to be well understood and accepted as part of the decision making process. Naturally, relying solely on metrics that have been used historically is unwise. Evaluations should include the metrics that are most relevant to the study. For example, if one designed a system for online appointment scheduling, the clinic where the system will be tested will most probably already keep track of the number of people who do not show up for appointments. Obviously, they will be interested in seeing the effects of the system on such a well known metric. In addition, it may be quite reasonable to measure the number of changes in appointments and the associated costs. A new system may not only affect no-shows but also the time needed for rescheduling existing appointments.

Below, a general approach to categorizing variables and measures is described. This is followed by a list of commonly used metrics. The metrics to be used are often determined by the field of study, the environment or the decision makers; however, it is important to remember that the choice of the metric also will affect the power of the study or how sensitive it is. Some metrics are better than others to show a significant effect even when used on the same dataset. For example, when an ordered list is being evaluated, rank order metrics, which take the order into account, show a higher effect size than all-or-none metrics, where only the presence of the correct answer counts [7].

Types of Variables

There is a broad choice of possible dependent variables. They have different advantages and disadvantages, and their popularity depends on the field of study. One way of looking at the types of variables is to categorize them according to the aspect of the information that is being evaluated. Goodman and Ahn [8] list five categories of measures: technical properties; safety; efficacy and efficiency; economic attributes or impacts; and legal, social, ethical or political impacts. Below are examples of many measurements that belong to the first three categories. The last two are beyond the scope of this book.

The development phase of a project affects the choice of dependent variable. There are several outcome measures that are suitable for use in multiple phases of the system's life cycle stage. However, there are other outcome measures that are particularly suited to early or late development phases. As Kushniruk and Patel [9] point out, usability evaluations are especially useful during the formative phases of the software. Waiting until the final development stages or implementation to test usability is not a good idea. Problems with the interface and expected interactions between the system and users need to be caught early. With current software tool-kits, user interfaces can be prototyped very early in the development cycle and tested with a variety of measures. During explorative, early phases of development, measures such as relevance, completeness of results, feasibility and risk will take center stage. Many other measures, such as cost savings or improved decision making, are usually better suited for later stages, when the system has reached maturity.

Once it has been decided what the dependent variable will be, it is necessary to choose a metric. Metrics are the measurement tools. For example, an already existing, validated survey instrument could be used to measure user preference. Another example is the use of a formula to calculate precision or recall. The metrics provide the concrete value for the chosen measure [10]. It is best to have a combination of measures for a study to have a balanced evaluation. The simplest and most straightforward approach is to use *single* or *base metrics*. However, sometimes *derived* or *composite metrics* are needed. This distinction also is referred to as *base* versus *synthetic metrics* [11].

For example, to determine user friendliness, one measure could be the subjective evaluation of system's user friendliness with a survey. However, in most studies, participants will be required to complete a task that allows additional metrics to be measured. For example, a complementary metric could be a count of the number of errors made when working on the tasks, which would capture objectively how user friendly the system was. When working with tasks, it is important that they are representative of the final intended usage of the system. Otherwise, the metrics will be irrelevant. For example, if a clinician is required to evaluate only one x-ray per half hour, the speed of loading the x-ray on the screen will not be that important. It should not matter whether it loads in 500 ms or in 1 s and a dependent variable measuring load time would be pointless in this case. However, when evaluating a decision support system where a few thousand images are loaded and clustered for a clinician to review, the time it takes to load them will be extremely important.

For study designers who have complete discretion over the choice of dependent variables, a good trio to consider is: effectiveness, efficiency and satisfaction. *Effectiveness* measures whether the information system does what it is supposed to do. Examples are the number of errors, counts of (relevant) events, precision, recall, and false positives and false negatives, among others. *Efficiency* measures whether the information system does its job in a suitable manner. Examples are whether tasks were completed, time taken to complete the task, run time and memory requirements, among others. An alternative view on these two measures is outcome versus performance measures. *Outcome measures* are used to evaluate the results of applying the information system, similar to effectiveness measures, while *performance measures* are used to evaluate the process itself, similar to efficiency measures. *Satisfaction measures* are more subjective and relate to the users' perception of a system. Example measures range from simple questions such as "Which system do you prefer?" (when comparing systems) to multi-item validated questionnaires.

Common Information Retrieval Measures

Precision, Recall and the F-measure are three outcomes that are among the most frequently used in information systems evaluations. Precision and recall are individual measures, while the F-measure is a composite value, providing a balanced number that combines both precision and recall. They are particularly popular in the evaluation of information retrieval systems. Yousefi-Nooraie et al. [12] use

precision and recall to compare three different PubMed search filters that are meant to help answer clinical questions. Kullo et al. [13] use precision and recall to evaluate algorithms that extract information from electronic medical records for use in genome-wide association studies.

Precision refers to how accurate a result set is. For example, when testing a search engine, precision indicates how many of the results returned in response to a query are relevant (see Eq. 2.1). *Recall*, on the other hand, refers to how much of the relevant information is contained in the result set (see Eq. 2.2). With a search engine evaluation, recall refers to the number of relevant items in the results set compared to all possible relevant items. Usually, a trade-off can be observed between precision and recall. When a system is tuned to be more precise, the recall goes down. When a system is tuned for higher recall, the precision goes down. Because of this trade-off, it is often difficult to compare information systems and the F-measure is sometimes preferred because it combines both measures (see Eq. 2.3).

$$Precision = \frac{\# \text{ retrieved and relevant items}}{\# \text{ retrieved items}} \quad (2.1)$$

$$Recall = \frac{\# \text{ retrieved and relevant items}}{\# \text{ relevant items}} \quad (2.2)$$

$$F - \text{measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2.3)$$

As noted above, the *F-measure* is a weighted average of precision and recall. In the best possible scenario, when both precision and recall are perfect, the F-measure's value is 1. In the worst possible scenario, when precision or recall is 0, the F-measure's value is 0. For example, assume there is a set of records and a subset of those is considered to be relevant to an information request. A search engine has been constructed to retrieve those relevant records from the entire set with a given query. When the query is executed, the search engine retrieves all relevant documents and no others. In this best case, precision is 100% and recall is 100%, resulting in an F-measure of 1. However, had the search engine retrieved no relevant documents, precision and subsequently the F-measure's value would be 0.

Classification Measures

Many algorithms and information systems are developed with the intent to automatically categorize or label people, records or other data items. They perform a type of prediction called classification. Based on a set of rules, a label is automatically assigned to each data point. The rules can be acquired with machine learning algorithms, based on codified expert knowledge or with statistical calculations.

There may be different kinds of labels that can be assigned. For example, algorithms can be trained to distinguish between two classes for brain images displaying a mass and label them as either benign growths or tumors.

Evaluating such a system entails finding out whether the label was assigned correctly. Several measures can be used for this. In the informatics community, *accuracy* is the most common measure used to evaluate the correctness of classification. Measuring accuracy requires that a gold standard is available to compare the algorithm outcome against the correct solution. Accuracy then refers to the percentage of items correctly classified in an entire set as compared against the gold standard (see Eq. 2.4). For example, if there is a set of mammograms with a subset known to display a tumor, then accuracy of an algorithm would be evaluated by calculating how many mammograms were correctly classified as containing a tumor or not. In medical informatics, accuracy is described using four more specific metrics: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). In addition, other derived measures are commonly used that form combinations of these four, namely specificity and sensitivity. However, this nomenclature is useful when only two classes are being distinguished. When there are more classes, a confusion matrix is a better choice. Each of these measures is described in detail below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.4)$$

The *True Positive*, *True Negative*, *False Positive* and *False Negative* classification can be used with a gold standard for evaluating system, algorithm and even expert classifications. For example, assume a neural network has been trained to distinguish between cancerous and normal tissue on images displaying masses. After training the algorithm, it is evaluated using a dataset for which the correct outcome is known. If such an algorithm classified an image as showing cancerous tissue, this is considered a True Positive (see Eq. 2.5) when that is a correct decision. However, if the image did not show cancerous tissue, the algorithm would be wrong and this would be called a False Positive (see Eq. 2.7). Similarly, if the algorithm correctly classified an image as not showing cancerous tissue, this is called a True Negative (see Eq. 2.6) if this was the correct decision. Again, if the tissue had been cancerous after all, then the algorithm was wrong and the decision would be called a False Negative (see Eq. 2.8). While the terms TP, TN, FP and FN are most commonly used in medicine, there are synonymous terms which are more commonly used in psychology: Hit instead of TP, Correct Rejection instead of TN, False Alarm instead of FP and Miss instead of FN. Table 2.1 shows an overview of this classification.

$$\text{True Positive (Hit)} = \text{an instance correctly labeled as} \\ \text{belonging to a group} \quad (2.5)$$

$$\text{True Negative (Correct Rejection)} = \text{an instance correctly labeled} \\ \text{as not belonging to a group} \quad (2.6)$$

Table 2.1 Demonstration of classification measures for two classes

Actual outcome	System predicted outcome		Total
	Diabetes	No diabetes	
Diabetes	TP	FN	
No diabetes	FP	TN	
Total			

$$\text{False Positive (False Alarm)} = \text{an instance incorrectly labeled as belonging to a group} \quad (2.7)$$

$$\text{False Negative (Miss)} = \text{an instance incorrectly labeled as not belonging to a group} \quad (2.8)$$

As noted above, in medical informatics the TP, TN, FP and FN are also combined into two additional, derived metrics. *Sensitivity* (see Eq. 2.9), also called the *detection rate*, refers to how well positive instances can be detected. *Specificity* (see Eq. 2.10) is a reference to how correctly that decision is made; in other words, how often the test correctly detects cancerous tissue.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2.9)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (2.10)$$

In medicine, these specific measures help evaluate systems and tests relative to their intended use. For examples, for a serious and life-threatening disease such as HIV, a diagnosis of that disease is devastating, stressful and will lead to a battery of follow-up tests and treatments. Therefore it is essential that the specificity is high to avoid unnecessary stress and treatment. On the other hand, missing such a serious disease is disastrous too, and so the test should be sensitive enough. In addition to differences based on the seriousness of the disease, the timing of diagnosis also plays a role. Sometimes, clinicians want to use tests with high sensitivity for screening or early detection and follow up with tests with high specificity to confirm a diagnosis. In addition to providing such evaluations, it is also common to compare new tests with existing tests for sensitivity and specificity, as was done, for example, for a genetic algorithm/support vector machine combined algorithm to find protein-protein interactions [14], scoring systems to predict blood stream infection in patients [15] or a search engine to search and identify cancer cases in pathology reports [16].

When there are only two classes possible, the TP, TN, FP, FN notation is sufficient. However, in machine learning algorithm evaluations, a different notation is used that can easily be extended to more than two classes: a *confusion matrix* or *contingency table*. A confusion matrix is a better tool to evaluate the results with

Table 2.2 Demonstration of classification measures for four classes

Actual outcome	System predicted outcome				Total
	Type 1 diabetes	Type 2 diabetes	Gestational diabetes	No diabetes	
Type 1 diabetes	X1	a	b	c	Y1
Type 2 diabetes	d	X2	e	f	Y2
Gestational diabetes	g	h	X3	i	Y3
No diabetes	j	k	l	X4	Y4
Total	Z1	Z2	Z3	Z4	

multiple classes. It can be seen as the extension of the foursome discussed above adjusted to multiple outcome classes. A confusion matrix provides accuracy numbers per class and also provides details on how errors are classified. When evaluating systems or algorithms, these measures provide an indication of how well algorithms can be used to complete tasks that are often labor intensive and boring for humans. It should be noted that the human classification process itself is seldom error free.

For example, assume a system has been trained to predict who will get diabetes based on data in electronic medical records. It is possible to have any of three types of diabetes or to have no diabetes at all. After running the classification algorithms on the training data (training the model), the outcome is evaluated using test data. The algorithm classification of this test data is compared against the actual outcome. A confusion matrix can provide a detailed evaluation. For example, if very many errors are made in classifying instances as gestational diabetes, one of the conditions, this would not be clear from reporting true positives and true negative but it would be apparent in the confusion matrix.

Table 2.2 shows how such an evaluation looks for a classification with four possible outcomes. The true positives in the case of four labels can be found on the diagonal (X1, X2, X3, X4). The other numbers in the matrix (a-l) show the errors and provide details of how the records are classified incorrectly. For example, there are *a* records classified as belonging to people with Type 2 diabetes that should have been diagnosed with Type 1 diabetes. The totals (Y1, Y2, Y3, Y4 and Z1, Z2, Z3, Z4) provide rates for each specific class. For example, Z1 records were predicted to belong to people with Type 1 diabetes; X1 of these were correct. And Y1 records belonged to people with Type 1 diabetes while the system predicted only X1 of these correctly.

N-Fold Cross-Validation

When the rules to classify or label instances, such as patients, images or records, are learned by algorithms without human intervention (machine learning), a dataset can be used multiple times. This is done by dividing the entire set into *n* subsets which are called folds. This is possible with machine learning algorithms because their memory can be erased from one dataset to another so that each evaluation can be seen as independent. With *n-fold cross-validation*, each data point is randomly assigned to one of *n* subsets. Once the dataset is divided, one set is set apart for

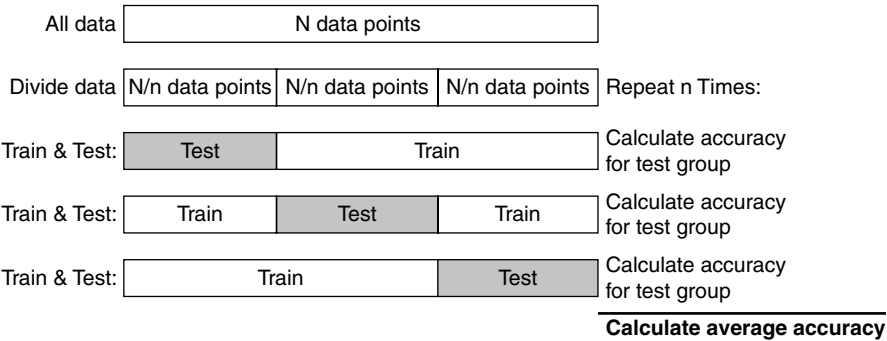


Fig. 2.2 N-fold cross-validation example (N=3)

testing and the remaining $(n - 1)$ sets are used for training the algorithm. Once training has been completed, the algorithm is tested on the set that was kept apart. The accuracy of the algorithm predictions is calculated for that test set. Note that the test set was not contained in the training data for the algorithm, and so this evaluation is for data that the algorithm will not have encountered during training. This process is repeated as many times as there are folds (n). The accuracy is then averaged over all folds. N-fold cross-validation cannot be used when rules are learned and encoded by a human because a person is unable to ignore the previous round of training and would be biased by previous interactions with the data.

As an example, Fig. 2.2 shows the process of evaluating a classification algorithm using 3-fold cross-validation ($n=3$). Assume a dataset is available that contains N mammograms. Each mammogram shows a mass which is known to be benign or not based on a biopsy that was conducted. The entire dataset is labeled with the correct answer: benign or not benign. The algorithm is developed to read those images and learn whether the mass is benign or not. The goal is to use such an algorithm as a second opinion for new mammograms showing a mass. With 3-fold cross-validation, the dataset is split into three equal groups: each subset has one-third of the mammograms. Training of the classification model¹ and subsequent testing is repeated three times. The model is first trained using the combined data from two subsets and tested on a third. This provides the first accuracy measure. This process is completed three times. Each fold or subset serves as the test set once. The final accuracy is the average of the three accuracy numbers.

There are several advantages to this approach. First of all, the evaluation is repeated n times and does not depend on a single outcome. If each evaluation results in high accuracy scores, this is a good indication that the algorithm is stable. The test dataset may contain many examples the algorithm was not trained for and so the results would be worse than expected with other datasets. It is also possible that the test dataset contains mostly examples that the algorithm is highly trained for and

¹Such training is done with supervised machine learning techniques also called classifiers. For example, a feedforward/backpropagation neural network (FF/BP), decision tree or other type of classifier could be used.

so the results will be better than expected with other datasets. To avoid reliance on one dataset, the common evaluation approach is to do n -fold cross-validation or n evaluations using the same dataset. In addition to a balanced evaluation, this approach is also useful for training the algorithm because it avoids over-fitting of the data.

The random assignment to n -folds may be adjusted to ensure that each subset is sufficiently representative of the entire dataset. This should be considered when there are classes in the dataset that appear very infrequently. If none of the examples belonging to this rare class was present in the training dataset, it would be impossible for the algorithm to learn its characteristics. For such datasets, stratified sampling would provide a more balanced approach of division into folds and would ensure that each subset has examples from each possible class.

For example, assume there is a dataset consisting of electronic health records (EHR) and researchers have developed an algorithm to predict which type of diabetes (if any) is associated with different characteristics. Each record in the dataset has a label associated with it indicating whether the person has diabetes and which type: Type 1 diabetes, Type 2 diabetes, gestational diabetes or no diabetes. This dataset is used to train the new algorithm to learn people's characteristics and if they are associated with diabetes. Since the number of people with gestational diabetes may be very small, the researchers should ensure that at least a few of these cases appear in each fold. Which ones will appear in a specific fold can be decided by random assignment to ensure that all records of women with gestational diabetes do not end up in one fold. If they were to end up in one fold, the algorithm would not be able to generalize information of this type in the evaluation. Software packages, e.g., Weka [17], usually include this option for evaluation and let the user choose the number of folds to be used.

Counts

Counts are a simple and often effective approach to evaluation. In many cases, several critical events can be counted that are indicative of the quality of a system. Some of these events are desired, while others need to be avoided. For example, with a decision support system, it is important to support correct decisions. In other cases, for example, a medical records system, it is important to reduce the amount of clicking and page visits needed to complete a record. Counting such events can contribute significantly to understanding why a system is accepted by users or not.

Sometimes, it will be necessary for the experimenter to observe the users' interactions with the system. However, be aware that few users will act in the same way when observed as when they are not. When people know they are being observed their behavior is not always completely natural. Luckily, counts often can be conducted by logging interactions with a system. For example, when conducting a study to evaluate a user interface it is possible to establish which links should or should not be followed because the assigned task and required outcome are known. In these cases, links followed in error can be tracked and they can be seen as indications that the interface is not intuitive.

Since counts are simple evaluation measures, they may not tell the entire story and it is best to complement them with other measures. The researcher also should consider how much emphasis should be put on these counts when explaining the study to the participants and when interpreting results. For example, with new software interfaces, many users will explore the new software and click many menus and options. This is not necessarily an error. If every click counts, the participants in the study should be aware of this so they focus on the task at hand without exploring.

Usability

Usability is an important characteristic of every information system. Today, an information system that could be considered perfect in all aspects but that was not usable or user friendly would not be considered acceptable. There are different approaches to measuring usability. A very popular approach in informatics is the use of survey based measures of usability. Validated surveys exist for this purpose. For example, the Software Usability Measurement Inventory (SUMI) developed by Kirakowski [18] contains 50 statements measuring five dimensions of usability: Efficiency, Affect, Helpfulness, Control and Learnability. Unfortunately, many researchers quickly put together a survey without taking any possible biases into account or without any validation. The conclusions that can be made using such an instrument are doubtful.

In addition to using surveys, usability also can be measured in an objective manner by counting events, counting errors or measuring task completion. Many different measures can be constructed in this manner. The more interesting measures compare different users on their training or task completion times. For example, when an information system is extremely usable, there should be little training time required. Some good examples of such systems can be found in museums where the goal is to have zero training time information systems. Visitors to the museum can walk up to a computer and use the information system without any training. Another good evaluation of usability is based on the comparison between novice and expert users. For example, REU or Relative User Efficiency (see Eq. 2.11), as described by Kirakowski [11], is the time an ordinary user would need to complete a task compared to the time needed by an expert user:

$$RUE = \frac{\text{Ordinary User Time}}{\text{Expert User Time}} * 100 \quad (2.11)$$

User Satisfaction and Acceptance

User satisfaction and user acceptance are generally related to each other, and both factors are often measured with a onetime survey. However, often users need to be satisfied with a system in the short term before the system will be accepted in the

long term. As a result, measuring acceptance becomes more meaningful when more time has passed for users to get acquainted with the system.

Almost every user study of information systems will contain a user satisfaction survey. Most of these surveys are filled out at the end of limited time interaction with the new system. Unfortunately, very few of these surveys have been validated. An exception is the Computer Usability Satisfaction Survey developed by Lewis [19]. It contains 19 items divided over 3 subscales: System Usefulness, Information Quality and Interface Quality. The 19 items are presented with a 7-point Likert scale ranging from “Strongly Agree” (score 1) to “Strongly disagree” (score 7), with a “Not Applicable” option outside the scale. The survey is fast and easy for study participants to complete [20, 21].

Many researchers and developers of information systems use their own surveys, but the usefulness, validity, objectivity and reliability of these is often questionable. It is very difficult to compose a survey that measures specific constructs. Wording of questions and answers will affect the results, biases will affect the results and many surveys will be incomplete or not measure what they intend to measure. It is therefore much better to use a survey that has been validated. This makes it possible to compare with other systems and be reasonably sure that the answers will be meaningful. Those researchers intending to develop their own survey should consult and learn the basics of psychometry, the field in psychology concerned with composing and conducting surveys. Evaluation of information systems is much more straightforward than psychological studies, since there deception is seldom necessary to measure constructs of interest. However, being knowledgeable on how to avoid biases, how to conduct a valid survey and how to avoid overlapping items measuring the same construct will improve any survey.

Processing Resources

Time and memory are two processing resources that are very suitable for evaluating individual algorithms. In algorithm development, there is a common trade-off between time and memory needed to complete a task. If all other factors are equal, a shorter processing time usually requires more memory, while using less memory will usually result in more processing time being needed.

A complexity analysis is a formal approach to evaluating an algorithm’s runtime or memory usage in comparison to the input given to the algorithm. *Big-O analysis* is a commonly used complexity analysis. It provides an evaluation of an algorithm independent of the specific computer or processor being used. This analysis is important since information systems may show very good evaluation results, but may be too complex to be used in a realistic setting. The “O” refers to “in the Order of ...” [22]. The analysis is used to define the worst case or average case of an algorithm’s hold on resources (time or memory) in relation to the given input (N). It is very useful when describing algorithms with varying performance. The analysis focuses on the loops in an algorithm and allows comparison of space or processing time in terms of order of magnitude.

For example, if the input datasets consist of N items, then an algorithm that runs in $O(N)$ time is an algorithm that runs in linear time: the amount of time needed to complete is directly related to the number of input items. However, an algorithm that runs in $O(N^2)$ needs much more time to complete. It needs $N \times N$ or N^2 to complete processing a dataset with N input items. This analysis provides a simple measure that is expressed as an order of magnitude. For example, it does not matter whether the time increase was 6x or 200x times the input X , both would be noted as $O(N)$. Similarly, if the algorithm requires $(x^2 + 10x)$ time for an input of X , it would be noted that the algorithm runs in $O(N^2)$. For a detailed description of how to conduct this analysis, the reader is referred to Nance and Naps [22] or other introductory computer science books covering algorithm and data structure analysis.

Although computer speed and memory have become a commodity for many simple applications, in medicine and biomedicine there are several applications where such analysis is essential, for example, visualization of protein folding or image analysis of moving organs. The analysis is essential in the development of many algorithms that will become part of sophisticated software packages. For example, Xiao et al. [23] describe the reduced complexity of such a new algorithm used in tomography, a technique used to reconstruct data for many types of medical scans. The algorithm complexity was reduced from $O(N^4)$ to $O(N^3 \log N)$.

Confounded Variables

Two variables are confounded when their effects cannot be separated from each other. When designing user studies, this problem is encountered when there is a variable other than the independent variable that may have caused the effect being studied. The variable causing the confounding reduces the internal validity of the study [24]: one cannot say for sure that the treatment, i.e., the independent variable, caused the effect. This variable changes with the experimental variable but was not intended to do so. As a result, the effect of the treatment cannot be attributed to the independent variable but may well have been caused by the other variable, the confounder. In some cases, confounded variables are difficult to avoid. Consider, for example, an experimenter effect. Most participants who voluntarily participate in user studies wish the researchers well and hope they succeed. If they know which condition the experimenters favor, they may evaluate it more positively.

To avoid having confounded variables, it is important to take possible bias into account, to make sure participants are assigned randomly to experimental conditions and to verify that the independent variable is the sole element that can be causing the effect. Consider the example of a weight loss support system that uses text messages and that is compared against another support system that does not use text messages. For practical reasons, the researchers may decide that it is easier to assign the text message condition to subjects who already possess a mobile phone because it makes it easier to run the study. Study participants without a mobile phone are assigned to the condition that does not use text messaging. With this design, it is very probable that the researchers have introduced confounded variables which

make it impossible to conclude that any differences in weight loss between the two groups can be attributed to the text messaging system. For example, compared to participants who possess a mobile phone, participants without mobile phones may belong to a less affluent demographic group with a different lifestyle, different access to health information and a different attitude to healthy living.

Several approaches can be taken to avoid losing validity due to confounded variables. Naturally, the best approach is to design the experiment such that confounding variables are avoided. When this is not possible, other precautions can be taken that would allow the researcher to draw valid conclusions. First, one can use demographic data to measure possible confounding. When conducting a user study, it is useful to collect additional demographic data so that expected confounding can be objectively evaluated. Systematic differences in such variables between conditions would indicate confounding variables. However, if experimental groups do not differ on these measures, the study is strengthened. For example, when evaluating the weight loss support system, researchers could collect information about education levels, reading comprehension and even attitudes toward healthy living. They could then compare whether there are systematic differences between the experimental groups with regard to these variables.

A second approach to avoid making conclusions based on confounded variables is to include complementary outcome measures in the study. When such complementary measures are used, contradictions in their outcomes may be an indication that there are confounded variables. In studies with the elderly, the author and her students observed such confounding between the experimental condition, a new versus an old user interface, and an experimenter effect. Usability was evaluated with subjective and objective measures. The study participants mentioned they enjoyed their time working with the graduate students and wanted to help them graduate. This was clearly visible in the participants' survey ratings; it could be concluded that the participants *loved* the new system compared to the old system. However, the objective measures used did not show any benefit of the new system over the old.

A third approach is to improve the design to avoid possible confounding. With new information technology, such as telemedicine or virtual reality, many more options to improve study designs and avoid bias are available. For example, in studies where communication styles or other similar characteristics are correlated with other personal characteristics, there could be confounded variables. Mast et al. [25] evaluated the impact of gender versus the communication styles of physicians on patient satisfaction. In most cases, the communication style is very much related to gender. As a result, it is extremely difficult to pinpoint which of the two affects patients' satisfaction. Information technology helped disentangle these variables. Using a virtual physician, the researchers were able to control each variable independently and measure the effects on patient satisfaction. The results showed it was the caring style, not the gender, which affected satisfaction.

Finally, other designs and analyses can take potential confounding into account and even use it. Multivariate statistical analysis can be used to take the effect of confounded variables into account [1]. In other cases, some controlled confounding

is sometimes integrated into the design to reduce the number of participants needed to complete a study. These complex study designs, which are more common in the behavioral sciences, are discussed, for example, in Chap. 13 of Kirk [2].

Bias Caused by Nuisance Variables

Nuisance variables are variables that add variation to the study outcome that is not due to the independent variables and that is of no interest to the experimenter. They introduce undesired variation that reduces the chance of detecting the systematic impact of the independent variable. Even if there is a true difference between the experimental conditions, it may be undetectable if there is too much variation unrelated to the experimental conditions. If this type of variation is unsystematic, it is called *noise*. When the variation is systematic, it is called *bias* [2]. If this bias also coincides with the levels of the independent variable, then the independent variable and the bias are confounded variables. Blocking, which can be used to counter some bias, is explained in Chap. 5 (blocking one nuisance variable) and Chap. 6 (blocking multiple nuisance variables). Other countermeasures to such bias are discussed in Chap. 13.

For example, assume a researcher is interested in the effects of caffeine on alertness in executive MBA classes, which are usually held in the evenings. It has been decided that the independent variable will be the number of cups of coffee: 0, 1 or 5. The dependent variable is the alertness in class and will be measured with a self-administered survey. All students in the class have agreed to participate in the study. The researchers realized that some students may have had a big meal before attending class, while others may hold out until after class. So, a nuisance variable in this study consists of eating (or not) a big meal before going to class. This variable will affect the outcome because participants will be less alert after that big meal. Thus, this nuisance variable needs to be controlled, for example, by giving all students a big meal before class. Then, any change in the measured alertness cannot be attributed to whether or not a meal was eaten.

Bias is a well studied topic and there are several famous experiments demonstrating bias. Many types of bias have received their own names over the years because they commonly appear in studies. Learning about these different types of bias will help the researcher design a better experiment by countering them as much as possible. Controlling the nuisance variables and the bias will increase the validity of the experiment and also the chances of discovering a true effect.

Subject-Related Bias

One subject-related bias is the *good subject effect*. This effect is the result of study subjects who act in a certain way because they know they are participating in a study. There are three types of good subject effects. The first type is the effect that is most often associated with being a good subject, which involves some form of

altruism. Such subjects are trying to give the researcher what he wants; they act in a way they believe is appropriate for the study and the treatment condition. There are two problems that result from this bias. The first is that the subjects do not behave naturally but act, i.e., they alter their behavior. The second is that the change in behavior is based on what the subjects believe or understand about the study. However, their understanding may be incomplete or wrong. With the emphasis on participants' satisfaction in many information systems' evaluations, this effect must be controlled. When the researcher is a doctoral student doing a dissertation study, subjects may be especially inclined to help the student with his research.

The second type of good subject effect is due to subjects who change their behavior as the result of a desire to comply with an *authority*. The subjects may feel that the researcher knows best. As a result, the study subjects may not feel qualified to argue, disagree or even voice an opinion. This is particularly true in medicine where the clinic staff is seen as the authority by patients, and so this effect may affect studies where patients or their caregivers are the subjects. In addition, there is also a clear hierarchy among the clinical staff members that may lead to the same type of bias and affect results in studies where the participants are clinical personnel.

Finally, a third type of good subject effect, the *look good effect* or the *evaluation apprehension effect*, is related to how the study subjects feel about themselves. Subjects who participate in experiments are often self-aware. They know they are being observed and they want to look good as a person. It is not clear how much this effect plays a role when evaluating software where behavior based measures, such as the number of errors made, are added to belief based measures, such as how good one feels or how much pain one feels. However, researchers should be aware of the effect and take it into consideration when forming conclusions.

Several studies have been done to evaluate and compare these biases. In the 1970s, a series of carefully controlled experiments was conducted to compare the different good subject effects [24, 26–28]. One of the goals of these studies was to tease apart the different origins of the good subject effect and discover the most influential reason. The evidence points mostly in the direction of a look good effect. When the look good and the altruism effect are competing factors, it seems that looking good becomes more important and is the main motivation of participants. A related study evaluated the effect of altruism in a medical context [29]. Subjects in three studies were interviewed about their reasons for participating in a study. There was no personal gain to participants in two of the three studies. The researchers concluded that the subjects' participation was mainly for the greater good. However, in each of these three studies, the participation may have been confounded by a personal look good feeling. There was no conflict in these studies between altruistic and look good feelings and so the differences between these two could not be measured. The authors also acknowledge this potential influence and refer to it as the potential to receive 'a warm glow' from participating. They refer to related work that looks at this type of altruism from an economic perspective via the donation of funds [30] instead of research participation.

A second subject-related bias is the *selection bias* or the *volunteer effect*. This bias may be encountered when the participants are volunteers [31] because there

may be traits common in volunteers that are different from people who do not volunteer. Naturally, this may influence the findings. For example, the healthy volunteer bias [1] refers to overrepresentation of healthier participants in a study. This bias is especially pronounced in longitudinal studies. In addition to health, other personal characteristics may be overly present in a group of volunteers. In a formal study of selection bias, Adamis et al. [32] found that the method of getting elderly patients' informed consent for a mental health study had an enormous influence on the size and characteristics of the sample of participants. A formal capacity evaluation procedure followed by informed consent was compared to an informal procedure, which was "the usual" procedure, where informed consent and evaluating capacity were mingled. The formal procedure led to a smaller group of participants with less severe symptoms who agreed to participate and who were considered capable of making that decision.

Finally, a third well recognized subject-related bias is the *authorization bias*. This is the bias found when people need to authorize the use of their data for an observational study that does not require active participation in the study. Variations have been found when the informed consent process included a request for people to agree having their data included in a study. Kho et al. [33] found that differences existed between groups who consented and those who did not, but there was no systematic bias across all studies reviewed.

In addition to these known biases, there are other study participant characteristics that may form a bias and influence the study. Although these are more difficult to control, measuring the relevant characteristics may be helpful to identify outliers. For example, language skills are important. It is vital that participants understand the questions asked in a question-answer task or the items presented in a survey. Physical characteristics also should be considered. Participants may have difficulty using a mouse or clicking on scroll bars due to poor eyesight or tremors. Religion and political belief also may influence attitudes during testing. Measuring the study participants' relevant characteristics may serve as a pre-selection tool and help exclude non-representative people from participating. For example, in a study to measure the impact of various writing styles on understanding of health educational pamphlets, Leroy et al. [34, 35] excluded people with any medical background. Because the information presented was at the layman's level, any medical knowledge would have influenced the study's measurements of understanding. Other examples are required physical abilities to conduct the study as intended. For example, when designing visualization tools using 3D displays or colors in visualization, it is necessary to test the ability of participants to perceive 3D displays or see different colors.

Experimenter-Related Bias

Experimenter-related bias or *experimenter effects* are a type of bias related to experimenter behaviors that influence the outcomes or data. In most cases these effects are unintentional, and good experimental design or use of information technology

can control many. There are two types of experimenter effects [24]: non-interactional and interactional experimenter effects. *Interactional experimenter effects* are those that lead to different behaviors or responses in the study participants due to the experimenter during the course of the study. The *non-interaction effects* are related to actions by the experimenter after the interaction with participants has been concluded.

Many interaction effects are not the results of dishonesty but of subtle cues that are given by the experimenter and picked up by the study participants. For example, an extra nod or more in-depth questions given during interviews may lead to better, longer or higher quality responses. An early and famous example is the Clever Hans effect. This effect is based on Clever Hans, a horse that could count. Oskar Pfungst determined that cues from his owner, who wasn't even aware of giving such cues, were responsible for the horse's abilities [36–38]. In the case of the horse, behaviors such as leaning forward when the count wasn't done yet and leaning backward when it was done helped the horse count. Current examples can be found in the many home videos of 'smart' pets. Another interactional experimenter effect is caused by personal characteristics of the facilitator that influence the participants. Gender, personal interaction styles, race, language skills and even personal hygiene are some of the many characteristics that may affect the outcome. Along the same lines, the topics or tasks covered may have an effect. People may be sensitive and prefer to avoid specific topics or may already be more or less biased before the experiment.

In addition to the biases that influence the interaction with participants, there exist *observer effects* or *non-interactional experimenter effects* that are the result of experimenter actions once the study has been executed. For example, different evaluation outcomes by different observers or evaluators demonstrate this effect. One evaluator may apply more lenient coding for the output of a new system. In most cases, this is unintentional.

Design-Related Bias

Some biases are inherent to the particular design used and cannot be attributed to subject or experimenter characteristics. One such design-related bias is the *placebo effect*, which is well known in medicine. It is an effect that can be attributed to participants believing that they are getting the treatment, even if they are not in reality getting any treatment.

In medicine, the placebo effect is significant when considering the importance of belief and mind over matter. Placebos are often used as the control condition in double-blind studies. Participants are given an inert pill, injection or treatment that looks the same as the experimental treatment. In some cases, the control condition has been found to have a positive effect even though no real treatment was provided. This placebo effect is therefore understood to be the effect of a control condition that is meant to be a placebo, without effect, but which has an effect after all. However, a note of caution is needed. The placebo effect found in medical studies

may sometimes be more than a response bias and may be based on actual change in the brain or body [39].

In informatics, the use of a placebo control condition is difficult to accomplish. When evaluating an information system, it is not easy to organize a placebo condition where all interactions with a system are the same except for some interactions provided by the new system. Therefore, in informatics a different control condition is generally used: a baseline to compare the next system against. Since the baseline is usually the existing situation and often includes an existing system, it is incorrect to speak of a placebo effect.

A second design-related bias is a *carryover effect* or *contamination effect* which can be found when there are effects from the control condition that carry over to the experimental condition. It shows clearly how no experimental design is completely free of bias and the importance of choosing the best design for each study. The carryover effect is often a worry with within-subjects designs where study participants participate in multiple experimental conditions. For example, consider a validation study for a survey where each participant fills out two versions: the existing paper version and the new computerized version. The results of both versions are compared. With a within-subjects design, participants first start with one version of the survey and then complete the second version. However, it is very possible that experience with the first version influences the results of the second. For example, participants may try to repeat the same answers without really reflecting on the survey questions. This bias can be countered by counterbalancing the orderings, which is discussed in Chaps. 5 and 6.

A third design-related bias is the *second look bias*. This term, used in particular in medicine, refers to an effect similar to a carryover effect. It is encountered when study participants view data or an information system more than once and each time under different experimental conditions [34]. This bias especially needs to be taken into account with studies adopting a within-subjects design. Participants have an initial interaction with the system and learn about using the system or form an opinion about it. This first interaction will influence the second interaction. When they have a second look at the system, they may already have a better idea of how to use it efficiently, they may be less inclined to search for functions and so be more efficient (or give up faster on a task) or they may believe the system to be useless. This bias also can originate when the same data is reused over different experimental conditions.

Hawthorne Effect

The Hawthorne effect is one of the most famous biases and its discovery resulted in a set of books, commentaries, articles and many criticisms. In short and in its most simple terms, the Hawthorne effect is a change in behaviors that is supposed to be due to the knowledge that one is being measured or monitored.

The origin of this well known effect lies in a series of studies, conducted over several years, 1927–1932, at the Hawthorne Works of the Western Electric Company

in Chicago [40, 41]. The studies started with five participants in the first few months, but soon many more workers, as many as 20,000 in total, participated and were interviewed. The experiments focused on worker conditions and efficiency by looking at changes such as rest pauses, shorter working days and wage incentives, among many other conditions. A general conclusion was that the changes in productivity were more due to the extra attention received and the knowledge that productivity was measured. However, these experiments were conducted in very different conditions from today: the tasks were monotone, the participants were females and many workers in those days had low levels of education. As such, there has been much debate over the years and caution is needed when generalizing these results [42]. It is doubtful the explanation is always as simple as a *measurement* or *attention effect*. Gale [43] shows how the context can help explain these effects.

The emerging use of informatics and technology in the realm of persuasion makes this effect a current topic again. It is said that 100% compliance, for example, with hand washing, can be accomplished with the placement of just one camera. Regardless of how simple or complex the effect may be, the term Hawthorne effect has stuck and is found frequently in educational and healthcare settings. For example, Leonard and Masatu [44] evaluate the impact of the presence of a research team on quality of care in Tanzania. They conclude that a Hawthorne effect is present with an increase of quality at the beginning of the team's presence, which over time gradually levels off to the same original levels. Conducting a randomized trial is no guarantee against a Hawthorne effect. Cook et al. [45] used self-reporting to measure the effects of an online versus print based diet and nutrition education program. They found significant improvements in both groups, regardless of the experimental conditions, and suggest this may be due to a Hawthorne effect.

Other Sources of Bias

There are other sources of variance and bias that cannot easily be categorized. One such source of variance that may result in bias is the availability of *identity information*. When doing studies, the identifying patient information is usually not present for privacy reasons. However, the identity of the treating physician may have an impact, especially when using historic cases. For example, when study participants consist of medical personnel, they may be acquainted with the treating physicians of the cases used in the study and they may put more or less trust in their own decisions for the case when seeing the decision by the known treating physician. This will influence how they work with each case and their willingness to make different decisions than the ones described in the case. Similarly, the study participants may have knowledge of the typical patients the treating physician works with and this may change their assumptions about the case and options for treatment.

The study environment factors are the characteristics of the environment that may affect the outcome of a study: characteristics associated with the room where the study is conducted, including noises, smells and temperature. Conducting experiments in a noisy room may prevent participants from concentrating and will affect

many types of outcomes. Smelling the kitchen from a local restaurant may lead to participants hurrying through a self-paced study if they were hungry. An experimenter's personal characteristics or habits may affect the study. Some people are not aware of a personal smell and participants may feel uncomfortable when in close proximity during the study. Others may click their pen continuously during a study, annoying the participants. These environmental factors may introduce additional variance and affect the potential to see an effect of the experimental treatment. For example, when evaluating a visualization algorithm of health text [46], the author found that results from a first pilot study did not look promising. When scrutinizing the comments made by participants in response to an open question requesting comments on the algorithm, one of the subjects remarked that there was too much noise in the room when conducting the study. This led to a close look at the data for each experimenter which revealed that weaker results were attained by one of the two experimenters. It was discovered that after explaining the purpose of the study, this experimenter would spend the duration of the study time chatting with friends.

The unwanted effects resulting from bias can have serious consequences. Biases of a similar nature across all conditions may prevent the study from showing any results. Such studies may lead to a halt in follow-up research because no effect was found. When the bias arises in one but not other conditions, the consequences may be more serious and erroneous conclusions may be reached. Different systems or algorithms may be developed based on results from such studies.

References

1. Starks H, Diehr P, Curtis JR (2009) The challenge of selection bias and confounding in palliative care research. *J Palliat Med* 12(2):181–187
2. Kirk RE (1995) Experimental design: procedures for the behavioral sciences, 3rd edn. Brooks/Cole Publishing Company, Monterey
3. Rosson MB, Carroll JM (2002) Usability engineering: scenario-based development of human-computer interaction. interactive technologies. Morgan Kaufman Publishers, San Francisco
4. Gorini A, Pallavicini F, Algeri D, Repetto C, Gaggioli A, Riva G (2010) Virtual reality in the treatment of generalized anxiety disorders. *Stud Health Technol Inform* 154:39–43
5. Sidani S, Miranda J, Epstein D, Fox M (2009) Influence of treatment preferences on validity: a review. *Can J Nurs Res* 41(4):52–67
6. Friedman CP, Wyatt JC (2000) Evaluation methods in medical informatics. Springer-Verlag, New York
7. Maisiak RS, Berner ES (2000) Comparison of measures to assess change in diagnostic performance due to a decision support system. In: AMIA Fall Symposium, AMIA, pp 532–536
8. Goodman CS, Ahn R (1999) Methodological approaches of health technology assessment. *Int J Med Inform* 56:97–105
9. Kushniruk AW, Patel VL (2004) Cognitive and usability engineering methods for the evaluation of clinical information systems. *J Biomed Inform* 37:56–76
10. Brender J (2006) Handbook of evaluation methods for health informatics (trans: Carlander L). Elsevier Inc, San Diego
11. Kirakowski J (2005) Summative usability testing: measurement and sample size. In: Bias RG, Mayhew DJ (eds) Cost-justifying usability: an update for the Internet Age. Elsevier, Ireland, pp 519–553

12. Yousefi-Nooraie R, Irani S, Mortaz-Hedjri S, Shakiba B (2010) Comparison of the efficacy of three PubMed search filters in finding randomized controlled trials to answer clinical questions. *J Eval Clin Pract* [Epub ahead of print]. doi:10.1111/j.1365-2753.2010.01554.x
13. Kullo IF, Fan J, Jyotishman Pathak, Savova GK, Zeenat Ali, Chute CG (2010) Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 17(5):568–574
14. Wang B, Chen P, Zhang J, Zhao G, Zhang X (2010) Inferring protein-protein interactions using a hybrid genetic algorithm/support vector machine method. *Protein Pept Lett* 7(9):1079–84
15. Apostolopoulou E, Raftopoulos V, Terzis K, Elefsiniotis I.(2010). Infection probability score, APACHE II and KARNOFSKY scoring systems as predictors of bloodstream infection onset in hematology-oncology patients. *BMC infectious diseases*, 26(10):135
16. Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney D (2007) The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. *J Am Coll Surg* 205(5):690–697
17. Witten IH, Frank E (2000) Data mining: practical machine learning tools and techniques with Java. The Morgan Kaufmann Series in data management systems. Morgan Kaufmann, San Francisco
18. Kirakowski J (1996) The software usability measurement inventory: background and usage. In: Jordan P, Thomas B, Weerdmeester B (eds) Usability evaluation in industry. Taylor and Francis, UK
19. Lewis JR (1995) IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J Hum Comput Interact* 7(1):57–78
20. Miller T (2008) Dynamic generation of a health topics overview from consumer health information documents and its effect on user understanding, memory, and recall. Doctoral Dissertation, Claremont Graduate University, Claremont
21. Leroy G, Chen H Med Textus (2002) An ontology-enhanced medical portal. In: Workshop on information technology and systems (WITS), Barcelona
22. Nance DW, Naps TL (1995) Introduction to computer science: programming, problem solving, and data structures, 3rd edn. West Publishing Company, Minneapolis/St. Paul
23. Xiao S, Bresler Y, Munson DC, Jr (2003). Fast Feldkamp algorithm for cone-beam computer tomography. In: 2003 international conference on image processing, IEEE, 14–17 September 2003, vol 813, pp II - 819–822, doi:10.1109/ICIP.2003.1246806
24. Rosenthal R, Rosnow RL (1991) Essentials of behavioral research: methods and data analysis. McGraw-Hill, Boston
25. Mast MS, Hall JA, Roter DI (2007) Disentangling physician sex and physician communication style: their effects on patient satisfaction in a virtual medical visit. *Patient Educ Couns* 68:16–22
26. Sigall H, Aronson E, Hoose TV (1970) The cooperative subject: myth or reality? *J Exp Soc Psychol* 6(1):1–10. doi:doi:10.1016/0022-1031(70)90072-7
27. Adair JG, Schachter BS (1972) To cooperate or to look good?: the subjects' and experimenters' perceptions of each others' intentions. *J Exp Soc Psychol* 8:74–85
28. Rosnow RL, Suls JM, Goodstadt BE, Gitter AG (1973) More on the social psychology of the experiment: when compliance turns to self-defense. *J Pers Soc Psychol* 27(3):337–343
29. Dixon-Woods M, Tarrant C (2009) Why do people cooperate with medical research? findings from three studies. *Soc Sci Med* 68(12):2215–2222. doi:doi:10.1016/j.socscimed.2009.03.034
30. Andreoni J (1990) Impure altruism and donations to public goods: a theory of warm-glow giving. *Econ J* 100:464–477
31. Ammenwertha E, Gräber S, Herrmann G, Bürkle T, König J (2003) Evaluation of health information systems—problems and challenges. *Int J Med Inform* 71:125–135
32. Adamis D, Martin FC, Treloar A, Macdonald AJD (2005) Capacity, consent, and selection bias in a study of delirium. *J Med Ethics* 31(3):137–143

33. Kho ME, Duffett M, Willison D, Cook DJ, Brouwers MC (2009) Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* 338:b866. doi:10.1136/bmj.b866
34. Leroy G, Helmreich S, Cowie J (2010) The influence of text characteristics on perceived and actual difficulty of health information. *Int J Med Inform* 79(6):438–449
35. Leroy G, Helmreich S, Cowie JR (2010) The effects of linguistic features and evaluation perspective on perceived difficulty of medical text. In: Hawaii international conference on system sciences (HICSS), Kauai, 5–8 January 2010
36. Baskerville JR (2010) Short report: what can educators learn from Clever Hans the Math Horse? *Emerg Med Australas* 22:330–331
37. Rosenthal R (1965) *Clever Hans: the horse of Mr. von Osten*. Holt Rinehart and Winston, Inc, New York
38. Pfungst O (1911) *Clever Hans (The Horse of Mr. von Osten): a contribution to experimental animal and human psychology*. Henry Holt and Company, New York
39. Price DD, Finniss DG, Benedetti F (2008) A comprehensive review of the placebo effect: recent advances and current thought. *Annu Rev Psychol* 59:565–590
40. Roethlisberger FJ, Dickson WJ (1946) *Management and the worker*, 7th edn. Harvard University Press, Cambridge
41. Landsberger HA (1958) Hawthorne revisited. *Management and the worker, its critics, and developments in human relations in industry*, vol IX. Corness studies in industrial and labor relations. W.F. Humphrey Press Inc, Geneva
42. Merrett F (2006) Reflections on the Hawthorne effect. *Educ Psychol* 26(1):143–146
43. Gale EAM (2004) The Hawthorne studies – a fable for our times? *QJM Int J Med* 97(7):439–449
44. Leonard K, Masatu MC (2006) Outpatient process quality evaluation and the Hawthorne effect. *Soc Sci Med* 63:2330–2340
45. Cook RF, Billings DW, Hersch RK, Back AS, Hendrickson A (2007) A field test of a web-based workplace health promotion program to improve dietary practices, reduce stress, and increase physical activity: randomized controlled trial. *J Med Internet Res* 9(2):e17. doi:doi:10.2196/jmir.9.2.e17
46. Miller T, Leroy G, Wood E (2006) Dynamic generation of a table of contents with consumer-friendly labels. In: American Medical Informatics Association (AMIA) Annual Symposium, Washington DC, 11–15 November 2006



<http://www.springer.com/978-0-85729-621-4>

Designing User Studies in Informatics

Leroy, G.

2011, XX, 260 p., Hardcover

ISBN: 978-0-85729-621-4