2 Experimental Design and Analysis

Colin Drury, Victor Paquet and Harrison Kelly

CONTENTS

INTRODUCTION

This chapter provides the ergonomics/human factors (E/HF) practitioner and researcher with practical advice on how to design studies and analyse the resulting data to achieve effectiveness and efficiency. In the 1930s, there was a major revolution in how experiments are designed. We moved from the traditional physical sciences model of 'vary one factor at a time; keep all other factors fixed' to a philosophy that emphasises varying multiple factors in the same experiment. What made this revolution possible was the development of sophisticated statistical techniques (e.g. analysis of variance [ANOVA]) that allowed for the parsing of the effects of each factor, and their combinations, and the testing of each effect against the normal variation experienced in any experiment. This is the model for experimental design and analysis we still use in E/HF.

Within this chapter, we use a number of terms as if they were already understood by researchers and practitioners. Examples are 'experiment', 'factor', 'effect' and 'statistical techniques'. We can do this because most E/HF professionals have had courses with titles such as 'design of experiments' (DoE) at some point in their careers. This chapter will cover some familiar ground, such as multifactorial experiment designs and analysis of variance, but will hopefully extend the topic in a number of ways. First, we will provide some definitions that broaden the topic to other E/HF studies. Next, we will delineate statistical techniques for data that does not fit the usual normality

assumptions, and finally, we will add material on power calculations, effect magnitudes and meta-analysis that will at least provide sources for further reading.

WHAT IS AN EXPERIMENT?

An experiment is 'a scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact' according to the Oxford Dictionary (www.oxforddictionaries.com, 2014). Experiments differ from all other techniques in that they directly change a system, normally in the form of a manipulation of the independent variables. These independent variables are derived from research hypotheses and enable the measurement of the degree to which each independent variable impacts and affects the dependent variable of interest. The research hypotheses are usually stated in the form of a null hypothesis, which states that the independent variable has no effect, and an alternative hypothesis, which states that the independent variable has an effect. The convention is to use statistical results to either 'reject' or 'fail to reject' the null hypothesis with a degree of certainty (sometimes also reported as supporting the alternative hypothesis). Hypotheses are discussed in more depth later in this chapter.

Alternative explanations of the results in terms of coincidence are most unlikely, because the experimenter determined when and how to change the system. Thus, experiments are able to detect and infer causality with some degree of confidence. A typical designed experiment is that of Drury et al. (2008) who studied the effects of workplace posture on performance and comfort in a security screening task. They used three postures at the airport workplace for screeners viewing x-ray images of baggage potentially containing threat objects. Screeners could be either standing, sitting on a high chair or sitting on a normal-height chair. The design of the experiment used three factors: posture, run order (to measure any learning/fatigue) and participant (12 experienced screeners). (The findings were no effects of posture on performance, but large effects of posture on discomfort and of run order on performance.) The results suggest that 'posture' can affect discomfort and 'order' can affect learning, but no conclusions can be reached about the effects of 'posture' on performance or 'order' on discomfort.

Experiments are not the only form of E/HF enquiry, as other chapters in this book amply demonstrate. However, they do have the most obvious impact on determining causality. Also, the same statistical design techniques used in experiments can also be used in other forms of study such as observation studies (where, as discussed in Chapter 1, we can sometimes impose an 'experimental paradigm'). We can choose which factors and which levels of these factors to observe, and apply similar statistical techniques to the data to understand the separate effects. As an example, Chang and Drury (2007) studied human interaction with doors by observing 1600 people as they used different doors. The doors (both push and pull types) were chosen to give four levels of physical difficulty, characterised by the restoring torque in N m needed to move the door. The people observed were characterised by their strength, determined by their gender and stature (observed against a set of marks on the door). The measurement was the use of body weight, rather than just one hand, to help open the door. Even though this was an observational study rather than an experiment, a statistical multifactorial design was used, one recognisable by anybody taking a DoE course. (Chang and Drury found, as expected, that task demands [door difficulty] and human capability [stature used as a proxy for upper body strength] jointly determined the likelihood of using body weight to open the door.)

As noted in Chapter 1, an experiment is NOT automatically a laboratory study. While laboratory studies are often experiments, so are field studies. For some classes of enquiry, such as examining the details of visual search or biomechanics, a laboratory may be the most convenient place to control the factors of interest and measure their effects precisely. But a valid experiment can take place in a field setting, for example measuring the effects of changes in workplace design or aircraft cockpit advances. In a typical field study, Latorella and Chamberlain (2001) asked experienced pilots to fly a twin-turbo-prop aircraft towards convective weather fronts to measure the effectiveness of three different weather displays on the pilots' decision points for avoiding the weather front. Note that the

experimenters made deliberate changes to the system (three displays, nine quantised distances from the weather front) and were in a position to measure both the pilots' decisions and their responses to weather questionnaires. This was not a simplified laboratory simulation, although much prior research derived from laboratory simulations was used in the detailed design of the experiment.

Any experiment changes the system under study. Thus, it is reactive on the system. Participants know the experiment is taking place and most details of how it will affect them. In any academic or research institution or government organisation, an Institutional Review Board (IRB) or ethics committee will have vetted the study before it is allowed to proceed. This helps ensure that the experiment is ethically designed and executed, protecting all participants to the extent possible, but it also emphasises that an experiment does impact the system and its participants. An IRB review might not always be used in an enterprise setting (business, industry) where the enterprise employs the experimenters and participants, although it should always be the norm.

Direct intervention can be both costly and potentially dangerous. Given that, what does this reactivity buy for the E/HF professional? In addition to the major advantage of determining causality, the three other advantages of a more highly reactive design are as follows:

1. The ability to be in the right place at the right time to observe. This is particularly important in human factors studies where the system behaviour observed is rare and unexpected (e.g. accidents or breakdowns).
2. The ability to use more obviously invasive, but information-rich, measurement techniques. For example, in inspection research, the response to each individual item inspected can be observed in considerable detail in an experiment (e.g. Drury and Sinclair, 1983), whereas in the real situation, only a simple accept/reject response is often given.
3. The ability to control or manipulate other variables not of primary interest that would otherwise alter the results of the investigation. Variables that are unaccounted for and contribute to variability in the measured effects of a study can mask the effects of the independent variable. In cases where such unaccounted variables are also correlated with the independent variable, the results of a study become 'confounded' by the unaccounted-for or spurious variable and it is impossible to determine the relative contributions of the independent variable and the confounding variable. Experiments, when designed properly, account for these variables so as to limit the spurious variability of the effects and minimise confounding. For example, in the security screening study, the order of presentation of conditions to screeners would be likely to have an effect, so this variable was controlled by explicitly including it as a factor in the study.

If the E/HF professional gains in experimental control and measurement detail by using highly reactive designs, what is lost? The major loss is in face validity. If we observe a system in its natural state, those associated with the system and possibly those who commissioned the study can be convinced that the study is realistic. An experiment, particularly one performed in a laboratory with artificial stimuli and non-representative participants, requires much more persuasion on the part of the E/HF professional to gain acceptance. The lead author was once involved in two studies of fork-lift truck control. One (Drury and Dawson, 1974) involved real drivers using real fork-lift trucks in a real warehouse to study lateral control behaviour. The other (Drury et al., 1974) involved real drivers controlling a toy train in a laboratory to study longitudinal control behaviour similar to Fitts' law tasks. It is obviously much easier to quote the former study to convince warehouse managers of its design implications.

FACTORS AND LEVELS

What scientists refer to as independent variables are characterised in DoE as factors. They are the things the experimenter varies in order to measure their effects. In the doors study example, the factors were 'physical difficulty', 'gender', 'push vs. pull' and 'stature'. For the weather display study,

the factors were 'display', 'proximity to weather front' and 'participant'. In the security screening study, the factors were 'posture', 'run order' and 'participant'. These are examples of factors designed into the experiment with specific values, called levels. For example, in the doors study, the factor 'gender' was at two levels: male and female. The factor 'physical difficulty' had four levels of door restoring torque: 30, 46, 55 and 72 N m. In the weather display study the factor 'display' was at three levels: aural only, out-the-window view plus aural and a graphic weather display plus aural. In that study also, there were six participants and the factor 'proximity to weather front' was at six levels: every 20 nautical miles from 120 down to 20. In the security screening study, 'posture' was at three levels: standing, high chair and normal chair, and so on. The combinations of levels studied in an experiment are called experimental conditions.

Statistical DoE emphasises factors and levels, almost to the exclusion of other ways of treating independent variables. In any study, there are a potentially infinite number of variables that could possibly affect the outcome. To reduce the normal variation between data points, the E/HF professional must deal with all of these variables, even though that sounds impossible in theory. Some independent variables become part of the experiment by including them as factors at a number of levels (>1), as in the earlier examples. The problem with allowing each variable to take on numerous levels is that in most experimental design solutions, the total number of trials is determined by the *product* of the number of levels of each factor. At the other extreme are what a physical scientist would call 'nuisance variables', which could affect the outcome if not controlled in some way.

The most useful ways of controlling these extraneous variables are to fix them at a single level or to use random assignment so that they do not *systematically* bias the outcome and conclusions. Fixing a variable at a single level is the most obvious way to control it. In the doors example, the whole study took place on a campus, eliminating some variability in age. For the weather display study, all participants had to be instrument-rated pilots with minimum experience defined by specific numbers of flights and flight hours to eliminate much variability due to skill and experience. Trained security screeners were used in the security study for the same reason. There is a price to be paid for fixing a variable at a single level in that the results may strictly apply only to that level of that variable. Thus, the doors study outcomes would not necessarily apply to, for example, people in senior citizen housing, in an elementary school or in a private home. In the weather display study, we could not generalise to, for example, novice pilots or military pilots.

Using randomisation to prevent, for example, all of the older participants being tested on a Monday is a powerful tool that prevents systematic bias by ensuring that any uncontrolled variability contributes only to the 'normal variation'. In statistical terms, this prevents bias to the mean, possibly at the cost of increasing the residual variance. Randomisation is a safe way to control unwanted variation, but the cost of the study might increase, because larger sample sizes are needed to reach the same level of certainty in the conclusions. The security screening study used three sets of baggage images to avoid screeners recognising particular bags: these three sets were presented in a different random order to each participant. A potential alternative to randomisation for reducing order effects (learning and/or fatigue) is to counterbalance the order in which factor levels are presented to each participant. It works best when any change over time or order is linear, which is rarely the case. Other order-balancing designs, such as Latin Squares, are presented later under DoE.

There is another way to treat an independent variable, part way between treating it as a factor at several levels and fixing or randomising it. If we know of an independent variable that is likely to affect the outcome of the experiment, we can measure that variable and treat it formally in the design as a covariate rather than a factor. As an example, in studies of inspection tasks such as that studied by Drury and Sinclair (1983), much evidence has been found that the cognitive ability of inspectors helps determine their effectiveness (Drury et al., 2009). One good measure is that of field independence, measured by an embedded figures test, first developed in the 1950s (Witkin, 1950; Jackson, 1956). Using the individual scores of field independence as a covariate to account

for the effects of different cognitive abilities when different groups of people are tested for different conditions removes some of the random error variance, thus increasing the power* of the study.

The E/HF professional designing a study must trade off the increased power coming from the reduced variance due to fixing a variable at a single level against the decreased applicability of the results beyond that single level. There is no rule for this; it depends upon the aims of the study. For example we typically sample participants from some population of interest (e.g. supermarket workers, retired women or current air traffic controllers). The list of all such potential participants is called the sampling frame. There are correct statistical procedures both for sampling safely and for treating the results so that we can generalise to the whole population of interest. The basic choice is between systematic sampling and random sampling. We start with the sampling frame and use either a systematic method (e.g. every fifth person) or a random method (e.g. choosing via random number table or programs) to determine the members of our sample (e.g. Section 3.3.3.2 of the *NIST Statistics Handbook* http://www.itl.nist.gov/div898/handbook/). In general, randomisation is statistically safer, although stratified sampling by randomly choosing from different defined strata of the sampling frame can be useful where more specific results are required. Occasionally, we are only interested in the specific participants who have taken part, particularly where there is a very small population from which to choose (e.g. active cosmonauts). This was the case in the fork-lift truck control study referenced earlier (Drury and Dawson, 1974) where only four drivers had been trained on all of the trucks tested and no more training was planned. In this case, the four participants were technically a fixed factor (at four levels of course) rather than the usual random factor. This changed the statistical treatment later in the analysis.

We have laid out some of the statistical and design issues in choosing factors and levels. However, the main considerations are more practical (i.e. design or resource driven) than statistical. DoE texts (e.g. Winer, 2012) or online expositions (e.g. *NIST's Statistics Handbook* http://www.itl.nist.gov/div898/handbook/) give general guidance on choosing factors, but cannot be specific when aimed at a diverse audience. Much traditional teaching of DoE uses data-driven insight, as if the process were truly a black box that needed to be understood from existing data or preliminary studies. Thus, exploratory data analysis is seen as a first step in process understanding. In E/HF we can provide more specific guidance. The key is E/HF insight, that is what is important in a situation or process. There is no substitute for E/HF knowledge, whether from textbooks, journal papers or applying E/HF insight into a specific process through direct observation. E/HF has over half a century of accumulated knowledge, on top of hundreds of years of insight from component disciplines such as psychology, physiology or occupational epidemiology. In DoE, we ignore this knowledge at our peril.

E/HF insight means understanding which variables affect individuals' and systems' behaviour, health and performance, and having some feel for the relative magnitude of the effects of these factors. Some factors are obvious, even to lay persons: there will be differences between individual participants in the experiment, there will be learning effects when a task is relatively new to the participant, larger participants will often be stronger, and so on. Some are more obvious to an E/HF practitioner: Task overload will often lead to degraded performance, non-neutral postures held for long periods will result in discomfort, performance may be worse on night shifts or with insufficient sleep.

Within this, some variables are almost mandatory and unavoidable. For example individual differences between participants or temporal changes within the study (learning, fatigue) could affect behaviour, health or performance. These have been discussed earlier, but we return to them later under DoE. Unless we deal actively with such variables, our conclusions will potentially be flawed. We cannot just ignore such factors and hope they will not affect our results. Study sponsors and journal editors will both catch such design problems, at the point where it is too late to deal with them. Other factors of importance are dependent on the particular study. Physical ergonomics studies would

* Power denotes the extent to which an experimental design is able to detect an effect that does indeed exist. A study with a power of 0.8 has an 80% chance of detecting the effect, if there is actually such an effect.

suggest that age, body size, strength and endurance are individual factors that could be important when assessing health (or physiological effects) and physical performance, and probably need to be dealt with by one of the techniques discussed earlier for reducing unwanted effects of individual differences. But other factors such as temperature, humidity and altitude would potentially be important too, as would workplace layout and task pace. We would expect some, but not huge, effects of cognitive variables on physiological effects and physical task performance, and no effects of such outlandish factors such as eye colour or moon phase. Similarly, cognitive ergonomics would suggest that age, cognitive style and task training would be very important to assess behaviour, health and performance during cognitive tasks, as would display layout strategies and task pacing. Physical factors would have less importance, but certainly measurable effects (see review in Drury et al., 2008).

In teaching experimental design, we have often found that the choice of sensible factors and number of levels is the most difficult aspect to communicate. Perhaps this is because traditional experimental design textbooks tend to ignore it, leaving more detailed knowledge to domain expertise. But the statistical and domain-specific aspects of DoE are rarely brought together in domain teaching (e.g. in E/HF subjects).

EFFECTS

The basic classification of E/HF-relevant effects comes from the definition of Ergonomics provided by the *International Ergonomics Association* (IEA) as:

> Ergonomics (or human factors) is the scientific discipline concerned with the understanding of the interactions among humans and other elements of a system, and the profession that applies theoretical principles, data and methods to design in order to optimize human well-being and overall system performance… (http://www.iea.cc/)

Clearly what E/HF professionals are 'about' are the twin groups of measures: system performance and human well-being. Just as in a road test of a car, there is no use knowing its speed and handling characteristics without also understanding its fuel consumption and reliability; in E/HF, we typically need to measure both the system performance and the cost to the human of achieving that performance.

'Performance' measures can involve the overall system, but are also possible for sub-systems even down to the individual human or group of humans. They can be simply classified into measures of speed and accuracy, or 'time and errors', although it is possible to argue that sub-standard speed performance is in itself an error so that we really only need to measure errors (e.g. Drury, 1994a). Speed measures are any measures with time in the numerator or denominator. In the numerator, they can be performance times, cycle times, reaction times at the individual, sub-system or system level, or even broader measures such as system down-time or systems availability. In the denominator, they are speed measures rather than time measures (e.g. output per shift, bits-per-second, rate of progress or miles per hour, as in driving). Accuracy is the positive aspect of performance (e.g. number of hits on target, quality level of a production process, percentage of driving task spent within the desired roadway). The negative aspect is errors (e.g. number of Methicillin-Resistant Staphylococcus Aureus [MRSA] infections per month at a hospital, fraction defective in a production process, percentage of missed threats and false alarms in a security process). Errors can either have a time-rate aspect (infections per month) or an event-rate aspect (fraction defective). If the time or event horizon is pre-specified and constant, perhaps the duration of the study, then the raw numbers of errors can be counted. As with the complementarity of performance and well-being measures, so speed and accuracy often need to be measured together. How can we praise the speed of a Formula 1 driver who has frequent crashes? How can we praise the output quality of a process operator who misses all deadlines?

Speed and accuracy do trade-off (although they may appear not to in current quality literature: Drury, 1997), often enough that many have studied the 'speed-accuracy trade-off' or SATO (e.g. Drury, 1994b). This is also known by the more accurate descriptor speed-accuracy operating

characteristic [SAOC] (e.g. Pew, 1969), because the plot of accuracy versus speed is in the form of a statistical operating characteristic (OC) curve. Typically, more accuracy demands less speed (e.g. in Fitts' Law; Hoffmann, 1992), or in visual search tasks (Drury and Forsman, 1996). Also different errors can trade off against each other, the most obvious example being correct detections (hits) against false alarms in a signal detection task (e.g. Fisher et al., 2012). Thus, choosing measures of performance needs good E/HF models of how the suite of performance measures might co-vary. It is assumed that there will be a suite of measures to capture both system and human performance (see Chapter 1 for the range of measures that might be selected).

Well-being measures can range from the 'soft', such as discomfort ratings, to the dramatic, such as fatalities. They tell of the positive and negative effects of the system, and the human role in the system, effects on the humans within the system, or even those whom the system impacts in use. Typical measures cover workload (e.g. NASA Task Load Index, TLX, scale), internal state of the operator (e.g. discomfort, fatigue), physiological/biomechanical stress, freedom from long-term diseases, health status, negative incidents (e.g. near misses), actual accidents and injury/equipment damage/fatalities. Most of these are covered elsewhere in this book (see Chapters 15, 18, 20 and 31 for examples).

The measures of effects must themselves measure up to scientific adequacy. No measure is useful unless it can be measured with sufficient reliability and actually represents a parameter of interest to those commissioning or reading the study (refer to Chapter 1 for a more in-depth discussion of reliability and other considerations when selecting methods). As with much of what is presented in this chapter, initial resolution of these issues with those who commission the study can save much grief when the study findings are finally presented.

HYPOTHESES

In designing an experiment, or other E/HF study, a critical challenge is turning a research hypothesis into one or several statistical hypotheses to guide future actions. With the revolution in statistical DoE and statistical testing came a profound change in how we interpret results. The philosophy is that we pre-specify which outcomes of the study will lead to which conclusions. Thus, the E/HF professional could in theory give the designed experiment to a competent subordinate and be assured that whatever the data outcome, the conclusions will be exactly as planned. (This rather rigid statistical approach of course contrasts with more grounded theory or emergent approaches as are described in Chapters 3 and 5.) Such a rigorous scientific methodology is often followed until the last step of conclusions, where experimenters have been known to hedge their bets somewhat when faced with conclusions they do not like. For this reason, we spell out the transformation of a research hypothesis into a testable statistical hypothesis and provide an example of following through to conclusions.

The example is the study of security screening outlined earlier (Drury et al., 2008). This did not start out as a study of security screening, but as an examination of 'the inter-relationship between physical ergonomics and cognitive performance'. The experimenters were in fact a graduate class performing a rigorous E/HF practicum (Drury et al., 2007). The aim was to review the extensive literature on the interactions of physical and cognitive work, and test key aspects of this experimentally. (The idea of using security screeners did not arise until later, when it was found to be a convenient domain for testing the research hypotheses.)

The process involves several steps, adapted from the text by Siegel and Castellan (1988), which are illustrated here for the security screening study, concentrating on one of the hypotheses tested.

1. *State the research hypothesis*: Here the research hypothesis was that there is a measurable effect of the posture enforced by the physical workplace on performance in a cognitive task. That is not particularly new, as some authors have found such an effect in the past, but others have not (e.g. Mozrall and Drury, 1996).
2. *State the null and alternative hypotheses*: For a statistical test, we must turn the research hypothesis into a null hypothesis that states 'nothing was found'. The negation of this null

hypothesis is the alternative or experimental hypothesis. The data will eventually show which hypothesis we can conclude is true. The use of a null hypothesis is important as we can find the sampling distribution of any test statistic fairly easily when there is no effect. Here the null hypothesis was that there is no difference in any performance measures (hits, false alarms, time per bag screened) between the three workplace postures tested, that is 'there is no effect of the workplace postures on the performance measures'.

3. *Choose the statistical test*: In a multifactorial experiment such as the screener study (where the factors were posture and participant), the most logical test is the F-test for many cases when the effects are measured on interval or ratio scales. A test statistic (such as t or F) is a dimensionless quantity that is typically calculated as the ratio of the size of an effect to the appropriate variability. The test statistic increases in absolute magnitude as the size of the effect increases and as the variability of the effect within a condition decreases. Size of an effect is represented by the difference between two means (or the variance between several means for an F test) and is thus a fact, although subject to sampling error. The appropriate variability for a test statistic is the standard error of the difference between two means, which in simple cases is

$$\mathrm{SE} = \frac{\text{Standard deviation}}{\sqrt{\left(\text{Number of data points}\right)}} = \frac{\mathrm{SD}}{\sqrt{\mathrm{N}}}$$

where SD is the standard deviation of a set of N data points. Thus the test statistic is

$$\text{Test statistic} = \frac{\mathrm{M1} - \mathrm{M2}}{\mathrm{SE}} = \frac{\left(\mathrm{M1} - \mathrm{M2}\right)\sqrt{\mathrm{N}}}{\mathrm{SD}}$$

where M1 and M2 are the means of the conditions being compared. For any statistical hypothesis, there are typically alternative tests with different assumptions. Non-parametric alternatives, which make fewer assumptions about the distribution, type and source of the measurement values, are possible. However, these tend to only be suitable for simple designs involving only one or two independent variables (e.g. Siegel and Castellan, 1988).

4. *Find the sampling distribution of the test statistic under the null hypothesis*: The F statistic's distribution is well known, tabulated in almost any statistics or DoE text and an integral part of most statistical software available for DoE and analysis. It does, however, rely on the assumption of normally distributed data, homogeneity of variance, independence of sampling and use of interval or ratio data. The assumptions and how violations of the assumptions impact results are also described in most DoE texts.
5. *Select the level of significance*: The level of significance or 'p value' is our threshold for concluding that the alternative hypothesis is true. This defines the likelihood of concluding that an effect exists when it truly does not exist, an erroneous conclusion. In statistical texts, this called a Type I error, and it is important in experimental research to ensure that this likelihood is small. In research, we typically choose a level which would rarely be found by chance (e.g. 1 in 10, 1 in 20, 1 in 100 or 1 in 1000). In this way, we limit our false alarm rate *when many tests are performed*. We do not eliminate false alarms, just make them rarer. If a researcher publishes 100 studies in a career, each with a single test at 1 in 20 ($p = 0.05$), then about 5 false alarms would be expected. A couple of notes are in order here. First, there is no generally accepted p value – it depends upon the circumstances. A p value of 0.01 may be needed if the consequences of false alarm are very high. A p value as low as 0.1 may be acceptable when there is a limited population to collect data from, but a decision is still required. Second, in theory, the person who commissions the study should

choose the significance level. This is fine when the commissioner is a government, scientific or medical agency. However, most managers, most lawyers and many public servants are quite unused to the idea of probability and balk at choosing a level because of their lack of statistical knowledge. Here, E/HF researchers must help them reach an informed opinion, just as they must help a participant understand the risks involved in study participation. In industry, which should be quite used to probabilities after years of Six-Sigma programs, E/HF professionals still get asked 'what is the right answer?' when they attempt to involve managers in the choice of level of significance. Taking some responsibility to help our colleagues from other fields is in fact helping to fit the task to the decision maker. In the screener study, we chose a level of significance of $p = 0.05$, a very traditional value in science, and the normal minimum level of significance adopted in E/HF journal publications, because rejecting the null hypothesis with only a 5% chance of a 'false alarm' would convince others in our profession that we really had found an effect.

6. *Determine the region of rejection*: If we know the sampling distribution of the test statistic and the probability we will accept for a false alarm, then we can split those values of the test statistic into a set for which we will accept the null hypothesis and a set for which we will reject the null hypothesis. Thus, we have mapped all possible outcomes (values of F in our example) onto the conclusions we will draw before running the experiment. From here on, the process is purely mechanical. The region of rejection of the F test in the screener study was those values of F beyond which lay only the upper 5% of the probability distribution. The exact F value depended upon the number of degrees of freedom* in the numerator (2 in this example, because there were three levels of the posture factor) and the denominator, which could vary, because there was more than one effects variable, and each had a different number of measurements.
7. *Run the study and determine the value of the test statistic*: This is the most time-consuming step, but collecting the data and calculating the F value are issues covered later. Setting up the screener experiment and collecting the data took several weeks of effort by all six authors of that study.
8. *Determine whether or not the data support the alternative hypothesis based on the criteria established in Step 6*: In the screener example, the null hypothesis was accepted for all performance measures.

What we have done is to force ourselves to pre-select our ultimate actions based on the statistical outcome. As with our discussion of factors, levels and the qualities of measures chosen, this discipline forces us as experimenters to think about the process of experimentation through to the final conclusions *before* we begin even recruiting participants or building equipment. In this way, there should be fewer studies that fail to meet the expectations of either client or experimenter.

EXPERIMENTAL DESIGN ALTERNATIVES

The Basis of Factorial Experimentation and Analysis

Any standard DoE text (e.g. Winer, 2012) will provide literally hundreds of potential designs for statistical experiments, with detailed instructions on the choices available and the correct analysis techniques. Clearly, such a treatment is inappropriate here, so we shall concentrate on issues of most interest to the E/HF practitioner. Most texts start with simple comparisons of two levels of a single factor using a t-test or non-parametric equivalent. Because we are using people as participants in our experiments, and differences between participants are non-trivial, we are rarely able to use such basic designs. Unless we wish to run the study on a single participant, we always have one factor

* Degrees of freedom are related to the number of values that can vary in a calculation.

of participant with differences between participants as a factor in our analysis. Even a comparison of two levels of a factor of interest will result in a multifactorial experiment, so that is where we must start.

Underlying all multifactorial designs is the analysis concept of ANOVA, again covered in depth in DoE texts. This in turn rests on the mathematical property of a variance: the variance of the sum of, or difference between, two means is equal to the sum of the variances of the separate means:

$$\mathrm{Var}(x_a + x_b) = \mathrm{Var}(x_a) + \mathrm{Var}(x_b) \quad \text{and} \quad \mathrm{Var}(x_a - x_b) = \mathrm{Var}(x_a) + \mathrm{Var}(x_b)$$

This simple assertion (again, proved in most statistical tests) extends to a linear combination of variables so that the variance of their sum is the sum of their variances. Thus, if we run a two-factor model with Factor A at a number of levels denoted by [i] and Factor B at levels [j], we can say that any single data point at level i of Factor A and level j of Factor B is composed of:

The overall mean of the whole data set
+ the difference between the overall mean and the true effect of Factor A at level i
+ the difference between the overall mean and the true effect of Factor B at level j
+ the random error of the combination of Factor A at level i and Factor B at level j

$$X_{ij} = \mu + A_i + B_j + \varepsilon_{ij}$$

This is the structural model of the experiment. Taking variances, which are additive, and noting that the variance of a constant such as μ is zero, we have

$$\mathrm{Var}(X_{ij}) = \mathrm{Var}(A_i) + \mathrm{Var}(B_j) + \mathrm{Var}(\varepsilon_{ij}),$$

provided that the error variance is the same for all [i,j] combinations of Factor A and Factor B. In this way, we can take the overall variance of all of the data ($\mathrm{Var}(X_{ij})$) and split it into components due to Factor A ($\mathrm{Var}(A_i)$), Factor B ($\mathrm{Var}(B_j)$), and the error variance ($\mathrm{Var}(\varepsilon_{ij})$). We can thus see how much variability is uniquely associated with each factor, called here a 'source of variance'. We can also compare this variance with the error variance to form a test statistic, as in the previous section, because a test statistic is an effect size divided by its appropriate variance. We have just partitioned variance into its components, so it is unsurprising that the technique is called ANOVA. If we can do it for two factors, then we can do it for any number of factors, provided that their effects are additive. Note that this paragraph has two uses of 'provided that', implying that to use ANOVA, we must make assumptions about homogeneity of variance and additivity of factor effects. We shall return later to alternative ways to analyse the data if these assumptions are not met.

Multifactorial experiments combine the factors at each level, so that the total number of combinations (i.e. conditions) studied must be the product of the number of levels of each factor. In the security screening experiment, we collected data on the three posture levels for each of the 12 participants, giving 36 combinations. Note that we did not use each run order for each screener, but balanced the three sets of images across participants so that each participant saw a different image set in their three postures.

Main Effects

For each experiment, there is a structural model, written as in the example earlier ($X_{ij} = \mu + A_i + B_j + \varepsilon_{ij}$). This example is a very simple model where the separate effects of A and B are strictly additive. That means that the effect of Factor A is the same at all levels of Factor B. This is called a 'main effect'. If we plotted a graph of our measure against the level of Factor A, then there would be parallel graphs for each level of Factor B.

Interactions

Experimental results are not usually limited to main effects. The effect of one factor may well depend on the level of another factor: the graphs need not be parallel. To take an example from physics, the combined gas law relates the pressure (P), the temperature (T) and the volume (V) of an ideal gas by

$$PV = kT$$

or

$$T = \frac{PV}{k}$$

Thus, the effects of P and V on T are not additive but multiplicative. If we plotted T against P at different values of V, we would get lines that were converging rather than parallel. This is a simple example where it is just a different operator relating the two factors P and V. In more complex situations, there may be different joint effects of the two factors. All non-additive combinations of factors are called interactions (which can be described as the effect of one independent variable on the *effect* of another independent variable on the dependent or measured variable), and are one of the main reasons for performing multifactorial experiments.

An example from the doors study mentioned at the beginning of the chapter is shown in Figure 2.1. It is obvious that the lines joining the data points at each level of participant stature are not parallel, although they do form a pattern that should be familiar to those with E/HF training. Compared to shorter people, taller (and presumably stronger) individuals do not need to use their body weight to open doors until a much higher level of door restoring torque.

In the design and analysis of multifactorial experiments, interactions can be treated statistically by including an extra term in the structural model to represent the combined effect of

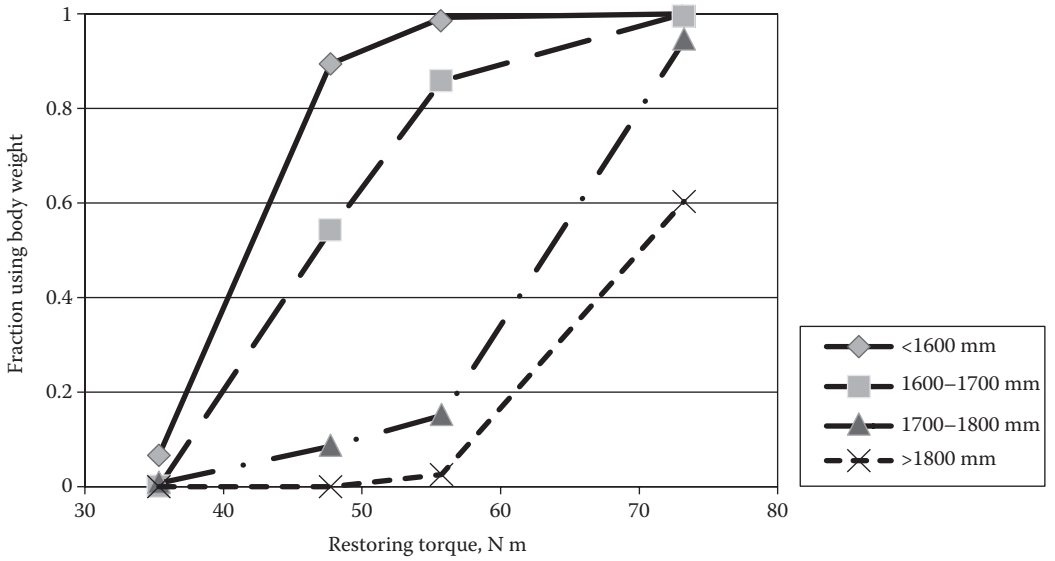

FIGURE 2.1 Example of an interaction from Chang and Drury (2007). The lines reflect four groups of people having different stature.

two factors that is not predictable from their individual additive effects. This is the term AB_{ij} in the following equations:

$$X_{ij} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ij}$$

$$Var(X_{ij}) = Var(A_i) + Var(B_j) + Var(AB_{ij}) + Var(\varepsilon_{ij})$$

Note that the structural model is still additive, so the variances are also still additive. The additional interaction term can now be tested for significance like all of the other terms against its appropriate error variance. Note also that we need some independent measure of the error variance. This is typically accomplished by repeating measurements under nominally identical conditions. In the doors study, multiple people were observed in each of the 16 combinations, but because the measure was only use/non-use of body weight (a nominal measure), it could not be easily used to estimate error variance. In the case of all three example experiments in this chapter, other assumptions were made in order to obtain estimates of error variance. The safest way is always direct replication, but this may not be possible for all experiments. For example, in the weather displays study also mentioned earlier in this chapter, the same weather front conditions could not be replicated so that only a single measure was possible for each combination of factor levels.

We can obviously extend the multifactorial idea to experiments with more than two factors. With each additional factor, we can find more interactions. For two factors, we can measure the effects of

$$A, B, A \times B.$$

When we add a third factor C, this becomes

$$A, B, A \times B, C, A \times C, B \times C, A \times B \times C.$$

Interactions are of great importance in E/HF, often because they represent multiple sources of stress on the human in a system, and each additional source may cause a more-than-additive effect as the limits of human capability are reached. Figure 2.1 is an example of this: at the lowest restoring torque, door opening is easy for all of the population, but as restoring torque increases, the impact is greater among the less strong members. Interactions also have importance in the theoretical underpinnings of our discipline. For example, in time-sharing between two tasks, the lack of an interaction implies that the two tasks must be processed serially, while the existence of an interaction implies that some parallel processing is possible (Wickens and Carswell, 2012). Also in visual search tasks, the rate of increase of reaction time with increasing background complexity depends upon the difficulty of discriminating a target from its background. Thus, if target/background discriminability is one factor and number of non-targets in the search field is another, there will be an interaction between these factors on reaction time (Treisman, 1986). In the extreme, if one target is pre-attentive, then the number of non-targets will have no effect on reaction time for that target, leading to an extreme interaction effect. Finding interactions, and finding situations where there are no interactions, are primary objectives of practical E/HF and can form the basis for DoE principles.

Design of Multifactorial Experiments

The basic design of a multifactorial experiment, known as a complete factorial, is to test all combinations of levels of all factors. We just add factors such as D, E, F, etc. beyond the aforementioned A, B and C. This was used in the doors study, where 4 levels of restoring torque were combined with 4 levels of participant stature to give the 16 conditions plotted in Figure 2.1. Also, in the weather display study, 6 pilots were combined with 6 weather front distance and 3 displays to give 108 conditions. In contrast, the security screening study did not use all combinations of 3 postures, 12 participants, 3 run orders and 3 image sets, using only a single image set for each run order. This gave a design comprising 3 postures × 12 participants and so could only find main effects of posture, participant and run order.

TABLE 2.1
Design Tableau of a 3 × 3 × 2 Complete Crossed Factorial Experiment with Three Replications

Factor A	Factor B	Participant	Replication 1	Replication 2	Replication 3
Level 1	Level 1	P1			
Level 1	Level 2	P1			
Level 1	Level 1	P2			
Level 1	Level 2	P2			
Level 1	Level 1	P3			
Level 1	Level 2	P3			
Level 2	Level 1	P1			
Level 2	Level 2	P1			
Level 2	Level 1	P2			
Level 2	Level 2	P2			
Level 2	Level 1	P3			
Level 2	Level 2	P3			
Level 3	Level 1	P1			
Level 3	Level 2	P1			
Level 3	Level 1	P2			
Level 3	Level 2	P2			
Level 3	Level 1	P3			
Level 3	Level 2	P3			

In E/HF, the factor of participant can require special treatment. An example of a complete factorial experiment, with 3 levels of Factor A, 2 of Factor B, 3 of participant and 3 replications (i.e. repetitions) of each combination, is given in Table 2.1. This of course requires a minimum of 3 × 3 × 2 × 3 = 54 measurements.

This design can be used to calculate the variance component of all three main effects (A, B, P), all two-way interactions (A × B, A × P, B × P) and the three-way interaction (A × B × P) as well as a true error variance. Note that each participant is tested at all three levels of Factor A and both levels of Factor B. This is known in E/HF as a 'within participants' (also known as repeated measures) design, as both factors are tested on the same participants (P1, P2 and P3). It is clearly a 'good' design if there are large inter-participant differences as the ANOVA can calculate the effects of A, B and A × B independent of participant. In other words, differences between participants do not contribute to the error variance associated with A, B and A × B effects. But what if we cannot test each participant more than once? An example is in comparison of learning technologies where a participant can only learn the task once.

So far, we have only considered designs where all combinations of levels were tested, known in DoE as crossed designs. There is another class of designs for complete factorial experiments known as nested designs, or in E/HF as 'between participants' (or independent samples) designs. In these designs, one factor (participant) is nested under other factors so that different participants are tested under different conditions. Table 2.2 shows a design equivalent to Table 2.1, but with participant nested under both Factor A and Factor B so that the 18 different combinations of A × B are tested with different participants rather than re-using the original 3 participants of the crossed design in Table 2.1. Such designs have the advantage of minimising confounding due to fatigue or learning (order effects), but differences between participants groups contribute to the error variance, making it more difficult to identify A, B and A × B effects.

Within-Participant versus Between-Participant Designs

The main criterion for choice between these two design structures is whether or not the participants change during the course of the experiment. Clearly a participant's strength will not change much over

TABLE 2.2
Design Tableau of a 3 × 3 × 2 Complete Nested Factorial with Three Replications

Factor A	Factor B	Participant	Replication 1	Replication 2	Replication 3
Level 1	Level 1	P1			
Level 1	Level 2	P2			
Level 1	Level 1	P3			
Level 1	Level 2	P4			
Level 1	Level 1	P5			
Level 1	Level 2	P6			
Level 2	Level 1	P7			
Level 2	Level 2	P8			
Level 2	Level 1	P9			
Level 2	Level 2	P10			
Level 2	Level 1	P11			
Level 2	Level 2	P12			
Level 3	Level 1	P13			
Level 3	Level 2	P14			
Level 3	Level 1	P15			
Level 3	Level 2	P16			
Level 3	Level 1	P17			
Level 3	Level 2	P18			

any experiment of reasonable duration, while the same participant's task knowledge and skill will surely change unless very experienced participants are chosen and tested using familiar conditions. Thus, in the doors study, we could have re-used participants, although we did not. Also in the weather display study, highly experienced pilots were used, so that changes between multiple trials on how to react to weather fronts would be unlikely to change, so again, the same participants can be re-used in a crossed design. However, in the security screening study, we found a definite learning across trials on the image sets used, so that even our experienced screeners did change over time. We used a partially within-participants design, but were able to measure and remove any change (learning) effects from our comparison of posture effects. To give other examples where different designs were used, consider the following

> Laughery and Drury (1979) used a between-participants design in a study of optimisation skills because it was suspected that techniques learned during the solution of one type of optimisation problem might transfer in an inconsistent manner to other problems, with an adverse effect on bias and variability. Thus five participants were used in each condition, which meant that any comparison between conditions had to be made against between-participant variability. The groups were kept reasonably homogeneous (engineering students) but this in turn limits the generalisability of the results.
>
> Drury et al. (1989a,b) studied the biomechanics and physiology of handle positions on boxes used ten participants, each performing a box holding task using ten handle positions. The within-participants design eliminated the influence of individual differences on the effects, allowing the effects of handle positions on boxes to be detected despite the limited sample size.

No changes to the participant were expected during the box holding experiment, but changes were expected in adaptation. Change occurs in humans in the short term as they fatigue and in the long term as they adapt or learn. With appropriate rest periods, no fatigue was expected (or found) in the box holding task and certainly an hour or two of experimentation on a well-practiced task is unlikely to change either a participant's body strength (adaptation) or box holding technique (learning). Hence, a biomechanical and physiologically limited task is unlikely to exhibit what Poulton's famous (1974) paper called asymmetrical transfer effects. The same cannot be said for

most intellectual skills. What you learn in first solving one optimisation problem is quite likely to affect your performance in solving the next. The transfer can be positive, if the same solution techniques are useful in both problems, or negative, if the solution to the first problem is inappropriate in solving the second. An optimisation task is a priori likely to be closer to an intellectual task than to a biomechanical one, hence the choice of a between-participants design. Any human functions, even anatomical ones, will adapt or change given sufficient time, but the key question is not whether or not change will occur but whether enough will occur to bias the experimental comparison. We can minimise change during an experiment by choosing participants already highly skilled, but as noted earlier that worked for pilots but not for security screeners. We can also provide extensive training in the task so that the typical negative exponential or fractional power law learning curve reaches enough of an asymptote to prevent further changes in task performance. Such techniques would allow the greater power of a within-participants experiment (less variance in comparisons). Extensive task training also helps when the participant pool is limited (astronauts) or non-existent (operators of an entirely novel system). Finally, we may be interested in the response of each individual participant rather than the overall distribution of performance, so that a within-participant design must be used.

There is no reason that an experiment must be entirely a crossed design or entirely a nested design. We can have useful designs that are partially crossed and partially nested, called mixed model designs. However, wherever different conditions are tested on the same participant in a factorial study, we must use a form of analysis called repeated measures ANOVA to correctly capture the contributions of the independent variables and interactions on the variability of the measurements.

The final word on between- vs. within-participant designs is that a between-participant design is always the safer alternative, but may not be practical within resource constraints. They are also subject to the risk that differences between individuals may mask influences of factors, particularly when the number of subjects within each condition is small. Within-participant designs need steps to ensure absence of carry-over effects and a different type of ANOVA but, if designed carefully, help to manage the potentially confounding effects of individual differences on the measurements.

Sample Size, Effect Size and Power

We have seen that the significance level chosen for any statistical test determines the probability of the test giving a false alarm, that is concluding that an effect exists when it truly does not (Type I error). As any E/HF professional can guess, it is not possible to discuss false alarms without also discussing the complementary error, that of failing to conclude that an effect exists when it truly does (Type II error). We have two hypotheses, null and alternative, so that if

$$\text{Significance level} = p(\text{conclude Alternative} \mid \text{Null is true}),$$

then its complement, power, is defined as

$$\text{Power} = p(\text{conclude Alternative} \mid \text{Alternative is true}).$$

Whereas we usually look for a very *low* significance level (0.10, 0.05, 0.01, 0.001), we would like a *high* value of power (0.90, 0.95 etc.).

There are four inter-related factors that need to be considered: significance level, power, effect size and sample size. The effect size is the magnitude of the difference between two means, or the variance between several means. The larger the effect size we are looking for, the easier the statistical testing is, so that for a large effect size, we can have both a low significance level and a high power. It is possible to manipulate the anticipated effect size with DoE. If, for example, age is related to a particular performance measure, a more powerful experimental design would be to

compare groups that are dramatically different in terms of age, as opposed to groups that are closer in age. Because a test statistic is the ratio of an effect size to its standard error, we can also reduce the size of the standard error for any given effect size by taking more samples. Recall that the standard error is calculated by dividing by the square root of the sample size, and it becomes obvious that increasing the sample size is quite an inefficient way to increase power.

In designing our experiment, we have so far not mentioned sample size, except when we have given examples of numbers of participants or numbers of replications per condition. But both of these experimental parameters must be chosen before we can proceed, and we can only logically do this on the basis of the other three underlying variables. If we need to know sample size, we must first decide on effect size, significance level and power of the test. The usual statistical textbook advice is to work with the study commissioner to find values of these parameters, but this is more taxing than asking a client for a significance level alone. One of the authors recently had to produce tableaus of effect sizes and sample sizes for an aviation security study so that the client could make better informed decisions about how to set up an experiment. Such an approach is facilitated by web-based software (e.g. http://homepage.stat.uiowa.edu/~rlenth/Power/) and advocated in related journal publications (Lenth, 2001).

The mechanics of calculating sample size given the other three parameters are not simple beyond the 'toy' designs of comparing two samples with no other sources of variation, but help is provided in many statistical packages. In addition to the web-based application noted earlier, one author has used the PASS (power analysis and sample size) software to determine sample sizes for experimental designs to be analysed by ANOVA. A much-used package (MINITAB) performs power/sample size calculations for many ANOVAs and some non-parametric statistical tests (e.g. chi-square) but not for complex mixed models. But in the airport security study described at the beginning of this chapter, only nominal data could be collected (e.g. number of threats found, number of false alarms) so that contingency tables and the chi-square test took the place of ANOVAs. The sample sizes were computed manually, although it was later found that they were part of the MINITAB statistical package.

Fractional Factorial Designs

The complete factorial design is a powerful and often-used tool in E/HF, but also a costly one. It can find important interactions, but if interactions are known (or assumed) not to exist, it is wasteful of resources. There are special designs that trade off knowledge of interactions for reduced experiment size. We have already seen this in the security screening study where the information on the posture × participant interaction was sacrificed to allow a study that would fit the available resources. More formal methods are available, known as fractional factorial designs. As their name implies, they only test a fraction of the combinations used in a complete factorial design.

The simplest fractional factorial for E/HF is probably the Latin Square design. It uses three factors, all at the same number of levels, and counterbalances their appearance so that only n^2 instead of n^3 combinations are tested. The Latin Square ensures that each combination of the levels of each factor occurs once and only once in the n × n tableau of the design. Clearly something must be lost for such a large saving in effort and indeed it is. All interactions are sacrificed so that only the three main effects can be calculated and the error variance term is merely the left-over variance when the three main effects are calculated. Even the term 'sacrificed' does not capture the whole loss: any interactions are confounded in a complex way with the main effects. A Latin Square should only be used where no interactions are expected or the interactions are known from prior research not to exist. This makes a long list of assumptions that should really be attached to the experimental conclusions, although they rarely are. The main use of Latin Squares in E/HF is in presentation order of different conditions to participants in a within-participant design. Thus, if we have six levels of the Factor (A, B, C, D, E, F) and six participants (1, 2, 3, 4, 5, 6), we can present them in the following trial order (Table 2.3):

In this Latin Square, each level of the factor (usually called a treatment) follows a different order for each participant, helping minimise the effects of unwanted transfer between treatments.

TABLE 2.3
Example of a 6 × 6 Latin Square Design to Eliminate the Potentially Confounding Effects of Trial Order

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
Participant 1	A	B	F	C	E	D
Participant 2	B	C	A	D	F	E
Participant 3	C	D	B	E	A	F
Participant 4	D	E	C	F	B	A
Participant 5	E	F	D	A	C	B
Participant 6	F	A	E	B	D	C

(The same pattern can be used for any even number of treatments and participants.) This use of a Latin Square is an excellent alternative to randomisation of treatment order across participants.

More complex fractional factorials have been advocated and used in industrial experimentation, often as part of a design with all factors at 2 levels, called 2^n designs. Like Latin Squares, they only require a fraction of the conditions to be tested, and similarly, they do not allow all interactions to be calculated. They are typically advocated in studying the response of a multivariate industrial process to a selection of the postulated variables in a very economical manner so that the most important variables can be determined efficiently. So far they have not been widely used in E/HF, although a few examples exist (e.g. Bishu et al., 1992; Lin and Radwin, 1998; Naugraiya and Drury, 2009). Rather than 2^n designs, these designs are called 2^{n-k} designs, where n factors are tested within the resources of an (n − k) design. The assumption behind fractional factorial designs of all types (see, e.g. Taguchi, 1965) is that higher-order interactions are inherently unlikely. Thus, we can confound these interactions with each other and not calculate them. Each fractional factorial design has a design operator, which is an identity equation showing which effects are confounded with which other effects. Typically, we use a design operator that confounds the main effects and lower-order interactions of interest with (unlikely) higher-order interactions.

The Naugraiya and Drury (2009) experiment was intended as a screening experiment to examine the significance of a large number of factors and interactions for a simulated process control task. It used a 2^{6-1} fractional factorial and examined six factors, each having two levels, with 32 cells rather than the 64 cells required for a full 2^6 factorial. Participants were assigned randomly with one to each of the 32 unique conditions tested, a most unusual procedure in E/HF experimental design, but one quite common in industrial experimentation. Each participant performed the simulation four times, 'producing' and 'shipping' 200 industrial parts under different quality challenges.

The six factors were

Operator expertise (E)
Operator training (T)
Process capability (Cp)
Challenge direction (D)
Challenge amount (A)
Cost criterion (C)

The design operator for the experiment was 1 = ETCpDAC. This ensures that main effects and low-order interactions are only confounded with the more unlikely higher-order interactions:

Main effects confounded with five-way interactions
Two-way interactions confounded with four-way interactions
Three-way interactions confounded with three-way interactions

Thus, we assume that main effects and two-way interactions are in practice unconfounded. We could try to estimate three-way interactions, but each is confounded with the interaction between the remaining factors (e.g. E × T × Cp is confounded with D × A × C) so that we cannot disentangle their separate effects.

With six main effects and 15 two-way interactions, much was learned about the task. Note however that the unusual experimental design had only a single degree of freedom for each effect tested, with 103 degrees of freedom for the error variance. Note also that there could be no 'between participants' effect calculated with a single participant per condition.

Sequential Experimentation

There is an alternative to fractional factorial experiments that can be useful in E/HF where a sequence of studies is performed instead of a single study. It has the potential to measure which interactions are important rather than merely assuming them away. Also it can deal with factors at more than the two levels assumed in most DoE texts for factorial experiments. With only two levels tested for each factor, we cannot find out much about the underlying response surface (i.e. the shape of the relationships between the independent variables and the effects). The strategy is to perform a 2^n complete factorial, measure which interactions are important, then perform a set of experiments with the desired levels of each factor, but only for factor combinations that have measurable interactions. As an example, suppose the design we would like to run has five factors at the following levels:

A at 5 levels
B at 2 levels
C at 3 levels
D at 3 levels
Replications at 2 levels to provide an error estimate

The full factorial will need 5 × 2 × 3 × 3 × 2 = 180 trials. It will allow calculation of

4 main effects: A, B, C, D
6 two-way interactions: AB, AC, AD, BC, BD, CD
4 three-way interactions: ABC, ABD, ACD, BCD
1 four-way interaction: ABCD
1 error term

Perhaps, we do not need all of these from a single grand design. We can use a 2^4 complete factorial with 2 replications as a screening experiment specifically to test for interactions. This will require 2 × 2 × 2 × 2 × 2 = 32 trials. Then run experiments on the significant interactions. For example if only A × B and C × D are significant, then we can run two additional smaller experiments:

A × B with 5 levels of A, 2 levels of B and 2 replications, requiring 5 × 2 × 2 = 20 trails
C × D with 3 levels of C, 3 levels of D and 2 replications, requiring 3 × 3 × 2 = 12 trails

We can thus measure all of the effects we were initially interested in with 32 + 20 + 12 = 70 trails instead of the original 180 trials. Of course, not all screening experiments produce the same interaction structure, so we could go all the way from every interaction being significant (requiring 180 trials) to no interactions being significant (requiring only (5 + 2 + 3 + 3) × 2 replications = 26 trails). We can always re-use data between experiments if that is logically possible, for example the 10 combinations in the A × B experiment include 2 × 2 = 4 conditions that have already been studied, leaving only 6 additional conditions with 2 replications each.

ANALYSIS ALTERNATIVES

Throughout this chapter, we have assumed that ANOVA will be the analysis method, primarily because it is well suited to multifactorial experimentation, and almost any E/HF study will have to be multifactorial. ANOVA uses the additive property of variances to decompose a total experimental variance into components associated with each variable and interaction. In simple cases, such as complete factorial designs with replications, we can have independent tests of each factor and interaction. With more complex, but less complete designs, such as fractional factorials or Latin Squares, we forgo some independence for experimental convenience or even for study feasibility. The familiar ANOVA table can be found in any DoE text and will not be repeated here. There are strict rules for how to compute the significance of effects based on their F statistic value, depending upon the structural model of the experiment. This tells the components of variance and thus what denominator to use in the F test.

Most E/HF professionals will not calculate variance components and F-values by hand, relying on statistical packages to perform the computations. These, such as SPSS (Statistical Package for the Social Sciences) or MINITAB, will require the user to input a structural model to control the computations. That is why understanding ANOVA and models remains important in times of automated computation. Failing to treat repeated measures correctly, confusing fixed and random effects or failing to check the ANOVA assumptions are all analysis errors that can be committed by experienced E/HF professionals. The use of packages where ease of use is a design feature should not reduce the diligence of the experimenter at the analysis phase. In this section, we enumerate various issues with analysis, examining what to do when there are multiple dependent variables, how to test assumptions (and what to do if they are not met), and sources for the many different statistical analysis packages available today.

Dealing with Data

We have discussed experimental design and ANOVA as if there were a single number or measure in each cell of the design, but we now need to expand beyond this. Each cell is the data from a single replication of each combination of factor levels in the design. Assuming that issues of reliability, validity, etc. have been addressed, the first thing to note is that we do *not* put the data into a spreadsheet and calculate means. Each data point must be kept separate: ANOVA procedures will provide table of means, variances and confidence intervals *ad nauseam*. Keeping data separate provides the full degrees of freedom for error variance, helping to ensure that the power of the tests is maintained. Second, many measures come in the form of a continuous variable recorded over a time interval, such as the record of car position in a lane while driving, or the continuous movement of the centre of gravity of a standing operator. In these cases, measures must be derived from the continuous records, for example root-mean-square error or average position in each dimension. It is this number which becomes the data for the ANOVA. These are quite simple matters to deal with and have usually been addressed during the initial design of the experimental study.

Quite often, however, there will be multiple measures per cell of different aspects of performance or well-being or both (i.e. multiple effects). In the security screening study, we measured four performance variables (hits, false alarms, time per image when a threat was found, time per image where no threat was found) and four well-being variables (two measures of body part discomfort, NASA TLX, number of non-work-related movements), giving 8 dependent variables. The obvious way to proceed is to perform an ANOVA on each dependent variable, but this raises two problems. First, because we are performing 8 tests for each factor or interaction, the likelihood of concluding that a statistically significant relationship between one or more independent variables and an effect variable exists when in fact it does not exist (i.e. the likelihood of Type 1 error) increases. Another way to think about this is that the significance levels become inflated. If we chose a 1 in 20 chance of false alarm for any single variable, then the likelihood of having at least one false alarm would be $(1 - (1 - 0.05)^8) = 0.33$, which is not what we had planned. Second, it might be that the dependent variables show similar patterns, so that even if one variable is not significant, the same pattern across several variables

may be significant. The standard way to proceed is to perform a multivariate analysis of variance, or MANOVA, across the complete suite of dependent variables. Then, if any factor or interaction proves significant, univariate ANOVAs can be run to determine which variables were responsible. This procedure does not inflate the significance levels, and provides an orderly exploration of the data.

Another technique that can be a powerful tool for reducing a large suite of data to a more manageable number of orthogonal and hence independent tests is factor analysis. Factor analysis (nothing to do with factors in DoE) groups together dependent variables with high inter-correlations. With a modification called Varimax rotation, it will produce a small number of new dependent variables called, confusingly, factors that are orthogonal to each other and summarise a large fraction of the total variability in the data set. It has a long history in the social sciences, and has been used many times in E/HF, from early papers (e.g. Drury and Daniels, 1980) to more recent studies (Ryan et al., 2009). In the security screening study, it was used to explore the inter-correlation matrix of all 8 dependent variables. We found three factors that met the usual criterion for significance: performance (4 variables), posture (3 variables) and workload (TLX only). Each could be analysed in the confidence that only three independent tests were being carried out, and that all of these new independent variables were orthogonal. To simplify the interpretation of the findings, the ANOVA results associated with each of the individual measures that contributed to each of the orthogonal factors were reported in the paper.

Dealing with Assumptions

Various assumptions have been made in this treatment of experimental design, assumptions that can often be tested directly from the data collected. We have already considered the additivity assumption and shown how intersection terms can extend the simple additive model of main effects to the generally more interesting interaction effects. There are other ways to deal with non-additivity in special cases. The combined gas law used as an example earlier has a multiplicative effect of P and V on T. Many human functions are multiplicative, particular sensory functions. In any E/HF study where effects look multiplicative, logarithms are a useful transformation tool to allow ANOVA while still preserving the model structure. It is simple to transform the common gas law into an additive function by taking logarithms of the equation:

$$\mathrm{Ln}(T) = \mathrm{Ln}(p) + \mathrm{Ln}(V) - \mathrm{Ln}(k)$$

Additivity is now satisfied. Transforms are an integral part of many science and engineering formulations, and are frequently used in measuring human performance. Examples are the use of the decibel scale in auditory perception and the same scale to study human tracking behaviour.

Transforms can also help with the ANOVA normality assumption. We can test this assumption by having the analysis package plot residuals (the difference between a data point and its expected value from the ANOVA model) as a cumulative normal distribution. Either visually or statistically, we can determine whether the normal distribution is a good fit to the data. If it is, then the ANOVA is valid for the normality assumption: if not, the pattern of deviations from the normal distribution function provides clues to suitable transforms.

Some measures of human performance and well-being are quite normally distributed, but task completion times are often not. Most performance time data will have a lower bound beyond which the human cannot react or move any more rapidly. But the upper bound is often unlimited, leading to time distributions that are positively skewed, with a longer ‘tail’ to the right. A lognormal distribution is often a good fit, so that transforming the raw time data to Ln(time) will produce normally distributed data suitable for use in ANOVA. In odd cases, even more skewness is expected: Search times in extended search tasks (such as security screening) are expected theoretically to follow a negative exponential distribution (Morawski et al., 1980). A ln(time) transform is usually sufficient to normalise the data, however. Most statistical texts include at least something on transformation to improve normality and homogeneity of variance or to meet additivity assumptions, removing interactions (e.g. Winer et al., 1991, pp. 354–358).

Probability or frequency data generated from repeated measures made on nominal or categorical scales, particularly the (0,1) form of data, are inherently non-normal. There is much discussion in the statistical and social science literature of the legitimacy of using ANOVA for categorical data, with some claiming it can be used with minimum danger, while others recommend arc-sine or logistic transforms before using ANOVA. Rather than take sides in this, articles such as Jeager (2008) should be consulted for the most recent findings.

This introduction of nominal data brings in the whole question of alternative forms of data analysis to ANOVA itself. Nominal data can often be best analysed using contingency tables and the chi-square or Fisher's exact tests. As texts on categorical data analysis (e.g. Agresti, 1996) explain, contingency tables can go far beyond the 2×2 example given in most statistics texts. One-way, two-way and even three-way tables can be analysed rather simply to give the equivalent of ANOVA for nominal data. One can even calculate standardised residuals to provide post hoc comparisons and determine which cells in the design have significantly high contributions to the overall chi-square statistic. The use of chi-square for contingency table analysis is particularly suitable for the relative small counts of rare events during experiments, such as errors. For many years (e.g. Drury and Daniels, 1975), one author has used ANOVA to analyse task completion times while using chi-square to analyse error frequencies.

The final assumption in ANOVA is homogeneity of variance. The homogeneity of variance assumption requires the variability of the data to be the same across the conditions tested with ANOVA. This is treated in most texts (e.g. Winer, 2012), and tests such as Bartlett's test and Scheffe's test are recommended to be run on the data to check that the variance is not different across cells in the design. If the test is not satisfied, the analysis can sometimes proceed with a transformation. For example if the variance in each cell is related to the cell mean, then a logarithmic or square-root transform will help homogeneity.

Dealing with Analysis Packages

One recent survey found over 40 available software packages for DoE and statistical analysis of the data. Online comparisons between features and costs are easily located using Internet search engines. As already noted, people do not perform ANOVAs manually any longer, although most E/HF practitioners learn this skill in elementary statistics courses. With the advent of data-based quality philosophies in industry (e.g. Six-Sigma) has come the wide demand for statistical analysis tools. These quality philosophies rarely take into account E/HF topics, so it is unsurprising that some statistical packages also omit the unique issues of experimenting with people, for example repeated measures designs. The range of software is huge, as is the range of costs. Some excellent software is free, such as 'R' which is integrated in the Internet-based *e-Handbook of Statistical Methods* available at the *National Institute of Standards and Technology* (http://www.itl.nist.gov/div898/handbook/index.htm) in the United States. Others cost many thousands of pounds/dollars, although academic licences are usually available to ease the cost to students. Some have free trial introductions, enabling users to test their suitability using their own data. Most packages work on the usual computer platforms of Windows, Mac-OS and Linux, although MINITAB and SAS/Stat appear (2014) to have stopped Mac-OS support. The authors have used MINITAB, SPSS, SAS, Number Cruncher Statistical System (NCSS) and PASS at different times, and so these are the basis for the following comments. Note that statistical software is updated frequently, so that the lack of a particular feature in one package may not be forever.

Three general observations are in order when considering statistical packages. First, the experimenter should be extremely cautious when attempting to use general-purpose spreadsheets (e.g. Excel) for statistical testing of experimental data. While such programs have statistical analysis routines, they are very simple and may not have much utility in E/HF experimentation. They may also make assumptions (e.g. on F-tests) that the experimenter may not have intended, and rarely have the ability to perform ANOVA with the appropriate main effect and interaction structure discussed previously in this chapter. Second, when using a statistical analysis software package, the experimenter should always perform a single analysis on the complete design (i.e. the full model), rather than

multiple one-way ANOVAs on each factor. While statistics texts never advocate the use of multiple one-way ANOVAs instead of using the full model, unfortunately, students or industrial users who have only had brief introductions to statistics seem more comfortable using such an approach. The complete ANOVA will calculate all available effects, and thus remove their variance contribution from the residual error, leading to much more powerful F-tests. Lastly, it is recommended that the experimenter, after designing the experiment, generate random data to populate the data array *before* running the experiment. In this way, the experimenter can see what the output will be like, which tests are possible and whether any factors are confounded. Running the analysis on random data first is a good final check on the design before data collection.

All of the packages mentioned perform ANOVAs to considerable levels of complexity, except for PASS which is a dedicated power and sample size calculation tool. All have moved many years ago from a command line interface to a graphical user interface. Data can be entered directly into a spreadsheet-style tableau, or pasted in from a file collected as part of the experiment. Import and export of data files between statistical packages is an important asset, allowing the strengths of different packages to be complementary. Most packages will accept standard spreadsheet files, such as .xls files from Excel, so that this intermediary can often be used if no automatic data transfer routine is available.

MINITAB grew from a very simple and small data analysis system, as its name implies. It is now a very complete package aimed at industrial users, with much support for process analysis and statistical process control. It works well for fractional factorial designs, both generating them and analysing the results. At times, the format of the structural model can be confusing in MINITAB, but it generally will cover your experiment if the model is specified correctly. NCSS has also grown considerably since it was first introduced as software for the computer industry. It includes complex ANOVAs, with options such as repeated measures designs useful to E/HF professionals. SAS, specifically SAS/STAT, is the statistical analysis component of a much broader range of business analytics. SPSS (Statistical Package for the Social Sciences), now part of a suite of statistical software from IBM, is more oriented to the social sciences as its name implies, with good support for ANOVA models where participants are a factor. It, like MINITAB, SAS/STAT and NCSS, will perform ANOVAs and MANOVAs, calculate components of variance and effect sizes, check for normality and homogeneity of variance and provide a variety of post-hoc comparisons between specific levels of a factor or interaction. In addition, all provide multivariate analysis such as factor analysis, non-parametric statistics and analysis of contingency tables.

This chapter has framed the traditional benefits and challenges of statistical DoE in terms of the work of the E/HF professional. Whether the issue is our basic science or more applied knowledge, the same principles apply: choice of dependent and independent variables, choice of sampling procedures and sample size, detailed design of the experiment to maximise effectiveness without excessive time and cost and the ethics of experimentation on humans. Without an experiment, it is difficult to impute causality to results. Without a designed experiment, it is difficult to explore the many causal factors known (or suspected) to influence human well-being and system performance. While excellent statistical texts provide many details of how to design and analyse experiments, they must use very general examples to make them applicable to a broad audience. In our discipline, we can use our specialist knowledge and insights to make more informed choices of what to vary, what to measure and what to control so as to remove sources of contamination or uncertainty. Experimental design is not performed in isolation: it is one more powerful means for us to make decisions about future actions.

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.

Bishu, R.R., Wei Wang, Hallbeck, M.S. and Cochran, D.J. (1992). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 36, 816–820. Santa Monica, CA: Human Factors and Ergonomics Society.

Chang, S.-K. and Drury, C.G. (2007). Task demands and human capabilities in door use. *Applied Ergonomics*, 38(3), 325–335.

Drury, C.G. (1994a). Function allocation in manufacturing. In: S.A. Robertson (Ed.), *Contemporary Ergonomics 1994*, Keynote address to the Ergonomics Society Meeting, Taylor & Francis Group, London, U.K., pp. 2–16.
Drury, C.G. (1994b). The speed-accuracy trade-off in industry. *Ergonomics*, 37, 747–763.
Drury, C.G., Cardwell, M.C. and Easterby, R.S. (1974). Effects of depth perception on performance of simulated materials handling task. *Ergonomics*, 17, 677–690.
Drury, C.G. and Daniels, E.B. (1975). Performance limitations in laterally constrained movements. *Ergonomics*, 18, 389–395.
Drury, C.G. and Daniels, E.B. (1980). Predicting bicycle riding performance under controlled conditions. *Journal of Safety Research*, 12(2), 86–95.
Drury, C.G. and Dawson, P. (1974). Human factors limitations in fork-lift truck performance. *Ergonomics*, 17, 447–456.
Drury, C.G., Deeb, J.M., Hartman, B., Woolley, S., Drury, C.E. and Gallagher, S. (1989a). Symmetric and asymmetric manual materials handling. Part 1: Physiology & psychophysics. *Ergonomics*, 32(5), 467–489.
Drury, C.G., Deeb, J.M., Hartman, B., Woolley, S., Drury, C.E. and Gallagher, S. (1989b). Symmetric and asymmetric manual materials handling. Part 2: Biomechanics. *Ergonomics*, 32(6), 565–583.
Drury, C.G. and Forsman, D.R. (1996). Measurement of the speed accuracy operating characteristic for visual search. *Ergonomics*, 39, 41–45.
Drury, C. G. (1997). Ergonomics Society Lecture 1996: Ergonomics and the Quality Movement. *Ergonomics*, 40(3), 249–264.
Drury, C.G., Holness, K., Ghylin, K.M. and Green, B.D. (September 2009). Using individual differences to build a common core dataset for aviation security studies. *Theoretical Issues in Ergonomics Science*, 10(5), 459–479.
Drury, C.G., Hsiao, Y.L., Joseph, C., Joshi, S., Lapp, J. and Pennathur, P.R. (2008). Posture and performance: Sitting vs. standing for security screening. *Ergonomics*, 51(3), 290–307.
Drury, C.G., Mirka, G.A. and Marras, W.S. (2007). *The Publishable Practicum*, *Contemporary Ergonomics 2007*, Taylor & Francis Group, London, U.K., pp. 247–252.
Drury, C.G. and Sinclair, M.A. (1983). Human and machine performance in an inspection task. *Human Factors*, 25, 391–399.
Fisher, D.L., Schweickert, R. and Drury, C.G. (2012). Mathematical models in engineering psychology: Optimizing performance. In: G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics*, John Wiley & Sons, Inc., Hoboken, NJ, pp. 962–989.
Hoffmann, E.R. (1992). Fitts' law with transmission delay. *Ergonomics*, 35(1), 37–48.
International Ergonomics Association, http://www.iea.cc/.
Jackson, D. (1956). A short form of the Witkin's embedded-figures test. *Journal of Abnormal and Social Psychology*, 53(2), 254–255.
Jaeger, T.F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.
Java applets for power and sample size, http://homepage.stat.uiowa.edu/~rlenth/Power/.
Laughery, K.R. and Drury, C.G. (l979). Human Performance as Strategy is a Two–Variable Optimization Task. *Ergonomics*, 22(l2), l325–l336.
Latorella, K.A. and Chamberlain, J.P. (2001). Decision-making in flight with different convective weather information sources: Preliminary results. *Focusing Attention on Aviation Safety. Proceedings of the 11th International Symposium on Aviation Psychology*, Columbus, Ohio.
Lin, M.L. and Radwin, R.G. (1998). Agreement between a frequency–Weighted filter for continuous biomechanical measurements of repetitive wrist flexion against a load and published psychophysical data. *Ergonomics*, 41(4), 459–475.
Lenth, R.V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55, 187–193.
Morawski, T., Drury, C.G. and Karwan, M.H. (1980). Predicting search performance for multiple targets. *Human Factors*, 22(6), 707–718.
Mozrall, J. and Drury, C.G. (1996). Effects of physical exertion on task performance in modern manufacturing: A taxonomy, a review and a model. *Ergonomics*, 39, 1179–1213.
Naugraiya, M. and Drury, C.G. (2009). A fractional factorial screening experiment to determine factors affecting discrete part process control. *Theoretical Issues in Ergonomics Science*, 10(1), 1–17.
NIST Statistics Handbook, http://www.itl.nist.gov/div898/handbook/.
Oxford English Dictionary, www.oxforddictionaries.com.
Pew, R.W. (1969). The speed-accuracy operating characteristic. *Acta Psychologica*, 30, 16–26.
Poulton, E.C. (1974). *Tracking Skill and Manual Control*, Academic Press, New York.

Ryan, B., Wilson, J.R., Sharples, S., Morrisroe, G. and Clarke, T. (2009). Developing a Rail Ergonomics Questionnaire (REQUEST). *Applied Ergonomics*, 40, 216–229.

Siegel, S. and Castellan, N.J. (1988). *Non-Parametric Statistics for the Behavioural Sciences*, 2nd edn., McGraw-Hill, New York.

Taguchi, G. (1986). *Introduction to quality engineering: Designing quality into products and processes*. Tokyo, Japan: Asian Productivity Organization.

Treisman, A. (1986). Properties, parts and objects. In: K.R. Boff, L. Kaufman and J.P. Thomas (Eds.), *Handbook of Perception and Human Performance*, Wiley, New York.

Wickens, C.D. and Carswell, C.M. (2012). Information processing. In: G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics*, John Wiley & Sons, Inc., Hoboken, NJ, pp. 117–161.

Winer, B.J. (2012). *Statistical Principles in Experimental Design*, Literary Licensing, LLC, New York.

Winer, B.J., Brown, D.R. and Michels, K.M. (1991). *Statistical Principles in Experimental Design*, McGraw-Hill Inc., New York.

Witkin, H. (1950). Individual perceptions of ease of perception. *Journal of Personality*, 19(1), 1–15.