The Earth is spherical ($p < 0.05$): alternative methods of statistical inference

KIM J. VICENTE* and GERARD L. TORENVLIET

Cognitive Engineering Laboratory, Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Rd., Toronto, Ontario, Canada M5S 3G8

Keywords: Data analysis; Methodology; Statistics; Analysis of variance; Null hypothesis significance testing; Statistical inference.

A literature review was conducted to understand the limitations of well-known statistical analysis techniques, particularly analysis of variance. The review is structured around six major points: (1) averaging across participants can be misleading; (2) strong predictions are preferable to weak predictions; (3) constructs and measures should be distinguished conceptually and empirically; (4) statistical significance and practical significance should be distinguished conceptually and empirically; (5) the null hypothesis is virtually never true; and (6) one experiment is always inconclusive. Based on these insights, a number of lesser-known and less-frequently used statistical analysis techniques were identified to address the limitations of more traditional techniques. In addition, a number of methodological conclusions about the conduct of human factors research are presented.

1. Introduction

In ergonomics science, the statistical analysis of data almost always relies on analysis of variance (ANOVA), which is a particular type of null-hypothesis significance testing (NHST). All have been taught these techniques and they are so commonly used and so widely accepted that they are frequently applied to data without a second thought. And, because the formulae for these statistical procedures have been embedded in easy-to-use software, their application is faster and less effortful than ever before. Having said that, consider the following quotations:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalised in the rote training of science students (Rozeboom 1997: 335).
>
> The physical sciences, such as physics and chemistry, do not use statistical significance testing to test hypotheses or interpret data. In fact, most researchers in the physical sciences regard reliance on significance testing as unscientific (Schmidt and Hunter 1997: 39).
>
> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories … is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology (Meehl 1978: 817).

These quotes are extreme, but the undeniable scientific point is that the statistical analysis techniques that are most familiar to, and most frequently used by, ergo-

* Author for correspondence. e-mail: benfica@mie.utoronto.ca

Theoretical Issues in Ergonomics Science ISSN 1463–922X print/ISSN 1464–536X online © 2000 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/1463922011003706 5

nomics scientists and practitioners have important limitations that could be overcome if we also relied on alternative methods of statistical inference.

Critiques of NHST and ANOVA go back at least to the 1960s (e.g. Rozeboom 1960, Bakan 1966, Meehl 1967, Lykken 1968), resurfaced periodically in the 1970s and 1980s (e.g. Meehl 1978, Hammond *et al.* 1986, 1987, Rosnow and Rosenthal 1989), and have appeared with increasing frequency and cogency during the past decade (e.g. Cohen 1990, 1994, Meehl 1990, Loftus 1991, 1993b, 1995, 2001, Loftus and Masson 1994, Hammond 1996, Thompson 1996, Harlow *et al.* 1997, Loftus and McLean 1997). These critiques have been met with rebuttals (e.g. Serlin and Lapsley 1985, Chow 1996, Abelson 1997, Hagen 1997, Harlow *et al.* 1997). The discussion has grown to the point where several journals have dedicated special sections to discussing the pros and cons of this issue (e.g. Thompson 1993, Shrout 1997, Chow 1998).

There is now a growing consensus that there are sound reasons to justify discontent with sole reliance on traditional methods of statistical data analysis. This dissatisfaction has led some journal editors to take significant actions to remedy the situation. As editor of *Memory & Cognition*, Loftus (1993a) strongly encouraged authors to adopt non-traditional data analysis and presentation methods. The editors of *Educational and Psychological Measurement* (Thompson 1994), *Journal of Applied Psychology* (Murphy 1997), and *Journal of Experimental Education* (Heldref Foundation 1997) went further, by requiring that authors report alternative statistical results. The editor of the *American Journal of Public Health*, the top journal in that discipline, even went so far as to ban statistical significance testing and any reference to *p*-values for a couple of years (Shrout 1997). More recently, the American Psychological Association Board of Scientific Affairs struck a Task Force on Statistical Inference consisting of a number of world-class researchers in both psychology and statistics 'to elucidate some of the controversial issues surrounding applications of statistics including significance testing and its alternatives; alternative underlying models and data transformation; and newer methods made possible by powerful computers' (Wilkinson *et al.* 1999: 594). There must be some substantive issues at stake for several scholars and organizations to take such strong actions.

The authors' experience has been that most ergonomics scientists are unaware of the controversy surrounding traditional methods of statistical inference, of the important limitations of these methods, and that alternative methods can be adopted to overcome some of these limitations (this opinion is empirically substantiated later, albeit informally). The purpose of this article is to discuss all of these issues in the context of ergonomics science. To be clear, the purpose is *not* to make an original technical contribution to this literature nor is it to dismiss the use of the traditional techniques. Instead, the aim is to bring the practical implications of this literature to the attention of the ergonomics science community, so that we can suggest some complementary ways of analysing data statistically.

2. Six issues in statistical inference

This literature review is organized into six sections, each of which identifies a major issue in statistical inference and a corresponding set of alternative methods. Although some of these points may seem self-evident, the review will show that they are frequently not heeded by ergonomics scientists. By making each of these points explicit, new ways of analysing data can be identified. These lesser-known statistical analysis techniques may, in turn, provide a different, and sometimes perhaps more valuable, set of insights into data.

Before proceeding, several caveats need to be mentioned. First, some of the limitations of ANOVA that are discussed are found only in more modern treatments and usages, and not in the original Fisherian formulation. But, since it is the former, rather than the latter, that is familiar to and generally adopted by most of the intended readers, it seems nevertheless worthwhile to discuss these limitations. Secondly, several of the alternative data analysis techniques discussed can be, but are not usually, derived from information generated by an ANOVA. However, regardless of how they are calculated, one of the main points is that alternative methods of statistical inference provide valuable information that is complementary to that which is usually reported using traditional techniques. Thirdly, in some cases, the limitations identified are not as much with ANOVA itself but rather with the way in which it is generally used. For example, there is no reason why the information usually provided by ANOVA cannot be supplemented by some of the complementary measures identified. The main point is that this practice is not usually followed in the human factors community, and that there are good reasons to change the way in which we currently analyse our data. Finally, we do not claim to have identified a panacea for the problems with traditional techniques. The alternative methods proposed, whilst useful and complementary, are not perfect. Furthermore, there is no substitute for having a clear idea of a study's objectives before determining the right mix of statistical techniques to apply to the data.

2.1. *Averaging across participants can be misleading*

We will begin by discussing an issue with which many researchers are familiar but that is, nevertheless, frequently overlooked. ANOVA involves averaging across participants. As a result, it is commonplace for ergonomics scientists to assess statistical significance at an aggregate level of group means. Yet, taking an average only makes statistical sense if the samples being aggregated are qualitatively similar to each other. Without looking at each participant's data individually, we do not know if the group average is representative of the behaviour of the individuals. In fact, it is possible for a group average to be a 'statistical myth' in the sense that it is not indicative of the behaviour of any single participant in the group.

Data from a 6-month longitudinal study conducted by Christoffersen *et al.* (1994) can be used to illustrate this point in a salient fashion. Figure 1 shows a learning curve illustrating the average time to complete a task as a function of experience. The curve is based on data averaged over six participants. A power law fit has been superimposed on the aggregate data. Based on visual inspection alone, it can be seen that there is a good fit between the data and the power law curve. A regression analysis showing a substantial r^2 value of 0.74 confirms this impression. One might conclude from this aggregate-level analysis that these data provide support for the power law of practice (Newell and Rosenbloom 1981). However, such a conclusion could be premature. Without looking at each participant's data it cannot be known whether the elegant, aggregate power curve fit would provide an equally good account of the skill acquisition of each individual.

Figure 2 shows the learning curve data for one of the six participants. Again, a best fit power law curve has been superimposed, but this time on the raw data of an individual, not the mean data of the group. The degree of fit between the power law of practice and this participant's data is obviously poor. Thus, to use the group

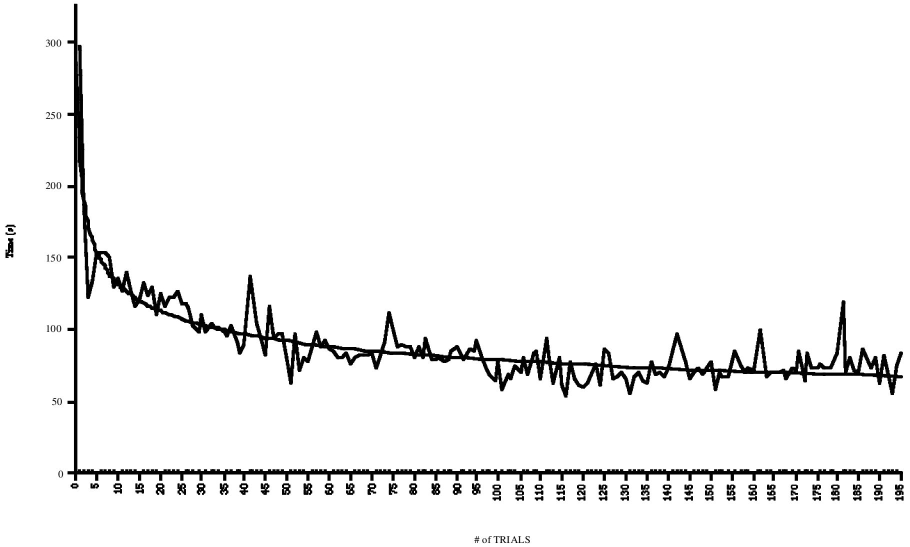

Figure 1. Learning curve averaged over six participants (Christoffersen *et al.* 1994).

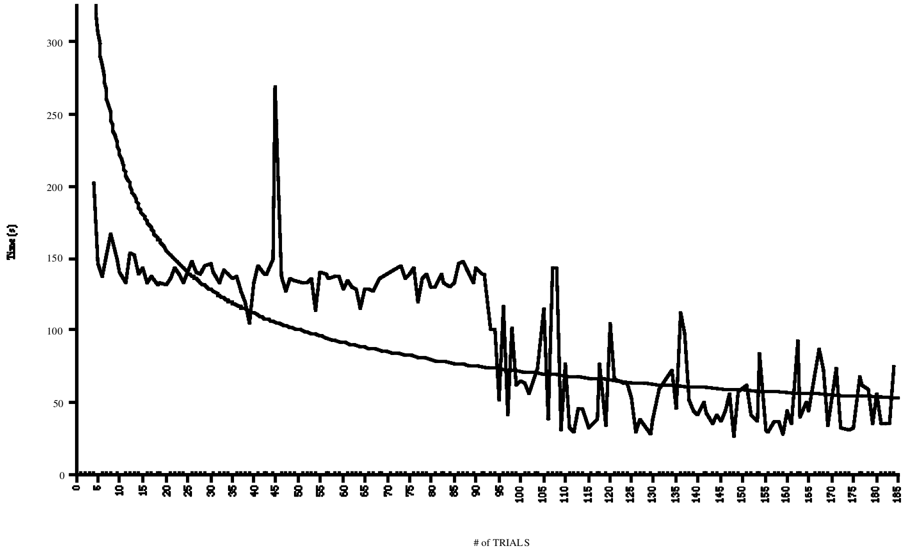

Figure 2. Learning curve for one of the six participants (Christoffersen *et al.* 1994).

average as a basis for generalizing to individuals would be quite misleading in this case.

Plateaus in learning curves and the dangers of aggregating data over participants are hardly new insights (Bryan and Harter, 1897, 1899, Woodworth 1938). Yet, as Venda and Venda (1995) pointed out, these insights are still frequently ignored by many, although by no means all, ergonomics scientists. It is believed that, in part, these oversights result from the fact that ANOVA encourages the aggregation of data over participants. Consequently, a special, added effort must be made to examine the data for each individual to see if what is true of the group is also true of the individual.

Taking the dangers of aggregating over participants to heart can actually lead to new and perhaps more compelling ways of analysing data. Several statistical methods can be used to address the aforementioned problems, but here only one is discussed. In cases where a within-participants design is adopted, each individual can be viewed as an experiment and see if the theoretical predictions being tested hold for each person. An example of this type of test is provided by Vicente (1992), who compared the performance of the same participants with two different interfaces, one labelled P and the other labelled $\mathrm{P} + \mathrm{F}$. There were theoretical reasons for hypothesizing that the $\mathrm{P} + \mathrm{F}$ interface would lead to better performance than the P. However, rather than just seeing if the group means of the two conditions differed, Vicente also conducted a more detailed analysis to see if the theoretical prediction held for each and every participant. The number of participants for whom the hypothesized relationship ($\mathrm{P} + \mathrm{F} > \mathrm{P}$) held was counted and then this count was analyzed statistically by conducting a sign test (Siegel 1956). In one analysis, the $\mathrm{P} + \mathrm{F}$ interface led to better performance than the P for 11 out of 12 experts, a statistically significant result.

This example is important for two reasons. First, in at least some applied situations, it may be more important for ergonomics scientists and practitioners to know how often an expected result is obtained at the level of the individual than at the level of an aggregate. For example, say the performance impact of an advanced control room for a nuclear power plant is being tested. Are we only interested in knowing whether the mean performance of the new control room is better than that with the old, or are we also interested in knowing the proportion of operators for which performance with the new control room is better? It seems that the latter is also valuable. After all, an ANOVA could show that the new interface leads to a statistically significant improvement in performance, but an analysis like the one conducted by Vicente (1992) might reveal that the new interface only leads to better performance for half of the operators (a non-significant result with a sign test). In this case, the aggregate level analysis is misleading, just as the aggregate data in figure 1 are. And, because of the potential hazard involved, designers might be wary about introducing a new control room that will result in a performance decrement for half of its operators. Secondly, this example also shows that non-parametric tests (e.g. the sign test and the c^2 test), that are statistically less powerful than parametric tests, can actually be used in innovative ways to test strong predictions. This topic is discussed in more detail next.

2.2. *Strong predictions are preferable to weak predictions*

Empirical predictions can be ordered on a continuum from strong to weak (Vicente 1998). At the strong end, there are *point* predictions. To take a hypothetical example from physics, a theory might predict that the gravitational constant, G, should be $6.67 \times 10^{-11}\mathrm{Nm}^2/\mathrm{kg}^2$. An experiment can then be conducted to see how well the data correspond to this point prediction. Slightly farther along the continuum, *interval* predictions are found. To continue with the same example, a different theory might only predict that $6 \times 1^{-11}\mathrm{Nm}^2/\mathrm{kg}^2 < G < 7 \times 10^{-11}\mathrm{Nm}^2/\mathrm{kg}^2$. An interval prediction is weaker than a point prediction because it is consistent with a wider range of results. Still farther towards the weaker side of the continuum, *ordinal* predictions are found. For example, a third theory might only predict the direction of the force of gravity. In this case, all one would know is that gravity pulls objects towards, rather than away from, the earth. Finally, at the weak end of the continuum, *categorical* predictions are found. For example, a very primitive theory might merely predict that the force of gravity on the earth is statistically significantly different from zero, regardless of its direction (i.e. that gravity exists).

Meehl (1967, 1978, 1990) has repeatedly pointed out that a mature science should strive to make predictions towards the strong end of this continuum, but that psychology has generally failed to do so. The same claim can generally be made for ergonomics science, although there certainly are exceptions. According to Meehl, one of the causes of this lack of maturity is that researchers have let the constraints of the statistical analysis techniques with which they are most familiar (i.e. ANOVA) govern the strength of the predictions they make. And, because ANOVA is usually used by behavioural researchers to determine if an effect is significantly different from zero (i.e. if the independent variable has no effect whatsoever), ergonomics scientists frequently restrict themselves to testing categorical predictions. This area is the weakest on the continuum and is, thus, indicative of a comparatively immature scientific practice. Because we are so accustomed to following this procedure, we may not even be aware that we are merely testing a categorical prediction. However, the hypothetical example cited above shows just how weak such a test really is. Merely predicting that gravity exists does not seem like an impressive scientific achievement. Granted, pairwise comparisons of means can be used to test ordinal predictions at an aggregate level, but this is still a far cry from the interval and point predictions located on the strong end of the continuum described above.

It could be argued that most areas of human factors research have not reached the level of theoretical maturity to make point or interval predictions. There is merit to this objection, but, even so, it does not follow that we cannot or should not be more ambitious than we have been in the past. Rather than letting familiar statistical analysis techniques keep us from achieving a mature science, we should instead seek out a different set of techniques that can be used to test stronger predictions, whenever they can be made. For a practical science like ergonomics, the value of quantitative prediction is particularly important. Engineering design always involves tradeoffs, so, in making the case for ergonomics science, it is invaluable to know how big an impact a particular design intervention will have on performance or safety (Chapanis 1967).

The innovative work of Hammond *et al.* (1987) provides an example of how ergonomics scientists can begin to make stronger predictions and how these can be tested using untraditional statistical analysis techniques. Hammond *et al.* were interested in comparing the efficacy of intuitive and analytical cognition in expert judgement. Accordingly, they conducted an experiment to investigate the impact of two independent variables, depth task characteristics and surface task characteristics, on the level of performance and the type of cognitive processing (i.e. intuition vs. analysis) of 21 professional highway engineers. There were three levels for the depth task characteristics dimension: (a) an aesthetics task that was intended to induce intuition; (b) an highway capacity calculation task that was intended to induce analysis; and (c) a safety judgement task that was intended to induce a hybrid of intuition and analysis. Each of these tasks was presented in three different formats, each with a different set of surface characteristics: (a) film strips that were intended to induce intuition; (b) formulae that were intended to induce analysis; and (c) bar graphs that were intended to induce a hybrid of intuition and analysis. Each of the 21 highway engineers experienced each of the nine combinations of depth and surface task characteristics.

From a traditional perspective, this experimental design fits neatly into a within-participants 3 × 3 randomized block factorial ANOVA. However, analysing the data in this fashion would only allow the experimenters to test null hypotheses. Such a test only amounts to an evaluation of a categorical prediction (equivalent to the fact that gravity exists). Furthermore, the ANOVA would only evaluate the results at an aggregate level of analysis, and, thus, could mask some important individual differences (see the previous section).

Hammond *et al.* (1987) addressed these deficiencies in three ways. First, instead of evaluating the NHSTs associated with ANOVA, they instead tested the prediction that the results from the nine experimental conditions should occur in a particular order predicted by the theory motivating their research. Note that this is a much stronger prediction. Instead of just hypothesizing that the effect was different from zero, Hammond *et al.* were committing to one specific ordering of their experimental conditions. Also, because there was a total of nine conditions in their experiment, there are many possible orderings that could conceivably occur ($9! = 362\,880$). Only one of these orderings is perfectly consistent with the prediction they were making. Secondly, instead of testing this ordinal prediction at the level of a group aggregate, they tested it individually for each of the 21 participants. That is, Hammond *et al.* (1987) predicted 'the exact order of appearance of a specific type of cognitive activity for each engineer separately, over a set of nine conditions, each of which included a sample of 40 highways. Thus, there were in effect 21 individual experiments, each of which tested the ... theory' (p. 769). Because of the level of specificity involved, the risk of being wrong is again greater than with ANOVA, thereby resulting in a stronger set of predictions. Thirdly, to test the predicted ordering on a participant-by-participant basis, Hammond *et al.* relied on correlational analysis and χ^2-based order table analysis. The technical details can be found in Hammond *et al.*'s article, but the basic rationale is similar to that for the Vicente (1992) study described in the previous section. Non-parametric tests were used to determine how often the predicted order of results was observed at the level of individuals rather than at the aggregate level of the group.

The study conducted by Hammond *et al.* (1987) provides a role model to show how the maturity of ergonomics science can be enhanced by using alternative statistical analysis techniques to test stronger predictions than those that are usually assessed using ANOVA alone.

2.3. *Constructs and methods for measurement should be distinguished conceptually and empirically*

Even if we were able to make and evaluate stronger predictions, the level of science is only as good as the empirical methods used. Of particular importance is the relationship between the constructs that are used to make predictions and the methods of measurement that are used to evaluate those predictions. This linkage is one of the key epistemological foundations supporting any kind of scientific activity, including ergonomics science (cf. Xiao and Vicente 2000). As Campbell and Fiske (1959) pointed out in their seminal article over 40 years ago, there are certain basic criteria that must be met before a pattern of experimental results can be interpreted in a meaningful fashion. *Reliability* refers to the extent to which similar results are obtained when the same construct is assessed using the same method of measurement under comparable conditions. If results cannot be replicated, then there is a lack of reliability. *Convergent validity* refers to the extent to which similar results are obtained when the same construct is assessed using different methods of measurement under otherwise comparable conditions. If different methods give different results, then the pattern of findings is contaminated, and, thus, difficult to interpret. Instead of observing the effects of the construct of interest, you are instead observing the effects of the way in which the construct was measured—a much less interesting phenomenon, unless you are a methodologist. Finally, *discriminant validity* refers to the extent to which distinct results are obtained when different constructs are assessed using the same measurement method under comparable conditions. If different constructs lead to similar results, then the pattern of findings is again contaminated, and, thus, difficult to interpret. Instead of observing differential effects across the various constructs of interest, you are instead observing similar effects caused by the method of measurement.

A few hypothetical ergonomics science examples can help make these abstract concepts more concrete. If an empirical investigation of the interaction between spatial ability and mental workload for a particular work context were being performed, how could the three criteria identified by Campbell and Fiske (1959) be operationalized? Beginning with the issue of reliability, whatever method used to measure each construct should lead to consistent results under comparable conditions. For example, the test for spatial ability should have a high test–retest correlation. Otherwise, we cannot have much confidence in our knowledge of one of the key constructs in the experiment. Moving on to convergent validity, different methods of measuring the same construct should lead to consistent results under comparable conditions. For example, if there were two different methods for measuring mental workload (e.g. a computer-based version and a paper-based version of the same subjective rating scale), it would be ideal if those methods were to give the same results for the same participant for a particular trial. If the two methods give different results, then the variance in the data is being caused by the method of measurement. In such a case, confident inferences cannot be made about the item of interest, namely the construct of mental workload. As for the third criterion of discriminant validity, the same measurement methods should lead to distinct results for different constructs of interest. For example, a computer-based test of spatial ability should be more strongly correlated with a paper-based test of spatial ability than with a computer-based assessment of mental workload. If this criterion is not met, then there is too high a correlation between tests that are intended to measure different constructs. Once more, such a result would provide a very shaky foundation for scientific knowledge.

In each of these three cases, the key objective is to determine whether the results observed can be safely attributed to the content of the constructs of interest rather than the form of the methods that are used to measure those constructs. Campbell and Fiske (1959) refer to the latter as ‘methods variance’. To make sure that methods variance is not contaminating the results, a way is needed to evaluate reliability, convergent validity, and discriminant validity empirically. To achieve this goal requires that any one experiment has at least two constructs and at least two methods of measurement. Using these insights, Campbell and Fiske proposed an analysis technique that can allow experimenters to determine if they are measuring the construct in which they are interested, rather than something entirely different. This technique, called the Multitrait-Multimethod Matrix (MTMM), was originally developed for the specific case of investigating individual differences (thus, the emphasis on traits). More recently, the technique was extended by Hammond *et al.* (1986) so that it can be applied to a much wider range of behavioral phenomena.

Campbell and Fiske (1959) used the MTMM technique to review the literature on individual differences. Their analysis painted 'a rather sorry picture' (p. 93) of the validity of the measures that had been used in that literature. Most of the results that had been generated were more likely to have been determined by the methods used for measurement than by the traits that had been hypothesized to account for the results. The MTMM technique provides a way of identifying such situations. However, as Hammond *et al.* (1986) pointed out, the technique is rarely used in experimental psychology. The same is true of ergonomics science; although some researchers have investigated convergent validity using other techniques, studies of all three threats to validity using the MTMM technique are exceedingly rare. Researchers tend to analyse their data using other more familiar techniques, such as ANOVA. However, those techniques do not provide an analytical means for evaluating reliability, convergent validity, and discriminant validity, as does MTMM. As a result, researchers cannot know if their results are being caused by methods variance. Hammond *et al.* make a very strong case that this situation makes it exceedingly difficult to develop a cumulative scientific knowledge base. Instead, the result is conflicting findings because researchers have not determined empirically that the preconditions for sound scientific knowledge have been satisfied in their experiments. The MTMM technique and its extensions provide a systematic means of remedying this situation.

Lee's (1992, Lee and Moray 1994) investigation of the relationship between operator trust, self confidence, and the use of automation is the only application of MTMM in the ergonomics science literature of which we are aware. As such, it can be used to illustrate the value of conceptually and empirically distinguishing between constructs and methods of measurement. In Lee's study, there were two constructs of interest, the operators' trust in the automation's ability to control a process and the operators' self confidence in their own ability to control a process. There were also two methods of measurement, ratings on a subjective scale and the frequency of operators' monitoring behaviour. The matrix shown in table 1 can be built from this experimental design. Note that Lee did not present the same conditions more than once, so it is not possible to assess the reliability values along the diagonal of table 1.

Nevertheless, it is possible to use MTMM to assess discriminant and convergent validity. Convergent validity is exhibited if different methods lead to similar results for the same construct under comparable conditions. There are two cells in table 1 that are relevant to assessing this criterion. The first is the cell in the second row and first column of table 1. One should expect to see a high correlation value in this cell (indicated by a '✔') because trust measured by monitoring behaviour should lead to results that are comparable to those obtained by measuring trust with a subjective scale. The second relevant cell is in the fourth row and third column of table 1. One should also expect to see a high correlation value in this cell because self confidence measured by monitoring behaviour should lead to results that are comparable to those obtained by measuring self confidence with a subjective scale.

Divergent validity is exhibited if the same or different methods lead to different results for different constructs under comparable conditions. The remaining four cells in the bottom left corner of table 1 are relevant to assessing this criterion.

Table 1. A multitrait-multimethod matrix relating trust and self confidence measured by subjective scales and frequency of monitoring behaviour for Lee's (1992, Lee and Moray 1994) study.

		Trust		Self-confidence	
		SS	MB	SS	MB
Trust	SS				
	MB	✔			
Self-confidence	SS	×	×		
	MB	×	×	✔	

✔: a high correlation is expected in that cell (i.e. convergent validity).
×: a very low correlation is expected in that cell (i.e. divergent validity).
SS: subjective scales.
MB: monitoring behaviour.

Table 2. A multitrait-multimethod matrix relating trust and self confidence measured by subjective scales and frequency of monitoring behaviour (Lee 1992, Lee and Moray 1994). The values are the means of *z*-transformed correlation coefficients of individual operators. Abbreviations are as in table 1.

		Trust		Self-confidence	
		SS	MB	SS	MB
Trust	SS				
	MB	0.15 (✔)			
Self-confidence	SS	0.42 (×)	0.04 (×)		
	MB	−0.07 (×)	−0.08 (×)	0.04 (v)	

✔: a high correlation was expected in that cell (a sign of convergent validity).
×: a very low correlation was expected in that cell (a sign of divergent validity).

One should expect to see lower correlation values (indicated by a ×) in these cells. For example, ratings of self confidence on a subjective scale and ratings of trust on a subjective scale should be weakly correlated, if at all, because they are measuring different constructs. If the data turn out to be strongly correlated, then one can infer that methods variance is at play (i.e. that the results are determined more by the fact that a subjective rating scale is being used as a method of measurement than by the constructs that are of real interest).

Table 2 shows the results that Lee (1992) obtained using the MTMM technique. A cursory examination shows that the criteria of discriminant and convergent validity were not consistently met in this study. For example, the highest correlation in table 2, 0.42, is that between two different constructs (trust and self confidence) when they were measured with a common method (subjective scales). One would expect to see a low correlation here because different constructs should lead to different results.

The fact that there is a comparatively large correlation suggests that methods variance is contaminating the results. As another example, there is a very low correlation, 0.04, between the two methods of measuring self confidence. One would expect to see a high correlation here because different methods for measuring the same construct should lead to the same results. The fact that there is a very low correlation suggests that methods variance is again contaminating the results.

This example provides a concrete illustration of how the MTMM technique can be used to evaluate discriminant and convergent validity in ergonomics science. Unless these criteria are satisfied, the results obtained from any study cannot lead to sound scientific knowledge. If the results obtained by Lee (1992, Lee and Moray 1994) and those reviewed by Campbell and Fiske (1959) and Hammond *et al*. (1986) are any indication, then the ergonomics science literature is likely to be full of results that are caused by methods variance rather than by the substantive, theoretical issues that motivated the research. The MTMM technique provides a means of identifying, and, thus, beginning to remove, such obstacles to scientific progress.

2.4. *Statistical significance and practical significance should be distinguished conceptually and empirically*

It is a truism in ergonomics science and practice that statistical significance is not the same as practical significance (Chapanis 1967). This truism has a sound basis in statistics (although, as will be discussed shortly, it is frequently ignored). For example, the NHSTs that are usually associated with ANOVA are measures of statistical significance, or, more precisely, the probability that the data could have arisen, given that the null hypothesis is true. This type of test does not tell us much that is likely to be very useful in determining the practical significance of a finding. To assess the latter, information is needed about magnitude, and it is useful to distinguish between two types: (a) measures of association strength; and (b) measures of effect size (Snyder and Lawson 1993). Using measures of magnitude and some criterion from the domain of interest regarding what magnitude is important for applied purposes, it is possible to assess the practical significance of a result. Such pragmatic information is of great importance to an intrinsically applied discipline, like ergonomics science. Nevertheless, it is much more common to see tests of statistical significance than tests of strength of association or effect size reported in the literature. In this subsection, the value added provided by data analysis techniques that provide magnitude information will be discussed.

Because of the central role that they play in multiple linear regression, measures of *association strength* are probably more familiar to readers, so will be discussed first (see Snyder and Lawson 1993 for more details). The most common statistic (and the simplest to calculate) is the proportion of the total variance explained by a particular effect, usually referred to in ANOVA as eta-squared. Despite the fact that eta-squared is easy to calculate from the information in an ANOVA table (it is simply a ratio of sums of squares), it is rare to see such information reported. Moreover, because the emphasis has been on significance tests, ergonomics scientists sometimes only report the result of the *F*-test and do not provide the information from the ANOVA table that could be used by other researchers to calculate the strength of association. This practice is unfortunate because an *F*- or *p*-value does not provide any information about the magnitude of an effect. In contrast, eta-squared provides an estimate of how strong the association is between the independent variable(s) of interest and the dependent variable chosen. When combined with criteria from a particular domain of interest, this statistic can help in making inferences about practical significance.

For example, if one is interested in the ergonomics science problem of worker selection, it is known from the individual differences literature that it is unusual for a particular selection test to account for say 20% of the variance in the data. Thus, if an eta-squared value is obtained that is greater than this benchmark value, then it is known that the result is practically significant (i.e. it may be used to develop a better basis for worker selection). Note that the result may or may not be statistically significant. If the sample size is small, it is possible to have a comparatively large eta-squared value (e.g. a selection test accounting for more than 20% of the variance) and results that are not statistically significant. Such a result could, nevertheless, be considered practically significant. Conversely, if the sample size is large, it is possible to obtain statistically significant results and yet have a comparatively very low eta-squared value (e.g. a selection test accounting for only 1% of the variance). Such a result would be of little practical value (i.e. it could not be used to develop a useful basis for worker selection). The bottom line is that measures of association strength, like eta-squared, provide a more complementary set of insights into the results than do the statistical significance tests that are typically reported with ANOVA.

There are various types of statistics that can be used to obtain information about association strength. For example, omega squared is an estimate of the population strength of association. It can be computed from knowledge of the *F*-statistic, number of treatment levels (p), and sample size (n) as follows:

$$\omega^2 = \frac{(p-1)(F-1)}{(p-1)(F-1)+np}.$$

Some sample statistics of association, like eta-squared, are biased estimates, meaning that they tend to overestimate systematically the proportion of variance explained. To be conservative, it is more appropriate to use unbiased estimates of strength of association that compensate for this tendency to overestimate. Snyder and Lawson (1993) describe several such statistics, and readers are referred there to obtain more details. But, regardless of the particular measure used, the fundamental point remains the same—measures of association strength provide information that complements that provided by statistical significance tests, and the former information is of greater interest in determining practical significance.

A similar argument holds for the second type of magnitude information. *Effect size* (Cohen 1988, 1990, 1994, Rosnow and Rosenthal 1989, Abelson 1995, Rouanet 1996) is a measure of the magnitude of an effect, and, thus, can also be used along with domain-specific criteria to indicate the degree of practical importance of ergonomics science results. Note that effect size and statistical significance provide complementary information: 'it is very important to realize that the effect size tells us something very different from the *p*-level. A result that is statistically significant is not necessarily practically significant as judged by the magnitude of the effect' (Rosnow and Rosenthal 1989: 1279).

In an applied science like ergonomics, effect size plays a critical role. As Chow (1996: 8) observed: 'a significant result may be a trivial one in practical terms. Alternatively, an important real-life effect may be ignored simply because it does not reach the arbitrary chosen level of statistical significance'. Despite this truism, an informal survey of the ergonomics science literature (see below) reveals that statistical significance is reported far more frequently than is effect size. Once again, it is believed that this is indicative of an over-reliance on NHST and ANOVA. Neither of these statistical techniques provides direct measures of effect size.

Because of the foundational importance of practical significance to ergonomics science, it is important that effect sizes are calculated in addition to assessing statistical significance. Several ways of calculating effect size have been proposed in the literature. For example, Cohen (1988) has proposed the standardized mean difference statistic, *d*, as a generalizable measure of effect size. Based on the results that are typically found in behavioural research, Cohen has suggested that $d = 0.2$ is indicative of a small effect, $d = 0.5$ is indicative of a medium sized effect, and that $d = 0.8$ is indicative of a large effect. These nominal values provide a starting point for evaluating the practical significance of research results.

Like the other points made earlier, the distinction between statistical significance and effect size is best conveyed by an example (adapted from Rosnow and Rosenthal 1989). Consider two hypothetical experiments, both conducted to evaluate the impact of two types of training programmes, T_1 and T_2, on human performance. In one experiment (with $n = 80$), T_1 is found to lead to significantly better performance than $T_2(t(78) = 2.21,\ p < 0.05)$. In another experiment (with $n = 20$), no significant difference between T_1 and T_2 is observed ($t(18) = 1.06, p > 0.30$). By relying solely on these tests, we might be tempted to conclude that the second experiment failed to replicate the results of the first. Such a conclusion would cast doubt on the practical impact of T_1 on human performance.

Calculating effect size adds new information that can help put the results in a more realistic light. In this hypothetical example, the magnitude of the effect is actually the same for both experiments ($d = 0.50$), despite the fact that the *p*-values for the two experiments differed considerably. How is this possible? Because the second experiment had a smaller sample size, the power to reject the null hypothesis at $\alpha = 0.05$ was very low, only 0.18. In contrast, the first experiment had a much larger sample size, and, thus, its power was 0.6—over three times greater than that in the second experiment. These results clearly show the difference between statistical significance and effect size, and, thus, why it is important to calculate effect size.

2.5. *The null hypothesis is virtually never true*

There is another reason for not relying solely on the results produced by NHST and ANOVA. As odd as it may sound, there are very good reasons to argue that the null hypothesis is almost never really true in behavioural research. This point has been made by many noted researchers (e.g. Meehl 1967, 1978, 1990, 1997, Cohen 1990, 1994, Loftus 1991, 2001, Abelson 1995, Thompson 1996, Steiger and Fouladi 1997), but its implications have not been taken as seriously as they should be in ergonomics science.

Consider a typical ergonomics experiment comparing the effect of two treatments (e.g. two interfaces, two training programmes, or two selection criteria) on human performance. One group of participants is given Treatment X, whereas another is given Treatment Y. The null hypothesis in such a study is that there is no difference whatsoever between the population means for the two treatment groups. Can we really consider such a hypothesis seriously? For example, can we realistically expect that the effects of two different interfaces are exactly the same to an infinite number of decimal points? Meehl (1967) was perhaps the first of many to point out that the answers to questions such as this one are almost sure to be 'no':

> Considering ... that everything in the brain is connected with everything else, and that there exist several 'general state-variables' (such as arousal, attention, anxiety and the like) which are known to be at least slightly influenceable by practically any kind of stimulus input, it is highly unlikely that any psychologically discriminable situation which we apply to an experimental subject would exert literally zero effect on any aspect of performance (p. 162).

One way to illustrate the implausible nature of the null hypothesis is to consider the insight that is gained by using NHST with very large sample sizes. Meehl (1990) describes a data set obtained by administering a questionnaire to 57 000 high school seniors. These data were analysed in various ways using χ^2 tables, with each analysis looking at the interaction between various categorical factors. In each case, the null hypothesis was that there was no interaction between the categories being compared. A total of 105 analyses were conducted. Each analysis led to statistically significant results, and 96% of the analyses were significant at $p < 0.000\,001$. As Meehl observed, some of the statistically significant relationships are easy to explain theoretically, some are more difficult, and others are completely baffling. To take another example, with a sample size of 14 000, a correlation of 0.0278 is statistically significant at $p < 0.001$ (Cohen 1990). Figures such as these show that the scientific knowledge that is gained solely by refuting the null hypothesis is minimal, at best. The same types of problems can occur in studies with low sample size as well (Chapanis 1967)

If the null hypothesis is almost always false, then the act of conducting a NHST means something very different than what we usually thinks it means. Rather than being a generator of scientific insight, the NHST instead becomes an indirect indicator of statistical power. For example, if a data set does not yield results that are significant at $p < 0.05$, then the likely interpretation is not that the alternative hypothesis is incorrect, but that the sample size of the experiment was too low to obtain an acceptable level of power. After all, as the Meehl (1990) and Cohen (1990) examples show, if one has the fortitude and resources to include enough participants in experiments, then virtually any null hypothesis can be rejected. Thus, the value of just conducting a NHST is minimal. As Cohen (1994: 1001) has pointed out, 'if all we ... learn from a research is that A is larger than B ($p < 0.01$), we have not learned very much. And this is typically all we learn'.

Accepting the fact that the null hypothesis is virtually never true in behavioural research, what are the implications for the statistical analysis of data? The short answer is that it would be useful to have other data analysis techniques that offer more insights than a NHST or ANOVA alone. Two related techniques have frequently been suggested to fulfil this role, power analysis and confidence intervals (Cohen 1990, 1994, Loftus 1993b, 1995, 2001, Loftus and Masson 1994, Abelson 1995, Meehl 1997, Steiger and Fouladi 1997).

Rather than using the results of a NHST as a surrogate measure of statistical power, researchers would be better off if they calculated power directly before an experiment is conducted to obtain a proper sample size. The resulting measure provides an explicit indication of the sensitivity of an experiment to detect an effect of interest. In addition to its preferred *a priori* role in determining the sample size for a planned experiment, the calculation of power is also valuable in a *post hoc* role where the failure to reject the null hypothesis is used as evidence to falsify a particular theory. In these situations, it is essential that statistical power be calculated. After all, the failure to reject the null hypothesis could simply be caused by the fact that too small a sample size was used to detect the effect of interest. Therefore, to keep ergonomics scientists from 'falsifying' theories simply by not including enough participants in their experiment, it would be useful to present calculations of power. Doing so would provide additional information over that obtained just by conducting a NHST or ANOVA.

Confidence intervals provide another data analysis technique that can be used to obtain greater insight into experimental results. Whereas the results of a NHST merely show the probability that the data could have arisen given that the null hypothesis were true, confidence intervals directly provide information about the range of values within which population parameters are likely to be found. As such, they have several advantages over NHST. First, confidence intervals provide a graphical representation of results rather than an alphanumeric representation (see the example, below). This format makes it easier for researchers to extract information from their data analysis. Secondly, the width of a confidence interval provides an indication of the precision of measurement. Wide confidence intervals indicate imprecise knowledge, whereas narrow confidence intervals indicate precise knowledge. This information is not provided by the *p*-value given by a NHST. Thirdly, the relative position of two or more confidence intervals can provide qualitative information about the relationships across a set of group means. If two confidence intervals do not overlap, then the means are significantly different from each other statistically, otherwise they are not. Whilst this information can be gained from a standard NHST, confidence intervals add information about the order of means across groups, information that cannot be found in, for instance, an ANOVA table. Finally, confidence intervals also allow us to assess the statistical significance of individual effects. If a confidence interval on a group mean includes zero, then the treatment did not have a significant effect. (To achieve a similarity between NHST and confidence intervals, the type I error rate for the NHST should be equal to 1— the confidence coefficient.) Therefore, the plotting of confidence intervals provides researchers with more insights into their data than could be obtained by NHST or ANOVA alone.

The informativeness of confidence intervals can be illustrated with a simple example borrowed from Steiger and Fouladi (1997). Figure 3 shows data from three hypothetical experiments, each consisting of two conditions. Thus, each confidence interval in the figure is for the difference between a pair of means. Each experiment was performed in the same domain and using measures with approximately the same amount of variability. Note that the confidence intervals from experiments 1 and 3 do not include zero. In these two cases, a NHST would indicate that the difference in means is significantly different from zero, leading to a decision to reject the null hypothesis. In experiment 2, the confidence interval includes zero. Thus, in this case, a NHST would indicate that the difference in means is not significantly different from zero. Thus, the confidence intervals in figure 3 provide all the information that can be obtained directly from a NHST, the difference being that that information is presented graphically.

However, additional information not directly available from a NHST can also be obtained from confidence intervals. For example, based on the results presented above, the NHST might lead us to believe that the results from experiment 2 do

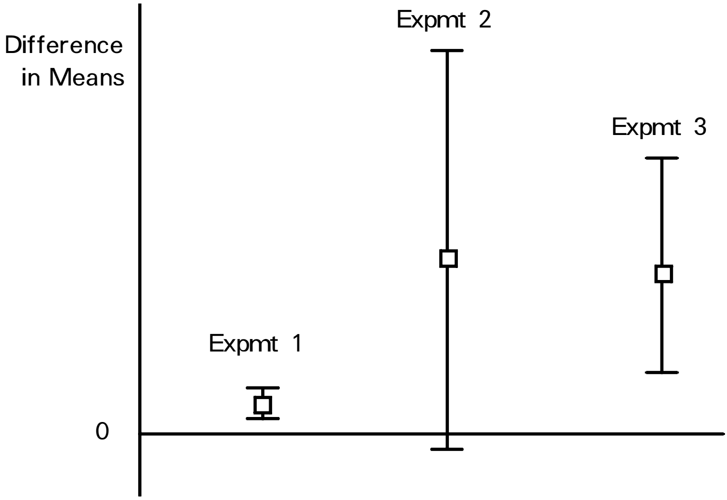

Figure 3. Hypothetical example showing how confidence intervals reflect different degrees of measurement precision (adapted from Steiger and Fouladi 1997).

not agree with those from the other two experiments. The confidence intervals provide a graphical basis for reaching a different interpretation. Experiment 1 had a very large sample size and a very high level of precision, resulting in a very narrow confidence interval band. However, precision should not be confused with magnitude. Figure 3 clearly shows that the effect size in experiment 1 is comparatively very small. The only reason why the null hypothesis was rejected was because the measurement precision was so great. Thus, the results from experiment 1 are precise but small in magnitude.

In contrast, experiment 2 has a very wide confidence interval band that indicates poor measurement precision. However, it could very well be that the magnitude of the difference in means in experiment 2 is larger than that in experiment 1, but that the power was just inadequate to detect that effect. Thus, the results from experiment 2 are imprecise, and, thus, it is not known with any certainty if they are large or small in magnitude.

Finally, experiment 3 also has a relatively wide confidence interval band indicating poor measurement precision. Nevertheless, this confidence interval does not overlap with that from experiment 1, indicating that the magnitude of the difference in means in experiment 3 is greater than that in experiment 1. Thus, the results from experiment 3 are comparatively imprecise but larger in magnitude.

The important point to take away from this hypothetical example is that confidence intervals provide much more information than do NHSTs alone. Furthermore, that information is provided in a graphical format, thereby making it easier for ergonomics scientists and practitioners to pick up meaningful patterns perceptually (e.g. width of bands, overlap across bands, inclusion of the zero point). In this hypothetical example, the added information leads to a very different interpretation than may have been obtained by reliance on NHST alone.

In summary, power analysis and confidence intervals are rarely-used, but very valuable, statistical analysis techniques. Together, they allow us to gain richer insights into data, and thereby allow us to go beyond merely rejecting the null hypothesis. Note that confidence intervals can be calculated for effect sizes and measure of strength of association as well, thereby combining the respective advantages of each of these techniques into one statistical procedure (Fowler 1985, Rosnow and Rosenthal 1989, Cohen 1990, 1994). In this way, information would be obtained about the precision of knowledge of effect size or strength of association, information that is surely to be of practical value in ergonomics science and practice (see previous subsection).

2.6. *One experiment is always inconclusive*

This final point cuts across the comparative advantages and disadvantages of any particular set of statistical analysis techniques. No matter how carefully it is designed, no matter how sophisticated the equipment, no matter how clever the researcher, and no matter what statistical analysis techniques are used, any one experiment alone can never provide definitive results. The origin of this limitation is a logical one. Empirical research relies on inductive inference, and as any philosopher or logician knows, induction provides no guarantees.

The same conclusion can be obtained empirically from the history of science. To take but one example, several times experimental results were obtained that supposedly falsified Einstein's special theory of relativity (Holton 1988). Each time, subsequent research revealed that it was the experiments and not the theory that were at fault. The important point, however, is that this conclusion was not apparent at the time that the results were generated. For example, 10 years passed before researchers identified the inadequacies of the equipment used in one of the experiments that had supposedly falsified special relativity. By implication, when an anomalous result is first obtained, only additional research can determine how best to interpret the result. In Einstein's words: 'whether there is an unsuspected systematic error or whether the foundations of relativity theory do not correspond with the facts one will be able to decide with certainty only if a great variety of observational material is at hand' (cited in Holton 1988: 253). In short, despite widespread belief to the contrary, there is no such thing as a 'critical experiment' because empirical knowledge is inductive and, thus, quite fragile when viewed in isolation (Chapanis 1967). Like the other points that are reviewed above, this insight is far from new, but it too has not been given the attention that it deserves.

As several authors have pointed out (e.g. Dar 1987, Rosnow and Rosenthal 1989, Cohen 1990, Thompson 1996, Rossi 1997, Schmidt and Hunter 1997), the way in which NHST and ANOVA are used in practice tends to cause researchers to overlook this epistemological limitation. In the extreme, the attitude is: 'if a statistical test is significant at $p < 0.05$, then the research hypothesis is true, otherwise it is not'. If valid, such an inferential structure would make life easier for researchers. Unfortunately, what NHST really evaluates is the probability that the data could have arisen given that the null hypothesis were true, *not* the probability that the null hypothesis is true given the data that were obtained (Cohen 1994). Although both of these quantities are conditional probabilities, they are logically very different from each other. NHST only allows us to make inferences of the first kind. Therefore, as surprising as it may sound, 'significance tests cannot separate real findings from chance findings in research studies' (Schmidt and Hunter 1997: 39), a statistical fact that should really give us considerable pause.

Researchers frequently ignore the fact that there is no objective, mechanical procedure for making a dichotomous decision to evaluate the validity of research findings (e.g. Chow 1996, 1998). This attitude can unwittingly have a devastating effect on a body of literature. A case study described by Rossi (1997) provides an incisive, if somewhat depressing, example. He reviewed the literature on a psychological phenomenon known as 'spontaneous recovery of verbal associations'. During the most intensive period of experimental investigation (1948–1969), about 40 articles were published on this topic. However, only about half of these studies led to a statistically significant effect of spontaneous recovery. Consequently, most textbooks and literature reviews concluded that the data were equivocal, and, thus, that the empirical evidence for spontaneous recovery was unconvincing. Eventually, the collective wisdom became that spontaneous recovery was an ephemeral phenomenon, and, as a result, research in the area was essentially abandoned.

Rossi (1997) conducted a retrospective analysis of the collective findings in this body of literature. Data from 47 experiments with an aggregate of 4926 participants were included in the analysis. Only 43% of these studies reported statistically significant results at $p < 0.05$. This low percentage of significant results led researchers to doubt the existence of the spontaneous recovery effect. However, when the experiments were analysed as a whole, there was statistically significant evidence in support of the spontaneous recovery effect ($p < 0.001$). Rossi also conducted an effect size analysis and a power analysis across these studies. The results indicate that the average effect size was relatively small ($d = 0.39$) and that the average power was quite low (0.38). Together, these findings explain why the significant effects were in the minority. Because researchers were dealing with a small effect and their studies had low power, many experiments failed to detect a statistically significant effect.

Together, these facts add up to a fascinating illustration of how naive attitudes about both statistical tests and the value of replication can have a deep impact on a body of literature. As Rossi (1997) pointed out, researchers did not report any effect sizes, so they did not know that they were dealing with a small effect. Similarly, no study reported power, so researchers were not aware that their experiments had low power. With this veil of ignorance as background, researchers (incorrectly) interpreted the results from each experiment using a dichotomous decision criterion: if $p < 0.05$, then the result is valid, otherwise it is not. However, as Rosnow and Rosenthal (1989: 1277) have observed, 'dichotomous significance testing has no ontological basis ... surely, God loves the 0.06 nearly as much as the 0.05' (see also Cowles and Davis 1982). Because of the combination of small effect and low power, 57% of the experiments did not generate results that passed the naïve (and indefensible) dichotomous decision criterion. This, combined with a lack of appreciation for the importance of replication across studies, led researchers to abandon what turned out to be a legitimate, albeit small, psychological effect.

What can be concluded from the spontaneous recovery case study? First, the case shows, once again, the value of calculating effect size and power so that researchers can better interpret their results. Secondly, the case also illustrates how misleading and unproductive it is to use the $p < 0.05$ criterion (or any other dichotomous decision rule) as the gatekeeper of scientifically acceptable knowledge. As Rossi (1997: 183) pointed out, 'the inconsistency among spontaneous recovery studies may have been due to the emphasis reviewers and researchers placed on the level of significance attained by individual studies ... A cumulative science will be difficult to achieve if only some studies are counted as providing evidence'. Thirdly, and relatedly, the spontaneous recovery case study also brings home the importance of replication across multiple studies. It is the pattern of results across studies that is most important for building scientific knowledge. In the words of Abelson (1995: 77), 'Research conclusions arise not from single studies alone, but from cumulative replication'. Even if no single result reaches statistical significance at the $p < 0.05$ value, the entire pattern of results can still be statistically significant when viewed as a whole. The converse point is equally valid: 'A successful piece of research doesn't conclusively settle an issue, it just makes some theoretical proposition to some degree more likely. Only successful future replication in the same and different settings ... provides an approach to settling the issue' (Cohen 1990: 1311).

How many cases like the one reviewed by Rossi (1997) are there in the ergonomics science literature? It is very difficult to answer this question. Nevertheless, there is one thing of which we can be sure: Making decisions on a dichotomous basis using NHST alone will only make it more likely for such problems to plague the ergonomics science literature. It is for this reason that an increasing number of noted researchers have felt the need to point to the importance of replication to building sound, cumulative knowledge (e.g. Rosnow and Rosenthal 1989, Meehl 1997, Schmidt and Hunter 1997). This lesson is perhaps the most important one of all amongst the ones that have been reviewed.

3. Is all of this obvious?

Seasoned ergonomics scientists might object that the six points in the previous section are obvious, and, thus, that this review does not make a significant contribution to the literature. If this is indeed the case, then one would expect that the vast majority of the empirical articles published in *Human Factors*, the flagship journal of the discipline in the US, would exhibit an awareness of most of these points. To test this hypothesis empirically, an informal review was conducted of all of the articles published in volume 40 of *Human Factors*.

3.1. *Method*

For each of the empirical articles in that volume, the number that reported: (a) an individual participant analysis of any kind (corresponding to point 1 in the literature review); (b) an analysis of a particular order of means, an interval magnitude, or a point prediction (point 2); (c) an MTMM analysis (point 3); (d) an analysis of association strength or effect size (point 4); and (e) power or confidence intervals (point 5) were counted. This procedure is not fool-proof (e.g. MTMM is not the only way to assess reliability, convergent validity, and discriminant validity), but it provides a more than adequate basis for an informal survey.

3.2. *Results*

Figure 4 illustrates the number of articles that used various data analysis methods in the sample. A number of patterns clearly stand out. First, ANOVA is, by far, the most frequently used method of data analysis. It was used in 33 of the articles that were surveyed, twice as frequently as the next most popular method of data analysis. Secondly, only one study reported an individual analysis of each participant's behaviour as opposed to relying just on group means, thereby showing that point 1 in this review is rarely recognized in practice. Thirdly, only a handful of articles used nonparametric tests, and, even in these rare cases, the tests were not used to make more stringent predictions, thereby showing that point 2 is not widely put into practice. Fourthly, no article reported a MTMM analysis, thereby suggesting that point 3 in this literature review is not put into practice. Fifthly, no article reported an analysis

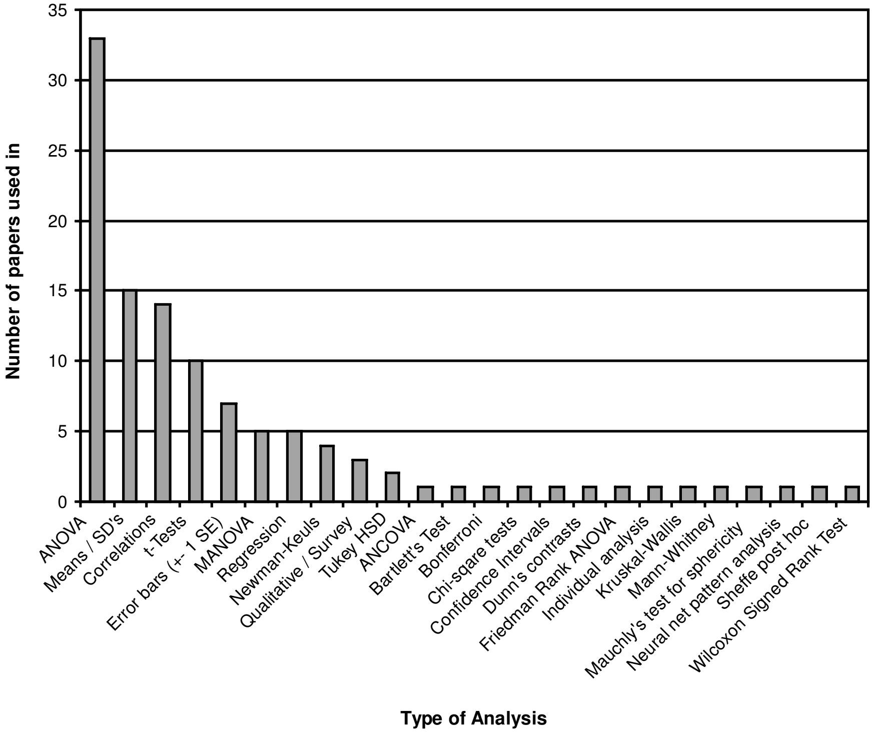

Figure 4. Histogram of the number of articles reporting various types of data analysis methods in volume 40 of *Human Factors.*

of effect size or association strength, thereby suggesting that point 4 is also not put into practice. Finally, only one article presented a graph of confidence intervals and no article presented a power analysis, thereby showing that point 5 is also rarely put into practice.

3.3. *Discussion*

Perhaps seasoned ergonomics scientists are indeed aware of all of the points in this review, but, if so, then this informal survey convincingly shows that there is a discrepancy between what is acknowledged to be true and the way in which we behave (the authors include themselves in this lot). The recommendations made in this article have very rarely found their way into ergonomics science. Even with all of the important limitations that are associated with it, ANOVA remains far and away the *de facto* standard for data analysis in ergonomics science.

4. Conclusions

The purpose of this literature review is not to point a finger at ergonomics scientists who have relied solely on NHST or ANOVA to analyse data. The authors are just as guilty of uncritically using traditional methods of data analysis as anyone else. After all, these are the techniques that have been taught, are well known by journal editors and reviewers, and are supported by software packages. Thus, there are numerous pressures that cause many to continue to use the traditional methods. Nevertheless, one of the main points of this article is that, by following these pressures, ergonomics scientists are—perhaps in many cases unknowingly—incurring a substantial scientific cost:

> When passing null hypothesis tests becomes the criterion for successful predictions, as well as for journal publications, there is no pressure on the ... researcher to build a solid, accurate theory; all he or she is required to do, it seems, is produce 'statistically significant' results (Dar 1987: 149).

The data analysis methods advocated here can lead to a more mature science of ergonomics, but they require one to follow a path of greater effort. For example, some of the methods reviewed require that experiments are designed differently. If we are going to conduct individual level analyses like Hammond *et al.* (1987) did, then we need to rely more on within-participants designs. If we are going to use the MTMM advocated by Campbell and Fiske (1959), we need to include multiple constructs and multiple methods in a single experiment. If we are going to be able to make point or interval predictions, we need to develop stronger theories to guide experimentation. If we are going to develop a more cumulative knowledge base, we need to engage in more replication than done in the past. Thus, a change in data analysis techniques is not a cosmetic modification to be taken lightly. Instead, it requires some deep changes in the ways in which ergonomics science is conducted.

Because of the enormity of this task, most ergonomists typically find it easier to stick to that with which they are most comfortable. Meehl (1990) describes a typical reaction to the critiques of NHST and ANOVA that he has made over the years:

> Well, that Meehl is a clever fellow and he likes to philosophize, fine for him, it's a free country. But since we are doing all right with the good old tried and true methods of Fisherian statistics and null hypothesis testing, and since journal editors do not seem to have panicked over such thoughts, I will stick to the accepted practices of my trade union and leave Meehl's worries to the statisticians and philosophers (p. 230).

In short, to effect a change in the way ergonomics scientists and practitioners analysze their research data will not be easy.

In this article, a step has been taken toward facilitating positive change by: (a) describing the limitations of traditional methods of analysis, especially ANOVA; (b) explaining why those limitations are relevant to ergonomics science and practice; (c) describing other methods of data analysis that address some of those limitations; and (d) citing many references that readers can consult to obtain the mechanical details on how to perform these less-familiar analyses (see Bailar and Mosteller (1988) and Wilkinson *et al.* (1999) for additional guidance and explanations). It is the authors' hope that ergonomics scientists and practitioners will consider using some of these techniques the next time that they conduct empirical research. Although the effort required will admittedly be higher than usual, through such incremental efforts we can progress toward more sound and cumulative scientific knowledge.

5. Postscript

'Much of what we have said has been said before, but it is important that our graduate students hear it all again so that the next generation of ... scientists is aware of the existence of these pitfalls and of the ways around them' (Rosnow and Rosenthal 1989: 1282)

Acknowledgements

This research project was sponsored by a contract from the Japan Atomic Energy Research Institute (Dr Fumiya Tanabe, contract monitor), and a research grant from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Ian Spence, Fumiya Tanabe, Bill Howell and especially the anonymous reviewers for their contributions and help.

References

Abelson, R. P. 1995, *Statistics as principled argument* (Hillsdale, NJ: Lawrence Erlbaum Associates).

Abelson, R. P. 1997, On the surprising longevity of flogged horses: why there is a case for the significance test, *Psychological Science*, **8**, 12–15.

Bailar, J. C., III and Mosteller, F. 1988, Guidelines for statistical reporting in articles for medical journals: amplifications and explanations, *Annals of Internal Medicine*, **108**, 266–273.

Bakan, D. 1966, The test of significance in psychological research, *Psychological Bulletin*, **66**, 423–437.

Bryan, W. L. and Harter, N. 1897, Studies in the physiology and psychology of the telegraphic language, *The Psychological Review*, **1**, 27–53.

Bryan, W. L. and Harter, N. 1899, Studies on the telegraphic language: the acquisition of a hierarchy of habits, *The Psychological Review*, **6**, 347–375.

Campbell, D. T. and Fiske, D. W. 1959, Convergent and discriminant validation by the Multitrait-Multimethod Matrix, *Psychological Bulletin*, **56**, 81–105.

Chapanis, A. 1967, The relevance of laboratory studies to practical situations, *Ergonomics*, **10**, 557–577.

Chow, S. L. 1996, *Statistical significance: Rationale, validity, and utility* (London: Sage).

Chow, S. 1998, Précis of statistical significance: rationale, validity, and utility, *Behavioral and Brain Sciences*, **21**, 169–239.

Christoffersen, K., Hunter, C. N. and Vicente, K. J. 1994, *Research on factors influencing human cognitive behaviour (I) (CEL 94-05)* (Toronto: University of Toronto, Cognitive Engineering Laboratory).

Cohen, J. 1988, *Statistical power analysis for the behavioral sciences*, 2nd edn (Hillsdale, NJ: Lawrence Erlbaum Associates).

Cohen, J. 1990, Things I have learned (so far), *American Psychologist*, **45**, 1304–1312.

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

Cowles, M. and Davis, C. 1982, On the origins of the .05 level of statistical significance, *American Psychologist*, **37**, 553–558.

Dar, R. 1987, Another look at Meehl, Lakatos, and the scientific practices of psychologists, *American Psychologist*, **42**, 145–151.

Fowler, R. J. 1985, Point estimates and confidence intervals in measures of association, *Psychological Bulletin*, **98**, 160–165.

Hagen, R. L. 1997, In praise of the null hypothesis statistical test, *American Psychologist*, **52**, 15–24.

Hammond, G. 1996, The objections to null hypothesis testing as a means of analysing psychological data, *Australian Journal of Psychology*, **48**, 104–106.

Hammond, K. R., Hamm, R. M. and Grassia, J. 1986, Generalizing over conditions by combining the Multitrait-Multimethod Matrix and the representative design of experiments, *Psychological Bulletin*, **100**, 257–269.

Hammond, K. R., Hamm, R. M., Grassia, J. and Pearson, T. 1987, Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment, *IEEE Transactions on Systems, Man, and Cybernetics*, **17**, 753–770.

Harlow, L. L., Mulaik, S. A. and Steiger, J. H. 1997, *What if there were no significance tests?* (Mahwah, NJ: Lawrence Erlbaum Associates).

Heldref Foundation 1997, Guidelines for contributors, *Journal of Experimental Education*, **65**, 95–96.

Holton, G. 1988, *Thematic origins of scientific thought: From Kepler to Einstein*, revised edn (Cambridge, MA: Harvard University Press).

LEE, J. D. 1992, Trust, self-confidence, and operators' adaptation to automation. Unpublished doctoral dissertation, Urbana, IL: University of Illinois at Urbana-Champaign, Department of Mechanical & Industrial Engineering.

LEE, J. D. and MORAY, N. 1994, Trust, self-confidence, and operators, adaptation to automation, *International Journal of Human-Computer Studies*, **40**, 153–184.

LOFTUS, G. R. 1991, On the tyranny of hypothesis testing in the social sciences, *Contemporary Psychology*, **36**, 102–105.

LOFTUS, G. R. 1993a, Editorial comment, *Memory and Cognition*, **21**, 1–3.

LOFTUS, G. R. 1993b, A picture is worth a thousand p values: on the irrelevance of hypothesis testing in the microcomputer age, *Behavior Research, Methods, Instruments, & Computers*, **25**, 250–256.

LOFTUS, G. R. 1995, Data analysis as insight: reply to Morrison and Weaver, *Behavior Research Methods, Instruments, and Computers*, **27**, 57–59.

LOFTUS, G. R. 2001, Psychology will be a much better science when we change the way we analyze data, *Current Directions in Psychological Science*, in press.

LOFTUS, G. R. and MASSON, M. E. 1994, Using confidence intervals in within-subject designs, *Psychonomic Bulletin & Review*, **1**, 476–490.

LOFTUS, G. R. and MCLEAN, J. E. 1997, Familiar old wine: great new bottle, *American Journal of Psychology*, **110**, 146–153.

LYKKEN, D. T. 1968, Statistical significance in psychological research, *Psychological Bulletin*, **70**, 151–159.

MEEHL, P. E. 1967, Theory testing in psychology and physics: a methodological paradox, *Philosophy of Science*, **34**, 103–115.

MEEHL, P. E. 1978, Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology, *Journal of Consulting and Clinical Psychology*, **46**, 806–834.

MEEHL, P. E. 1990, Why summaries of research on psychological theories are so often uninterpretable, *Psychological Reports*, **66**, 195–244.

MEEHL, P. E. 1997, The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky predictions, in L. L. Harlow, S. A. Mulaik and J. H. Steiger (eds), *What if there were no significance tests?* (Mahwah, NJ: Lawrence Erlbaum Associates), pp. 175–197.

MURPHY, K. R. 1997, Editorial, *Journal of Applied Psychology*, **82**, 3–5.

NEWELL, A. and ROSENBLOOM, P. S. (1981, Mechanisms of skill acquisition and the law of practice, in J. R. Anderson (ed.), *Cognitive skills and their acquisition* (Hillsdale, NJ: Lawrence Erlbaum Associates), pp. 1–53.

ROSNOW, R. L. and ROSENTHAL, R. 1989, Statistical procedures and the justification of knowledge in psychological science, *American Psychologist*, **44**, 1276–1284.

ROSSI, J. S. 1997, A case study in the failure of psychology as a cumulative science: the spontaneous recovery of verbal learning, in L. L. Harlow, S. A. Mulaik and J. H. Steiger (eds), *What if there were no significance tests?* (Mahwah, NJ: Lawrence Erlbaum Associates), pp. 175–197.

ROUANET, H. 1996, Bayesian methods for assessing importance of effects, *Psychological Bulletin*, **119**, 149–158.

ROZEBOOM, W. L. 1960, The fallacy of the null-hypothesis significance test, *Psychological Bulletin*, **57**, 416–428.

ROZEBOOM, W. W. 1997, Good science is abductive, not hypothetico-deductive, in L. L. Harlow, S. A. Mulaik and J. H. Steiger (eds), *What if there were no significance tests?* (Mahwah, NJ: Lawrence Erlbaum Associates), pp. 335–391.

SCHMIDT, F. L. and HUNTER, J. E. 1997, Eight common false objections to the discontinuation of significance testing in the analysis of research data, In L. L. Harlow, S. A. Mulaik and J. H. Steiger (eds), *What if there were no significance tests?* (Mahwah, NJ: Lawrence Erlbaum Associates), pp. 37–64.

SERLIN, R. C. and LAPSLEY, D. K. 1985, Rationality in psychological research: the good-enough principle, *American Psychologist*, **40**, 73–83.

SHROUT, P. 1997, Should significance tests be banned? Introduction to a special section exploring the pros and cons, *Psychological Science*, **8**, 1–2.

SIEGEL, S. 1956, *Nonparametric statistics for the behavioral sciences* (New York: McGraw-Hill).

SNYDER, P. and LAWSON, S. 1993, Evaluating results using corrected and uncorrected effect size estimates, *Journal of Experimental Education*, **61**, 334–349.

STEIGER, J. H. and FOULADI, R. T. 1997, Noncentrality interval estimation and the evaluation of statistical models, in L. L. Harlow, S. A. Mulaik and J. H. Steiger (eds), *What if there were no significance tests?* (Mahwah, NJ: Lawrence Erlbaum Associates), pp. 221–257.

THOMPSON, B. 1993, Foreword, *Journal of Experimental Education*, **61**, 285–286.

THOMPSON, B. 1994, Guidelines for authors, *Educational and Psychological Measurement*, **54**, 837–847.

THOMPSON, B. 1996, AERA editorial policies regarding statistical significance testing: three suggested reforms, *Educational Researcher*, **25**, 26–30.

VENDA, V. F. and VENDA, V. Y. 1995, *Dynamics in ergonomics, psychology, and decisions: Introduction to ergodynamics* (Norwood, NJ: Ablex).

VICENTE, K. J. 1992, Memory recall in a process control system: a measure of expertise and display effectiveness, *Memory & Cognition*, **20**, 356–373.

VICENTE, K. J. 1998, Four reasons why the science of psychology is still in trouble, *Behavioral and Brain Sciences*, **21**, 224–245.

WILKINSON, L. and TASK FORCE ON STATISTICAL INFERENCE, APA BOARD OF SCIENTIFIC AFFAIRS 1999, Statistical methods in psychology journals: guidelines and explanations, *American Psychologist*, **54**, 594–604.

WOODWORTH, R. S. 1938, *Experimental psychology* (New York: Holt).

XIAO, Y. and VICENTE, K. J. 2000, A framework for epistemological analysis in empirical (laboratory and field) studies, *Human Factors*, **42**, 87–101.

About the authors

Kim J. Vicente received his BASc (1985) in Industrial Engineering from the University of Toronto, his MS (1987) in Industrial Engineering and Operations Research from the Virginia Polytechnic Institute and State University, and his PhD (1991) in Mechanical Engineering from the University of Illinois at Urbana-Champaign. During 1987–1988, he spent 1 year as a visiting scientist in the Section for Informatics and Cognitive Science of the Risø National Laboratory in Roskilde, Denmark. During 1991–1992, he was on the faculty of the School of Industrial and Systems Engineering at the Georgia Institute of Technology. Since 1998, he has been professor of Mechanical & Industrial Engineering at the University of Toronto, and founding director of the Cognitive Engineering Laboratory there. He is also cross-appointed to the Institute of Biomaterials & Biomedical Engineering and the Department of Computer Science at the University of Toronto, an adjunct professor of psychology at Miami University, Ohio, and a registered Professional Engineer in the province of Ontario. Currently, Kim serves on the editorial boards of the *International Journal of Cognitive Ergonomics*, *Human Factors*, and *Theoretical Issues in Ergonomics Science*, and on the Committee for Human Factors of the US National Research Council/National Academy of Sciences. He is also a Senior Fellow of Massey College. Kim is the recipient of several research awards, including the Premier's Research Excellence Award, valued at $100 000. He has authored or co-authored 60 journal articles, and over 75 refereed conference papers. He is the author of *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-based Work*, published by Lawrence Erlbaum Associates.

Gerard L. Torenvliet received his BASc (1997) in Industrial Engineering and his MASc in Mechanical and Industrial Engineering (1999), both from the University of Toronto. From 1994–1995, he was employed as a User Support Specialist at Netron, Inc. in Toronto, Canada. During his graduate studies, he was affiliated with the Cognitive Engineering Laboratory at the University of Toronto, where he worked as a research assistant. After finishing his Master's degree, he was employed by Cognos, Inc. in Ottawa, Canada as a User Interface Designer, specializing in the design of data modelling and data administration products. Since the autumn of 2000, he has been employed by Watchfire, Inc. in Kanata, Canada, where he is leading the company's efforts in User Interaction Design.